Aus dem Adolf-Butenandt-Institut
Lehrstuhl Molekularbiologie im Biomedizinischen Centrum
Institut der Ludwig-Maximilians-Universität München
Medizinische Fakultät
Vorstand: Prof. Dr. rer. nat. Peter B. Becker

Computational methods for exploratory analysis of proteomics data

Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von

Wasim Aftab
aus
Howrah

2021

Mit Genehmigung der Medizinischen Fakultät

der Universität München

Betreuer:	Prof. Dr. rer. nat. Axel Imhof
Zweitgutachterin:	Prof. Dr. Maria del Sagrario Robles Martinez
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	28. Juni 2022

Table of content

Summary

With the emergence of the rapidly developing fields of omics and informatics, the 21st century has seen extraordinary advances in biological sciences. Omics technologies (genomics, tran-scriptomics, proteomics, metabolomics, metagenomics, phenomics, etc.) have made it possible to measure biological molecules of different classes in a high-throughput manner. Informatics has enabled us to see the big picture within the biological systems by interpreting those large datasets. As proteins are closer to a gene's function than the gene itself, proteomics has evolved as a major omics technology that enabled the analysis of all proteins in a mixture. Over the last decade, proteomic data has accumulated at an unprecedented rate, posing a slew of computational problems. In this thesis, I have attempted to familiarize readers with three such challenges that demand significant amount of computational effort. The opening chapter of my thesis serves as an introduction to the three challenges whereas the following chapters describe the methods I developed to overcomes these hurdles.

One of the key questions when studying protein function is to discover with which other proteins a protein of interest interacts with (alt: forms a complex with). Protein complexes mediate virtually all biological functions within a cell. During my thesis I developed a platform to comprehend the conceptual and technical underpinnings of protein complex discovery. Existing computational methods for predicting protein complexes rely mainly on traditional machine learning, which require multiple experiments in various modalities to generate a large amount of labeled data. Nevertheless, the complexomic experiments entail preparing, measuring, and analyzing numerous samples in the mass spectrometer, which comes at the expense of significant measurement time. Moreover, to the best of my knowledge, no full-fledged graphical user interface (GUI) based application existed in the literature that enables users to explore such complex datasets, perform sanity checks, and provide high-quality dynamic visualization of extracted information with ease. Since the majority of the researchers

in the proteomics community are wet lab scientists, a GUI is always favored over a command line application. Therefore, I proposed a desktop application ComplexMiner that promises a user friendly and aesthetic frontend to explore protein complexes in datasets generated by native liquid chromatography (LC) followed by mass spectrometry (MS). We employed one shot learning which is a variant of deep learning paradigm that aims to discover protein complexes with fewer complexomic experiments. However, ComplexMiner is still in the development and testing phase therefore, releasing it soon as an open-source software is my prime vision.

Other challenges inherent in the comprehensive analysis of high-throughput protein interaction data is the problem of batch effects, missing values, lack of effective visualization resources, and the issue of dealing with only few replicates in a given proteomic experiment. When I started my PhD, no workflow existed that combined robust statistical inference with dynamic visualization of the protein interaction network for the purpose of differential enrichment analysis of proteomics data. Therefore, I developed an easy-to-use pipeline in R, that facilitated users to analyze high-throughput proteomic datasets and to improve visualization of protein interaction network, I developed a software forceNetwork++ using R and JavaScript. Furthermore, using forceNetwork++, I designed a web application MiGENet (https://migenet.shinyapps.io/migenet/) which enabled researchers to extract spatial information regarding connectivity and molecular mechanisms. As a proof-of-concept, this web application has been used to visualize a large bait-prey interaction network regulating mitochondrial gene expression.

Finally, I also developed a software (ImShot) that allows a reliable identification of peptides from MALDI imaging and a novel strategy for integrating datasets from imaging mass spectrometry (IMS) with shotgun proteomics. The new ImShot software combines information from IMS and shotgun proteomics (LC-MS) measurements of serial sections of the same tissue.

It takes advantage of a two-group comparison to determine the search space of IMS masses after an unbiased hierarchical clustering aided deisotoping of the corresponding spectra. Ambiguity in annotations of IMS peptides when comparing with LC-MS datasets is eliminated by introduction of a novel scoring system (MLP score) that identifies the most likely parent protein of a detected peptide in the corresponding IMS dataset.

All the software pipelines, web, and desktop applications that I developed as a part of my PhD are open-source and freely available on my GitHub account (https://github.com/wasimaftab?tab=repositories). I hope that my efforts will inspire readers to pursue future research in this very exciting and rapidly growing field of study.

Zusammenfassung

Die technologische Entwicklung der letzten Jahrzehnte hat es möglich gemacht, biologische Moleküle verschiedener Klassen im Hochdurchsatzverfahren zu messen. Die enorme Datenmenge, die durch diese sogenannten -omics Technologien (Genomik, Transkriptomik, Proteomik, Metabolomik, Metagenomik, Phenomik usw.) generiert wurden, führte zur Etablierung eines neuen Forschungsbereichs im Rahmen der Lebenswissenschaften: der Bioinformatik. Die Bioinformatik hat es uns ermöglicht aus dieser großen Datenfülle ein molekulares Gesamtbild der biologischen Systeme zu erkennen und führte zur Entdeckung neuer unerwarteter Zusammenhänge. Die Fokussierung meiner Arbeiten auf die Proteomik beruht auf der Erkenntnis, dass das Endprodukt eines Gens wesentlich komplizierter und funktionsnäher ist als das Gen selbst. In dieser Arbeit habe ich versucht, die Leser mit drei großen Herausforderungen zur Interpretation und Darstellung proteomischer Daten vertraut zu machen, die einen erheblichen Rechenaufwand erfordern. Das Eröffnungskapitel meiner Arbeit dient als Einführung in die drei Herausforderungen, auf die ich in den folgenden Kapiteln ausführlicher eingehe.

Eine der wichtigsten Fragen bei der Untersuchung der Funktion von Proteinen ist die Frage, mit welchen anderen Proteinen ein bestimmtes Protein interagiert (d. h. einen Komplex bildet). Proteinkomplexe vermitteln praktisch alle biologischen Funktionen innerhalb einer Zelle. Die Charakterisierung von Proteinkomplexen in ihrer Gesamtheit ist unerlässlich, um die Geheimnisse dieser komplexen zellulären Maschinerie zu entschlüsseln. Während meiner Doktorarbeit habe ich eine Plattform entwickelt, um die konzeptionellen und technischen Grundlagen der Entdeckung von Proteinkomplexen zu verstehen. Bestehende computergestützte Methoden zur Vorhersage von Proteinkomplexen stützen sich hauptsächlich auf traditionelles maschinelles Lernen, das mehrere Experimente in verschiedenen Modalitäten erfordert, um eine große Menge an markierten Daten zu erzeugen. Die Komplexomik-

Experimente erfordern jedoch die Vorbereitung, Messung und Analyse zahlreicher Proben im Massenspektrometer, was mit einem erheblichen Zeitaufwand für das Massenspektrometer verbunden ist. Darüber hinaus gibt es meines Wissens in der Literatur keine vollwertige Anwendung auf der Basis einer grafischen Benutzeroberfläche (GUI), die es dem Benutzer ermöglicht, solche komplexen Datensätze zu erforschen, die Korrektheit zu überprüfen und die extrahierten Informationen auf einfache Weise dynamisch zu visualisieren. Um eine solche Analyse auch für Wissenschaftler ohne bioinformatischen Hintergrund zu ermöglichen, habe ich die Desktop-Anwendung ComplexMiner entwickelt, die eine benutzerfreundliche und ästhetisch ansprechende Oberfläche zur Untersuchung von Proteinkomplexen in Datensätzen bietet. Dabei habe ich eine One-Shot-Learning Methode eingesetzt, eine Variante des Deep-Learning-Paradigmas, das darauf abzielt, Proteinkomplexe mit weniger biologischen Experimenten zu entdecken. ComplexMiner befindet sich allerdings noch in der Entwicklungs- und Testphase, wobei ich hoffe es bald als Open-Source-Software veröffentlichen zu können.
weitere Herausforderungen bei der umfassenden Analyse von Proteininteraktionsdaten mit hohem Durchsatz sind das Problem der Batch-Effekte, der fehlenden Werte, des Mangels an effektiven Visualisierungshilfsmitteln und des Umgangs mit nur wenigen Replikaten in einem bestimmten Proteomikexperiment. Meines Wissens gab es (zu Beginn meiner Promotion) keine Software, de eine robuste statistische Inferenz mit einer dynamischen Visualisierung des Proteininteraktionsnetzwerks zum Zweck der differenziellen Anreicherungsanalyse von Proteomikdaten kombiniert. Daher habe ich eine einfach zu bedienende Pipeline in R entwickelt, die es den Nutzern erleichtert, proteomische Hochdurchsatzdatensätze zu analysieren, und um die Visualisierung von Proteininteraktionsnetzwerken zu verbessern, habe ich die Software forceNetwork++ mit R und JavaScript entwickelt. Darüber hinaus habe ich mit Hilfe von forceNetwork++ eine Webanwendung entwickelt (MiGENet (https://migenet.shinyapps.io/migenet/)), mit der Forscher räumliche Informationen über

Interaktion und molekulare Mechanismen extrahieren können. Als Machbarkeitsnachweis wurde diese Webanwendung verwendet, um ein großes Bait-Prey-Interaktionsnetzwerk zu visualisieren, das die mitochondriale Genexpression reguliert.

Schließlich habe ich auch eine Software (ImShot) entwickelt, die eine zuverlässige Identifizierung von Peptiden aus der MALDI-Bildgebung und eine neuartige Strategie zur Integration von Datensätzen aus der bildgebenden Massenspektrometrie (IMS) mit der Shotgun-Proteomik ermöglicht. Die neue ImShot-Software kombiniert Informationen aus IMS- und Shotgun-Proteomics-Messungen (LC-MS) von aufeinanderfolgenden Schnitten desselben Gewebes. Sie nutzt die Vorteile eines Zwei-Gruppen-Vergleichs, um den Suchraum der IMS Massen nach einer Deisotopisierung der entsprechenden Spektren (mit Hilfe von unvoreingenommenem hierarchischen Clustering) zu bestimmen. Mehrdeutigkeit in den Annotationen von IMS-Peptiden beim Vergleich mit LC-MS-Datensätzen wird durch die Einführung eines neuartigen Scoring-Systems (MLP-Score) beseitigt, das das wahrscheinlichste Vorläuferprotein eines detektierten Peptids im entsprechenden IMS-Datensatz identifiziert.

Alle Software-Pipelines, Web- und Desktop-Anwendungen, die ich im Rahmen meiner Promotion entwickelt habe, sind Open-Source und auf meinem GitHub-Konto frei verfügbar (https://github.com/wasimaftab?tab=repositories). Ich hoffe, dass meine Bemühungen die Leser dazu inspirieren werden, zukünftige Forschungen in diesem sehr spannenden und schnell wachsenden Forschungsbereich zu betreiben.

1. Introduction

Proteins implement nearly all cellular functions. Therefore, studying proteins can improve our understanding of cell as an integrated system. The study of proteins on a large scale is known as proteomics (Blackstock and Weir, 1999; Anderson and Anderson, 1998). Mass spectrometry (MS) that measures the mass-to-charge ratio of molecules in a sample has become an essential tool in proteomics research. Currently, two different types of mass spectrometry techniques are used in proteomics: top-down and bottom-up approaches. The use of top-down proteomics is mostly employed to determine distinct proteoforms (Smith et al., 2013). In contrast, bottom-up approaches quantify small peptides derived from proteins prior to MS analysis via protease-mediated cleavage (Eidhammer et al., 2013). The computational methods presented in this thesis are based on bottom-up MS datasets, hence I will focus on this proteomics approach solely.

1.1 Experimental approaches

1.1.1 General proteomics

In bottom-up approach, cleavage of protein is mainly done using trypsin as it produces peptides that are 6-25 amino acids long, which is ideal for the mass spectrometer. A typical workflow of bottom-up proteomics is shown in Fig. 1.1. As the mixture of peptides generated from all proteins in a cell is highly complex (E.coli cell lysate, for example, contains approximately 2.5k-5k proteins (Eidhammer et al., 2013)), the peptide sample must be separated further using reversed phase chromatography (RPC) directly coupled to the mass spectrometer. RPC separates peptides primarily on the basis of their hydrophobicity and after ionization, the sample is ionized using Electrospray ionization (ESI) and injected into the mass spectrometer. Within the mass spectrometer, the ions are separated based on their mass to charge ratio (m/z) by the mass analyzer and detected in using an ion detector.

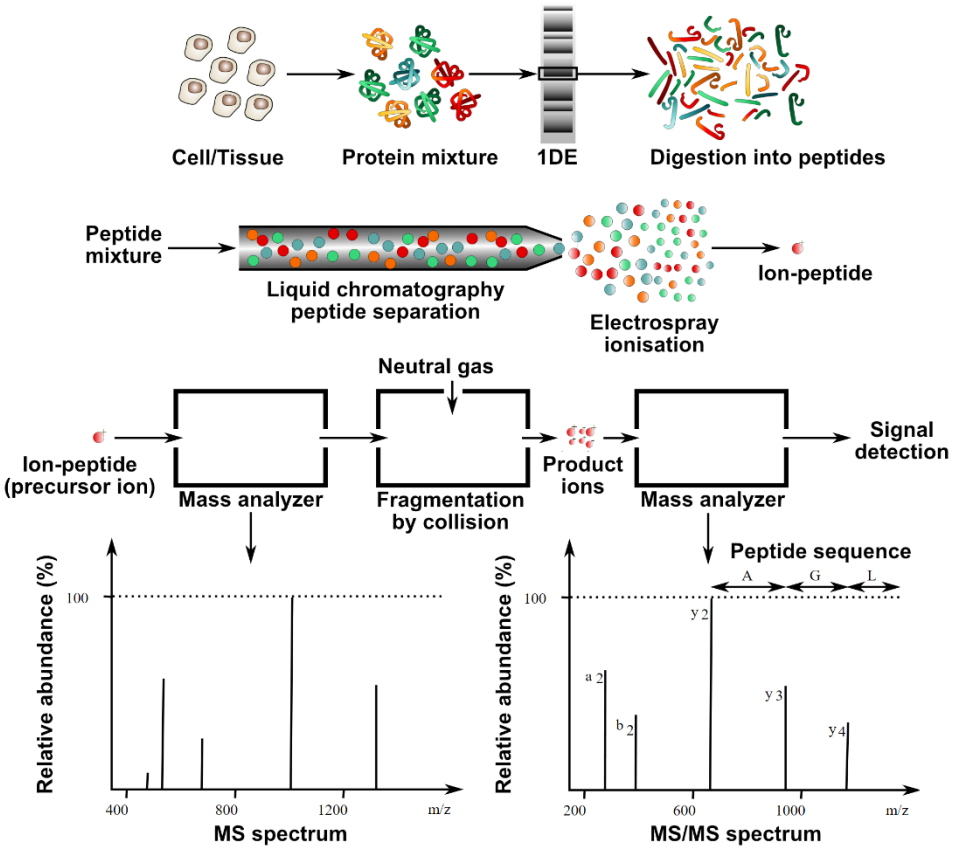


Fig. 1.1: A typical workflow of shotgun proteomics [Image adapted from (Hupé, 2012); Licensed under CC BY-SA 3.0].

When the ions collide with the detector, a mass spectrum is generated, and the corresponding data is stored in files with a proprietary format defined by the MS instrument's vendor. After the MS data processing software has identified the peptides in the sample they are mapped to the corresponding proteins.

MS experiments have two primary objectives: peptide identification and quantification. Typically, identification is accomplished by comparing the MS2 spectrum to a database using MS data processing software. However, quantification approaches can be classified into two categories based on the researcher's resources and/or objectives: label-based quantification and label-free quantification.

1.1.1.1 Label based quantification (LBQ)

In this peptide quantification approach, peptides are labeled with stable isotopes that have a defined mass shift, such that their observed mass in the MS1 or MS2 spectrum is shifted relative to the unlabeled peptide. Two of the more frequently used techniques are stable isotope labeling with amino acids in cell culture (SILAC) and isobaric tag for relative and absolute quantitation (iTRAQ).

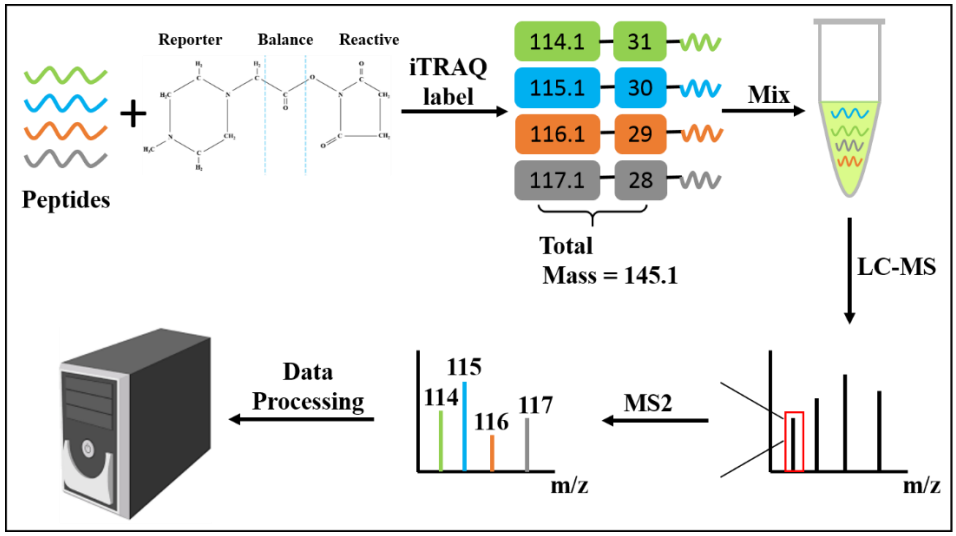


Fig. 1.2: iTRAQ reagent-based shotgun proteomics using iTRAQ-4-plex as an example [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

i. iTRAQ

It is an isobaric labeling method for measuring the amount of protein in multiple samples in a single experiment and can be used to compare 4 or 8 samples at once. The iTRAQ reagents are composed of three groups: a reporter, a reactive, and a balancer (See Fig. 1.2). The reactive group of an iTRAQ reagent is used to label peptides in multiple samples by covalently binding to their free amines, which are typically found at the N-terminus and lysine side chain. There are several different types of iTRAQ reagents, but they all have the same cumulative mass of different groups. Their chemistry has been optimized to ensure that all labeled peptides elute simultaneously from the liquid-chromatography (LC) system and can be quantified using so-called reporter ions (See Fig. 1.2). This article contains a more detailed description of how to

use iTRAQ reagents for protein complex and profiling studies (Zieske, 2006). There exist several tools (Tyanova et al., 2016a; Matrix-Science, 2021; Chen et al., 2021; SCIEX-ProteinPilot, 2021; Röst et al., 2016) for analyzing reporter-based data. However, MaxQuant (Tyanova et al., 2016a) is most popularly employed.

ii. SILAC

It is based on metabolically incorporating stable isotope labeled amino acids into the entire proteome. In this technique, cells are labeled with lysine and arginine, which retain stable, non-radioactive isotopes. In SILAC, two distinct populations of cells are cultured in two separate mediums. The light medium contains amino acids with the natural isotopes and the heavy medium contains stable isotope labeled amino acids. All the proteins from the cells cultured in heavy medium have amino acids in the heavy state after a sufficient number of cell divisions. In quantifying SILAC, we compare the ratio of introduced isotope-labeled peptides to unlabeled peptides. The signal intensities from light and heavy samples allow for a quantitative comparison of their relative abundance in the mixture.

To investigate protein-protein interactions using SILAC method, protein complexes are immunoprecipitated from a mixture of labeled cell lysates. Using SILAC, it is possible to efficiently distinguish specifically interacting proteins from non-specific background proteins. The abundance of specific interaction partners purified from the bait sample is significantly higher than the one from the control sample, resulting in quantified ratios much higher than one. In comparison, the abundance of non-specific background proteins should be comparable between the bait and control samples, resulting in a ratio close to one. When investigating exogenous, endogenous, or inducible PPIs, quantitative proteomics based on SILAC can be used to identify the proteins that interact specifically (Chen et al., 2015). Although, there are multiple software (Röst et al., 2016; Mortensen et al., 2010; Matrix-Science, 2021; Tyanova et

al., 2016a) available for processing SILAC data, MaxQuant (Tyanova et al., 2016a) is the most popular software for processing raw files from SILAC experiments.

1.1.1.2 Label free quantification (LFQ)

Labelling can result in a defined and measurable mass shift so that you can know, based on the mass, which sample you're working with. However, there is another type of quantification known as label-free quantification, and one of the more recently developed methods that works exceptionally well is intensity-based absolute quantification. As one might expect, label-free means that no mass tag or stable isotope labeling is required to obtain quantitative data. LFQ has a number of advantages over label-based proteomics, including its low cost and lack of the need for costly labeling reagents. Additionally, label-free quantitative proteomics is more time efficient than some label-based techniques, which include extensive labeling stages (Abdallah et al., 2012). However, the drawback here is high variability as LFQ approach do not control it internally. Here, I will elaborate on the LFQ approaches that are relevant to the computational methods presented in this thesis.

i. Intensity Based Absolute Quantification (iBAQ)

The underlying principle of iBAQ is quite straightforward. When developing a quantitative metric for the level of expression of a particular protein P_i in a mixture, the first step is to determine which identified peptides can be mapped to P_i. The cumulative intensity of those peptides is then divided by the number of theoretically observable peptides based on prior knowledge of P_i's sequence and the specificity of the digestion enzyme used, which is typically trypsin. This is to address the issue of larger proteins naturally generating more peptides due to their size.

The concept is similar to that of mRNA sequencing data analysis, in which the sequencing reads of a transcript are divided by the transcript length to account for the fact that longer transcripts simply generate more fragments. This corrected value is referred to as iBAQ score

of P_i, and it can be calculated for any protein of interest. It has been demonstrated that the iBAQ score correlates very well with the initial amount of protein injected into the mass spectrum (Schwanhäusser et al., 2011). A protein's iBAQ score can be defined mathematically as,

$$iBAQ = \frac{\sum_{j=1}^{n} I_j}{N} \quad (1.1)$$

Where $I_j \rightarrow$ is the intensity value of the jth peptide

N is the total number of theoretically observable peptides

n is the total number of observed peptides

This calculation is provided as an option in MS data processing software programs like MaxQuant (Tyanova et al., 2016a; Cox and Mann, 2008).

Another very useful feature of iBAQ is that if you analyze the proteome thoroughly, i.e., if you quantify the levels of almost every protein in the mixture and keep track of how much protein was in your sample and how much of that protein was actually input into the mass spectrometer, you can then estimate the absolute copy numbers or absolute concentrations of proteins in your original sample based on the quantified levels. Thus, the fraction of iBAQ scores for a particular protein relative to the sum of all iBAQ scores is proportional to the fraction of protein in your initial sample. The iBAQ score can be used to estimate the total number of copies in a cell. Schwanhäusser and colleagues used it to estimate the absolute copy number per cell for a variety of proteins in a fibroblast cell line in an incredibly impressive manner (Schwanhäusser et al., 2011).

When it comes to identifying as many proteins as possible in a sample, shotgun or discovery proteomics is the method of choice. Occasionally, however, the objective is to consistently identify and precisely quantify the same set of proteins under a variety of conditions. Then targeted proteomics may be the optimal technique, and in this section, we will discuss the most popular targeted proteomics approach in recent years: Selected Reaction Monitoring (Fig. 1.3)

serves as a foundation for comprehending SWATH-MS, a mass spectrometry technique that we used in our complexomics study.

ii. Selected Reaction Monitoring (SRM)

SRMs are typically determined using triple quadrupole mass spectrometers, with the first quadrupole Q1 acting as a mass filter to isolate a single peptide. The second quadrupole Q2 acts as a collision chamber, fragmenting the peptide selected. The third quadrupole Q3 performs the same function as the mass filter, but this time it filters the fragment ions of the selected peptide that hit the detector. Finally, a SRM measurement records the pairs (precursor, fragmentation-ion) over time to generate a chromatographic trace, also referred to as an SRM trace (Fig. 1.2).

By utilizing this SRM trace, software programs can accurately identify and quantify individual proteins, which is frequently not the case with conventional shotgun proteomic approaches. Readers will find this tutorial (Lange et al., 2008) very interesting as it discusses the application of SRM to quantitative proteomics.

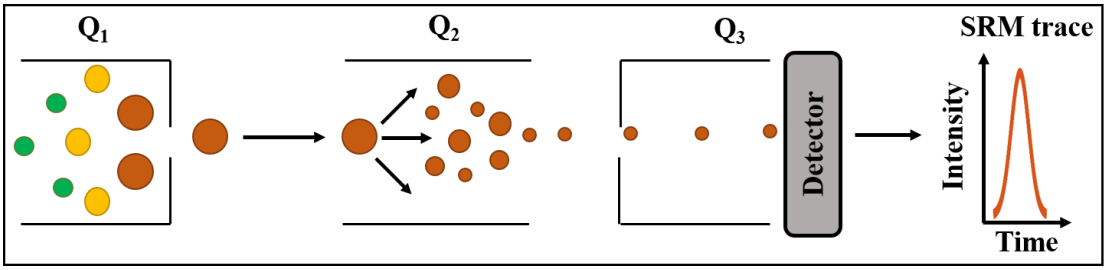


Fig. 1.3: SRM workflow; performed on a triple quadrupole mass spectrometer [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

Datasets obtained employing data-dependent acquisition (DDA) based bottom-up proteomics approach contains many missing values (Fig. 1.4). In DDA, during the MS1 scan, the instrument randomly selects the *n* most abundant peptides for MS2. As a result, you obtain snapshots of MS2 spectra that correspond to the specified time point in MS1 space. The critical point here is that the precursor space is sampled discontinuously in both mass and retention

time (t_r) dimensions, and because the on-the-fly heuristic is to select only the few most abundant peptides, quantification is biased toward high abundance species. As a result, numerous intensities corresponding to proteins/peptides are missing. In SRM, the MS instrument's role is to sample the precursor space deterministically based on the target peptides. The precursor space is sampled discontinuously in mass dimensions but continuously in the t_r dimensions in this case. However, the instrument monitors a relatively small number of precursors per run, despite the instrument's accuracy and consistency in quantification.

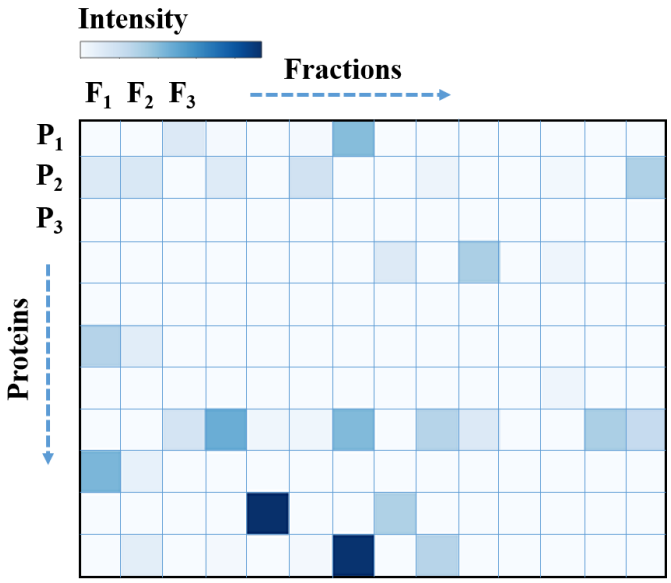


Fig. 1.4: Quantitative peptide/protein matrix- Columns are fractions, or the conditions and the rows are typically proteins/peptides. The colors in each cell represents intensity of a protein/peptide in the corresponding fraction. Higher intensities are shown using darker color. Missing values are depicted in white colors. [Image adapted from: (Aftab and Imhof, 2021); Licensing information: Appendix B].

So, is it possible to achieve SRM-like accuracy and consistency in quantification while covering almost the entire precursor space? The answer is yes, if we use a technique called sequential window acquisition of all theoretical mass spectra (SWATH-MS), which is one of the most advanced data independent acquisition (DIA) technologies available. Precursor selection is deterministic in DIA, and the instrument selects only the specified precursors. On

the other hand, data acquisition is completely untargeted/unbiased, i.e., you are not required to specify which peptide you are interested in. Proteomics researchers seek consistent and precise quantification of all peptides in multiple samples to address a variety of biological questions, and SWATH aims to accomplish just that.

iii. SWATH-MS

SWATH-MS divides the precursor space into chunks of small m/z precursor isolation windows over the measurable m/z range. A swath is an ensemble of fragment ion spectra acquired over the chromatographic range for a specified isolation window. The transitions (precursor, fragmentation-ion pairs) are then recorded using a high-resolution Q-TOF mass spectrometer. However, unlike SRM, SWATH does not target specific peptides (Fig. 1.5).

The resulting MS/MS signals from SWATH are continuous in both mass and time dimensions, allowing for a more comprehensive coverage of the proteome. This comes at the cost of extremely complex spectra that cannot be analyzed using conventional tools. Skyline (MacLean et al., 2010), PeakView (SCIEX-PeakView), Spectronaut (Bernhardt et al., 2012), and OpenSWATH (Röst et al., 2017) are all popular programs for deciphering complex SWATH spectra. Skyline and OpenSWATH are both free to use. Skyline has a graphical user interface and can also be used from the command line. In comparison, OpenSWATH only has a command line interface, making it more cumbersome to use. Additionally, third-party software packages are required to use OpenSWATH, as it makes use of a variety of tools from other sources to process DIA datasets.

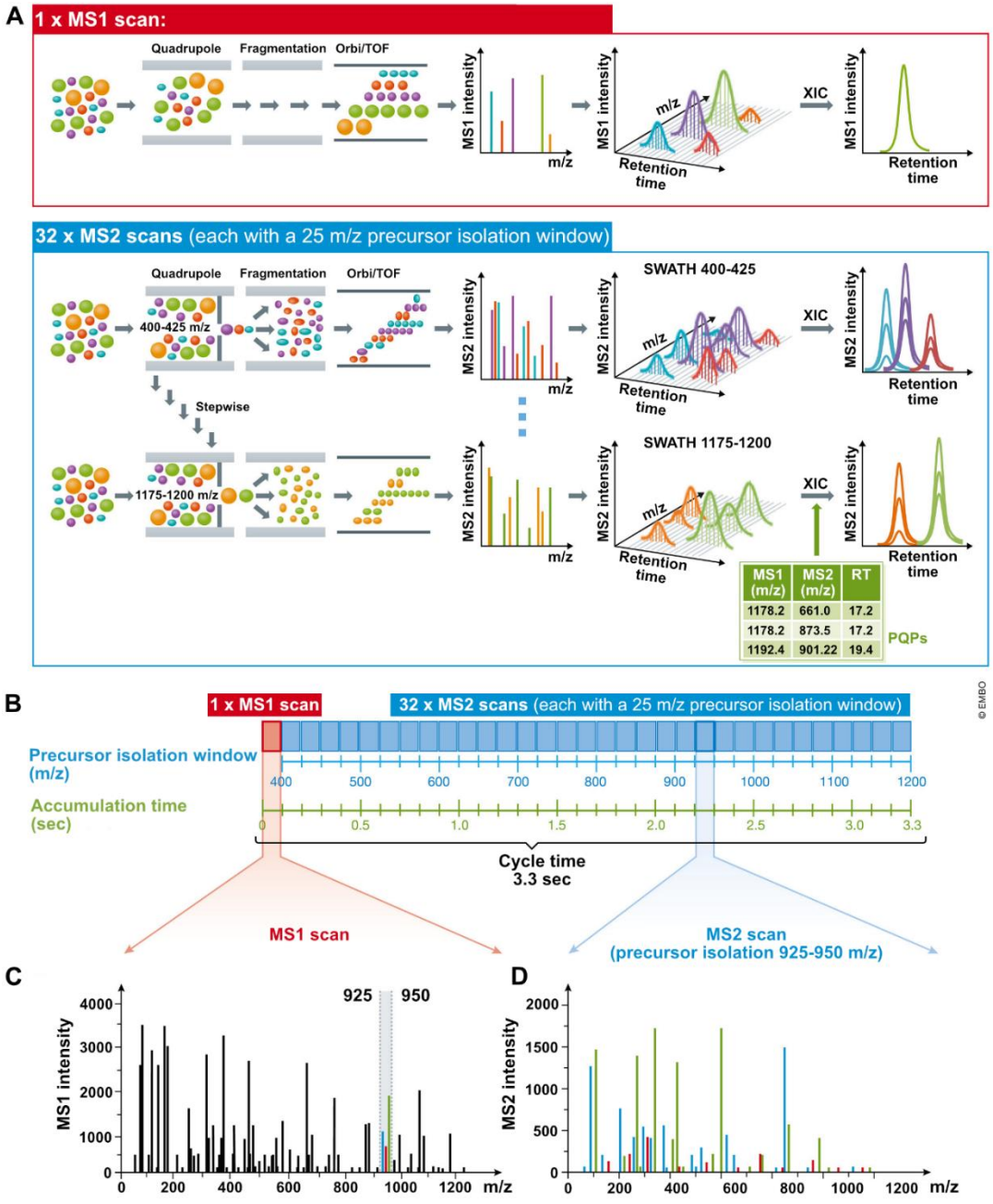


Fig. 1.5: Concept of SWATH based MS- (A) SWATH-MS measurements are performed using a quadrupole as first mass analyzer and a TOF or Orbitrap as second mass analyzer. **(B)** SWATH data acquisition scheme. **(C)** The MS1 full scan detects all peptide precursors eluting at a given time point. For example, in the mass range from 925 to 950 m/z, three co-eluting peptide species are detected (green, red, and blue). **(D)** The corresponding MS2 scan with a precursor isolation window of 925–950 m/z represents a mixed MS2 spectrum with fragments

of all three peptide species [Image adapted from: (Ludwig et al., 2018); Licensed under CC BY 4.0]**.**

SWATH-MS is more suitable for the current scenario as it generates fuller datasets by significantly reducing protein quantification uncertainties. But this comes at the cost of an additional computational burden. In order to set up a SWATH data analysis pipeline, the first step is to create a spectral library. In order to identify and quantify proteins in a sample, DIA data processing software utilizes information (non-redundant peptide transitions, t_r, etc.) stored in a spectral library and correlate with the corresponding information from the peptides in a sample.

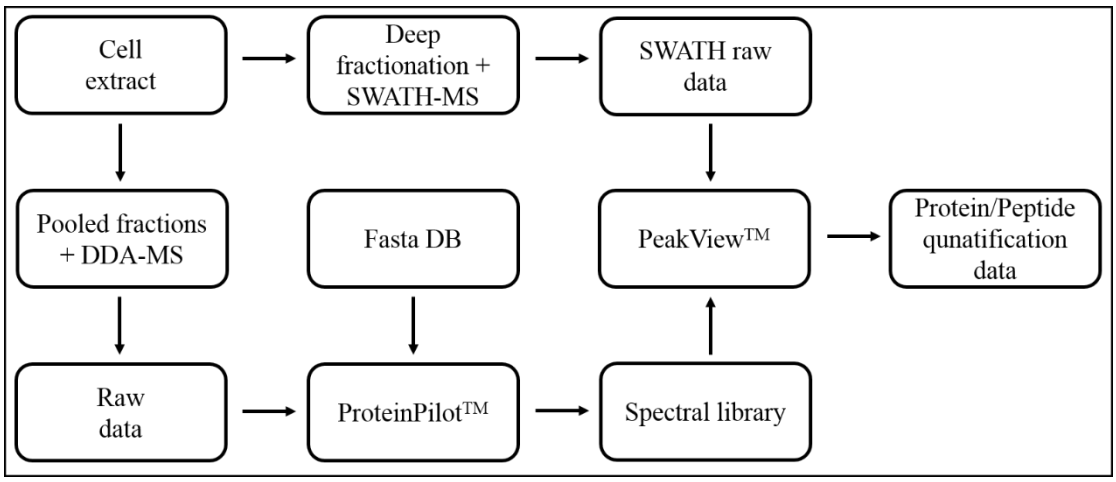


Fig. 1.6: Steps to generate spectral library for processing DIA-SWATH data [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

Generation of a spectral library: The most frequently used method to generate a spectral library is to pool a small aliquot of the cell extract for DDA measurements. After analyzing this DDA data, one can create a curated, annotated, and unique collection of fragment ion (MS2) spectra which is known as spectral library. It is critical to align the t_r of the library peptides with the t_r of the sample peptides. This is accomplished by spiking in the peptides with covering a broad range of t_r in both the library and in the samples. Biognosys standard peptides (Bernhardt et al.) are frequently used for this purpose. The steps to generate a spectral library

using ProteinPilot (SCIEX-ProteinPilot) and PeakView (SCIEX-PeakView) are depicted in Fig. 1.6.

Tsou CC and colleagues recently proposed a tool called DIA-Umpire (Tsou et al., 2015) that can generate spectral libraries directly from DIA data, saving time by omitting the experiments required to generate a classical spectral library. Additionally, deep learning (DL) approaches have resulted in methods for constructing theoretical spectral libraries by predicting the ion intensities of peptide fragments (Gessulat et al., 2019; Tiwary et al., 2019). These theoretical libraries have the potential to augment or even replace traditional spectral libraries. However, more research is needed to determine the extent to which these theoretical libraries outperform their classical counterparts in solving actual biological problems.

Generation of protein/peptide quantification dataset: PeakView receives the raw SWATH data, as well as the spectral library generated in the preceding step, and processes it (Fig. 1.6). As a result of that a final protein/peptide quantification matrix is produced, in which each row represents a protein profile over a range of fractions (labeled in columns). To discover protein complexes, it is imperative to cluster these elution profiles, which is a challenging task because of the inherent noise in the bottom-up proteomics data.

1.1.2 Proteomic profiling for protein complex discovery

1.1.2.1 Protein complexes drive almost all functions in a cell

Proteins hardly ever exist as single subunits but rather interact with other proteins to form bigger protein assemblies or complexes. These complexes are in charge of executing a multitude of functions within cells including formation of cytoskeleton, transportation of cargo, metabolism of substrates to produce energy, replication of DNA, protection and maintenance of the genome, transcription, and translation of genes to gene products, maintenance of protein turnover, and protection of cells from internal and external damaging agents (Srihari et al.,

2017). Thus, discovering the composition of protein complexes is critical to understand the cell as an interconnected system (Guruharsha et al., 2011).

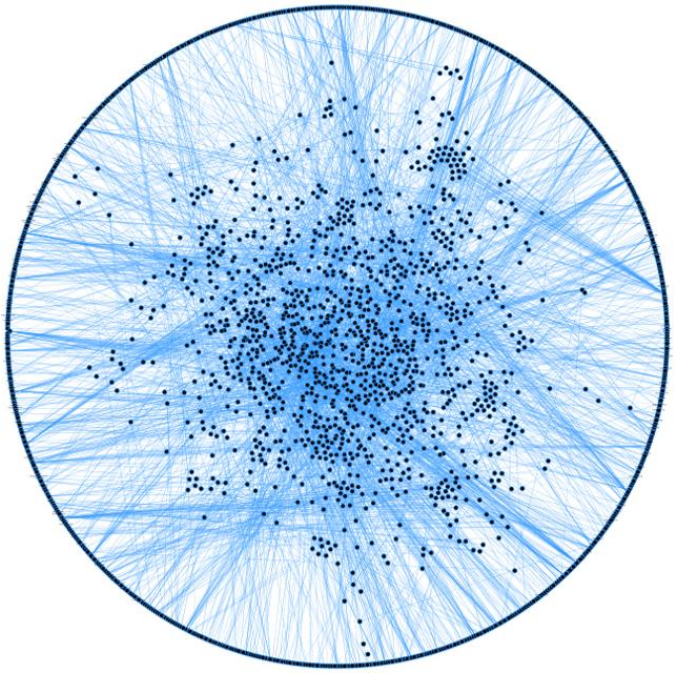


Fig. 1.7: Cells contain highly connected protein networks- The tiny blue dots represent protein complexes and the interactions among them are shown by light blue lines [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

A densely connected network of proteins/protein complexes exists in all cellular systems (Fig. 1.7). When somewhere in these densely linked networks a mutation occurs, the effect(s) spreads across the whole network via several PPIs. Thus, when a gene is mutated in such a way that the corresponding amino acid change impairs its interaction with one or more partners, a phenomenon known as edgetic perturbation (disturbing many edges within a network) occurs (Dreze et al., 2009). This mutation may result in a variety of diseases (Dreze et al., 2009). Because of the high degree of connectivity in the network, such perturbations can have a significant impact on a large portion of the protein network and, consequently, on the physiology of a cell (Diss and Lehner, 2018). When such critical changes occur in a cell, one way to investigate and predict their consequences is to create a map of protein complexes and then compare the network maps of the wild type and the disturbed (mutant) system.

Understanding the molecular biology of protein interaction networks may be made easier because of such investigations. Moreover, the MS-based protein correlation profiling that I present in this thesis may help increase our understanding of proteins whose functions are unknown (Crozier et al., 2017; Webb-Robertson et al., 2015; Havugimana et al., 2012). Here, I present an integration of experimental and computational techniques to find protein complexes in a cell extract. The methods entail fractionating native protein complexes using native LC followed by a tryptic digest and quantitative reversed phase chromatography (RPC) coupled to mass spectrometry to quantify protein complexes, as well as sophisticated computational methods to extract information from complex datasets (Fig. 1.8).

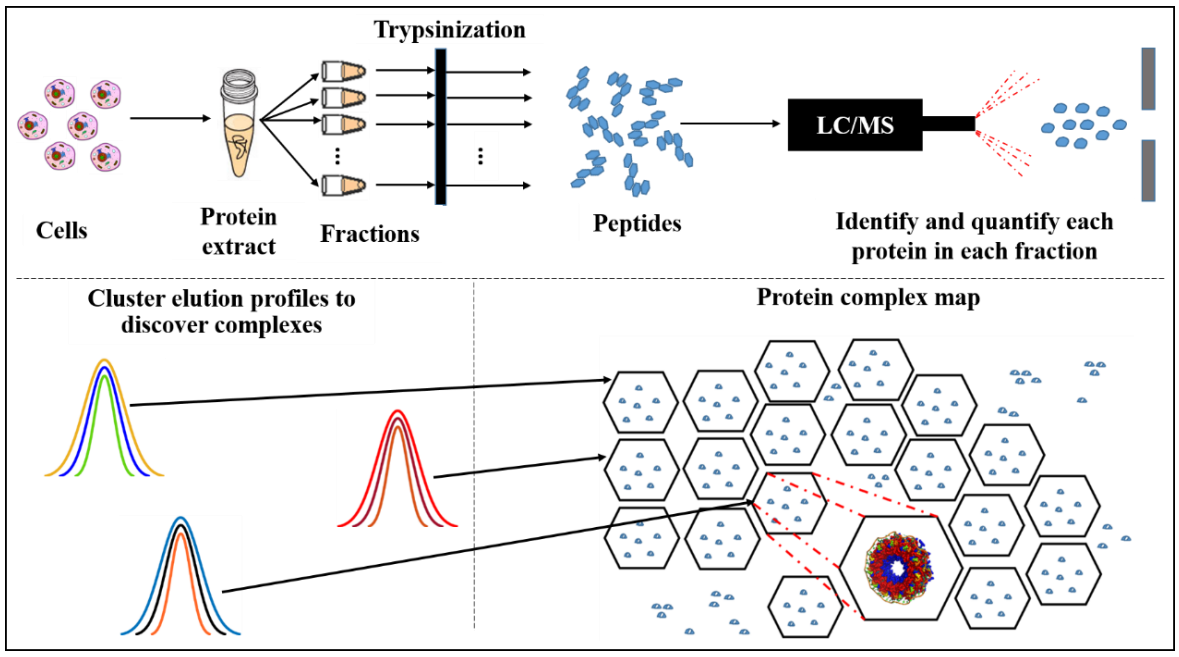


Fig. 1.8: Workflow to detect protein complexes in a cell extract- the protein extract is obtained from a population of cells and is further fractionated using liquid chromatography to separate protein complexes from each other [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

1.1.2.2 Role of LC and MS in protein complex discovery

To successfully isolate protein complexes from complex mixtures, the quality of the extracts is critical. To accomplish this, the mixture's complexity should be determined using a

traditional DDA shot gun proteomics experiment. If the number of proteins identified is consistent with expert knowledge (typically between 5000 and 10,000 proteins can be identified from whole cell extracts isolated from higher eukaryotes), the extract can be further fractionated. One of the most significant difficulties in this procedure is the relatively narrow dynamic range of many MS-based experiments. As a result, only extremely abundant protein complexes or subunits of them are detected. Havugimana et al. (Havugimana et al., 2012) suggests very deep fractionation by using multiple orthogonal modalities to deal with it. Chromatographic separation techniques based on size and charge have been widely used to address this problem (Havugimana et al., 2012; Webb-Robertson et al., 2015; Crozier et al., 2017).

i. Size Exclusion Chromatography (SEC)

SEC is a chromatographic technique that separates protein molecules based on their size differences as they elute from an SEC column. Spherical particles with defined pore sizes fill the column. Protein molecules larger than the pore sizes of the beads are unable to diffuse into them and thus elute first. Molecules ranging in size from very large to extremely small can penetrate the pores to varying degrees depending on their size. A molecule that is smaller than the smallest pore in the resin can enter the total pore volume. Eluted last are molecules that enter the total pore volume.

ii. Ion-exchange chromatography (IEX)

IEX is based on the reversible electrostatic interaction of protein with separation matrix. Beads which are either negatively charged (Strong cation exchange, SCX) or positively charged (Strong anion exchange, SAX) are packed into the chromatography column, which is also attached to a solvent support, such as glass or polystyrene, to yield protein complexes that are separated using increasing concentrations of salt ions such as Na^+ for SCX or Cl^- For SAX. Protein complexes with a higher net negative charge require a higher concentration of anions

hence eluting later than the complexes with a lesser net negative charge and vice versa is true for cationic exchange chromatography. Some factors, such as plate numbers, resin's peak capacity, and the type of analyte all contribute to determining the resolution (i.e., how well can very similar protein complexes be separated) of a native liquid chromatography column. (Madadlou et al., 2011; Loughran and Walls, 2011). In highly complex samples such as whole cell lysates there is simply not enough resolution to specifically separate these complexes. In order to achieve a higher resolution, measurements must be more time-consuming, which quickly becomes impractical. The trade-off between high resolution and the capacity to measure a variety of physiological variables can be performed by doing replicate tests on various modalities and then examining how consistently the apexes of a pair of proteins are aligned across all replicates. A Co-Apex score (defined later) can be generated from the above-mentioned procedure and could potentially be used to penalize/boost every pair of elution profile.

Once the protein complexes are separated (under native conditions) into different fractions, it is essential to identify and quantify proteins in each fraction. This is done using the bottom-up MS approach (Fig 1.1, Sec 1.1). Although most methods in the literature (Skinnider M, 2021; Havugimana et al., 2012; Guruharsha et al., 2011; Crozier et al., 2017) relies on DDA MS approach to discover protein complex, in recent past a method named CCprofiler (Heusel et al., 2019) (based on SEC followed by SWATH-MS) has demonstrated that DIA is a good alternative to resolve protein complexes present in an extract. One of the primary objectives of MS in this context is to maximize protein identification and quantification in the sample. SWATH-MS has the potential to provide the most exhaustive data in this regard, as it can theoretically cover the entire precursor space, resulting in very few missing values. Due to the stochastic peak selection process used in DDA experiments, missing values for proteins present in samples are a significant issue in quantitative proteomics using DDA.

1.1.2.3 Capturing transient interaction

The experimental strategies outlined so far are useful in providing a broad perspective of PPIs in a cell extract/system. However, they lack spatiotemporal control, which means it is difficult to know which proteins interact with the bait at a specific point in time, particularly, when the bait interacts with a certain interactor or when it is a subunit of a specific protein complex (Kenkel, 2018). Moreover, it is challenging to detect weak/transient protein interactions which are typically important in biological processes. Thus, scientists have attempted complementary approaches to detect dynamic protein interactions based on in-vivo labelling (Sears et al., 2019; Roux et al., 2012; Roux et al., 2013; Schopp et al., 2017; Varnaitė and MacNeill, 2016). BioID is the first and most widely employed method (Roux et al., 2013) in this category is described next.

i. BioID

Proximity-dependent Biotin identification, also referred to as BioID, is a new method for detecting weak or transient protein-protein interactions (PPIs). Biotin, often known as Vitamin H, is a co-enzyme that is covalently linked to the active site of a specific group of proteins termed Biotin dependent carboxylases and decarboxylases. The attachment of Biotin to its target protein occurs in two steps, both of which are mediated by the Biotin protein ligase (Fig. 1.9A). The biotin molecule is first activated by ATP, and then linked to a free amine group of the target protein, often a lysine amino acid. Biotin protein ligases have a very high specificity for their natural targets. Even though biotinylation occurs in almost all species, fewer than five proteins per creature are spontaneously biotinylated. BirA is the name of the biotin protein ligase in E. coli that naturally biotinylates just one protein. Kwon and Beckett (Kwon and Beckett, 2000) circumvented BirA's selectivity by engineering a mutant R118G capable of prematurely releasing biotin in its active state. BirA* is a mutant of BirA that was utilized to create the BioID method. Biotin in its active state is very reactive and either rapidly biotinylates

adjacent free amine groups or is quickly hydrolyzed. Streptavidin beads aid in the purification of biotin-containing molecules because the biotin molecule has a strong non-covalent interaction with streptavidin.

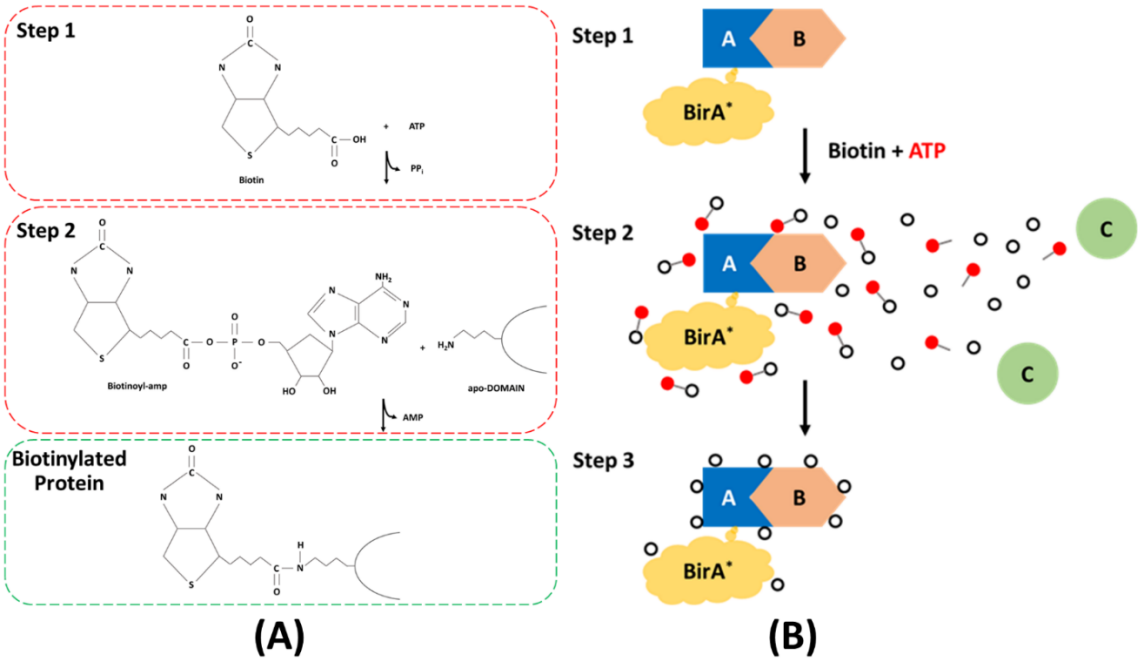


Fig. 1.9: Mechanisms of BioID- Showing **(A)** biotin protein ligase (BPL) reaction, which is a two-step (shown in red dashed boxes) reaction catalyzed by the BPL. In the first step, addition of ATP activates the biotin molecule. In the second step activated biotin molecule gets attached to the free amine group of a target protein. Product of biotinylation reaction is shown in green box. [Image adapted from (Henke and Cronan, 2014); Licensed under CC BY] **(B)** Biotinylation of a protein is a three-step process, in step 1, bait protein is fused with mutant protein BirA*. In the next step, biotin and ATP added to cell culture. BirA* and ATP convert biotin into its active form. Finally, the activated biotin molecules biotinylate nearby proteins by getting attached to their lysine group.

The idea of BioID was suggested in 2004 (Choi-Rhee et al., 2004) and developed in 2012 by Roux and colleagues (Roux et al., 2012). It relies on the fusion of a bait protein shown as protein A (Step 1, Fig. 1.9B) with BirA*. However, the fusion of protein A with BirA* must be performed in such a way that the properties of protein A remain unaltered. i.e., protein A

must be able to interact with its ligands, such as protein B, in the same way that it would normally do in the cell. Biotin and ATP are added to the cell culture media (Step 2, Fig. 1.9B) to distinguish between proteins that interact with or are nearby protein A in the cell and proteins that are further away (protein C in Fig. 1.9B). Biotin is converted to its active form in the presence of BirA* and ATP. The activated Biotin molecule either binds to a lysine of a nearby protein or is hydrolyzed, making it inactive. As a result, the closer a protein is to BirA*, the more likely it is to be biotinylated (Step 3, Fig. 1.9B). Biotinylated proteins are subsequently isolated using affinity capture with streptavidin beads and identified using mass spectrometry.

There are several benefits of using BioID method:

1. Only proteins that were in proximity or interacting with the bait at the time of labelling get biotinylated and can be easily enriched using streptavidin affinity capture.
2. BioID enables both the bait and the prey to interact in their natural environment. This minimizes the risk of environmental incompatibility that could occur if another method was used.
3. BioID is able to detect weak and transient interactions.

However, the method does have its disadvantages:

1. The level of biotinylation of a protein does not necessarily reflect the level of interaction of that protein with the bait. Perhaps that protein just has either more or less lysine groups to be biotinylated.
2. Low abundant proteins may be difficult to detect with the BioID technique, whereas highly abundant proteins can get artificially biotinylated.
3. Biotinylation is a permanent modification that has the potential to alter the behavior of certain proteins.

Although BioID is a relatively new technology, it has been extensively used in recent years. For example, Kyle J Roux's team used it to study Lamin A, a critical structural component of

the nuclear envelope of mammalian cells (Roux et al., 2012). Another group of scientists used BioID to discover proteins that interact with HIV-1 Gag, a structural polyprotein that mediates virus assembly of HIV by trafficking to the plasma membrane (Le Sage et al., 2015). Lambert et al. conducted a comparative study of chromatin-associated protein complexes using BioID and Affinity purification coupled with mass spectrometry(AP-MS) and discovered that BioID-MS recovered histone-associated proteins with less abundance bias than AP-MS (Lambert et al., 2015). Khan et al. compared BioID to their previously published AP-MS proteomic dataset using HopF2b as bait in the model plant Arabidopsis thaliana. In addition to many common protein interactions, they found several novel ones using BioID (Khan et al., 2018). Using proximity-based biotinylation based proteomics, a complex multi-layered structure of the chromocenter in *Drosophila* has recently been discovered (Kochanova et al., 2020). We recently used BioID based proteomics to understand how the translational activators: Cbp3-Cbp6, Cbs1, Cbs2, and Cbp1 regulate the translation of COB mRNA and we discovered a novel feedback loop by which these translational activators control the translation COB mRNA (Salvatori et al., 2020b). In another recent study we combined individual BioIDs of 40 baits to generate a large bait-prey network which revealed several key factors involved in mitochondrial gene expression (Singh et al., 2020).

ii. How BioID is compared with other PPI capturing experiments?

In past decades the proteomics community has been very active in developing experimental methods to explore PPI and this is evident by the existence of enormous number of articles on this domain in PubMed. As mentioned above proteins seldom function alone, rather by interacting with other proteins in different protein complexes and sometimes same protein shown additional functions (also known as a moonlighting protein) in altered cell conditions, under the influence of a stimuli or in pathological states and so on (Jeffery, 2018).

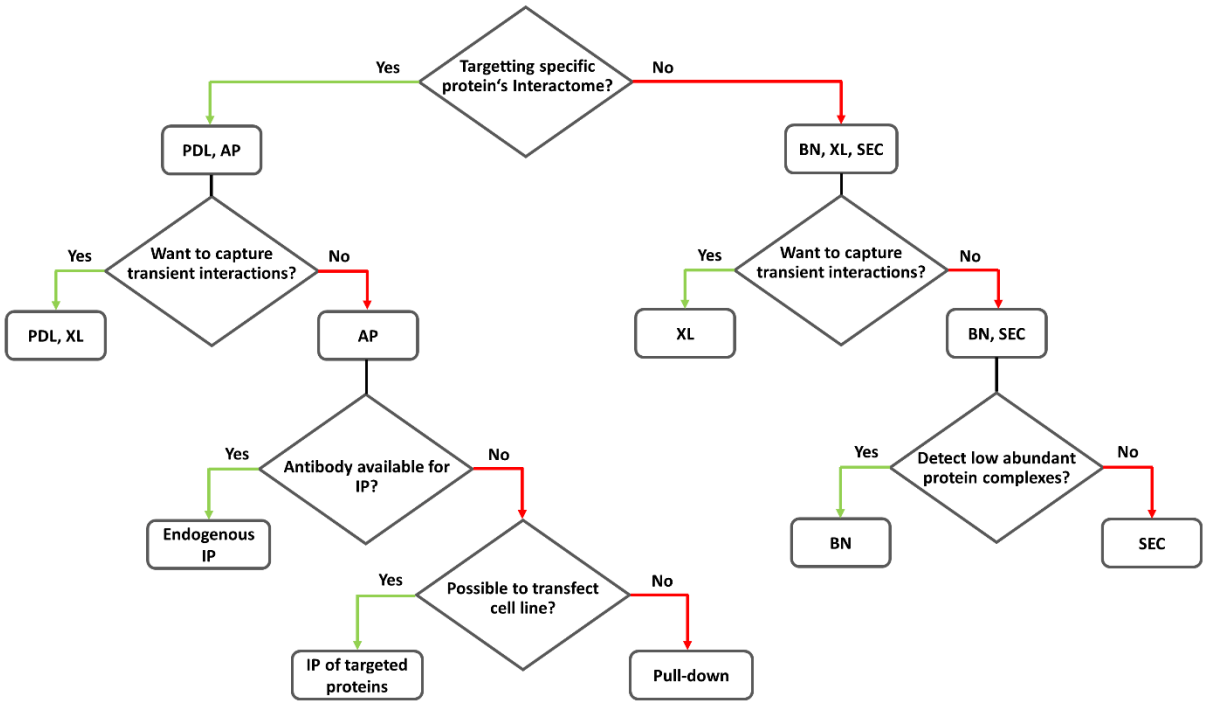


Fig. 1.10: Depending on the aim of the investigator this flowchart serves as a practical guide to select an appropriate experimental approach to study protein interaction- PDL → proximity dependent labelling; AP → affinity purification; BN → blue native gel; XL → cross-linking; SEC → size exclusion chromatography; IP → immunoprecipitation [Image adapted from (Iacobucci et al., 2020); Licensing information: Appendix A)]

Therefore, the understanding of a biological process is tightly related to the resolution of the interactome of a protein of interest. Depending on the aim(s) of an investigator, the biochemical procedures that have been developed to capture PPIs can be broadly classified as *targeted* or *untargeted* (Fig. 1.10). BioID based proximity labeling approach is applicable in vivo and is able to capture specific dynamic interactions, mainly due to the ability to biotinylate proteins quickly within a small radius of bait. Other methods in comparison have some drawbacks: like Pull-down assays work only in-vitro. Co-immunoprecipitation requires highly specific antibodies which necessitates tweaking the protocols for each target. Cross-linking experiment's major drawback is highly tedious downstream data analysis. Moreover, the search space grows exponentially as the number of peptides increases (Leitner et al., 2010; Yu

and Huang, 2018). Size exclusion chromatography suffers from low resolution. BN requires enormous amount of MS data analysis. Although BioID too has some disadvantages (Sec. 1.1.2.3), it allows the investigator to obtain a high resolution interactome that contains transient interactions and to reduce the complexity of downstream data analysis. I present the challenges associated with analyzing BioID proteomics data and provide computational methods to tackle them in chapter 3.

1.1.3 Spatial proteomics

Proteomic studies over the years have aimed at understanding the functional landscape of cells by optimum mapping of protein profiles at steady state and following a variety of perturbations in space and time. In addition to conventional LC-MS, emergence of Imaging mass spectrometry (IMS) has added a new dimension enabling observation of protein profiles in situ. IMS is a new chemical imaging technique that enabled us to obtain more information about biological samples than ever before possible (Cornett et al., 2007; McDonnell and Heeren, 2007; Stoeckli et al., 2001). It has emerged as a powerful tool allowing label free detection of numerous biomolecules in situ. In contrast to shotgun proteomics, proteins/peptides can be detected directly from biological tissues and correlated to its morphology, providing critical clinical information. Currently, the two ionization procedures that have revolutionized the use of mass spectrometers are matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), which in turn has made it possible for scientists to analyze biological substances much more easily (El-Aneed et al., 2009). MALDI based IMS is more popular because it produces singly charged peptides and proteins thus minimizing the complexity of IMS spectra. Therefore, in the next section I will elaborate on principles of IMS in the light of MALDI technique.

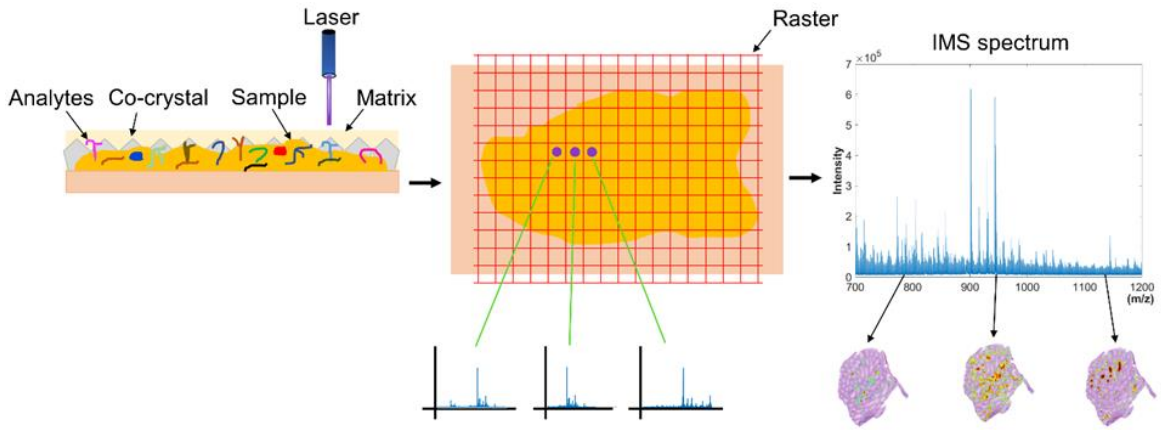


Fig. 1.11: MALDI Imaging mass spectrometry- Steps involved in an MALDI-IMS experiment.

1.1.3.1 MALDI-IMS

In a nutshell, MALDI-IMS enables direct imaging of biomolecules (proteins, peptides, metabolites, etc.) on a tissue surface, providing a spatial map which is crucial in tissue-based clinical research. To comprehend MALDI-IMS, it is critical to first and foremost understand the underlying technology, mass spectrometry (Fig. 1.1, sec. 1.1.1). In order to preserve the structural and chemical integrity of the tissue, the specimen is coated with a matrix solution. The matrix which frequently composed of organic acid or, more recently, nanoparticles is a critical component of MALDI-IMS. Spraying is the most common application method, as it results in the formation of a homologous matrix layer atop the sample. A consistent layer is necessary for the generation of unbiased data. The matrix facilitates mass spectrometry by localizing the sample, analytes and providing an ionization environment. Once deposited, matrix particles crystallize with the analytes, stabilizing them and preparing them for laser distribution (Left panel, Fig. 1.11). The Matrix selectively crystallizes with molecules of certain size, range, or characteristics, thus varying its composition can narrow down the window to find your molecules of interest. Following that, laser is used to vaporize and ionize the analyte simultaneously. Typically, a solid-state laser with ~355 nm is used. The sample space is divided into grids, with each square referred to as a raster (Middle panel, Fig. 1.11).

Within each raster, the laser beam ionizes a spot, and the resulting ions are passed through a mass spectrometer as described earlier, resulting in the generation of a mass spectrum for each spot in the sample. The data for a particular analyte can be extracted using its distinct mass-to-charge peak, and an image can be created using the intensity at each spot (Right panel, Fig. 1.11) (Hanrieder et al., 2011).

Traditionally, resolution of MALDI IMS has been limited by the size of the crystal. Once hit with laser, the entire crystal will be vaporized and sent to a mass spectrometer, making it the smallest resolvable detail in the image. Degradation of the organic matrix also worsens the signal to noise ratio. However, the recent use of nanoparticle matrix has alleviated these limitations (Sugiura and Setou, 2010; Banazadeh et al., 2018; Kratochvíl et al., 2021). Now, resolution is only dependent on the size of the laser dot, which can range from 200 μm – sub μm. However, decreasing the size of the laser dots will lengthen the analysis time as more spots need to be analyzed. Thus, there is a tradeoff between resolution and imaging speed. It is also possible to only image a portion of the sample to increase the speed. MALDI IMS has a lower resolution than common imaging modalities such as immunohistochemistry but compensates with its feature of multi-analyte detection.

Since its inception, IMS studies have successfully mapped molecular profiles to tissue morphologies in the disease context. Discovery of biomarkers and biological categorization of relevant diseases was also possible through IMS based investigations (Meistermann et al., 2006; Rauser et al., 2010; Gustafsson et al., 2011; Balluff et al., 2017; de San Roman et al., 2017; Pauker et al., 2019). However, most of these studies included complementary validation of IMS data as direct identification of molecules was not possible. Although metabolic profiles of tissues can now be identified in a relatively better way than before (Marcotte et al., 1999; Palmer et al., 2017), proteins/peptides cannot be identified yet by IMS leading to sub-optimal understanding of functional profiles of different cell types within tissues. Lack of high-

throughput MS2 modalities for peptides within the current IMS setups, especially MALDI-IMS result in this reduced identification rates. Further, due to lack of defined fragmentation modes, MS2 of peptides generates a mixture of fragment ions leading to ambiguous downstream identification. Therefore, complementation of MALDI-IMS with orthogonal shotgun proteomics has been adopted as a feasible approach in the recent past (Alberts et al., 2018; Longuespée et al., 2019; Schober et al., 2012; Huber et al., 2018; Groseclose et al., 2007; Franck et al., 2009).

1.2 Computational approaches

1.2.1 Computational methods for general proteomics

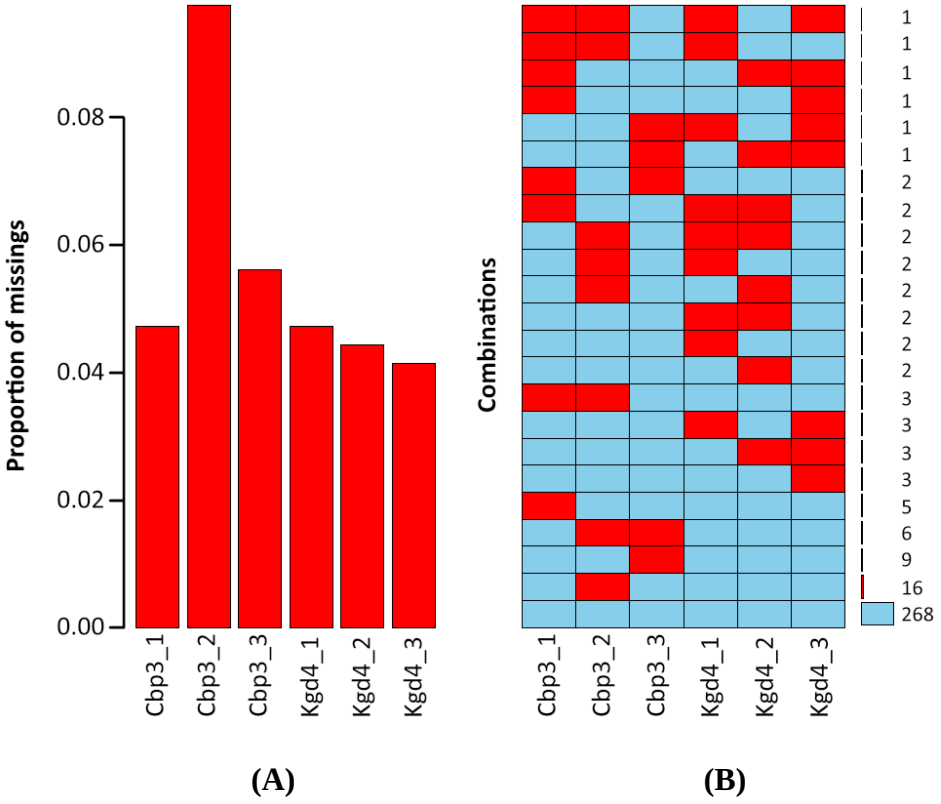


Fig. 1.12: Missing values in proteomics dataset- Showing **(A)** percentage of missing values in each replicate, at least 4% of values are missing in every replicate. **(B)** different combinations of missing values across replicates. Notice 268 proteins were identified and quantified in all the replicates, while missing values were found in at least one replicate for the remaining proteins.

A primary goal in mass spectrometry-based proteomics is to detect significant changes in protein abundance. This is especially important when studying subjects belonging to treatment/ control, mutant/wild-type, or diseased/healthy groups. Therefore, critical statistical data analysis tools are needed to prevent inaccurate conclusions. This inspired computational biologists to develop methods for extracting hidden patterns from large and complex proteomic datasets. These computational tools are typically released in the form of packages written in the R programming language, which includes a large number of statistical libraries that aid with data analysis. Nevertheless, knowledge discovery from proteomic datasets is challenging because of multiple factors, viz. batch effects, missing values, small number of replicates, lack of resource(s) for effective visualization of protein networks etc.

1.2.1.1 Challenges in high-throughput proteomic data analysis

i. Missing values

Missing values in proteomics dataset is a common scenario which occurs when a protein is quantified in some of the replicates but not all (Fig. 1.12A). Missing values can interfere with the statistical tests, one needs to handle them appropriately. The best would be working with full dataset i.e., with no missing values. However, as shown in Fig. 1.12B, there is no protein in which one more replicate is missing a value. Since it is common to apply t-test in the downstream analysis of proteomics data to assess the significant changes in protein abundances between groups, therefore, if missing values are not imputed properly then it can give rise to a set of proteins with artificially high fold change which might contribute to the set of false positives.

ii. Batch effects

Batch effects can occur when subsets of data are collected in a manner that systematically differs from the ways the other subsets of data are collected. The systematic differences refer to the innate differences between batches that may occur in time, place, instrument, calibration

of instruments or persons doing the experiment and many more. For example, starting blocks of animals on separate weeks, assaying samples in different runs, or euthanizing some of the animals at a time. Batches aren't themselves necessarily a problem and are sometimes unavoidable in many cases, but it takes planning to make sure they are organized correctly.

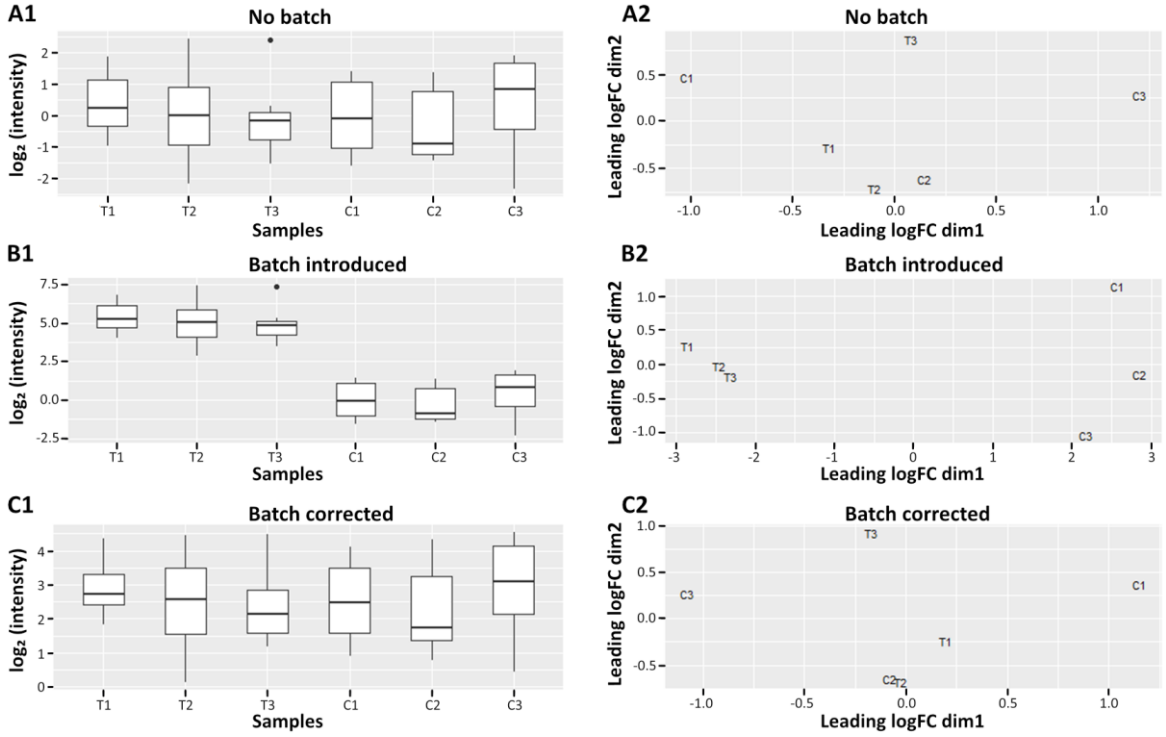


Fig. 1.13: Sanity checks to detect batch effects are important while analyzing proteomics data- Data is simulated from a normal distribution to demonstrate batch effect. Showing **(A1)** Boxplots of the replicates from two groups. **(A2)** Corresponding Multidimensional scaling (MDS) plots. **(B1)** Boxplots of the replicates from two groups after introducing batch effect **(B2)** MDS plot after introducing batch effects. **(C1)** Boxplots of the replicates from two groups after correcting batch effect. **(C2)** MDS plot after batch correction.

Batch effects can cause confounding if the treatments are in some way related to the batches. In the extreme, this can happen if all of one treatment was processed in one batch and all of a different treatment was processed in another batch. We can avoid confounding treatment effects with batch effects by balancing treatments within batches. This helps ensure that we don't create an association between the time, place or measurement method and our outcomes

of interest. If each batch is a little different then there is some variability in the data that is caused by the batches. In some cases, the variability from batch could be just as big or bigger than the treatment effects. But if we balance treatment groups across batches, that variability is balanced across treatments as well. Then we can account for batches using a statistical model. It is always preferable to perform sanity checks to ensure that no batch effect exists in the data. In case there is batch effect in data then it is critical to deal with it, generally an MDS plot is an easy and potent way to detect batch effect (Fig. 1.13). There are many software packages in R that provide functions for batch effect correction (Leek JT, 2021; Ritchie ME, 2021). However, it is important to keep in mind that these methods will lead to exaggerated confidence in downstream analysis if the batch-group design is unbalanced (Nygaard et al., 2016).

iii. Small number of replicates

For some of the proteomics experiments the number of replicates could be very small. For example, in the clinical setup often there is not enough sample to go for many replicates which can affect the statistical inference during the downstream data analysis when using traditional methods like t-test. Bayesian estimation for 2 groups comparison has been shown to supersede the classical t-test (Kruschke, 2013) and for small sample sizes the inference using Bayesian approach is far more stable.

iv. Challenges in visualizing protein interaction networks

High throughput proteomic experiment such as BioID often yields hundreds of protein interactions which are typically visualized as static network graphs. However, when such a static network contains large number of nodes and edges then interpreting it becomes extremely difficult. Also, the process of generating data structure compatible to network visualization software from the dataset is a tedious job. The choice of appropriate graph/network layout while visualizing the protein interaction network is another crucial factor to consider. The idea of network layouts emerged to address the challenges pertaining to generic graph visualization:

Given: A graph $\Gamma = (\beta, \varepsilon)$

Find: Legible and uncomplicated drawing of Γ

The force-directed (FD) network layout organizes nodes and edges in the graph in a unified and aesthetically pleasing manner. Therefore, it is well suited for visualizing large protein interaction networks. It works according to a force model which exploits Coulomb's law and Hooke's law to implement an energy-based node placement algorithm (Eades, 1984; Fruchterman and Reingold, 1991). In FD layout, nodes are represented as electrically charged particles in a Euclidean plane that repel each other whereas the edges bridging them attract adjacent particles mimicking a spring-force. The algorithm iteratively repositions nodes so that sum of all the forces become zero, pushing the system to attain an equilibrium.

While R packages (Leek et al., 2012; Gandrud, 2017; Ritchie et al., 2015) exist to address most of the aforementioned challenges independently, unavailability of a comprehensive data analysis pipeline (covering meaningful analysis and visualization) makes the data analysis task even more overwhelming, especially for an investigator with limited computational experience. Tyanova et al. provided a GUI named *Perseus* (Tyanova et al., 2016b) that offers a set of statistical approaches to help wet lab scientists. However, the resulting plots are not publishable and require additional graphics editing tools. A limitation of *Perseus* while performing t-test is that it does not compute fully moderated t statistics (See Tools sec. 2.1.1.4), which has been shown to outperform the standard t statistics (Kammers et al., 2015). Moreover, it is not possible to include batch information in the model during differential enrichment analysis in *Perseus*. Rather, it offers a two-step approach to deal with batch effects: First, the batch-effects are removed from the data and then the resulting data is used for t-test. But this two-step approach can lead to wrong conclusions in some cases (See 1.2.1.1). Another GUI, *PANDA-view* (Chang et al., 2018), written in Python, is quite similar to *Perseus* in terms of features and also suffers from the similar issues. Additionally, these desktop apps lack the ability to create

interactive network plots, which are critical for deducing biological implications from the data (See sec 3.1.2 in Results**)**. Therefore, to address these concerns we proposed an easy-to-use pipeline (See Tools sec. 2.1) that includes data pre-processing, stable statistical inference and, ability to visualize and interact with the protein interaction networks.

1.2.2 Computational methods for protein complex prediction

Researchers had employed sophisticated computational approaches to discover protein complexes. Using Drosophila melanogaster as a model organism, Guruharsha et al., had applied a hypergeometric distribution error model to score the PPIs before clustering them using Markov clustering algorithm (Guruharsha et al., 2011). Havugimana et al. used machine learning based approach to filter noise from protein interaction datasets in multiple steps prior to clustering using a soft clustering method (discussed later in this section) (Havugimana et al., 2012). Crozier et al., used machine learning approach proposed by Havugimana et al., to discover novel complexes in the parasite Trypanosoma brucei (Crozier et al., 2017). Heusel et al., employed protein correlation profiling to discover complexes in HEK293 cell line (Heusel et al., 2019). However, they did not use machine learning because the focus was on basic detection and quantification rather than de novo complex prediction.

It is interesting to note that there is no precise formula or set of rules that can be used to detect protein complexes from the complexomics datasets. Rather, it is more suitable to model the protein complex prediction problem as a pattern recognition challenge, where one uses labeled data to train an algorithm to capture regularities in the data. Machine learning, a branch of artificial intelligence, is an excellent fit for this task. Especially, if the aim is to discover novel protein complexes. As mentioned in sec. 1.1.2.1 (Fig. 1.8), here the computational challenge is to cluster the protein profiles (obtained using LC followed by bottom-up MS). Nevertheless, to cluster the data with extremely high sensitivity, the data must be treated with a series of

sophisticated computational steps that have shown promising results in previous attempts (Crozier et al., 2017; Webb-Robertson et al., 2015; Havugimana et al., 2012).

1.2.2.1 Noise modelling and missing value imputation

The quantitative protein/peptide matrix generated by DDA-MS has a high proportion of zeros and/or low spectral counts. While these zeros/low values frequently correlate extremely well with one another, they contribute little in building good predictors for machine learning. Havugimana et al. proposed adding noise to the data matrix artificially to address this issue (Havugimana et al., 2012). Researchers have approached missing value imputation in a variety of ways. Tyanova et al. proposed using a truncated normal distribution located near the lower tail of the original data distribution to generate random values (Tyanova et al., 2016b). Karpievitch et al. proposed a method for random selection based on the censoring probability calculated from the ANOVA model parameters (Karpievitch et al., 2012). Webb-Robertson et al. conducted a review of several imputation methods, which readers are encouraged to read (Webb-Robertson et al., 2015). Noise modelling step is not essential for SWATH-MS data. However, if there is some missing values then appropriate imputation algorithm should be employed before feature extraction step which is described next.

1.2.2.2 Extraction of PPI features from experimental dataset(s)

In order to predict protein complex, it is imperative to predict each binary interaction that comprise the complex. In other words, to solve protein complex prediction challenge one need to first solve the PPI prediction problem. One of the major problems while applying machine learning in this scenario is the need of *good features* that can discriminate between a PPI and non interaction. Some of the important features that can be extracted from data are described below.

i. Pearson correlation

An elution profile contains the quantified intensities of a protein in each fraction. A Pearson correlation can be computed for each pair of protein elution profiles, which serves as a feature for the PPIs. Pearson correlation (r) is defined mathematically as follows for the elution profiles of two proteins x and y:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (1.2)$$

Where, $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

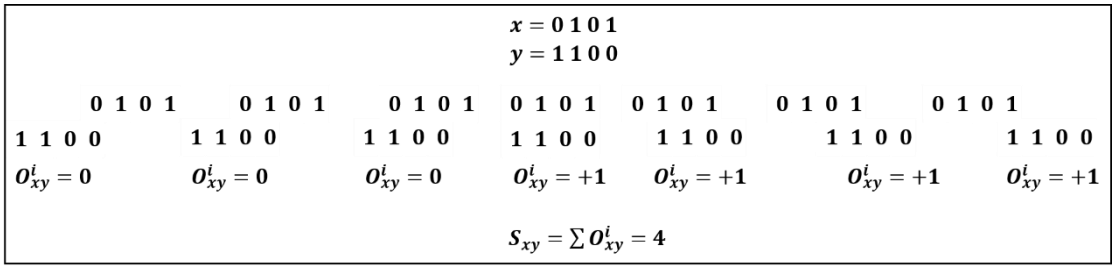


Fig. 1.14: Example showing similarity between two binary vectors using wcc- O^i_{xy} is the intermediate correlation between x and y [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

ii. Weighted cross correlation (wcc)

The Pearson correlation coefficient defined above does not consider changes in the elution profiles. The wcc can be used to compare the intensity of a protein within a single fraction to the values in another protein's chromatographic neighborhood. It is also possible to account for relative shifts in the elution profiles by weighing different fractions in the neighborhood. Havugimana et al. ranked the features they used in their exploratory analysis of protein complexes in human, and wcc was a top-ranking feature (Havugimana et al., 2012). The mathematical definition of similarity based on wcc is as follows: (de Gelder et al., 2001),

$$S_{xy} = \frac{\int W(\Delta) C_{xy}(\Delta) d\Delta}{\sqrt{\int W(\Delta) C_{xx}(\Delta) d\Delta \int W(\Delta) C_{yy}(\Delta) d\Delta}} \quad (1.3)$$

Where, Δ implies relative shift between the elution profiles

C_{xx} and C_{yy} are the auto correlation functions

W is the weighting function.

Fig. 1.14 demonstrates wcc using two sample vectors using weighing equal to one (Bodis, 2007).

iii. Co-apex score

It quantifies the degree to which the apexes of a pair of elution profiles align across replicates. Thus, the co-Apex score for a PPI is a ratio of the number of replicates in which the apexes of a pair of elution profiles align to the total number of replicate experiments. These techniques are frequently used to extract features from datasets (Phanse et al., 2016; Havugimana et al., 2012; Crozier et al., 2017).

1.2.2.3 Extraction of PPI features from the literature

Additionally, incorporating genomic and proteomic evidence from the literature can aid in the discovery of protein complexes (Havugimana et al., 2012; Webb-Robertson et al., 2015). Typically, these features are usually stored in a database as evidence codes for PPIs. STRING (Szklarczyk et al., 2018) and HumanNet (Hwang et al., 2019) are two widely used databases. HumanNet (v2) is the most recent version, and it contains a probabilistic gene network of 17,929 validated protein-coding genes spanning 525,537 interactions. HumanNet is constructed using a modified Bayesian integration of 21 different types of 'omics' data, with each data type weighted according to the degree of functional interaction between human genes. Each interaction in HumanNet is weighted according to a log likelihood score, which quantifies the probability that a functional interaction is true or false.

STRING has imported many PPIs from other databases in addition to the ones they predicted. The PPIs in STRING forms a Bayesian network, which means each link has a probability associated with it. The most recent version of STRING (v11) contains PPIs from 5090 organisms, totaling 3,091,648,416 edges. However, if only high-scoring (>0.7) interactions are

considered, that number drops to 152,484,793. Some of the frequently used features from the literature is described below.

i. Conserved Neighborhood

If the neighborhood of two genes encoding proteins is conserved across multiple genomes, there may be a possibility of functional interaction between the proteins. (Dandekar et al., 1998).

ii. Gene Fusion

The Rosetta Stone method has revealed that two or more interacting proteins in one genome can occasionally fuse into a single protein in another genome. (Marcotte et al., 1999).

iii. Co-Evolution

The concept here is interacting proteins tend to co-evolve. This feature has previously been used to predict PPIs in humans and bacteria. (Tillier and Charlebois, 2009; Pazos and Valencia, 2001).

iv. mRNA Co-Expression

It is well established that there is a positive correlation between mRNA expression patterns and interacting proteins, and that these patterns are frequently conserved throughout evolution. (Eisen et al., 1998; Marcotte et al., 1999; Lee et al., 2004).

v. Protein domain co-occurrence

It has been demonstrated that two proteins can interact via some of their shared/co-occurring domains. (Wang et al., 2011).

vi. Text mining

Text mining methods aims is to extract facts and relationships between entities from text. In this context, the term "entity" refers to proteins. It processes text in a variety of ways, one of the most important of which is natural language processing (NLP). The published literature is a treasure trove of PPI data. STRING mines PubMed for PPIs. One of the most significant

challenges in extracting PPIs from the literature is the enormous variety of possible gene names. Su(var)205 and HP1a, for example, should recognize the same named entity. Named-entity recognition is an active area of NLP research and in biomedical domain it poses even more challenges due to inconsistent usage of gene names in the literature. As a result, text mining is not as confident feature for a PPI as other features. Readers might find this introduction to biomedical NLP quite exciting (Cohen and Demner-Fushman, 2014).

1.2.2.4 Application of Machine learning to predict PPI

The Fig. 1.15 depicts a pipeline that utilizes ML for predicting PPI which is an important step towards solving protein complex detection problem. As previously described, ML is an appropriate tool for solving the current problem since it facilitates the detection of patterns even when no explicit rules describing those patterns exist. After extraction of features from the dataset, next step is to label the PPI.

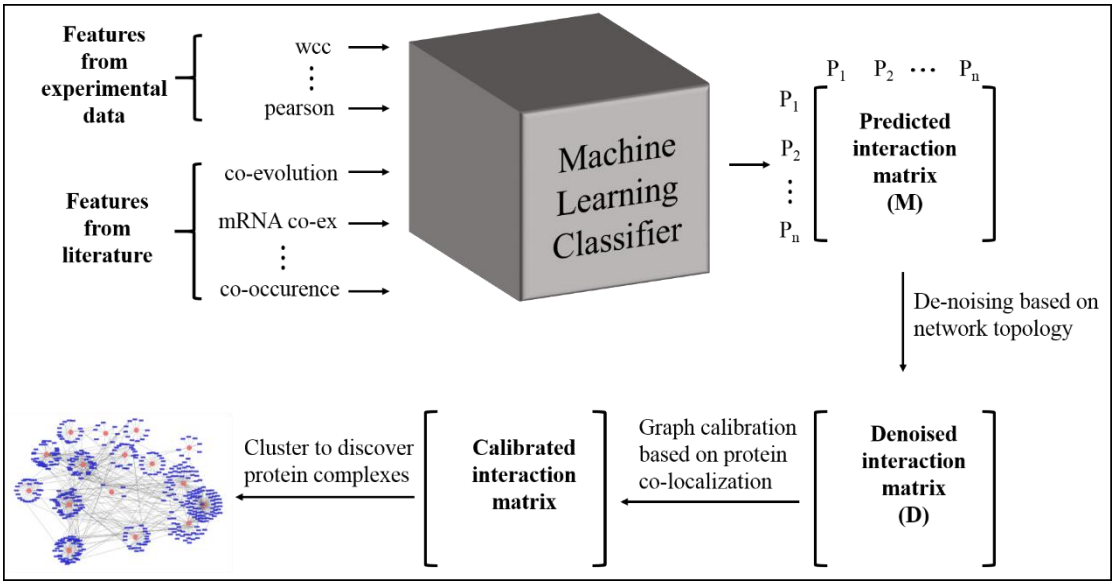


Fig. 1.15: Computational pipeline to discover putative protein complex using machine learning- The top panel depicts how the extracted features are used to train a machine learning model that is then used to predict the PPIs. The bottom panel demonstrates the multistage filtering of protein interactions followed by clustering of PPIs to detect of protein complexes. [Image adapted from: (Aftab and Imhof, 2021); Licensing information: Appendix B].

i. Ground truth

Researchers employed mainly supervised machine learning approaches for protein complex prediction (Havugimana et al., 2012; Crozier et al., 2017). In supervised machine learning, we must provide a labeled feature matrix to the machine learning model. In this context, a feature matrix is a collection of extracted features that represent either PPI or non interaction. Thus, each row in the matrix implies a PPI/non interaction, and each column implies a feature. The feature matrix is then divided into three parts: the training set, the test set, and the validation set. Then, for each PPI in the training and validation sets, we label it as interacting/non-interacting or, more frequently, as 1/0 or positive PPI/negative PPI. We provide no labels for the test set.

Readers will soon realize that this is the most challenging part of the machine learning-based complex prediction pipeline. The critical question here is: from where do we obtain true or positive PPIs? And how do you define a negative PPI? As you will see shortly, the latter is significantly more difficult to answer than the former.

Numerous databases contain true PPIs (positive PPIs in the present context). For instance, BioGrid (Oughtred et al., 2018) contains inferences about physical and genetic interactions based on numerous high-throughput experiments. STRING (Szklarczyk et al., 2018) is a collection of functional and physical interactions originating from a variety of species. The Centre for Cancer Systems Biology (CCSB) catalogs protein interactions from humans, yeast, viruses, and plants (Yu et al., 2011; Guo et al., 2008; Rolland et al., 2014). (Pagel et al., 2004) created the MIPS Mammalian Protein-Protein Interaction Database (MPPI) to store manually curated protein interactions from yeast and mammals. CORUM (Giurgiu et al., 2018) is a database of mammalian protein complexes that have been manually annotated. HumanNet (Hwang et al., 2019) is a database of human protein interactions. These already-identified PPIs are also referred to as ground truth in machine learning terminology. However, because the

majority of experiments are conducted solely to infer true PPIs, the resource for negative interaction is quite limited. Only 2171 protein pairs that are less likely to interact are cataloged in the Negatome Database 2.0 (Blohm et al., 2013). As a result, people developed heuristics to generate negative interactions. Havugimana et al. defined a negative interaction pair by associating two proteins that were previously annotated as being in two distinct complexes (Havugimana et al., 2012). Fig. 1.16 illustrates this concept, where negative PPI implies a pair of proteins that are less likely to interact. As you may have guessed, the validity of this heuristic is contingent upon previously identified interactions, and the absence of an interaction does not imply a negative interaction. In another study, Crozier et al. defined negative pairs (also known as negative PPIs) by randomly selecting a pair of proteins from the set of proteins present in their ground truth dataset (Crozier et al., 2017). This heuristic also has the potential to introduce false negative interactions. Due to the possibility that some of the randomly sampled PPIs are in fact true interactions.

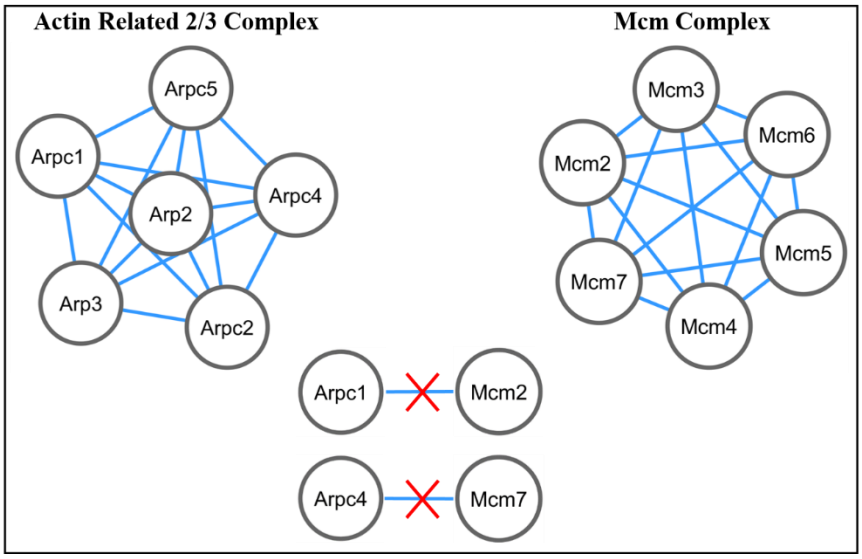


Fig. 1.16: A heuristic approach towards negative PPI generation [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

ii. Classification Algorithm

Numerous researchers have used random forests to solve the PPI classification problem (Crozier et al., 2017; Havugimana et al., 2012). Additionally, some investigators have made use of support vector machines (Webb-Robertson et al., 2015; Guo et al., 2008).. Recently, as deep learning has gained popularity, many researchers have turned to deep neural network-based architectures to solve this problem (Sun et al., 2017; Tian et al., 2016; Hashemifar et al., 2018). The advantage of these deep architectures is that the classifiers automatically select features. However, these classifiers require a large amount of training data to produce accurate results. My recommendation is to start with a simple classifier and gradually increase its complexity until the desired performance is achieved.

iii. Feature Selection

It is critical to comprehend the significance of features in machine learning classification. Using irrelevant features may result in longer training times and may cause the model to become overfit (Bermingham et al., 2015). Researchers have employed innovative ways for improving classification accuracy through feature selection. Before training a random forest classifier for PPI interaction prediction, Havugimana et al. used a greedy stepwise feature selection algorithm (Havugimana et al., 2012). There are numerous additional methods for selecting features, and while a discussion of them is beyond the scope of this chapter. Readers can find detailed information about them in this reference (Guyon and Elisseeff, 2003).

iv. Training a ML classifier

While choosing machine learning to solve a problem, the first thing one should determine is the type of training. There are two types of training that are generally available: supervised and unsupervised. Due to the fact that the supervised approach is the most frequently used method for solving the current problem, we will confine our discussion to it.

One can use the training set to train a single machine learning algorithm or an ensemble of algorithms. Training a machine learning classifier includes fine-tuning a number of parameters,

which is beyond the scope of this thesis to explore in depth. However, the primary goal of training for the PPI classification task is to minimize classification error/loss, which is a critical parameter. A type of error that many machine learning algorithms attempt to minimize is sum squared error (SSE), which is defined as,

$$SSE = \sum(y_i - \hat{y_i})^2 \quad (1.4)$$

Where, $y_i \rightarrow$ actual label of a PPI and $\hat{y_i} \rightarrow$ predicted label for that PPI. We check the SSE in each training epoch and continue training until the SSE falls below a predefined threshold after a predetermined number of epochs. If the SSE is unable to reach the desired threshold despite extensive training, the model is said to be underfit.

Once the SSE is sufficiently small, we turn our attention to model validation. Here, we examine the SSE using the validation dataset. If the difference between the training and validation errors is too large, the model is said to be overfit.

Overfitting and underfitting are two major issues that must be addressed during the training phase of any machine learning algorithm. As a result, a cycle of training-validation-training is frequently conducted. We can be confident in the training only when the training error is small and the gap between the training error and the validation error is also small. Only after we are sufficiently confident in our training should we proceed to testing.

When you feed feature vectors corresponding to a set of unlabeled PPIs (test dataset) into the classifier during testing, the classifier outputs a probabilistic score for each PPI. If the score for a PPI is zero or close to zero, we consider it to be a false interaction; however, if the score is close to one or one, we consider it to be a true PPI.

v. Imbalanced Dataset Problem

When the number of samples in one class is much greater than the number of samples in other, the dataset is said to be imbalanced. Assume that a proteomics experiment identifies and

quantifies *n* proteins from a sample of a particular species. Then all possible pairs of PPIs that those *n* proteins could generate are,

$$n_{C_2} = \frac{n!}{2!(n-2)!} \quad (1.5)$$

Plugging in n = 3000, a realistic expectation for today's mass spectrometers, results in 4.5 million theoretical PPIs, of which only a few will be labeled as positive interactions. Most public databases that host true PPIs, such as BioGrid (Oughtred et al., 2018), contain ~0.5 million PPIs for Homo sapiens, ~76 thousand for Drosophila melanogaster, and ~30 thousand for Mus musculus. As a result, the labeled dataset (after feature extraction) will contain many more false interactions than true ones. Class imbalance is a significant issue, because if a machine learning-based binary classifier is trained without addressing the class imbalance, the model will be completely biased toward the class with the most samples. The Synthetic Minority Oversampling Technique (SMOTE) has been used by machine learning researchers to attempt to address this issue (Chawla et al., 2002). Others have attempted to train a classifier ensemble using an equal number of negative and positive PPIs (Crozier et al., 2017). Nonetheless, it is critical to recognize that class imbalance is inherent in the PPI prediction problem and must be effectively addressed.

1.2.2.5 Denoising a predicted PPI matrix

The amount of noise in the predicted interaction matrix can be further reduced by removing the edges lacking support from the network topology. Havugimana et al. (Havugimana et al., 2012) determined connectivity using a multi-step diffusion procedure, which can be defined mathematically as,

$$C = e^{\lambda M} - \lambda * M \quad (1.6)$$

Where, M→ predicted PPI matrix

λ→ Inverse of the largest eigenvalue of M.

In order to denoise the PPI interaction matrix, edges with connectivity less than a threshold value τ are deleted from the matrix M. Let us refer to this denoised matrix as $M^{\dagger}$, which can then be calibrated based on the information about protein co-localization in the sample. Here the idea is to penalize any PPI that results from two proteins located in different cellular compartments by computing a score using the PPI prediction scores in $M^{\dagger}$ and the GO-CC scores (Consortium, 2004). Therefore, the combined score matrix (R) is defined mathematically as,

$$R = (1 - M^{\dagger}) * (1 - \frac{S}{S_{max}}) \quad (1.7)$$

Where, S → maximum pairwise similarities matrix, each cell (S_{ij}) of which is the maximum pairwise similarity between the two groups of GO-CC terms to which protein i and protein j are annotated.

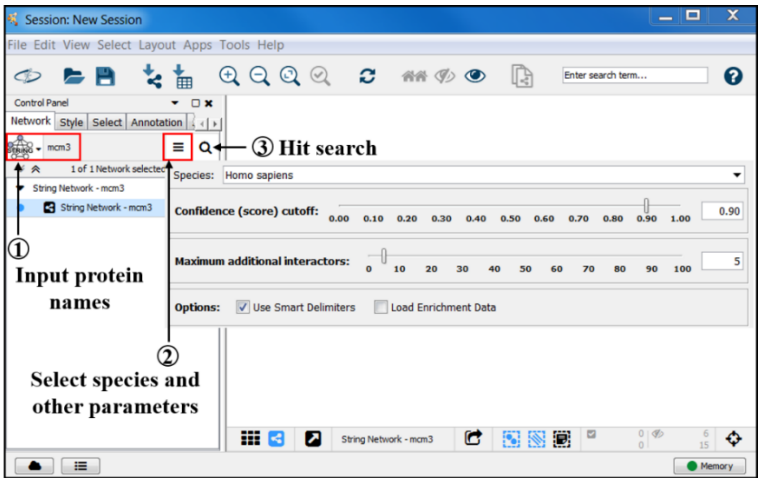
S_{max} → A normalizing factor can be used as the maximum value among all the semantic similarity scores.

In order to get a thorough understanding of the scoring presented in Eqn. (1.7) readers are encouraged to read this article (Yang et al., 2012).

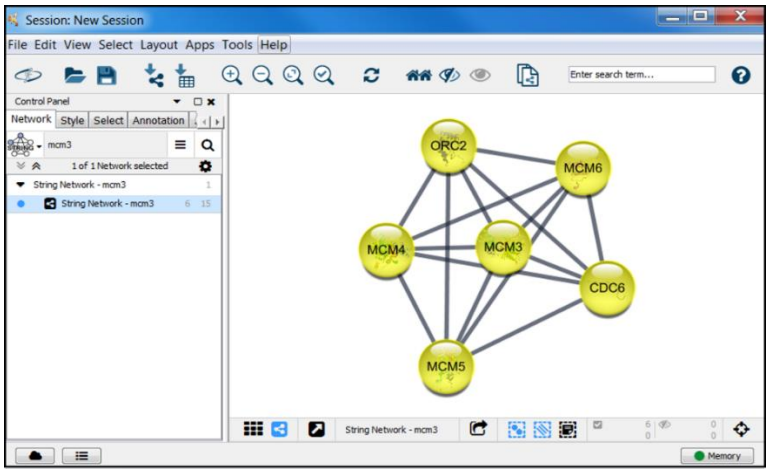
1.2.2.6 Cluster the denoised PPI matrix

To identify the densely connected areas in the denoised PPI matrix, we need to cluster the PPIs. There are two main types of clustering methods: hard clustering and soft clustering. A data point is never assigned to more than one cluster when using hard clustering. The K-means clustering algorithm is a widely used hard clustering algorithm. Soft clustering, on the other hand, assigns a score (membership probability) to each data point, allowing it to belong to several different clusters. Soft clustering methods are more appropriate in our situation because a protein can be found in a variety of different complexes at the same time. In other words, it is possible that protein complexes will overlap. Shuye and colleagues developed a comprehensive CYC2008 catalogue, which contains 408 heteromeric protein complexes in S.

cerevisiae that were manually curated (Pu et al., 2008). 207 proteins out of 1628 proteins in CYC2008 are found to be involved in multiple complexes (Yang et al., 2012). This clearly indicates that we require a clustering algorithm that allows a protein to participate in multiple clusters. ClusterONE (Yang et al., 2012), developed by Nepusz et al., addresses these issues by detecting overlapping protein complexes from PPI datasets. ClusterONE has been widely used in a number of studies with the goal of detecting protein complexes from PPI data (Havugimana et al., 2012; Crozier et al., 2017; Webb-Robertson et al., 2015).



(A)



(B)

Fig. 1.17: stringApp can be used to import PPIs from STRING into the Cytoscape environment- (A) From a Cytoscape session, search interaction partners of a protein (mcm3 in this example) in the STRING database **(B)** The extracted MCM complex from STRING

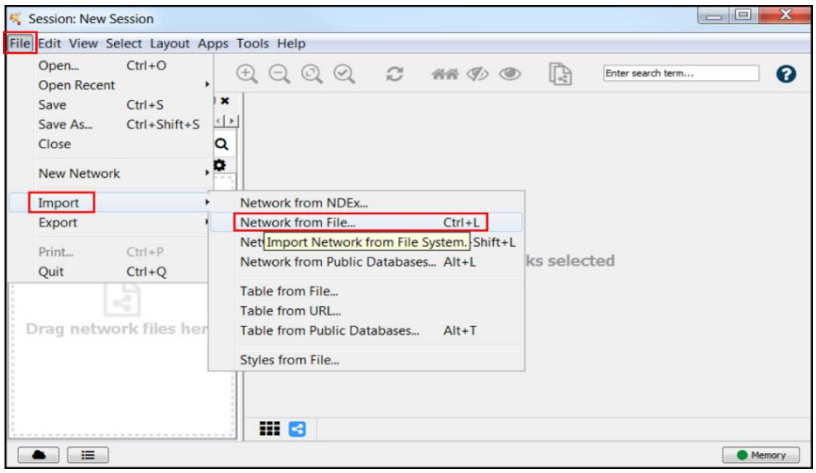
database becomes available in the current Cytoscape session via stringApp [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

1.2.2.7 Network analysis using Cytoscape

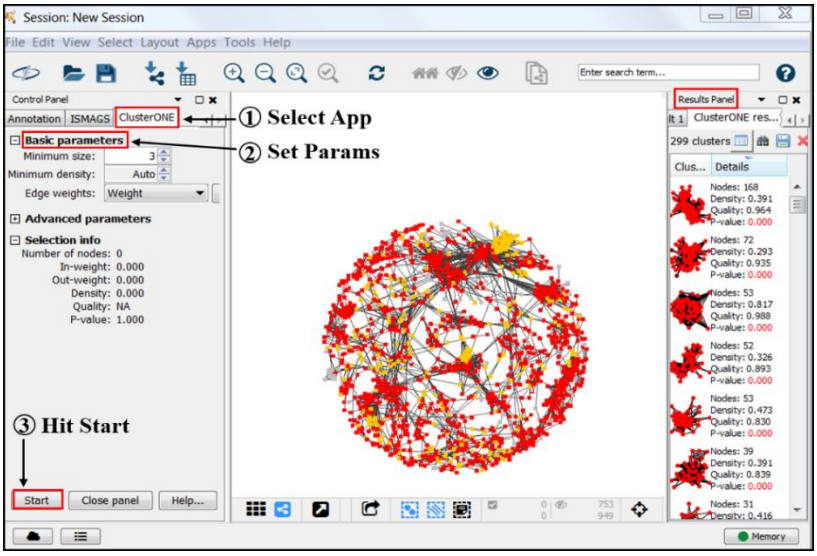
Cytoscape is an open-source software for visualizing and analyzing data in the form of a two-dimensional matrix (Shannon et al., 2003). It was originally intended for knowledge discovery from biological experiment datasets, but over time, users added plugins to solve problems in other domains such as social network analysis, semantic web, and so on. Cytoscape's most recent version can be downloaded from the download page (https://cytoscape.org/download.html). It is an excellent workbench for network data analysis, with over 200 plugins. Additionally, it is possible to integrate PPIs from the user's own laboratory measurements with PPIs from multiple databases. For instance, users can import PPI from STRING DB and augment it in Cytoscape using stringApp, as illustrated in Fig. 1.17. Cytoscape also includes an application for the ClusterONE algorithm discussed previously. In Fig. 1.18, we demonstrate how to use it in Cytoscape. Readers are encouraged to read the tutorials (https://github.com/cytoscape/cytoscape-tutorials/wiki) that have been officially released to get started with the Cytoscape.

Researchers mainly employed DDA based proteomic approaches (except Heusel et al. (Heusel et al., 2019)) for protein complex prediction. However, as mentioned earlier, DDA based shotgun proteomics results in datasets with many missing values which induce uncertainties in the quantified proteins. Moreover, the approaches of Guruharsha et al. (Guruharsha et al., 2011) and Havugimana et al. (Havugimana et al., 2012) necessitate undertaking either a significant number of IP or LC (orthogonal modalities) experiments. This results in a significant increase in MS instrument usage time, which creates a bottleneck in a multiuser environment. Moreover, not each published computation method is available for the user to analyze his/her data. Among the previously mentioned computational approaches, only Heusel et al. (Heusel et al., 2019)

and Skinnider et al. (Skinnider M, 2021) has provided codes, but they are available as MATLAB scripts/R packages/command line interfaces (CLIs) with no or little GUI support. Since most of the scientists in the proteomics community are researchers having mainly wet lab expertise, a full-fledged GUI is always preferred over CLIs. Therefore, to fill this gap and to enable an investigator to discover bona fide as well as novel protein complexes from a relatively small number of datasets, I developed a desktop application ComplexMiner (Wasim Aftab)(Chapter 2).



(A)



(B)

Fig. 1.18: Cytoscape ClusterONE app- (A) Loading a network file **(B)** Running ClusterONE [Image source: (Aftab and Imhof, 2021); Licensing information: Appendix B].

1.2.3 Computational methods for spatial proteomics

1.2.3.1 IMS and LCMS data integration

The combination of the aforementioned (sec 1.1.3.1) orthogonal MS technologies have led from poor to substantial identification of proteins in a contextual manner. The main reason for this variability was the lack of appropriate strategies that could effectively combine data from these two platforms into an efficient screening module of proteins in situ. Although these studies could successfully identify important disease associated molecules, most of the attempts involved considerable manual curation leading to very limited number of identified discriminative masses (Alberts et al., 2018; Longuespée et al., 2019).

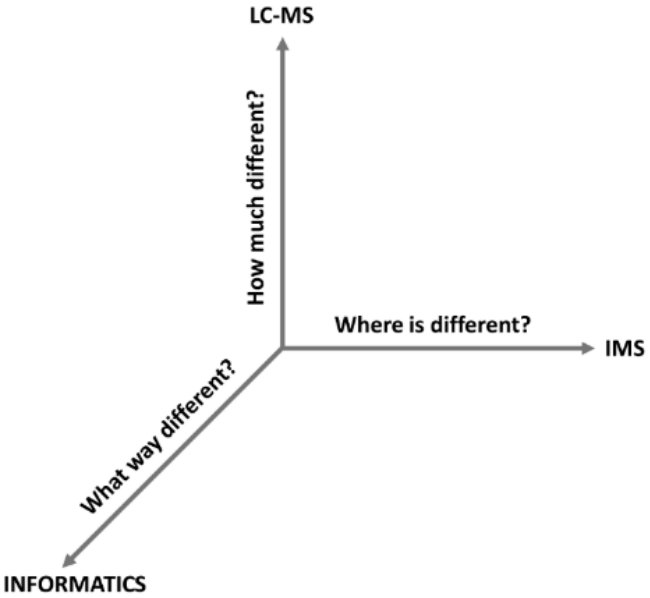


Fig. 1.19: Rationale behind IMS and LC-MS data integration- from IMS experiment we determine the differential spatial regions between the diseased and healthy tissues; by performing LC-MS experiment we quantify the differences; informatics allows us to demonstrate what way they are different by combining the information from the two orthogonal MS technologies.

More successful approaches were associated with measurements of in situ tryptic peptides with very high mass accuracy (comparable to LC-MS) leading to the analysis method being

particularly resource intensive (Schober et al., 2012; Huber et al., 2018). In addition, these approaches were exclusive of the two-group comparison scenario (healthy vs. diseased) thereby providing very limited biological insights as part of a data analysis pipeline. None of the approaches developed so far has a defined integrated 'one-in-all' workflow/software leading to the tedious task of combining multiple platforms with substantial manual input.

To address these concerns, we introduce ImShot, a conveniently designed software that can be deployed as a screen for probabilistic identifications of proteins in situ in a disease vs healthy context. ImShot is based on a systematic matching of spatially resolved peptide masses from MALDI-IMS with their corresponding identified proteins in LC-MS. The rationale behind IMS and LC-MS data integration in ImShot is depicted in Fig. 1.19. We provide this bioinformatics pipeline in the form of a desktop application built using web technologies on Electron framework (Electron, 2013) aimed at seamless plug-n-play operation. Electron, which is an open-source software framework, streamlines the development process with the help of a web environment. It builds an application that feels native and at the same time looks aesthetically pleasing. Conventional desktop apps in comparison appear outdated as interface elements, charts and plots are not as captivating as in an Electron application. In addition, as Electron apps run on a browser, thereby allowing it to render interactive graphics. This allows us to observe unforeseen patterns and trends in otherwise hidden information within the data. Although in recent years Electron apps have gained large popularity in other domains (OpenJS-Foundation, 2021), to the best of our knowledge, no proteomic data analysis platform has used it so far in spite of the need for convenient GUI and graceful visualization of data. While developing ImShot, we tried to establish its applicability as a general tool that would be useful in different situations and to different users. It is largely independent of the measurement platforms as the algorithm depends on general data format instead of proprietary ones. In the

current version it can integrate shotgun proteomics and MALDI-IMS datasets to address questions involving two group comparisons, like diseased vs healthy or high-calorie vs low-calorie diets etc. In addition, the software can operate in dual mode, i.e., its functions and features can individually be used to solve research problems in shotgun proteomics alone. In addition, we also provide an R package to facilitate ImShot's command line mode of operation. The ImShot desktop app and R package can be installed and run on all major operating systems (Windows, Linux and macOS).

1.3 Aims of the thesis

The inundation of proteomic data over the last decade has given rise to a multitude of computational challenges and the methods to solve them have progressed along with proteomics technologies (Srihari et al., 2017). My objective in this thesis is to familiarize readers with some of these challenges with a particular emphasis on cutting-edge computational methods.

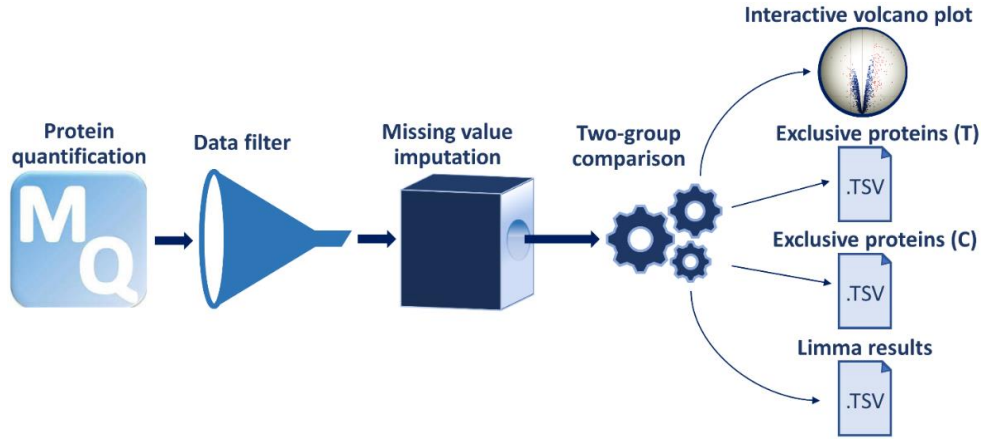
Today's shotgun proteomic experiments aimed at capturing biological differences between two groups (viz. diseased/healthy) generate enormous, multi-dimensional datasets comprising information about biological, technical, and experimental variables. Due to the lack of a comprehensive data analysis pipeline, it has been incredibly difficult to derive insight from these complex datasets. Therefore, an easy-to-use pipeline is developed to facilitate differential enrichment analysis and effective visualization of massive protein interaction data.

Almost every cellular process is executed by protein complexes. To understand how the cellular machinery works, it is critical to identify and characterize all protein assemblies. Due to the lack of precise deconvolution rules, discovering novel protein complexes is a challenging task. In addition, due to the absence of an effective GUI (capable of comprehensive analysis), exploring complexomics datasets has been frequently intimidating. As a result, a desktop application called *ComplexMiner* is proposed.

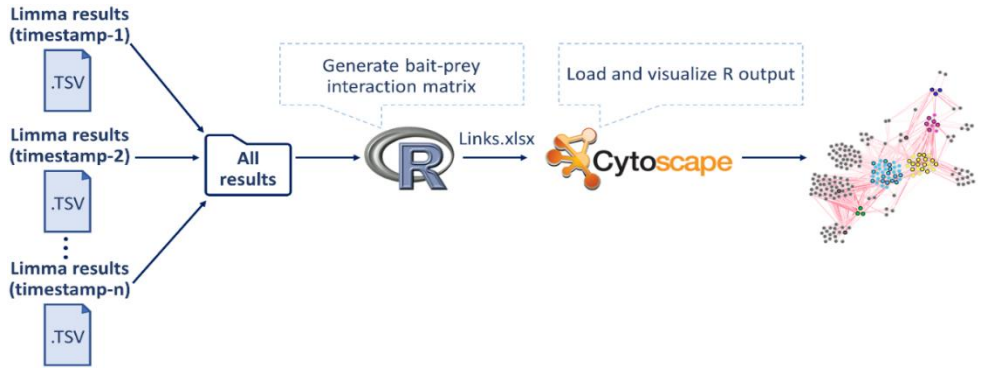
Most tissues, where the diagnostic information of diseases is intact, are very complex and heterogeneous with respect to variety of cell types present. Understanding spatial protein profiles has therefore become imperative for better interpretation of disease effects and possible mechanisms. While IMS can conserve the spatial distribution of molecular species on tissue, the inability to identify peptides directly (in situ) has been a limitation of MALDI-IMS based spatial proteomics. To overcome this, ImShot, a user-friendly desktop application that combines information from MALDI-IMS and LC-MS data, is designed.

2. Tools

In this chapter, I will describe the tools that I have developed to tackle the computational challenges mentioned in the Introduction.



(A)



(B)

Fig. 2.1: Limma pipeline to analyze proteomic datasets- (A) Illustrating the BioID proteomics data analysis pipeline using results from MQ search as input. It has different modules for data cleaning and detecting significant changes in protein abundances in two groups. Output from pipeline comes in the form of an interactive volcano plot and tab-separated values (TSV) files listing exclusive proteins in treatment, control, and results of the two-group comparison. [Image adapted from (Salvatori et al., 2020a); Licensed under CC BY 4.0] **(B)** Automating the generation of bait-prey interaction network (from the Limma results) using R and visualizing it in the Cytoscape.

2.1 Computational pipeline for differential enrichment analysis and effective visualization of high throughput proteomics datasets

We present the data analysis pipeline in the light of BioID proteomics approach, but it is also applicable to any proteomic experiment that aims to capture the biological differences in two groups (viz. diseased/healthy). To detect significant changes in protein abundance, we need to perform two-group comparisons. For this purpose, I have developed a data analysis pipeline (Fig. 2.1A) in R that employs moderated t-test (described later) to perform differential enrichment analysis. But first, we must filter and prepare the data for statistical analysis via a number of pre-processing steps (Fig. 2.1A) that are elaborated in the next section.

2.1.1 Statistical quantification and generation of volcano plots

Input to the pipeline is the proteingroups.txt file obtained from MaxQuant (Tyanova et al., 2016a) search. The statistical analysis comprises of the following steps:

2.1.1.1 Data filter

The data cleaning is done in the following four steps:

i. First, the contaminant proteins are removed from the dataset. To do so, code investigates the columns: *Reverse*, *Potential contaminant* and *Only identified by site* and removes all rows that contains the symbol +.

ii. Then the script finds and removes any blank proteins, i.e., bait and control replicates have all zeros for such proteins.

iii. Then it finds the proteins for which either all control replicates or all replicates in treatment group has only zeros. It extracts such ‘exclusively enriched’ proteins from the dataset and puts them into separate files as they do not participate into two-group comparisons. The reason for excluding such proteins is as follows: In a two-group comparison, the null hypothesis (H0) is that the replicate means of the proteins in each group are equal and we need statistical tests to see if this is indeed true, when

H0 is not true we reject it and accept the alternate hypothesis (H1). But in the case of 'exclusively enriched' proteins, no tests needed as all the proteins in one group has zero and in other group has non-zero means.

iv. Code further removes the proteins based on **k out of N criteria:** where N is the number of replicates in one group (treatment/control) and for each protein, k implies the desired number of non-zero values out of N replicates. The code applies this criterion individually for each group and keeps only those proteins that satisfy it in both the groups simultaneously.

2.1.1.2 Log transformation and missing value imputation

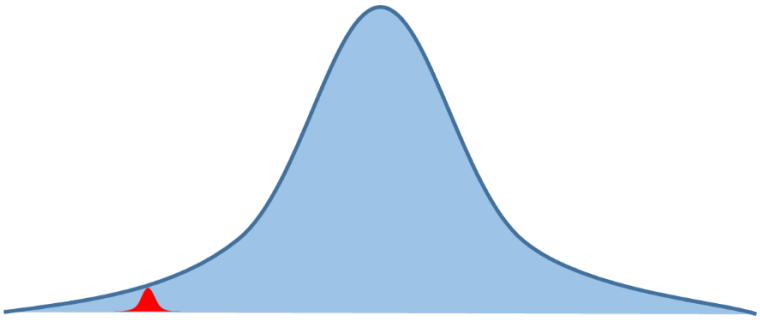


Fig. 2.2: Concept of missing value imputation algorithm - Original data distribution after $\log_2$ transformation is shown in blue and the tiny normal distribution (obtained by shifting and shrinking the original distribution) from where missing values are drawn randomly is shown in red.

Then the script log transforms the filtered dataset so that the outcome is normally distributed, and statistical tests become applicable. However, this transformation produces many undefined values (NaNs) also popularly known as missing values. We used the imputation algorithm proposed in (Tyanova et al., 2016b) because it allows us to randomly draw values from a distribution meant to simulate values below the detection limit of the MS instrument. To achieve that, the script creates a tiny normal distribution by shrinking and downshifting the original data distribution and then imputes the missing values randomly from that tiny normal distribution as depicted in Fig. 2.2.

2.1.1.3 Normalization

While running the pipeline, users will be asked if they want to normalize data prior to two-group comparison. There are two modes of normalization supported.

Normalize by subtracting median: This method normalizes the protein intensities in each experiment by subtracting the median of the corresponding experiment.

Column wise median normalization of the data matrix: Assume the protein quantification matrix is called *y*. Then, this approach can be implemented using following three lines of R code:

```
row_avg <- rowMeans(y)
y3 <- matrixStats::colMedians(y - matlab::repmat(row_avg, 1, ncol(y)))
y4 <- y - matlab::repmat(y3, nrow(y), 1)
```

In words, it first subtracts from the columns (experiments) of *y*, the row average of *y*. Then computes median of each experiment (column wise) and saves the results in another variable *y3*. Finally, the normalized matrix *y4* is obtain by subtracting *y3* from *y*. This procedure ignores missing values and assumes that the bulk of rows remained unchanged.

After the pipeline executes successfully, it will create volcano plot in html format and will also save 'exclusively enriched' proteins (if any) with corresponding LFQ/iBAQ values.

2.1.1.4 Two-group comparison (H0: means are equal)

In order to determine the proteins that are statistically significant in two groups, our pipeline employs Limma (Linear Models for Microarray Data) moderated t-test statistics as proposed in (Kammers et al., 2015) over standard t test. Limma was originally developed to find differentially enriched genes in microarray-based experiments and since many years it is a state of the art to analyze data from gene expression experiments such as RNA-seq. It employs empirical Bayes approach that uses the entire dataset to shrink the estimated sample variances for each gene towards a pooled estimate (Lönnstedt and Speed, 2002; Smyth, 2004). This

statistical approach results in much more stable and powerful inference compared to ordinary t statistics mainly when the number of replicates is small (Smyth, 2004; Yu et al., 2011). Very often proteomics datasets come with small replicates/sample sizes (See sec. 1.2.1.1) where such Bayesian treatment is appropriate and has therefore gained some popularity within the proteomics community over time (Brusniak et al., 2008; Salvatori et al., 2020b; Schwammle et al., 2013; Ting et al., 2009; van Ooijen et al., 2018). It comprises of following two steps,

i. First, a code module fits multiple linear models for every protein.

ii. Then another module uses that linear model fit information and by employing an empirical Bayes method it computes the moderated t-statistics, which results in shrinkage of a protein's variance towards a pooled estimate, thereby providing stable inference.

The two-group comparison module runs in two modes: either use full data or remove exclusive proteins before Limma analysis. Code will ask you to provide treatment and control names. It will guide by printing the instructions on your RStudio console.

2.1.2 Automation of bait-prey interaction network generation

For each two-group comparison, the pipeline (Fig. 2.1A) saves the moderated t-test results in a TSV file (with a timestamp in the filename). Therefore, when you have multiple baits/treatments, you'll end up creating many such timestamped files. The bait-prey interaction information is hidden within these files. So, to avoid time consuming and error prone manual data mining we have extended the pipeline of Fig. 2.1A to include automatic generation of bait-prey interaction network table as illustrated in Fig. 2.1B. The Rscript that automates this process, accepts the path of the folder containing multiple timestamped TSV files as input. It asks users first, to specify a log fold change cutoff. Then, for every TSV file in that folder, it will print a list of iBAQ/LFQ column names and ask the user to enter a bait name from that list. Here assumption is that iBAQ/LFQ columns will contain bait names. Using that

information, the code will create a virtual list of bait-prey interactions that have fold-change ≥ 1.5 and $p < 0.05$. After that, the code will print a message if that file is processed successfully. Finally, when all TSV files are processed, code will concatenate those virtual lists into one, print top 10 rows that list and save all the proximity interactions as *Links.xlsx* in the same directory where the TSV files are present. Users can then visualize the bait-prey interaction network by loading the excel file in Cytoscape (Shannon et al., 2003), an open-source software for visualizing complex networks and integrating them with information from other sources. However, the network graphs are rendered in static manner which is problematic for large protein interaction network (See sec. 1.2.1.1).

2.1.3 Effective visualization of large protein interaction networks

To deal with the data visualization challenge mentioned in the Introduction (Sec. 1.2.1.1), we developed a network visualization routine using R and JavaScript (JS) that allow users to interact with the network and extract information from it. The software was built by modifying the source code of networkD3 R package (Gandrud, 2017) which lacked certain desired functionalities (Discussed later in sections 3.1.2 and 4.1). We created an improved HTML widget forceNetwork ++ by significantly modifying the source codes of forceNetwork and associated JS subroutines from networkD3 package. We then used forceNetwork++ to develop a dashboard called *MiGENet* (Singh et al., 2020) which served as a platform for the user to interact with large protein interaction network.

2.2 Computational approaches to discover protein complexes

We employed Superose® 6 SEC followed by SWATH-MS on Drosophila embryonic extract and quantified ~1400 proteins spanning across 42 fractions. The dataset is shown in Fig. 2.3 where the protein elution profiles (in rows) are clustered based on hierarchical clustering method to visualize the presence of clusters (sanity check).

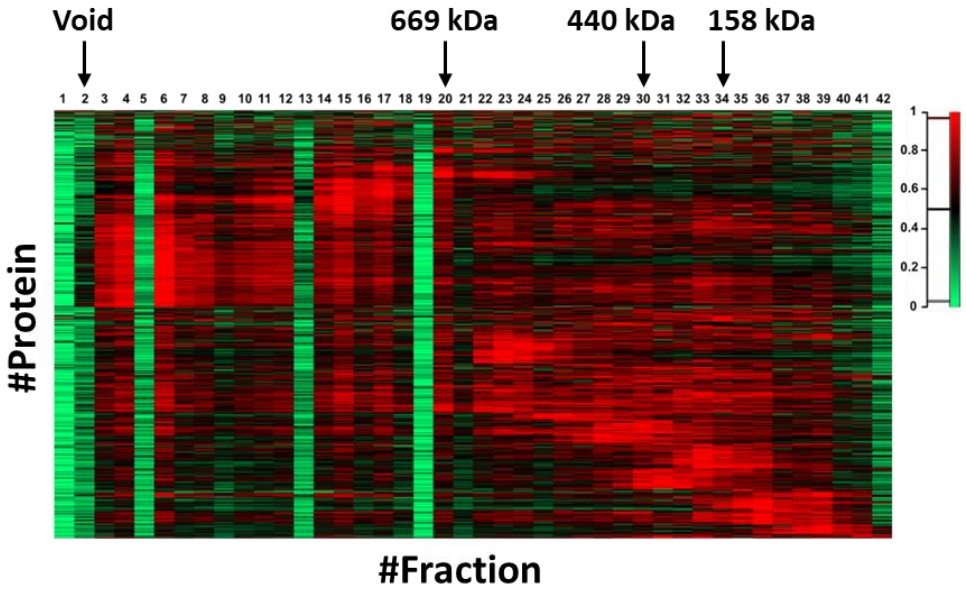


Fig. 2.3: Superose® 6 size exclusion chromatography dataset- Contains ~1400 protein elution profiles spanning over 42 fractions. Min-max normalization was applied prior to hierarchical clustering of elution profiles. Elution of standard proteins are marked with arrows.

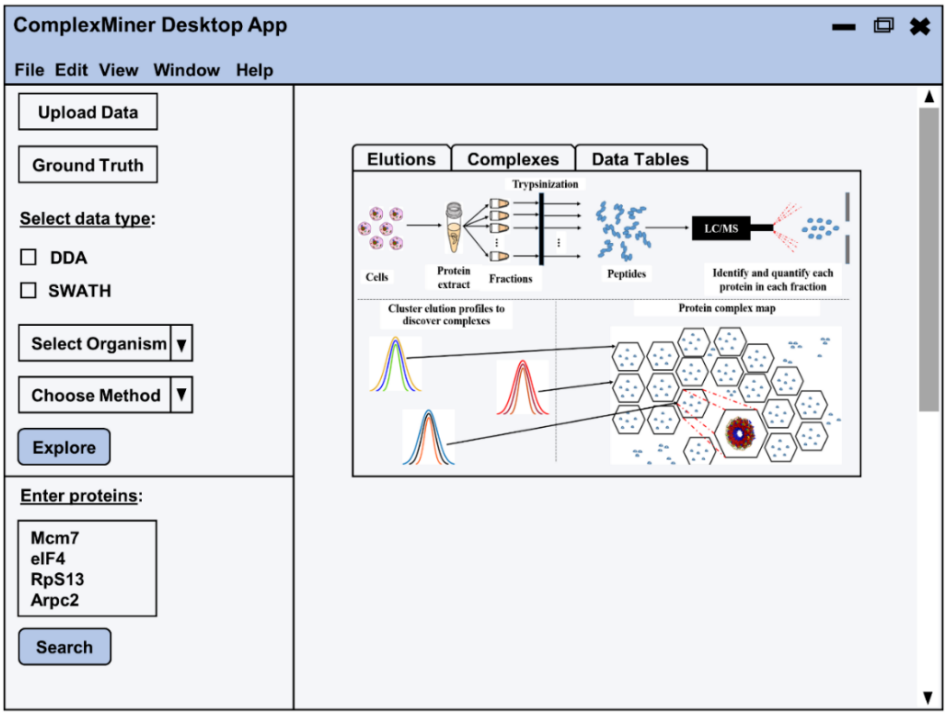


Fig. 2.4: A wireframe of ComplexMiner desktop application

As we mentioned in the introduction that *ComplexMiner* (written using Python, R, JS, HTML and CSS) is conceived to help users to analyze the complexomics datasets. However, *ComplexMiner* is not fully developed yet but initial testing has shown some promising result

(See sec. 3.2). It is an easy-to-use desktop application for protein complex discovery aimed at providing the bench scientist a support in analyzing complexomic datasets (See wireframe in Fig. 2.4). Here, the complex discovery is governed by the Siamese neural network (SNN) architecture depicted in Fig. 2.5. A Siamese network employs two or more identical subnetworks with the same architecture, parameters, and weights. Two subnetworks with the purpose of extracting features accept as input a pair of elution profiles (X1, X2 in Fig. 2.5). Then the difference between the extracted features (F1, F2 in Fig. 2.5) is feed to a fully connected network which acts as a binary classifier. The output from the classifier is converted to a probability (between 0 and 1) value with the help of sigmoid scaling operation. Finally, the scaled value is feed to a layer that computes training loss (L) which is proportional to the binary cross-entropy between the between the predicted and the true label values:

$$L = -y * log(\hat{y}) - (1 - y) * log(1 - \hat{y}) \quad (2.1)$$

Where, y, $\hat{y}$ imply the true and predicted labels respectively.

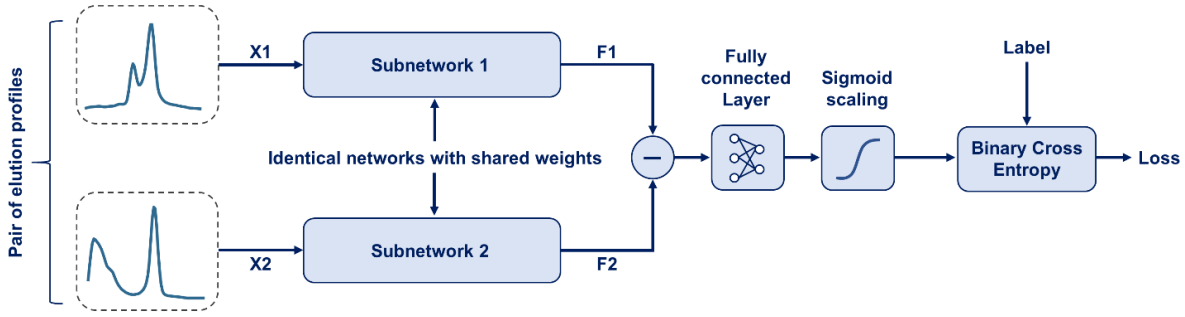


Fig. 2.5: Architecture of a Convolutional neural network to solve PPI classification task- X1, X2 are the two elution profiles that are passed through two subnetworks sharing identical weights yielding the corresponding feature vectors F1, F2. [Image adapted from https://www.mathworks.com/help/deeplearning/ug/train-a-siamese-network-to-compare-images.html]

While developing *ComplexMiner*, I programmed a MATLAB command line software, *CoreClust*. It was developed with the goal of rapidly discovering protein complexes from a

single dataset. Although it is less sensitive than machine learning-based methods, it can quickly identify bona fide protein complexes and hence can be used as a sanity check module alongside ML-based approaches. The *CoreClust* algorithm is described below.

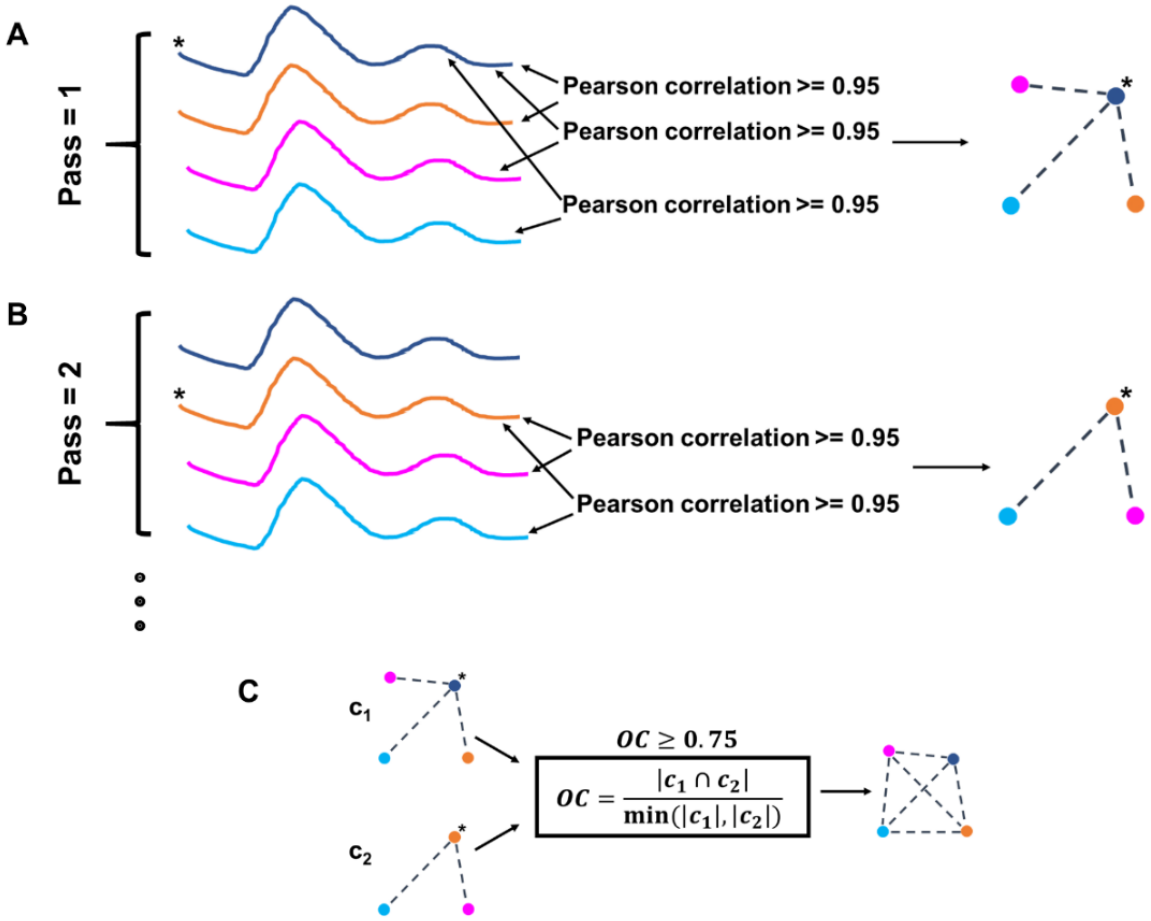


Fig. 2.6: Concept of *CoreClust* algorithm- (A-B) Shows the cluster exploration phase, only first steps of two passes of the cluster exploration phase is shown. **(C)** Shows the cluster fusion phase.

2.2.1 CoreClust algorithm

It works in the following two steps:

Step1: Given *n* elution profiles, this step comprises of *n-1* passes. In each pass *p* (*p* = 1, 2, …n-1) it selects p^{th} elution profile and computes pairwise Pearson correlation with the other *n-p* profiles and clusters the elution profiles that have high correlation (>= 0.95) with the p^{th} profile (Fig. 2.6A-B). However, this step creates clusters that have many redundant entries. To

understand this issue, consider p^{th} profile has high correlation with k other profiles that are part of cluster C_p and consider that $(p+1)^{th}$ profile has high correlation with m other profiles that are part of cluster C_{p+1}. The clusters C_p and C_{p+1} will have at least k elements in common. To reduce this redundancy, *CoreClust* algorithm fuses highly redundant clusters in step 2.

Step 2: In this step, *CoreClust* generates a vector containing overlap coefficient (OC) for each pair of complexes then, iteratively fuses clusters with high overlap (OC >= 0.75) as depicted in Fig. 2.6C. OC is defined mathematically as follows,

$$OC = \frac{|C_1 \cap C_2|}{\min(|C_1|, |C_2|)} \tag{2.2}$$

2.3 Integrating shotgun proteomics and MALDI-IMS datasets to directly identify proteins in situ

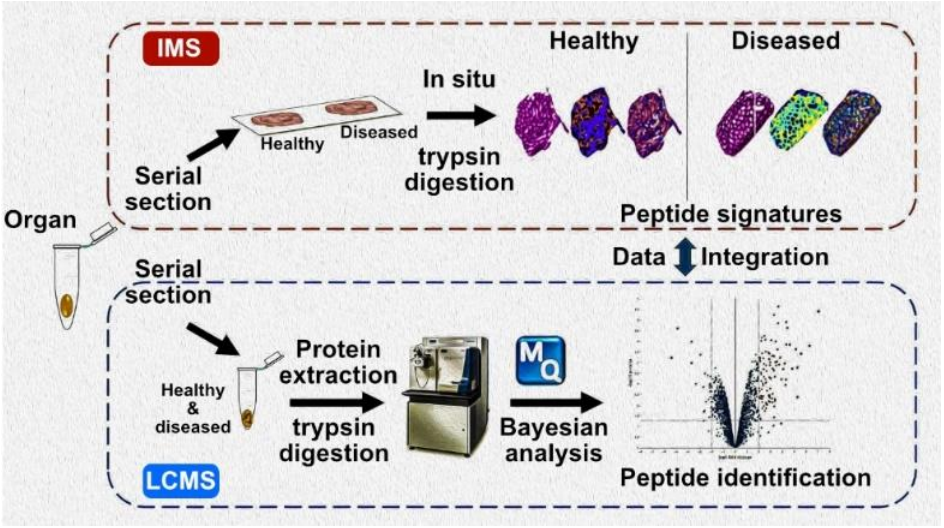


Fig. 2.7: IMS and LC-MS data integration pipeline

The datasets to develop ImShot comes from our previously published study (Lahiri et al., 2021) where we used the serial sections of mice testes and performed in situ trypsin digestion prior measuring them in the imaging mass spectrometer. This way we obtained the spatial distribution of the peptides within the healthy and diseased tissues to get an idea about where the diseased and healthy tissues differ (Fig. 2.7). In a parallel, using the serial sections from same mice testes we conducted LC-MS/MS experiments to identify and quantify the protein

from where those peptides originated. This gave an idea about what is different between the tissues and by how much (Fig. 2.7).

Here, I will elaborate on the computational method and readers are encouraged to read our previous study (Lahiri et al., 2021) for details about experimental setup.

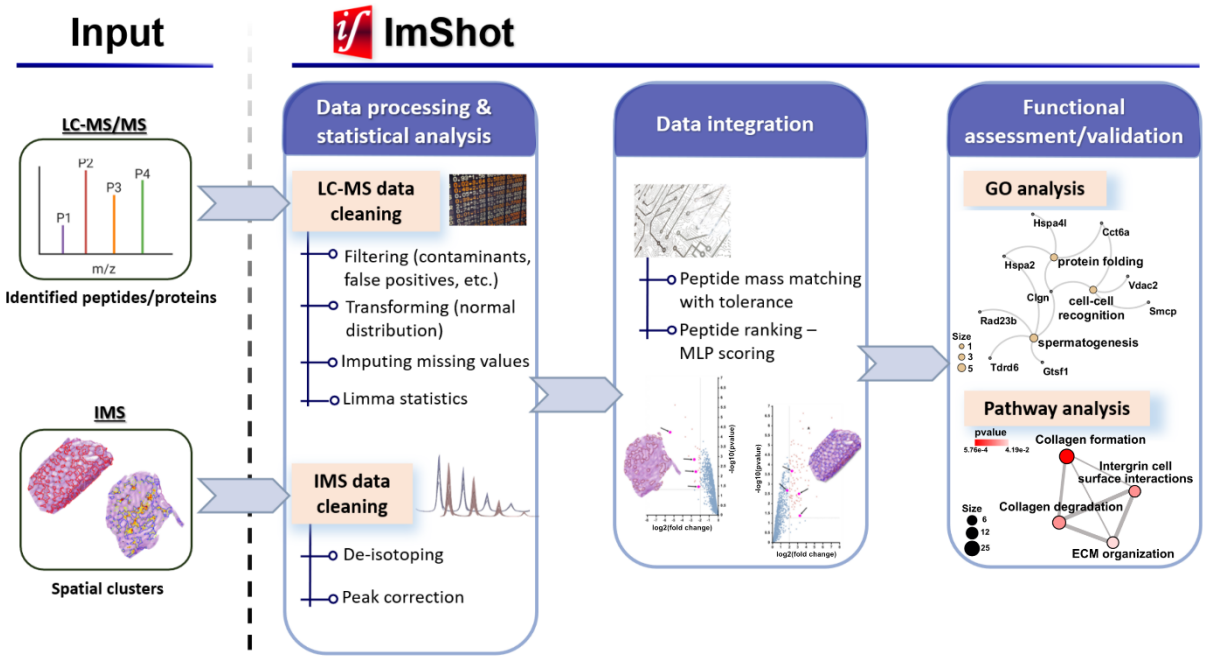


Fig. 2.8: ImShot modules and data integration pipeline - Panel to the left of dotted vertical line (Input) shows that ImShot accepts datasets from both the IMS (peptide clusters) and LC-MS (MaxQuant output: proteingroups.txt/peptides.txt) experiments as input. The ImShot panel consists of 3 segments: i) Data processing and statistical analysis. This is responsible for transforming LC-MS and IMS datasets in a format that is compatible for data integration module. ii) Data integration module. This segment identifies the parent protein of each IMS peptide by associating it to an LC-MS peptide based on mass matching and MLP scoring. iii) Functional assessment/validation module. This serves as a validation tool for the MLP scoring by integrating information from the literature through GO and Pathway enrichment analysis. The arrows show information flow between the modules to actualize the data integration pipeline.

2.3.1 Overview of ImShot algorithm

The ImShot software employs an algorithm that identifies peptides from imaging mass spectrometry (IMS) datasets based on comparison with corresponding proteomic data followed by a scoring method. It initially processes data from both IMS and LC-MS to filter for experimental, analytical, and isotopic contaminants. The individual mass lists thus created from the two complementary techniques are matched within a user-specified tolerance that depends on the measurement accuracy of the mass spectrometry platforms used. The resulting ambiguity arising from one-to-many mass annotations is subsequently resolved by ranking the peptide masses according to a novel scoring system. To further validate the likelihood of peptide identification from IMS, the software has functional validation tools like GO and Pathway analysis that associates biological processes to the most likely region within a tissue specimen. ImShot has been developed using a modular structure that allows the user and/or the developer to customize their individual needs. It also therefore enables a user to use this software for analyzing LC-MS data separately. Finally, we have developed this whole package into an open source, convenient, user-friendly desktop application. The software operates in three modes (Fig. 2.8), viz. Data processing & statistical analysis, data integration and Functional test/validation.

2.3.2 Data processing

Data processing in the backend is done in R. Each block depicted in Fig. 2.8 is subdivided into modules that individually carry out the desired tasks for the user. The details on how those modules operate are as follows.

2.3.2.1 LC-MS data cleaning

Output text files (proteingroups.txt and peptides.txt) from the MaxQuant (MQ) search can be used as input files for this module. This module filters the identified protein list according to the following steps:

i. In a first step, this module removes proteins classified as contaminants from the dataset. Within ImShot the algorithm searches the columns *Reverse*, *Potential contaminant* and *Only identified by site* for positive entries and removes all rows that contain the + sign.

ii. Following that, the algorithm takes care of any blank proteins that is included in the MQ list of identified proteins. The program removes rows (proteins) from the dataset that contain only zeros for all replicates in all conditions.

iii. The filtered dataset is log transformed to ensure its normal distribution for subsequent performance of statistical tests. As mentioned before that, this transformation results in missing values. Missing values in proteomics dataset is a common scenario which occurs when a protein is quantified in some of the replicates but not all. Since such missing values can interfere with the statistical tests, one needs to handle them appropriately. Missing values are imputed from a normal distribution (Fig. 2.2).

iv. Finally, as an added feature of this software, users can extract gene names with a regular expression that can be displayed in the resulting interactive volcano plot.

2.3.2.2 IMS data cleaning

The main aim of this module is to create monoisotopic IMS mass lists from peptide measurements. Owing to the lack of physio-chemical separation of the peptides generated on tissues, a serious problem of overlapping isotopic envelopes arise in almost all the spectral files. Peaks at isotopic positions are often masked by peaks belonging to entirely different peptide(s) (left top panel of Fig. 2.9). Deisotoping of imaging mass spectra has therefore been an unresolved challenge in the field so far.

In this algorithm, we took advantage of the fact that distinct peptides (at isotopic positions) from a tissue display distinct distribution patterns (left middle panel of Fig. 2.9). Subsequently, we have used the unbiased hierarchical clustering algorithm of SCiLS Lab (SCiLS) to segregate

the entire IMS dataset into its component spatial clusters. These clusters are therefore characterized by peptides having identical distribution pattern (left middle panel of Fig. 2.9). Since spatial distributions of isotopic peaks of the same peptide are supposed to be identical, we applied the deisotoping algorithm on the mass lists that distinguished one cluster from the other. We did not encounter any isotopic envelope violating the above-mentioned condition in our IMS datasets (Lahiri et al., 2021).

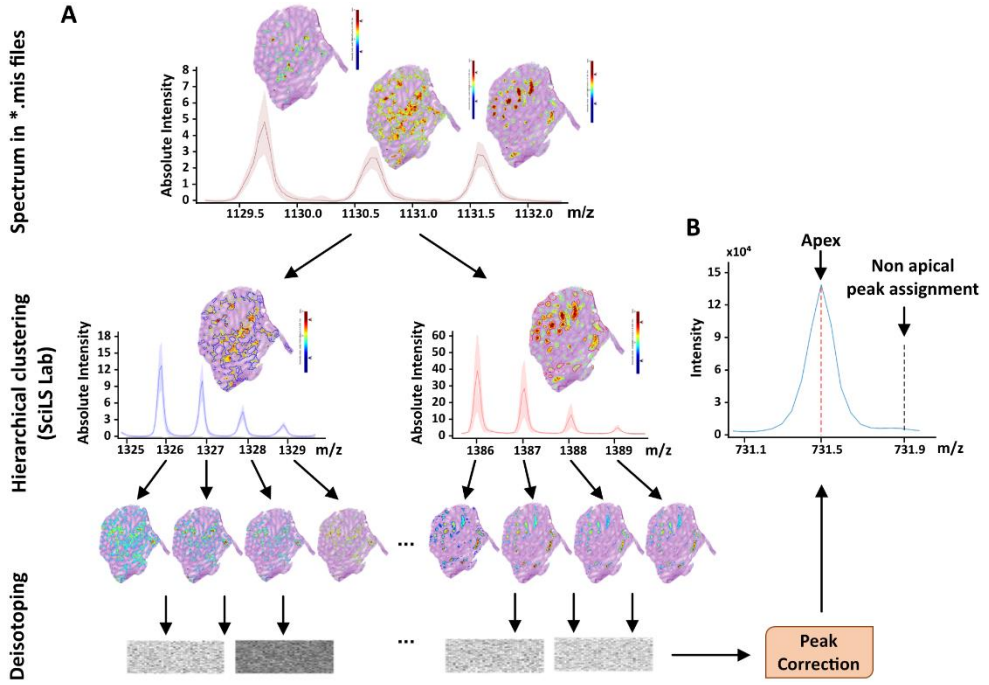


Fig. 2.9: Generating mono-isotopic IMS mass lists by filtering overlapping spectra – (A) Upper panel (Spectrum in *.mis files) shows a supposed isotopic envelope extracted from a MALDI IMS spectrum. The supposed isotopic peaks show very distinct spatial distribution patterns (HE tissue images alongside the peaks). The middle panel (Hierarchical clustering (SCiLS Lab)) shows the segregation of the complex spectra into those of distinct peptides (m/z = 1326 and 1385.75 in this case). The spatial distributions of masses at isotopic positions are identical in this case (ion images below the isotopic envelopes). Deisotoping algorithm is applied on these spectra to generate the mono isotopic IMS mass lists (bottom panel). **(B)** Non apical peak assignments in some cases by proprietary software e.g., SCiLS lab. It highlights the importance of the peak correction module in ImShot.

i. Deisotoping

In general, the most intense peak (usually the 1st) of an isotopic envelope is considered as the monoisotopic peak that comprises of the naturally occurring most abundant elemental isotopes. Deisotoping was performed using standard tolerances which for MALDI was ± 0.15, 50% for m/z and intensity respectively, i.e., for any m/z =m we removed all the m/zs (m_i) falling within the interval: m+0.85 ≥ m_i ≥ m+1.15. Following the deisotoping step, we observed non-apical assignment of a small fraction (~20-25%) of all the monoisotopic m/z values in a cluster (right panel of Fig. 2.9). For example, the apex of for a peptide peak is at m/z = 731.5 but SCiLS Lab assigns the value at m/z = 731.9, since it deals with m/z intervals rather than m/z peaks. Since our aim is to identify peptides from IMS by comparing it with LC-MS data, we applied the following peak correction algorithm to the 'incorrectly' assigned m/z values.

ii. Peak correction

To correct a peak corresponding an m/z = m, ImShot scans a 1 m/z window that contains m. If the intensity of m is not the highest within that window then, it updates m with the m/z value corresponding to the highest peak there. The rationale here is, for MALDI based ionization method there can be only one peak within a 1 m/z window. Following the peak correction on the monoisotopic mass list, we generate the final IMS mass list as an output of the IMS data cleaning module. This list is subsequently compared with LC-MS data to identify the corresponding parent proteins.

2.3.3 Statistical analysis and data integration

2.3.3.1 Bayesian statistics for LC-MS data

This module enables users to determine the significantly enriched proteins by employing Limma moderated t-test statistics as described in section 2.1.1.4. ImShot employs the Limma R package (Ritchie et al., 2015) in the backend to computationally compare two groups

(healthy vs. diseased) in proteomics datasets. For every protein, the t statistics t_{ord} is computed using the mathematical formula presented in eq. (2.3). Where, lfc implies difference between means of the two groups in $\log_2$ scale and σ, $\sigma unscaled$ imply residual standard deviation and unscaled standard deviation respectively.

$$t_{ord} = \frac{lfc}{\sigma_{unscaled} * \sigma} \tag{2.3}$$

$$t_{mod} = \frac{lfc}{\sigma_{unscaled} * \sigma_{posterior}} \tag{2.4}$$

$$\Delta_{\sigma} = \frac{\sigma - \sigma_{posterior}}{\sigma} * 100 \tag{2.5}$$

The Limma moderated t statistics t_{mod} is computed using the mathematical formula presented in eq. (2.4). Where, $\sigma_{posterior}$ imply posterior value of σ which is obtained by applying empirical Bayes method on the entire data. Therefore, for any $lfc \neq 0$ if $\sigma > \sigma_{posterior}$ then $|t_{ord}| < |t_{mod}|$. Thus, the processed and analyzed (for enrichment) LC-MS mass list is created (at the peptide level, in this case) for the sub-sequent step of data integration.

2.3.3.2 Data integration

This module performs the task of combining and comparing IMS and LC-MS datasets with an aim to identify the parent proteins of the peptides detected in IMS. To accomplish this task, the software uses the following two sub-modules:

i. Tolerance search

The monoisotopic mass list for every spatial cluster is searched within either diseased or healthy set of enriched LC-MS peptides depending on the occurrence of the respective cluster. Since the accuracy of measurement differs according to the measurement platform (ion source, mass analyzer, etc.) the search is performed within a certain tolerance (τ). This part of the module has been kept flexible (user specified input) keeping in mind the wide variety of measurement platforms that the users might use.

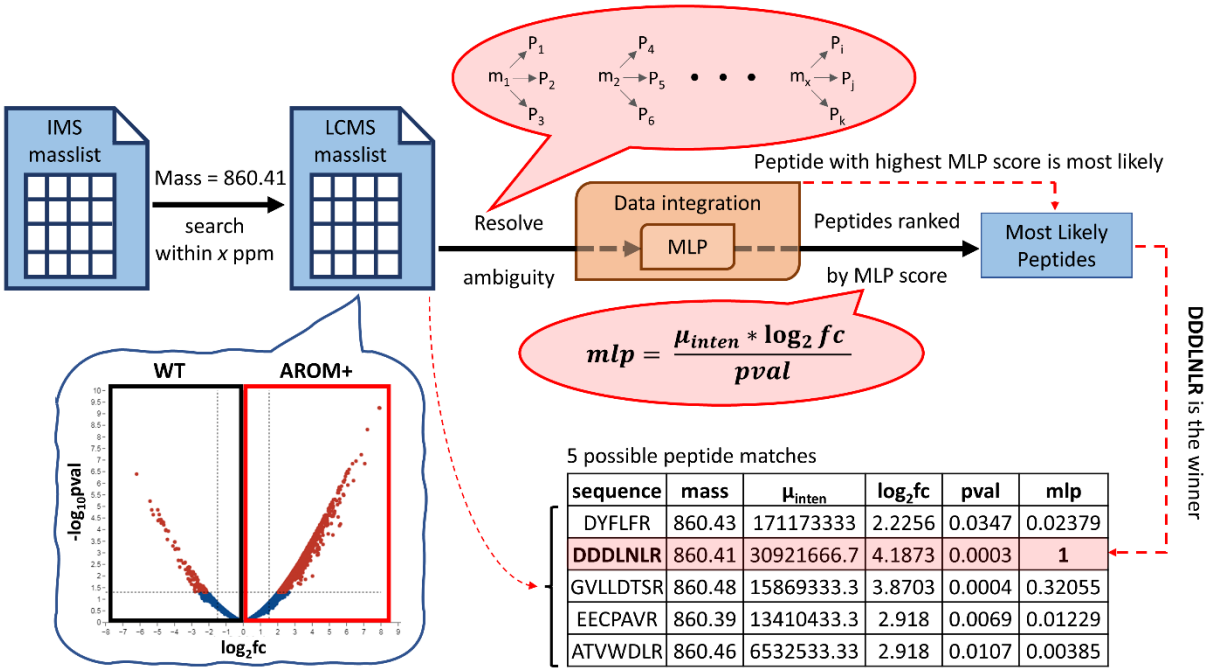


Fig. 2.10: Data integration challenges and solutions- When IMS masslist is searched inside LC-MS masslist (corresponding to one of the groups) within a tolerance it results in ambiguity as shown in red speech balloon. The Data integration module of ImShot resolves the ambiguity by ranking peptide by MLP scores where peptide with highest MLP score is most likely. When an IMS peptide with mass 860.41 Da is searched inside LC-MS masslist it results in 5 possible matches however, as DDDLNLR has the highest MLP score it is the most likely peptide among the five.

ii. MLP scoring

Owing to the relatively low accuracy of IMS as compared to conventional shotgun proteomic measurements, the tolerance search yields a 1:many mapping between IMS and LC-MS peptides, i.e., one IMS peptide mass is annotated to multiple LC-MS peptides originating from different parent proteins (Fig. 2.10). To resolve this ambiguity, we devised a novel scoring method that ranks the identity of peptides (as being part of the parent protein) based on the following equation:

$$MLP = \frac{\mu * \log_2 fc}{p_{mod}} \tag{2.6}$$

where μ is the mean intensity of a peptide across the replicates in either diseased or healthy group, $pmod$ is the Limma moderated p-value of the same peptide and $\log_2 fc$ implies the fold change between the diseased and healthy groups, which is defined as follows,

$$\log_2 fc = \log_2 \left[\frac{\mu_{diseased}}{\mu_{healthy}}\right] \quad (2.7)$$

The intensities used here can be raw intensity, iBAQ or LFQ values, depending on the need of the user. Likelihood of a peptide to belong to its corresponding identified parent protein was correlated to increasing MLP score for that peptide based on the following reasoning:

a. Peptides of relatively higher abundance are preferably detected in MALDI-IMS mainly due to the lack of any separation technique and competitive co-crystallization of matrix/biomolecules (μ). Hence peptides (belonging to certain proteins) having a higher μ value among the multiple possibilities are most likely the ones that are detected in IMS.

b. The search space for a peptide belonging to a cluster detected in IMS measurements is narrowed down to either healthy or diseased LC-MS data depending on their occurrence in the corresponding tissues ($\log_2 fc$). This increases the likelihood of a peptide belonging to a particular protein with very high confidence.

c. Inclusion of the moderated p-value in the scoring system is used to increase the likelihood of a peptide belonging to a given parent protein even further (p_{mod}). Lower the p_{mod}, higher the score and higher the probability of an IMS peptide to belong to its corresponding identified protein.

Therefore, peptides from spatial IMS clusters with top MLP scores are regarded as belonging to the corresponding parent protein identified in LC-MS. Fig. 2.8 shows the interaction between the modules of ImShot to actualize the data integration.

2.3.4 Functional assessment/validation

The IMS peptides identified in the previous step can be screened here based on the correlation of their identity to occurrence of specific biological processes in the most likely tissue compartment. To perform this task ImShot uses the following two modules:

2.3.4.1 GO analysis

This module allows users to associate a common theme to the genes/proteins of interest that can help answering the biological question. Gene ontology (GO) provides annotation for genes or gene products at different domains: cellular component, molecular function, and biological process that are organized in the form of directed acyclic graphs (DAGs) data structure. It is possible that proteins could be annotated to multiple GO nodes. Moreover, due to the nature of DAG data structure a gene annotated to a particular node also inherits annotation from the ancestors of that node. Therefore, in order to find out if a GO term enriched in specified list of genes not by chance, ImShot calculates p-values as proposed in this study (Boyle et al., 2004):

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} * \binom{N-M}{n-i}}{\binom{N}{n}} \quad (2.8)$$

Where, N is the number of genes/proteins in background list, M is the number of genes within that list that have direct/indirect annotation to the GO node of interest, n represents the length of the list corresponding to the genes of interest and k is the number of genes within that list which are annotated to the node. ImShot uses an R package called ClusterProfiler (Karpievitch et al., 2012) in the backend to perform the GO over representation test. For the background use in ImShot, the user can either use the global background provided in ClusterProfiler or can use a gene/protein list of their own, which contains customized background list for user specific needs. The results are displayed in the form of a network graph and a table. Network graph

(rightmost panel of Fig. 2.8(C)) shows association between gene and GO terms, an edge is drawn between a gene and a GO term if and only if gene is found enriched in that term.

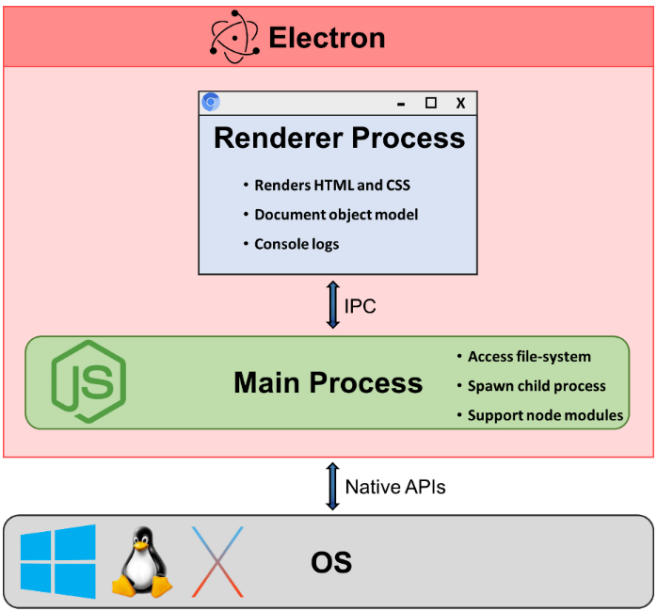
2.3.4.2 Pathway analysis

Pathway analysis module allows users to associate a common theme to the genes/proteins of interest by annotating (statistically significant manner) them to the biological pathways. Often knowledge of affected (due to treatment) biological pathways can help answering the biological question. Here, we used the R package ReactomePA (Yu and He, 2016) to discover biological pathways in which the genes/proteins of interest participate. Like GO analysis module, the ReactomePA package use hypergeometric distribution model to calculate P-values to determine whether any pathways in Reactome database annotate a specified list of genes at a frequency greater than that would be expected by chance. In addition, ImShot also supports pathway enrichment analysis using KEGG database, thereby increasing the applicability of the software for the community. Just like GO analysis, results are displayed in the form of a network graph and a table. Network graph (bottom graph in rightmost panel of Fig. 2.8(C)) shows association between Pathway terms, an edge is drawn between two Pathway terms if at least one common gene is found to be enriched the two Pathways.

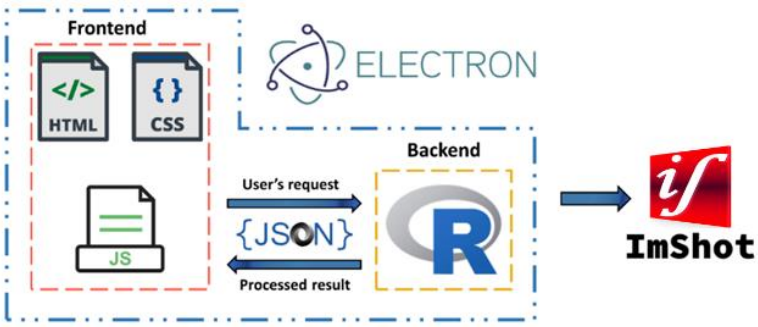
2.3.5 Development of ImShot desktop application

ImShot desktop application was developed using the open-source software framework Electron that allow users to build desktop applications by integrating web technologies such as JavaScript (JS), HTML and CSS. It does so by combining Chromium rendering engine and the Node.js runtime. Fig. 2.11A depicts the multi-process architecture of Electron, which consists of two types of processes: the main process and renderer process. The main process's prime task is to start the application and respond to its lifecycle events such as creation and destruction of renderer process. It is also responsible for communicating with OS via system APIs. The

Renderer process uses Chromium engine (the core code module of the Chromium, open-source version of Google's Chrome web browser) to render a web page as an independent process.



(A)



(B)

Fig. 2.11: Software architecture- Shows **(A)** the architecture of Electron framework, which combines the core web browsing component of Chromium with the low-level system access of Node.js to create platform independent applications that have great UI like web applications and high performance like native applications. **(B)** the architecture of ImShot software. Frontend comprises of HTML, CSS, and JavaScript (JS) and backend is programmed in R. JS intermediates between the two ends using inter-process communication. User's request and processed results are exchanged in the form of JSON objects. Electron packages the whole code base into a single executable that can run in standalone mode in all major OS platforms.

It handles fetching and rendering HTML, loading any referenced CSS and JavaScript, styling the page accordingly, and executing the JavaScript. However, unlike typical web pages, it has privilege to interact with OS by-means-of the Node.js runtime. The Node.js runtime uses Google's open source V8 engine to interpret JavaScript and provide APIs for accessing the filesystem, loading code from external modules, and communicating with other programming language. We used the child_process module of node.js with the help of js-call-r package (codejie) to call functions written in R programming language that perform data wrangling (in the backend) as depicted in Fig. 2.11B. The child_process module provides the ability to spawn subprocesses who can easily communicate with each other with a messaging system. We used the child_process.spawnSync() function which spawns child process in a synchronous manner that blocks the event loop until the spawned process either exits or is terminated. i.e., we wait until the R code that is being executed on a child process is finished execution. The data is then passed between JS and R in the form of a JSON (JavaScript Object Notation) object (JSON). JSON is a lightweight data-interchange format that helps us to achieve minimal communication latency. Moreover, data in JSON format is easy for humans to read and write and is easy for machines to parse and generate. We JSONify (to encode in JSON format) the data to be sent to R and in R environment we un-JSONify (to decode from JSON format) the received data so that R code can use them. Finally, we JSONify the R output and send to JS where it gets un-JSONified for displaying in the frontend.

3. Results

3.1 Biological implications derived by applying Limma proteomics pipeline and web application MiGENet

3.1.1 Limma proteomics pipeline aided the discovery of a novel pathway in yeast mitochondria that regulates translation of a specific mRNA

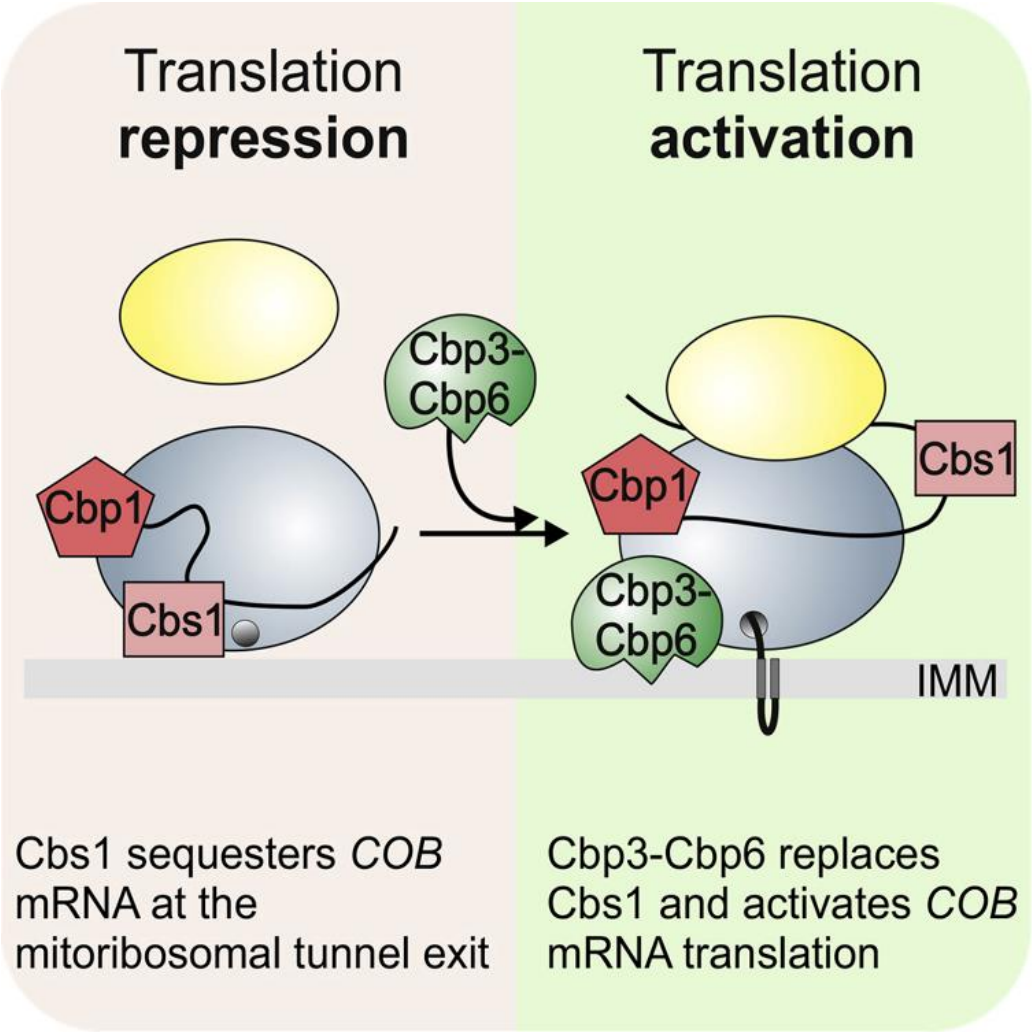


Fig. 3.1: Cbp1, Cbs1, Cbs2 and the Cbp3-Cbp6 complex co-ordinate in a feedback loop to regulate the translation of COB mRNA. [Image reused from (Salvatori et al., 2020a) under author's rights in Elsevier's proprietary journals; Licensing information: Appendix C]

The mitochondrial respiratory chain is composed of proteins that are encoded by both nuclear and mitochondrial genes. To ensure effective assembly, the two-expression systems must

coordinate closely. Translational activators the nuclear encoded proteins that interact with their client mRNA (mitoribosome) regulates the translation of mitochondrial mRNAs. As a result, the activators exert strict control over the production of mitochondrial proteins. Translational activators Cbp1, Cbs1, Cbs2, and the Cbp3-Cbp6 complex are required for cytochrome b synthesis the core subunit of complex III. By employing the data analysis pipeline presented in Tools (Sec. 2.1), we discovered a biochemical pathway via which these translational activators regulate COB mRNA translation. This feedback loop (Fig. 3.1) is dependent on the alternating binding of Cbs1 and Cbp3-Cbp6 at the mitochondrial ribosome's polypeptide tunnel exit. When COB mRNA translation is inhibited, Cbs1 binds to the polypeptide tunnel exit and holds COB mRNA, preventing it from being translated. When Cbp3-Cbp6, that has been released from its complex III assembly intermediate, interacts with the polypeptide tunnel exit, translation is triggered. Cbp3-Cbp6 induces the migration of Cbs1, resulting in the availability of COB mRNA for translation.

3.1.2 MiGENet enabled mining of spatial information about connectivity and molecular mechanisms regulating mitochondrial gene expression

Mitochondria have their own machinery for gene expression, which requires several proteins for transcription, RNA processing, translation, and assembly of the newly synthesized subunits. By using the data analysis pipeline and the web application MiGENet (developed using forceNetwork++ widget, see sec. 2.1.3 in Tools), we identified a large network of factors involved in biogenesis of mitochondrial proteins in baker's yeast. The interactive graphic system of MiGENet empowered users to interact and extract information even from the densest regions of the network (Fig. 3.2) which enabled them to gain knowledge about the mitochondrial gene expression machinery in a fast and efficient manner.

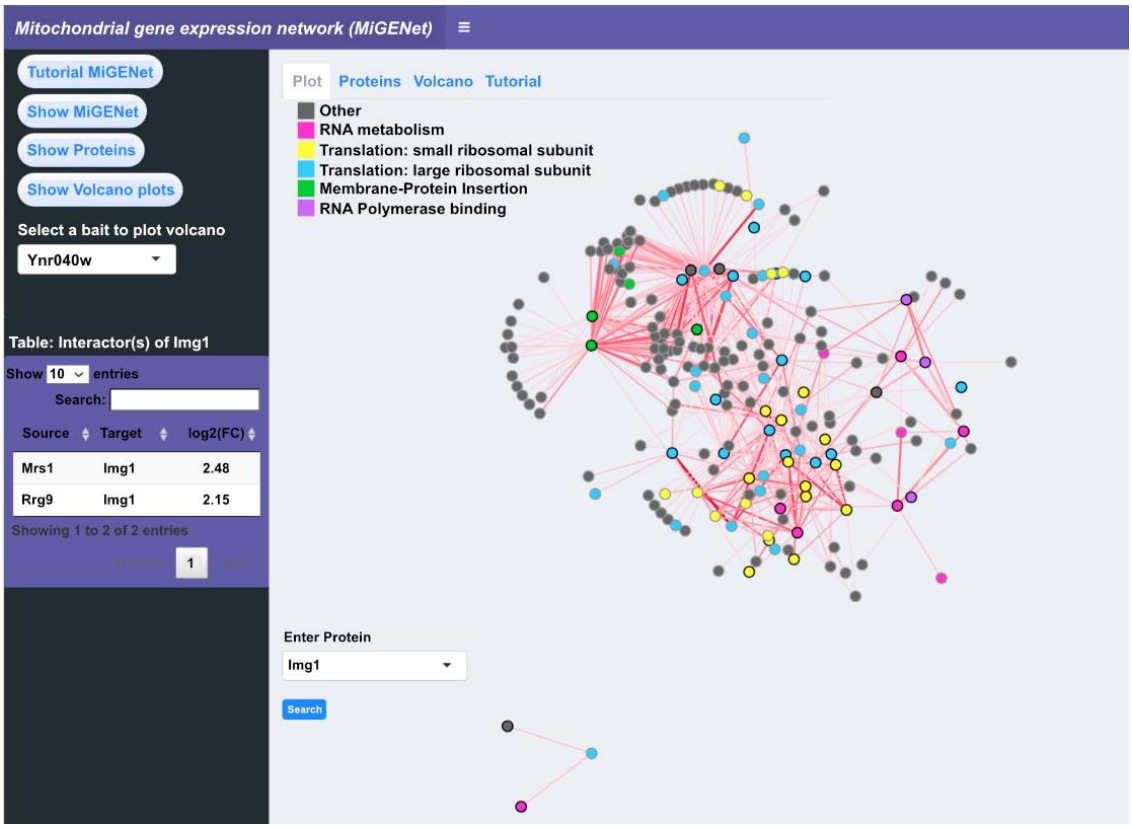


Fig. 3.2: MiGENet is a resource to visualize and interact with large protein interaction networks: Showing a snapshot of interactive network graph and demonstrating features of MiGENet that enable users to extract neighbors of a searched protein.

3.2 Results obtained by applying *CoreClust* and SNN methods

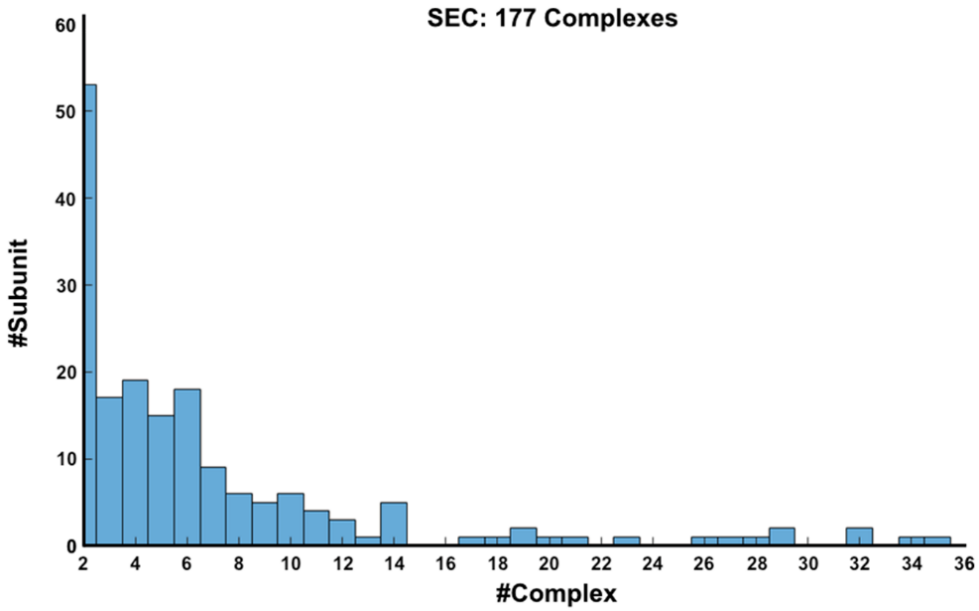


Fig. 3.3: *CoreClust* algorithm discovered 177 protein complexes in our SEC dataset.

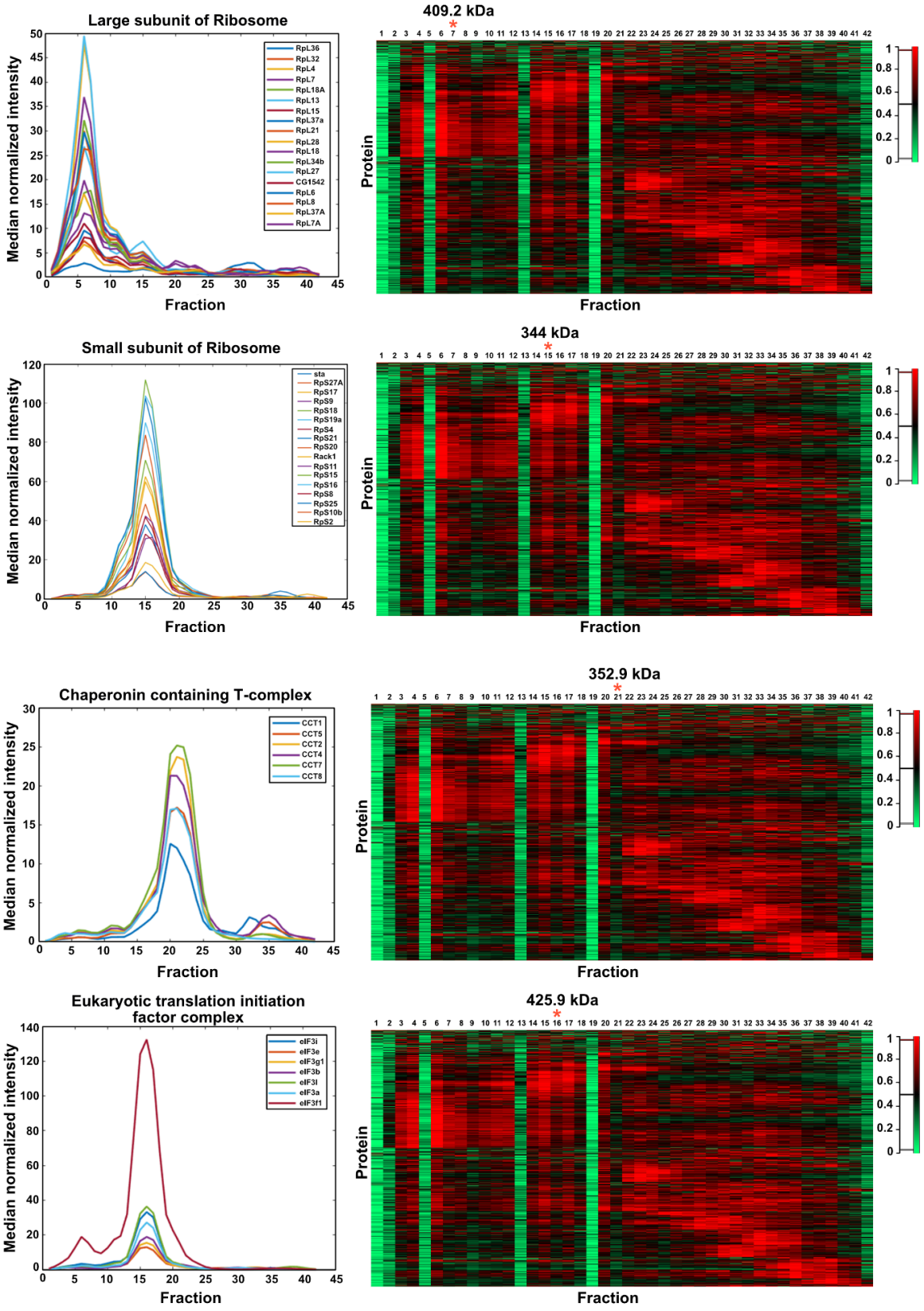


Fig. 3.4: Some of the bona fide complexes that were discovered by applying *CoreClust* algorithm on our SEC dataset- Elution of the subunits of protein complex is marked with red asterisk.

After applying *CoreClust*, on our SEC dataset (Fig. 2.3 in Tools) we discovered 177 clusters/protein complexes (Fig. 3.3). Locations of the elution profiles for the 3 standard proteins with molecular weights 669 kDa, 440 kDa and 158 kDa were noted (marked with arrows in Fig. 2.3 in Tools), using this information we validated the output of *CoreClust* algorithm. The rationale is, if sum of the molecular weights of the subunits of a protein complex is within the milestones set by the standard proteins, then we consider that protein complex as a true complex. Using this heuristic, we found some of the bona fide protein complexes such as large and small subunit of ribosome, Chaperonin containing T-complex, Eucaryotic translation initiation complex etc. as depicted in Fig. 3.4. *CoreClust* is a quick and dirty way to detect protein complexes, but due to inherent noise in the complexomics datasets, one should employ machine learning as it can even recognize not so obvious patterns in the data. Thus, increasing the sensitivity of protein complex (putative) discovery pipeline. *ComplexMiner* is developed with this goal in mind.

The prime goal of the Siamese subnetworks (See sec. 2.2, Fig. 2.5) in *ComplexMiner* application is to create a pair of feature vectors that are easily classifiable in the fully connected layer. If the profiles are very similar, then after sufficient number of iterations the subnetworks generate very similar feature vectors; otherwise, they produce very dissimilar feature vectors as demonstrated in Fig. 3.5. Fig. 3.5A shows elution profiles of two subunits from mcm complex, as they are part of a protein complex, they elute together from a chromatography column. When they are fed to the subnetworks then, after ~1000 iterations the SNN generates the corresponding feature vectors as shown in Fig. 3.5B. Notice they almost overlap with each other. On the other hand, when two non interacting protein pairs went through the same operation as before, the corresponding feature vectors look quite different (Figures 3.5C-D). So, in essence, SNN tries to maximize the differences (in the feature space) between dissimilar input pairs and minimize the same in case of similar pairs.

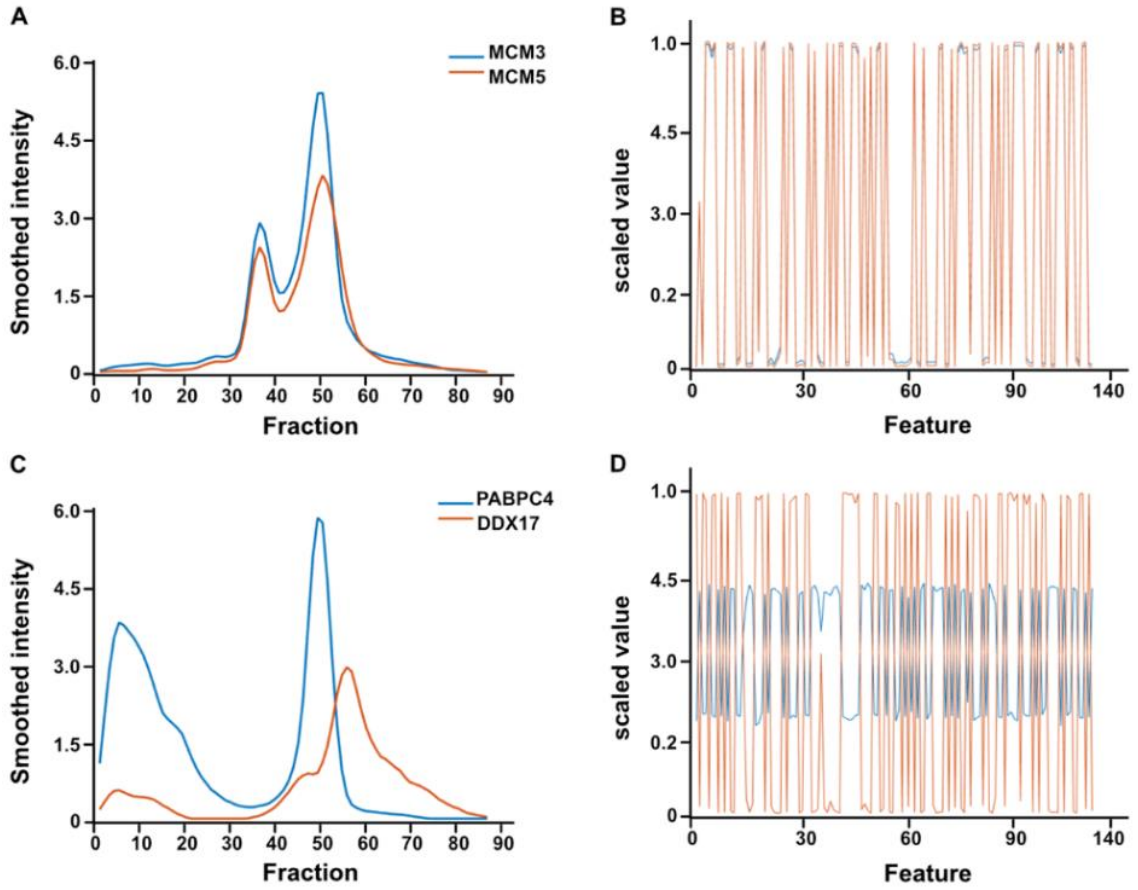


Fig. 3.5: Feature space during training- (A) Shows a pair of elution profiles of interacting proteins MCM3, MCM5 (subunits of MCM complex). **(B)** Shows the feature vectors of the profiles after ~1000 iteration. **(C)** Shows a pair of elution profiles of non-interacting proteins. **(D)** Shows corresponding feature vectors after ~1000 iteration.

3.3 ImShot to facilitate spatial proteomics

3.3.1 Moderated t-test yields more significant and biologically relevant proteins

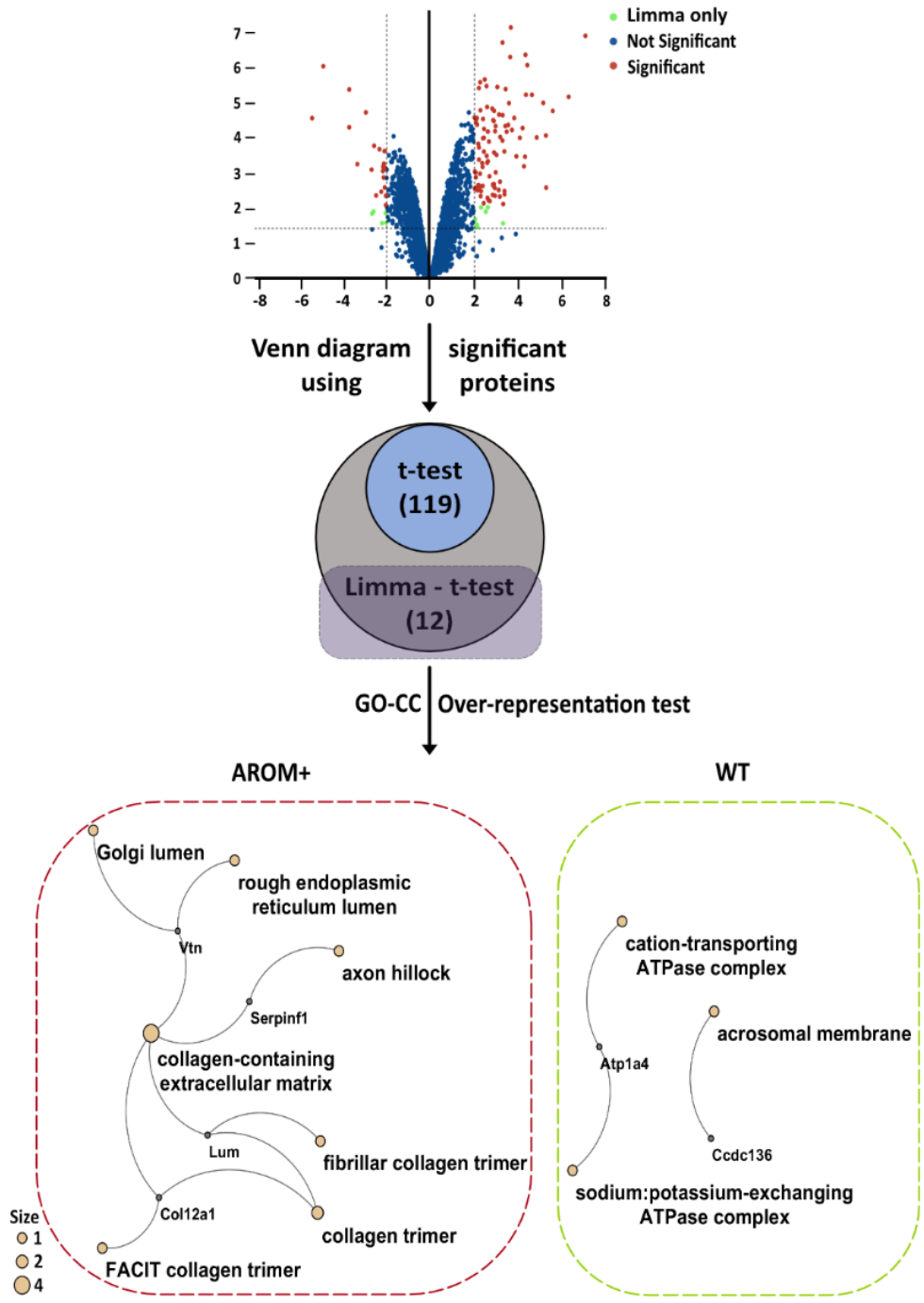


Fig. 3.6: Limma moderated t-test provides more powerful inference - Top panel shows a volcano plot after two group comparisons using Limma statistics where dots with red colors

correspond to statistically significant proteins having $(lfc > 2 \,||\, lfc < -2)$ & $pvalue < 0.05$ that are overlapping between standard t-test and LIMMA. The green dots in the plot show the proteins that become statistically significant only when LIMMA is applied. The Venn diagram in the middle panel demonstrates that Limma statistics yields 12 more proteins than ordinary t-test. These were used in GO-CC enrichment analysis whose results are depicted in the bottom panel in the form of a gene-GO term network. Two distinct GO clusters are observed: the red rounded dashed rectangle displays the cluster of terms enriched in AROM+ proteins, whereas the green rounded dashed rectangle highlights the cluster of terms enriched in WT proteins.

Table 1: Statistically significant proteins determined exclusively using Limma statistics

gene	lfc	t_{ord}	p_{ord}	t_{mod}	p_{mod}	σ	$\sigma_{posterior}$	Δ_σ
Fsip2	-2.58786	-2.54893	0.063377	-3.17401	0.017473	1.243452	0.998568732	19.69
Afm	2.304437	2.729307	0.05248	3.371293	0.013497	1.034089	0.837170525	19.04
Vtn	2.282418	2.318201	0.081297	2.88337	0.025814	1.20584	0.969483584	19.6
Hist1h1e	2.10801	2.072954	0.106872	2.581464	0.039149	1.245457	1.000120384	19.7
Lum	2.611408	2.718441	0.053072	3.377887	0.013382	1.176523	0.946837126	19.52
Serpinf1	2.453236	2.521249	0.06527	3.134472	0.018413	1.191706	0.958562629	19.56
Ccdc136	-2.11621	-2.22431	0.090179	-2.7628	0.030447	1.165223	0.938113925	19.49
Col12a1	3.304637	2.097001	0.104005	2.639595	0.036104	1.93006	1.533317489	20.56
Lypd4	-2.61134	-2.53249	0.064494	-3.1553	0.017911	1.262879	1.01360433	19.74
Fmo2	2.138849	1.976454	0.119285	2.466548	0.045988	1.325376	1.062028449	19.87
Tex33	-2.00396	-2.1484	0.098158	-2.66628	0.034791	1.142407	0.920511851	19.42
Atp1a4	-2.0369	-2.51776	0.065513	-3.1028	0.019206	0.990835	0.804009894	18.86

lfc → log fold change; $\mathbf{p_{ord}}$ → p-values corresponding to the t-statistics ($\mathbf{t_{ord}}$); $\mathbf{p_{mod}}$ → p-values corresponding to the moderated t-statistics ($\mathbf{t_{mod}}$); **σ** → sample standard deviations for each gene/protein; $\boldsymbol{\sigma_{posterior}}$ → posterior values for **σ**; **Δσ** → percentage shrinkage

Applying Limma based moderated t-test in the dataset from our recent study on aromatase induced male infertility (Lahiri et al., 2021) we found that 12 more proteins (Fig. 3.6, green

points on the volcano plot; Table 1) turned out to be statistically significant using Limma moderated t-test when compared to the ordinary t-test. For all these proteins, we observe that the percentage shrinkage (Δ_σ) in sample variance (computed using Eqn (2.5)) is always positive (Table 1) and $|t_{mod}|$ is always greater than $|t_{ord}|$. As higher t-value is associated with smaller p-value, we observe that Bayesian modelling yields more statistically significant proteins when compared to ordinary t-test by shrinking the sample variance towards a pooled estimate.

However, to learn if these additional statistically significant proteins are at all relevant biologically, we performed GO enrichment analysis using the GO analysis module of ImShot. In GO-CC enrichment network we notice that most of the proteins that are significantly upregulated upon aromatase overexpression, are involved in regulation of the extracellular matrix (ECM) (Fig. 3.6, lower red panel). Interestingly, it has been shown that components of the ECM are upregulated in men suffering from infertility (Adam et al., 2012; Alfano et al., 2019). In case of WT, we observe the prominence of acrosomal membrane and protein complexes required for high-energy cellular processes (as expected in case of normal spermatogenesis) (Fig. 3.6, lower green panel). Together, these results justify the application of Bayesian modelling over ordinary t-test: we not only get more proteins that are statistically significantly different between two conditions but also can get more information that describes the condition biologically.

3.3.2 Deisotoping and peak correction prevents false positive inclusion and loss of information

Re-analysis of our previously published data (Lahiri et al., 2021) and further analysis on the same dataset revealed a substantial reduction in the number of peaks after deisotoping (Table 2). Since the isotopic peaks of the same peptide has identical spatial distribution, deisotoping could get rid of false positive peaks. The peak correction module allowed us to reduce the false positives further by including apical m/z values, thereby preventing loss of

information that could otherwise have affected the matching of IMS data with that of the LC-MS measurements. Together, these modules reduced the false positives by 46.23%. Without the deisotoping and peak corrections we would lose many more peptides and some among them could be crucial in understanding the biological differences between the disease and healthy tissues. Moreover, retaining more legitimate peptide peaks could also help better integration with the LC-MS data because, during the data integration step, peptides identified in IMS are searched in the set of peptides identified by LC-MS experiment.

3.3.3 Localization of proteins and pathways in situ validates MLP scoring

Results of the tolerance search shows that ~63% of m/z values (spanning over the IMS clusters mentioned in Table 2) bear the 1:many correspondences with the identified peptides in LC-MS. As described before, we apply a novel scoring method (MLP scoring) to resolve this ambiguity. However, validation of our reasoning behind the scoring approach is required to impart further confidence in the applicability of this scoring method, in general. In addition to validating experimentally the distribution pattern of a subset of peptides identified in IMS measurements (Lahiri et al., 2021), we attempted to further validate the scoring here by using modules from ImShot itself and available public data as a proof of concept. The GO and Pathway analysis modules of the Functional assessment/validation section (Fig. 2.8 right panel in Tools) are used here to assess the functional relevance of the peptides by dint of their spatial localization.

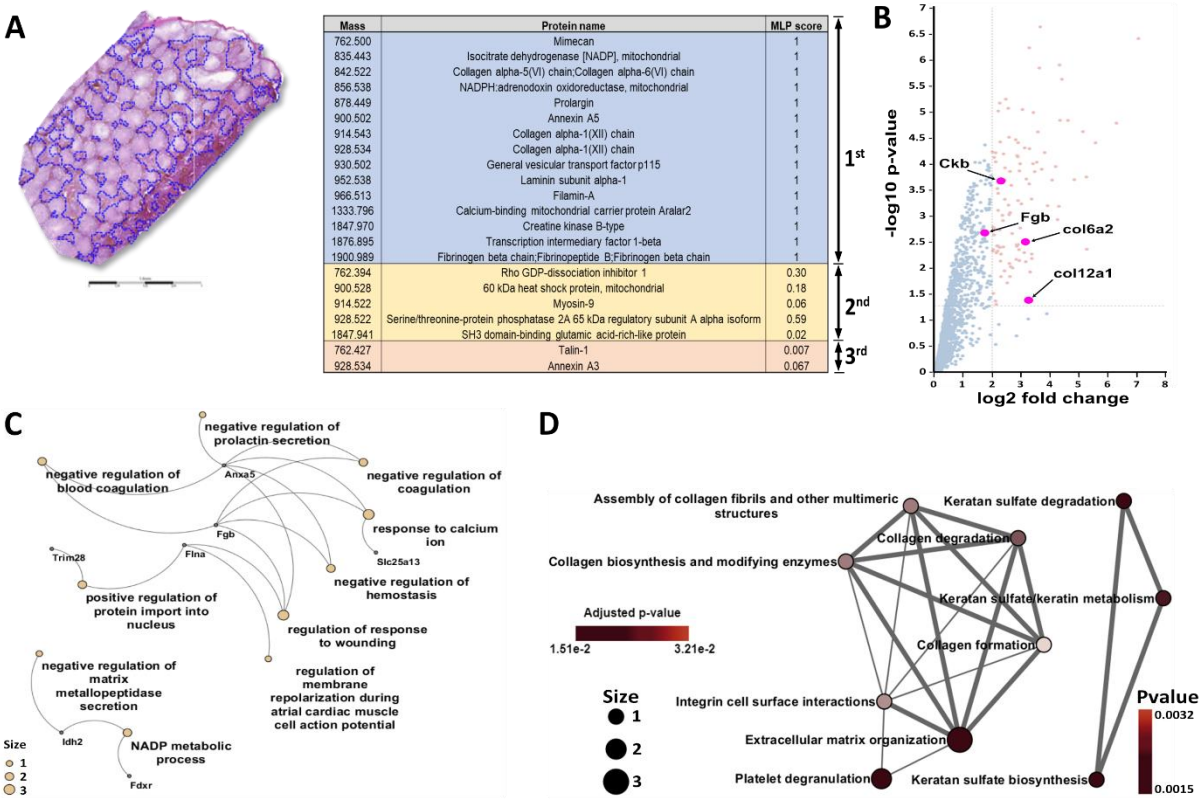


Fig. 3.7: Validating MLP scoring computationally - (A) HE stained image of AROM+ mouse testis. The deep blue pattern within the tissue represents an interstitial cluster detected exclusively in AROM+. The peptides from this cluster were searched in corresponding LC-MS data and the results after MLP scoring are shown in different colors according to the ranks. **(B)** Some proteins having peptides with top MLP scores are highlighted (in Pink) in the volcano plot, showing that they were also highly enriched in the LC-MS data corresponding to AROM+ mice. **(C)** Gene-GO term network after over-representation test using proteins from the table annotated with highest MLP scores (1st group). Tiny fixed sized grey nodes in the network represent genes and larger light-colored nodes (variable sizes, see size legend) represent GO terms. An edge between a gene node and GO term node indicate that the term was not enriched by chance. **(D)** Gene-Pathway network after over-representation test using proteins from the table annotated with highest MLP scores (1st group). Nodes in the network represent Reactome pathways. Two pathways are joined with an edge if they share enriched (not by chance) genes. Thickness of an edge is proportional to the number of common genes. Nodes are colored

according to p-value of over-representation test and the color gets darker as the p-value decreases.

3.3.4 Computational validation

We first opted to validate the parent proteins of peptides (with highest MLP score) from a cluster observed exclusively in the interstitial spaces of AROM+ testis (Fig. 3.7A). Here we observe that peptides from proteins like mimecan, different chains of collagen and prolargin, which have previously been shown to be involved in extracellular matrix assembly and regulation (Mayer et al., 2016; Lahiri et al., 2021; Alfano et al., 2019) have acquired the highest MLP scores (Fig. 3.7A). In addition, we find that the interstitial cluster is enriched in biological processes related to connective tissue formation involving ECM components (Fig. 3.7C) and hemostasis. This is in line with the observation that interstitial spaces and the ECM components involved therein are severely affected in the AROM+ phenotype (Adam et al., 2012; Alfano et al., 2019). At the same time, these proteins are also observed to be highly enriched in AROM+ LC-MS data (Fig. 3.7B). Applying expert knowledge and database mining on the peptides identified with highest, second highest and third highest MLP scores for the above cluster (Fig. 3.7A), we could say that the biological relevance decreases as the MLP scores decreases. Therefore, the MLP scoring based ranking method is providing us with probable protein identification in situ with reasonable accuracy and minimum false positives. Consistent with the GO analysis and other findings, from pathway analysis we see that extracellular matrix organization through collagen synthesis and assembly (Fig. 3.7D) and immune responses characterize the interstitial cluster in AROM+ testis. This provides further timber to our ranking method for identifying peptides in situ.

3.3.5 Literature based validation

As a further step of validating the identity of the IMS peptides by MLP scoring, we proceeded to compare our findings with publicly available data. To avoid a specific disease

model, we chose peptides localized to the seminiferous tubules of the WT testis. The 'identified' proteins in the WT seminiferous tubules are all responsible for healthy development and functioning of testis. We generated ion images (Fig. 3.8) for those peptides from our previously published (Lahiri et al., 2021) dataset, which clearly demonstrates their tubular localization in the WT mouse testis. Comparing the distribution with human testes cross sections (Khan et al., 2018), we observe that the proteins identified by MLP scoring indeed localizes to the testicular tubules (Fig. 3.8). This serves as an additional and strong validation of our MLP scoring method.

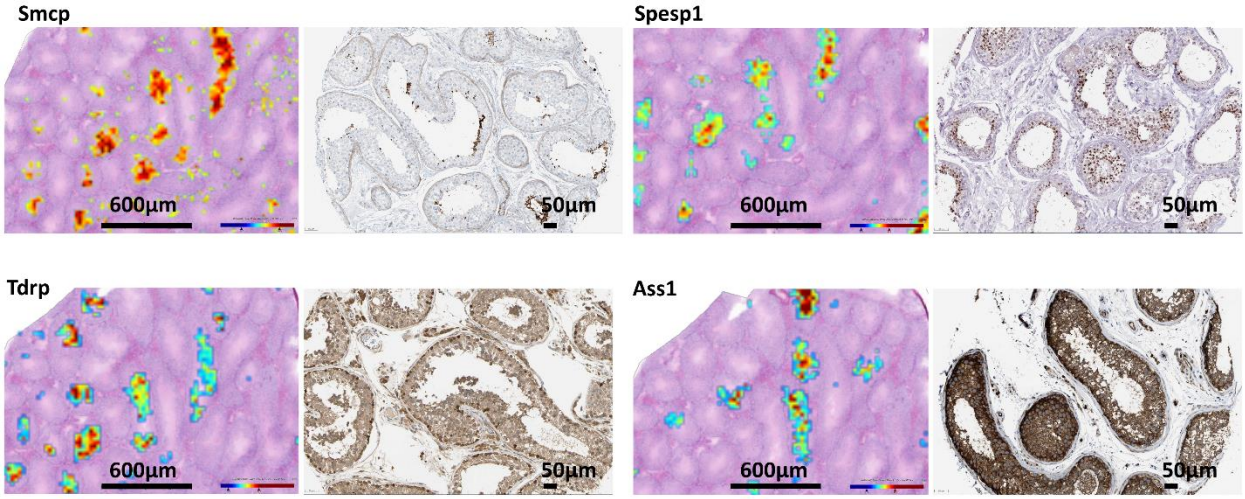


Fig. 3.8: Validating MLP scoring based on literature- Distribution pattern of some important proteins in IMS measurements (identified using MLP scoring) in WT tissue. Ion images (zoom: 600µm) from our IMS experiments indicate these proteins are distributed in tubular regions of WT mice testis and the distribution patterns in the corresponding immunohistochemical staining (IHC) images from human tissue atlas (Tissue-Atlas, 2021) also corroborate with this finding.

3.3.6 ImShot: The desktop application and GUI

ImShot GUI has two parts: sidebar and main panels (Fig. 3.9A). The sidebar contains the different modules of the software as dropdown menus. For the two-group comparison (LIMMA), it renders high resolution interactive volcano plot along with a numeric input box

for fold change adjustment that allow users for desired data thresholding (Fig. 3.9B). Users can also save the plot in the PNG file format. ImShot also shows the Limma moderated t-test results in the form of a searchable table which can be exported as excel or csv format (Fig. 3.9C).

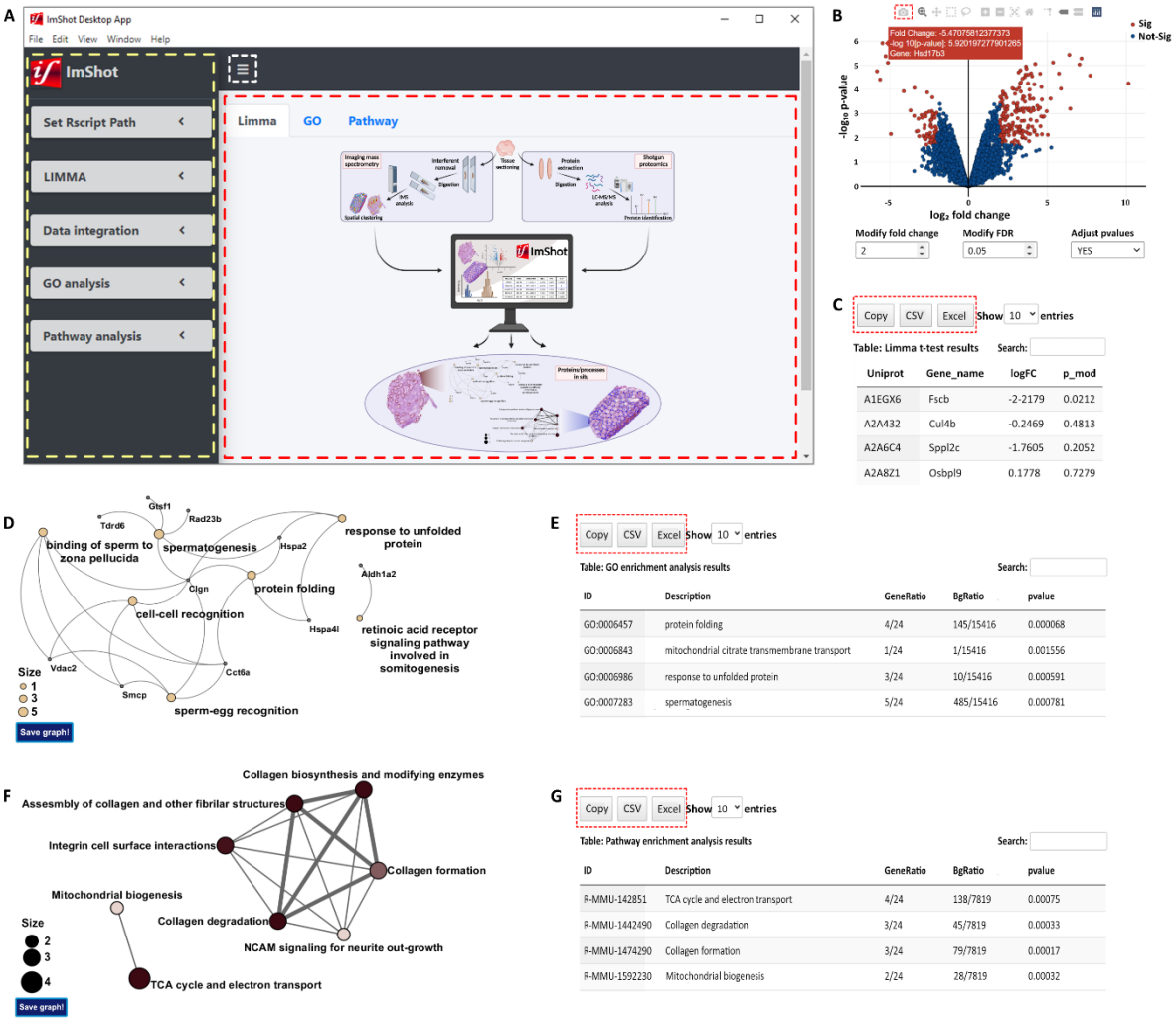


Fig. 3.9: ImShot GUI - (A) ImShot GUI sidebar (yellow dashed rectangle) and main panel (red dashed rectangle). The sidebar panel can toggle upon clicking on the icon enclosed in white dashed rectangle. **(B)** Interactive volcano plot. Reddish dots indicate statistically significant proteins after LIMMA moderated t-test ($(lfc > 2 \,||\, lfc < -2)\ \&\ pvalue < 0.05$) and q-value based FDR control. The plot updates automatically when the log fold change/FDR is tuned using the input boxes provided. Additionally, the plot can be updated according to whether or not a p-value adjustment is desired. Placing the cursor on a data point of the volcano plot provides information about the protein identity, its fold change and p-value (reddish

rectangle). **(C)** Searchable table after LIMMA moderated statistics. **(D)** Protein-GO term interaction network after over representation test. Tiny dark colored, fixed sized nodes represent proteins and light colored, variable sized node represent GO terms and their sizes are proportional to the numbers of proteins involved in them. A protein is connected to a GO term via an edge if and only if the term is enriched (pvalue is adjusted) in that protein. **(E)** Searchable table after GO analysis. **(F)** Pathway-Pathway interaction network after over representation test. Two pathways are connected via an edge if and only if both are enriched (pvalue is adjusted) in at least one common protein. Size of a node is proportional to the number of proteins involved in it and the color is proportional to the adjusted pvalue, lower pvalue maps to darker color. **(G)** Searchable table after Pathway over representation test. Contents from all the tables can be copied into the clipboard or exported as CSV/EXCEL by clicking corresponding buttons on top the tables (shown in dashed red rectangle).

For the GO and Pathway enrichments analyses, it creates high-resolution plots of GO-gene and Pathway-Pathway interaction networks using the top 10 most significant GO and Pathway terms respectively (Fig. 3.9D and Fig. 3.9F). The plots can be exported as PNG image files. These plots in the GUI are zoomable and the nodes are highly flexible allowing the users to select nodes of their choice for rearranging them freely (see video tutorial) to create a network map according to their convenience and need. ImShot also shows the over representation test results in the form of a searchable table (for top 10 most significant GO terms/Pathways) below the network plot which can be exported as excel or csv format (Fig. 3.9E and Fig. 3.9G)).

In addition, ImShot maintains an operation log that allow users to record all the steps along with names of the files and version of R used. This paves the way for the user to reproduce the data analysis later without any ambiguity.

4. Discussion and Outlook

4.1 How Limma proteomics pipeline and forceNetwork++ facilitates knowledge discovery form high-throughput proteomics datasets?

To highlight how our approach aids a researcher confronting with high-throughput proteomic data analysis, I will discuss the salient features of forceNetwork++ (Fig. 4.1) and Limma proteomics pipeline.

Features	forceNetwork	forceNetwork++
Search network	✗	✓
Extract sub-network	✗	✓
Available node colors	Limited	Unlimited
Node border colors	Limited	Unlimited
Mouse hover shows	Only the selected node name	Selected node & immediate neighbors
Node name colors	Limited	Unlimited

Fig. 4.1: Comparing forceNetwork and forceNetwork++ in the context of effective visualization of protein-protein interaction network- The improved forceNetwork++ function contains features (that aid in the effective visualization of PPI networks) that were either limited or absent in the original forceNetwork function from networkD3 R package.

4.1.1 Improved statistical inference

Proteomic studies often use t-tests to identify differentially expressed proteins. In section 3.2.1, I have shown that how applying moderated t-statistics from the empirical Bayes approach can improve outcomes. In order to demonstrate the power of Limma statistics, the pipeline outputs two volcano plots: one using the classical t-test and the other employing moderated t-test. Since the plots are interactive and the pipeline also provides the associated data, it is easy to compare and assess the power of Limma statistics over ordinary t-test. In this way, our easy-to-use proteomics data analysis pipeline enables an investigator to extract insights from data even with limited computational experience.

4.1.2 Extraction of exclusively enriched proteins

The Limma proteomics pipeline allows user to extract list(s) of 'exclusively enriched' proteins (See sec. 2.1.1.1). These *exclusive proteins* are often of prime biological importance. This is a critical aspect of our pipeline since it enables the user to deduce biological implications from the list(s) of *exclusive proteins* via GO and Pathway analyses.

4.1.3 Improved interaction with the large network plot

The MiGENet app developed using forceNetwork++ allows users to search a protein inside large network and extract its immediate neighbors in a separate plot (Fig. 4.1). Moreover, a table listing the interactors of the searched protein with $\log_2$ fold change is also provided as an output (Fig. 3.2). This is an extremely important feature because users can quickly gain knowledge about the biological system. By facilitating effective visualization of complex protein-protein interaction landscapes, forceNetwork++ has enabled researchers understand the network biology of the biological systems (Singh et al., 2020; Lukacs et al., 2021).

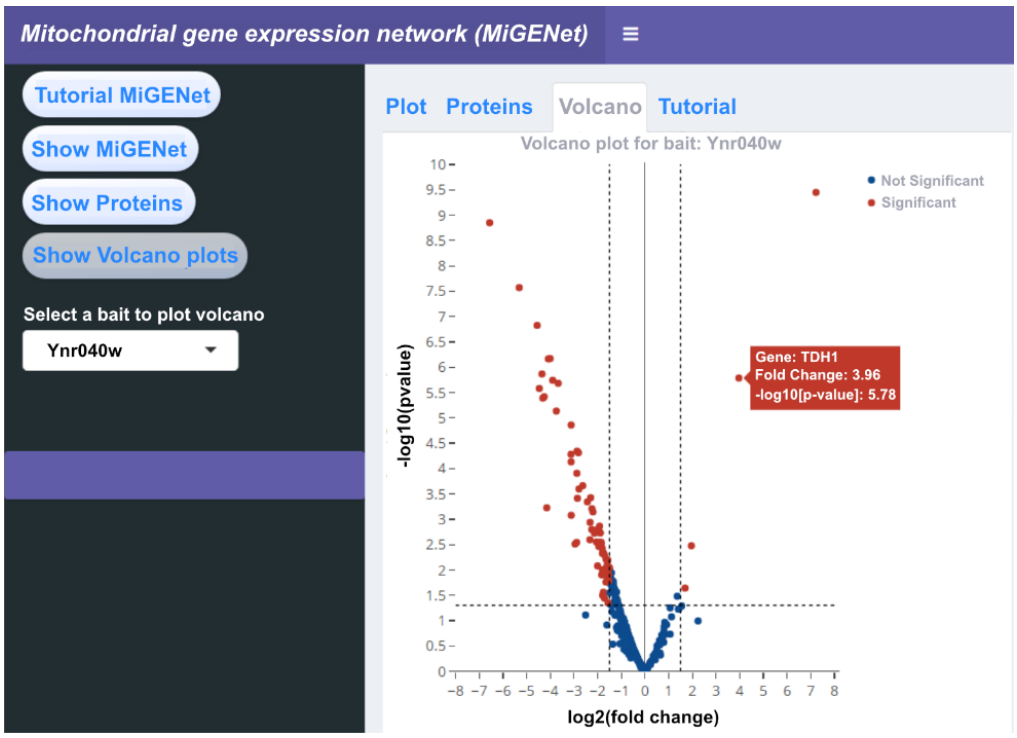


Fig. 4.2: MiGENet offers interactive visualization of volcano plots for each two-group comparison.

4.1.4 Interactive visualization of volcano plots

For every bait, MiGENet app incorporates the volcano plots after Limma based moderated t-test. Therefore, on a single platform, users can combine information from volcano plots and bait-prey interaction network plot, greatly speeding up interpretation and knowledge discovery (Fig. 4.2).

4.1.5 Code reusability

Given that the Limma proteomics pipeline's code for data pre-processing and statistical analysis is written in R, it makes sense to design a graphics system in R. Therefore, to develop an improved graphics system, I choose to reuse codes from networkD3 R package because it uses HTML widgets (Web-page-htmlwidgets, 2021) to render D3 like interactive network in a force-directed layout from R environment. HTML widgets offer a platform to generate R bindings to JavaScript libraries so that calling JavaScript functions within R environment becomes feasible. The benefit of using HTML widgets is that they can render in different context viz. in the shiny apps (Web-page-shiny, 2021), R console and R Markdown. This approach of reusing existing software to create new software is referred to as code reuse or software reuse, and it is one of the best practices in software engineering. Since the source code of forceNetwork++ is written in a modular fashion therefore, it can be reused further in another software. This is what we did when we used forceNetwork++ to build MiGENet app.

Currently, the data analysis pipeline and the visualization software (forceNetwork++) are distributed separately, requiring the user to perform two installations. However, it would be more convenient to have them integrated into a single piece of software. My future outlook is to convert the entire data analysis pipeline into a full-fledged GUI application capable of directly communicating with network visualization software such as Cytoscape. This gives the user the flexibility of visualizing their data in a variety of ways/layouts.

4.2 How *ComplexMiner* will aid in protein complex discovery?

In *ComplexMiner*, we employed one-shot learning which is a subset of machine learning to improve protein complex prediction. The Siamese network based one-shot learning architecture (Fig. 2.5) enables the discovery of protein complexes with fewer datasets. Siamese networks are commonly employed to learn relationships between two comparable entities in several problem domains viz. image recognition, signature verification, paraphrase identification (Bromley et al., 1993; Yin and Schütze, 2015; Koch et al., 2015) etc. However, to the best of my knowledge, it has never been used to solve protein complex prediction problem.

ComplexMiner offers a computational platform with several benefits. Possible uses of the software include exploring complexomics datasets, visualizing discoveries, and passing output to other software for additional analysis. For instance, it can send the network table containing cluster information directly to a Cytoscape session. This way users familiar with Cytoscape can alter the appearance of the network graph and perform additional analysis. In addition, *ComplexMiner* allows user to query a list of proteins in the dataset and visualize their elution profiles. Another striking feature of *ComplexMiner* is the ability to visualize data in the form of an interactive heatmap. With the advancement of MS technologies, now a typical complexomic experiment can quantify 4000-5000 elution profiles across dozens of fractions. Interactive visualization of such a dataset is a real challenge. ComplexMiner tries to solve this by enabling GPU to handle visualization and interaction. User will be able to drag select a cluster from the heatmap and query for PPI in the String database and the novel connections will be shown in different colors than connections found in the String database. In addition to complex discovery, ComplexMiner provides several sanity-checks to test the quality of the dataset. We will integrate *CoreClust*, a standalone command line utility, into ComplexMiner to speed up the discovery of bona fide protein complexes. Because ComplexMiner is still in

the development and testing stages, my primary goal is to make it available as open-source software as soon as possible.

4.3 How ImShot facilitates spatial proteomics?

ImShot is the first software of its kind to provide an end-to-end analysis of diseased vs healthy systems by integrating two orthogonal MS technologies. The software can be used in any two-group comparison i.e., animal models, patient samples etc. allowing user flexibility in terms of experimental context. The software elegantly deals with both the IMS and LC-MS data and integrates them through a conveniently designed GUI that does not require either proteomic or computational expertise to operate.

While dealing with IMS data, ImShot performs a very crucial task of deisotoping the peptide spectra based on spatial data segregation. In absence of deisotoping, the resulting IMS spectra would be biased towards an overestimation of the number of peptide peaks and will also include ambiguous annotations of peptide masses when comparing with LC-MS data. To the best of our knowledge, this is the only software that deisotopes IMS peptide spectra to get rid of false positives. The novel method of ranking of IMS peptides in case of multiple annotations (based on our proposed MLP scoring) associates most likely biological pathways with the most probable areas of the tissue. Computational, experimental and literature based validation of the ranking method has imparted sufficient confidence in our scoring approach, which can now be applied to any type of tissues for two-group comparisons.

The GUI based software for LC-MS data analysis mainly either have been desktop applications running only on Windows platform (Tyanova et al., 2016b; Rigbolt et al., 2011) or web applications (Weiner et al., 2018; Gallant et al., 2020; Efstathiou et al., 2017). Though web apps have many benefits, their stability depends on the state of the server running the application, number of users accessing it, network bandwidth etc. Moreover, often these web apps are written using shiny R package (Weiner et al., 2018; Gallant et al., 2020) which

provides easy to use APIs to render output of R script(s) in web page(s) (al.). However, as the shiny documentation mentioned *"Debugging Shiny applications can be challenging"* (RStudio-Inc., 2021). No software is free from bug. Therefore, effective debugging will not only boost development time but will also enable software developers to be more creative with the design and implementation of their ideas. Generally, breakpoints are employed for debugging a software. A breakpoint is a point in a program where it is intentionally stopped or paused for debugging purposes. However, in shiny apps this is not so flexible as the documentation states: *"Unfortunately, breakpoints aren't helpful in all situations. For technical reasons, breakpoints can only be used inside the shinyServer function. You can't use them in code in other .R files."* (RStudio-Inc., 2021). One can use browser function instead of a breakpoint whenever the code execution needs to stop. This will activate the debugger irrespective of the file containing the command. But the drawback is that the developers must remember to remove the browser function calls every time they want to commit code to a repository. On the other hand, the desktop applications so far have been lacking the aesthetics in the charts and plots and often users need to write additional scripts or use graphics editing software to make publication quality figures. However, in ImShot we have tried to incorporate the best of the two worlds i.e., it produces high quality graphics like a web app and at the same time runs natively on user's computer. Thereby, ImShot GUI allows the user to analyze data in an independent manner free of external influencing factors, viz. internet connection/bandwidth, cloud computing limitations etc.

ImShot feels almost like a native web app that can read and write data besides accessing computer's file system. Moreover, Electron framework saves time by providing a large pool of Application Programming Interfaces (APIs) which the developers can integrate into their desktop apps quite easily. It is an opensource software under MIT license, which therefore practically allows anyone to view and modify its source codes to adapt or extend it to use in

more customized environments. The front and backends of ImShot operate in an independent manner. The major benefit of this approach is that implementation of new features become very time efficient. Other advantages of this mode that can be used by developers for further improvement/customization are code optimization, modularity, faster deployment, and flexibility in switching frameworks. Modular architecture of ImShot's codebase in another critical advantage, where each function performs a specific task. Therefore, the modules are available to use independently. Since ImShot performs lots of statistical computations in the backend, the use of R makes a perfect choice and usage of HTML, CSS, JS in the front-end make the software extremely flexible and feature rich. In addition, it also records the R code runtime which allows a software developer to monitor and optimize (if needed) the backend. ImShot desktop app provides question mark icons with hover effects next to every interface element (input, dropdown box, file upload wizard etc.) to guide users about the meaning of the input making the app quite easy to use. Moreover, ImShot generates the tables, plots and graphs in time efficient manner and they're of publication quality already.

ImShot is freely available in the GitHub (https://github.com/wasimaftab/ImShot) with detailed instructional material on its various use cases. I am also planning to augment many more functionalities based on user feedback in the future releases of ImShot.

List of abbreviations

ANOVA - Analysis of variance

API – Application Programming Interface

AP-MS - Affinity purification coupled with mass spectrometry

AROM+ - Mice overexpressing human P450 aromatase

ATP - Adenosine triphosphate

BioID - Proximity-dependent biotin identification

BN - Blue native gel

BPL - Biotin protein ligase

ClusterONE - Clustering with overlapping neighborhood expansion

COB - cytochrome b

DAG - Directed acyclic graph

DDA - Data dependent acquisition

DIA - Data independent acquisition

DL - deep learning

ECM - Extracellular matrix

ESI - electrospray ionization

FD - Force-directed

GO – Gene ontology

GPU - Graphics processing unit

GUI - Graphical user interface

HIV - Human immunodeficiency virus

HTML - HyperText Markup Language

iBAQ - Intensity based absolute quantification

IEX - Ion exchange

IHC - Immunohistochemical staining

IMS - Imaging mass spectrometry

IP - Immunoprecipitation

iTRAQ - Isobaric tag for relative and absolute quantitation

JS – Java script

JSON - Java script object notation

KEGG - Kyoto encyclopedia of genes and genomes

LBQ - Label based quantification

LFQ - Label free quantification

Limma - Linear models for microarray data

MALDI-IMS – Matrix assisted laser desorption/ionisation imaging mass spectrometry

MCM - Minichromosome maintenance protein complex

MDS - Multidimensional scaling

ML – Machine learning

MLP – Most likely peptide

MQ – MaxQuant

MS – Mass spectrometry

NLP - Natural language processing

OC - Overlap coefficient

OS – Operating system

PDL - Proximity dependent labelling

PNG - Portable Network Graphics

PPI – Protein-protein interaction

RPC - Reversed phase chromatography

SAX - Strong anion exchange

SCX - Strong cation exchange

SEC - Size exclusion chromatography

SILAC - Stable isotope labeling with amino acids in cell culture

SMOTE - Synthetic minority oversampling technique

SNN - Siamese neural network

SRM - Selected Reaction Monitoring

SSE - Sum squared error

SWATH-MS – Sequential window acquisition of all theoretical fragment ion spectra mass spectrometry

TOF – Time of flight

TSV - Tab-separated values

UI - User interface

WT – Wild type

XL – Cross linking

References

ABDALLAH, C., DUMAS-GAUDOT, E., RENAUT, J. & SERGEANT, K. 2012. Gel-based and gel-free quantitative proteomics approaches at a glance. *International journal of plant genomics,* 2012.

ADAM, M., URBANSKI, H. F., GARYFALLOU, V. T., WELSCH, U., KÖHN, F. M., ULLRICH SCHWARZER, J., STRAUSS, L., POUTANEN, M. & MAYERHOFER, A. 2012. High levels of the extracellular matrix proteoglycan decorin are associated with inhibition of testicular function. *International journal of andrology,* 35**,** 550-561.

AFTAB, W. & IMHOF, A. 2021. Discovery of Native Protein Complexes by Liquid Chromatography Followed by Quantitative Mass Spectrometry. *Separation Techniques Applied to Omics Sciences.* Springer.

AL., W. C. E. *shiny: Web Application Framework for R* [Online]. Available: https://cran.r-project.org/web/packages/shiny/index.html [Accessed 2021/07/05].

ALBERTS, D., POTTIER, C., SMARGIASSO, N., BAIWIR, D., MAZZUCCHELLI, G., DELVENNE, P., KRIEGSMANN, M., KAZDAL, D., WARTH, A. & DE PAUW, E. 2018. MALDI imaging-guided microproteomic analyses of heterogeneous breast tumors—a pilot study. *PROTEOMICS–Clinical Applications,* 12**,** 1700062.

ALFANO, M., PEDERZOLI, F., LOCATELLI, I., IPPOLITO, S., LONGHI, E., ZERBI, P., FERRARI, M., BRENDOLAN, A., MONTORSI, F. & DRAGO, D. 2019. Impaired testicular signaling of vitamin A and vitamin K contributes to the aberrant composition of the extracellular matrix in idiopathic germ cell aplasia. *Fertility and sterility,* 111**,** 687-698.

ANDERSON, N. L. & ANDERSON, N. G. 1998. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis,* 19**,** 1853-1861.

BALLUFF, B., HANSELMANN, M. & HEEREN, R. M. A. 2017. Mass spectrometry imaging for the investigation of intratumor heterogeneity. *Advances in cancer research,* 134**,** 201-230.

BANAZADEH, A., PENG, W., VEILLON, L. & MECHREF, Y. 2018. Carbon nanoparticles and graphene nanosheets as MALDI matrices in glycomics: a new approach to improve glycan profiling in biological samples. *Journal of The American Society for Mass Spectrometry,* 29**,** 1892-1900.

BERMINGHAM, M. L., PONG-WONG, R., SPILIOPOULOU, A., HAYWARD, C., RUDAN, I., CAMPBELL, H., WRIGHT, A. F., WILSON, J. F., AGAKOV, F. & NAVARRO, P. 2015. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports,* 5**,** 10312.

BERNHARDT, O. M., SELEVSEK, N., GILLET, L. C., RINNER, O., PICOTTI, P., AEBERSOLD, R. & REITER, L. 2012. Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. *Biognosys. ch*.

BLACKSTOCK, W. P. & WEIR, M. P. 1999. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in biotechnology,* 17**,** 121-127.

BLOHM, P., FRISHMAN, G., SMIALOWSKI, P., GOEBELS, F., WACHINGER, B., RUEPP, A. & FRISHMAN, D. 2013. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic acids research,* 42**,** D396-D400.

BODIS, L. 2007. *Quantification of spectral similarity: towards automatic spectra verification.* ETH Zurich.

BOYLE, E. I., WENG, S., GOLLUB, J., JIN, H., BOTSTEIN, D., CHERRY, J. M. & SHERLOCK, G. 2004. GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics,* 20**,** 3710-3715.

BROMLEY, J., BENTZ, J. W., BOTTOU, L., GUYON, I., LECUN, Y., MOORE, C., SÄCKINGER, E. & SHAH, R. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence,* 7**,** 669-688.

BRUSNIAK, M.-Y., BODENMILLER, B., CAMPBELL, D., COOKE, K., EDDES, J., GARBUTT, A., LAU, H., LETARTE, S., MUELLER, L. N. & SHARMA, V. 2008. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC bioinformatics,* 9**,** 542.

CHANG, C., XU, K., GUO, C., WANG, J., YAN, Q., ZHANG, J., HE, F. & ZHU, Y. 2018. PANDA-view: an easy-to-use tool for statistical analysis and visualization of quantitative proteomics data. *Bioinformatics,* 34**,** 3594-3596.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. & KEGELMEYER, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* 16**,** 321-357.

CHEN, C.-T., WANG, J.-H., CHENG, C.-W., HSU, W.-C., KO, C.-L., CHOONG, W.-K. & SUNG, T.-Y. 2021. Multi-Q 2 software facilitates isobaric labeling quantitation analysis with improved accuracy and coverage. *Scientific reports,* 11**,** 1-12.

CHEN, X., WEI, S., JI, Y., GUO, X. & YANG, F. 2015. Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics,* 15**,** 3175-3192.

CHOI-RHEE, E., SCHULMAN, H. & CRONAN, J. E. 2004. Promiscuous protein biotinylation by Escherichia coli biotin protein ligase. *Protein science,* 13**,** 3043-3050.

CODEJIE. *js-call-r* [Online]. Available: https://www.npmjs.com/package/js-call-r [Accessed 2021/04/24/13:41:16].

COHEN, K. B. & DEMNER-FUSHMAN, D. 2014. *Biomedical natural language processing*, John Benjamins Publishing Company.

CONSORTIUM, G. O. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research,* 32**,** D258-D261.

COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology,* 26**,** 1367-1372.

CROZIER, T. W., TINTI, M., LARANCE, M., LAMOND, A. I. & FERGUSON, M. A. 2017. Prediction of protein complexes in Trypanosoma brucei by protein correlation profiling mass spectrometry and machine learning. *Molecular & Cellular Proteomics,* 16**,** 2254-2267.

DANDEKAR, T., SNEL, B., HUYNEN, M. & BORK, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences,* 23**,** 324-328.

DE GELDER, R., WEHRENS, R. & HAGEMAN, J. A. 2001. A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. *Journal of Computational Chemistry,* 22**,** 273-289.

DE SAN ROMAN, E. G., MANUEL, I., GIRALT, M. T., FERRER, I. & RODRÍGUEZ-PUERTAS, R. 2017. Imaging mass spectrometry (IMS) of cortical lipids from preclinical to severe stages of Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Biomembranes,* 1859**,** 1604-1614.

DISS, G. & LEHNER, B. 2018. The genetic landscape of a physical interaction. *Elife,* 7**,** e32472.

DREZE, M., CHARLOTEAUX, B., MILSTEIN, S., VIDALAIN, P.-O., YILDIRIM, M. A., ZHONG, Q., SVRZIKAPA, N., ROMERO, V., LALOUX, G. & BRASSEUR, R. 2009. 'Edgetic'perturbation of a C. elegans BCL2 ortholog. *Nature methods,* 6**,** 843.

EADES, P. 1984. A heuristic for graph drawing. *Congressus numerantium,* 42**,** 149-160.

EFSTATHIOU, G., ANTONAKIS, A. N., PAVLOPOULOS, G. A., THEODOSIOU, T., DIVANACH, P., TRUDGIAN, D. C., THOMAS, B., PAPANIKOLAOU, N., AIVALIOTIS, M. & ACUTO, O. 2017. ProteoSign: an end-user online differential proteomics statistical analysis platform. *Nucleic acids research,* 45**,** W300-W306.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. & BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences,* 95**,** 14863-14868.

ELECTRON. 2013. *Electron | Build cross-platform desktop apps with JavaScript, HTML, and CSS* [Online]. Available: https://www.electronjs.org/ [Accessed 24th Apr. 2021].

FRANCK, J., EL AYED, M., WISZTORSKI, M., SALZET, M. & FOURNIER, I. 2009. On-tissue N-terminal peptide derivatizations for enhancing protein identification in MALDI mass spectrometric imaging strategies. *Analytical chemistry,* 81**,** 8305-8317.

FRUCHTERMAN, T. M. J. & REINGOLD, E. M. 1991. Graph drawing by force-directed placement. *Software: Practice and experience,* 21**,** 1129-1164.

GALLANT, J. L., HEUNIS, T., SAMPSON, S. L. & BITTER, W. 2020. ProVision: a web-based platform for rapid analysis of proteomics data processed by MaxQuant. *Bioinformatics,* 36**,** 4965-4967.

GANDRUD, C. networkD3 https://github.com/christophergandrud/networkD3 [Accessed on 29th April 2021]

GESSULAT, S., SCHMIDT, T., ZOLG, D. P., SAMARAS, P., SCHNATBAUM, K., ZERWECK, J., KNAUTE, T., RECHENBERGER, J., DELANGHE, B. & HUHMER, A. 2019. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods,* 16**,** 509.

GIURGIU, M., REINHARD, J., BRAUNER, B., DUNGER-KALTENBACH, I., FOBO, G., FRISHMAN, G., MONTRONE, C. & RUEPP, A. 2018. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research,* 47**,** D559-D563.

GROSECLOSE, M. R., ANDERSSON, M., HARDESTY, W. M. & CAPRIOLI, R. M. 2007. Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry. *Journal of Mass Spectrometry,* 42**,** 254-262.

GUO, Y., YU, L., WEN, Z. & LI, M. 2008. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research,* 36**,** 3025-3030.

GURUHARSHA, K., RUAL, J.-F., ZHAI, B., MINTSERIS, J., VAIDYA, P., VAIDYA, N., BEEKMAN, C., WONG, C., RHEE, D. Y. & CENAJ, O. 2011. A protein complex network of Drosophila melanogaster. *Cell,* 147**,** 690-703.

GUSTAFSSON, J. O. R., OEHLER, M. K., RUSZKIEWICZ, A., MCCOLL, S. R. & HOFFMANN, P. 2011. MALDI imaging mass spectrometry (MALDI-IMS)—application of spatial proteomics for ovarian cancer classification and diagnosis. *International journal of molecular sciences,* 12**,** 773-794.

GUYON, I. & ELISSEEFF, A. 2003. An introduction to variable and feature selection. *Journal of machine learning research,* 3**,** 1157-1182.

HANRIEDER, J., LJUNGDAHL, A., FÄLTH, M., MAMMO, S. E., BERGQUIST, J. & ANDERSSON, M. 2011. L-DOPA-induced dyskinesia is associated with regional increase of striatal dynorphin peptides as elucidated by imaging mass spectrometry. *Molecular & Cellular Proteomics,* 10**,** M111. 009308.

HASHEMIFAR, S., NEYSHABUR, B., KHAN, A. A. & XU, J. 2018. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics,* 34**,** i802-i810.

HAVUGIMANA, P. C., HART, G. T., NEPUSZ, T., YANG, H., TURINSKY, A. L., LI, Z., WANG, P. I., BOUTZ, D. R., FONG, V. & PHANSE, S. 2012. A census of human soluble protein complexes. *Cell,* 150**,** 1068-1081.

HENKE, S. K. & CRONAN, J. E. 2014. Successful conversion of the Bacillus subtilis BirA Group II biotin protein ligase into a Group I ligase. *PLoS One,* 9**,** e96757.

HEUSEL, M., BLUDAU, I., ROSENBERGER, G., HAFEN, R., FRANK, M., BANAEI-ESFAHANI, A., VAN DROGEN, A., COLLINS, B. C., GSTAIGER, M. & AEBERSOLD, R. 2019. Complex-centric proteome profiling by SEC-SWATH-MS. *Molecular systems biology,* 15.

HUBER, K., KHAMEHGIR-SILZ, P., SCHRAMM, T., GORSHKOV, V., SPENGLER, B. & RÖMPP, A. 2018. Approaching cellular resolution and reliable identification in mass spectrometry imaging of tryptic peptides. *Analytical and bioanalytical chemistry,* 410**,** 5825-5837.

HUPÉ, P. 2012. *File: Mass spectrometry protocol.svg - Wikimedia Commons* [Online]. Available: https://commons.wikimedia.org/wiki/File:Mass_spectrometry_protocol.svg [Accessed 10th Aug. 2021].

HWANG, S., KIM, C. Y., YANG, S., KIM, E., HART, T., MARCOTTE, E. M. & LEE, I. 2019. HumanNet v2: human gene networks for disease research. *Nucleic acids research,* 47**,** D573-D580.

JEFFERY, C. J. 2018. Protein moonlighting: what is it, and why is it important? *Philosophical Transactions of the Royal Society B: Biological Sciences,* 373**,** 20160523.

JSON. *JavaScript Object Notation* [Online]. Available: https://www.json.org/json-en.html [Accessed 20th Jun 2021].

KAMMERS, K., COLE, R. N., TIENGWE, C. & RUCZINSKI, I. 2015. Detecting significant changes in protein abundance. *EuPA open proteomics,* 7**,** 11-19.

KARPIEVITCH, Y. V., DABNEY, A. R. & SMITH, R. D. 2012. Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics,* 13**,** S5.

KENKEL, B. 2018. *Split-BioID: An Improved Method for Studying Protein-Protein Interactions* [Online]. Available: https://blog.addgene.org/split-bioid-an-improved-method-for-studying-protein-protein-interactions [Accessed 20th Jul. 2021].

KHAN, M., YOUN, J.-Y., GINGRAS, A.-C., SUBRAMANIAM, R. & DESVEAUX, D. 2018. In planta proximity dependent biotin identification (BioID). *Scientific reports,* 8**,** 1-8.

KOCH, G., ZEMEL, R. & SALAKHUTDINOV, R. Siamese neural networks for one-shot image recognition. ICML deep learning workshop, 2015. Lille.

KOCHANOVA, N. Y., SCHAUER, T., MATHIAS, G. P., LUKACS, A., SCHMIDT, A., FLATLEY, A., SCHEPERS, A., THOMAE, A. W. & IMHOF, A. 2020. A multi-layered structure of the interphase chromocenter revealed by proximity-based biotinylation. *Nucleic acids research,* 48**,** 4161-4178.

KRATOCHVÍL, J., PRYSIAZHNYI, V., DYČKA, F., KYLIÁN, O., KÚŠ, P., SEZEMSKÝ, P., ŠTĚRBA, J. & STRAŇÁK, V. 2021. Gas aggregated Ag nanoparticles as the inorganic matrix for laser desorption/ionization mass spectrometry. *Applied Surface Science,* 541**,** 148469.

KRUSCHKE, J. K. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General,* 142**,** 573.

KWON, K. & BECKETT, D. 2000. Function of a conserved sequence motif in biotin holoenzyme synthetases. *Protein Science,* 9**,** 1530-1539.

LAHIRI, S., AFTAB, W., WALENTA, L., STRAUSS, L., POUTANEN, M., MAYERHOFER, A. & IMHOF, A. 2021. MALDI-IMS combined with shotgun proteomics identify and localize new factors in male infertility. *Life Science Alliance,* 4**,** e202000672.

LAMBERT, J.-P., TUCHOLSKA, M., GO, C., KNIGHT, J. D. R. & GINGRAS, A.-C. 2015. Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *Journal of proteomics,* 118**,** 81-94.

LE SAGE, V., CINTI, A., VALIENTE-ECHEVERRÍA, F. & MOULAND, A. J. 2015. Proteomic analysis of HIV-1 Gag interacting partners using proximity-dependent biotinylation. *Virology journal,* 12**,** 1-5.

LEE, I., DATE, S. V., ADAI, A. T. & MARCOTTE, E. M. 2004. A probabilistic functional network of yeast genes. *science,* 306**,** 1555-1558.

LEEK, J. T., JOHNSON, W. E., PARKER, H. S., JAFFE, A. E. & STOREY, J. D. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics,* 28**,** 882-883.

LEEK JT, J. W., PARKER HS, FERTIG EJ, JAFFE AE, ZHANG Y, STOREY JD, TORRES LC ComBat: Adjust for batch effects using an empirical Bayes framework https://rdrr.io/bioc/sva/man/ComBat.html [Accessed on 27th June 2021]

LEITNER, A., WALZTHOENI, T., KAHRAMAN, A., HERZOG, F., RINNER, O., BECK, M. & AEBERSOLD, R. 2010. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular & Cellular Proteomics,* 9**,** 1634-1649.

LONGUESPÉE, R., LY, A., CASADONTE, R., SCHWAMBORN, K., KAZDAL, D., ZGORZELSKI, C., BOLLWEIN, C., KRIEGSMANN, K., WEICHERT, W. & KRIEGSMANN, J. 2019. Identification of MALDI Imaging Proteolytic Peptides Using LC-MS/MS-Based Biomarker Discovery Data: A Proof of Concept. *PROTEOMICS–Clinical Applications,* 13**,** 1800158.

LÖNNSTEDT, I. & SPEED, T. 2002. Replicated microarray data. *Statistica sinica*, 31-46.
LUDWIG, C., GILLET, L., ROSENBERGER, G., AMON, S., COLLINS, B. C. & AEBERSOLD, R. 2018. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular systems biology,* 14.
LUKACS, A., THOMAE, A. W., KRUEGER, P., SCHAUER, T., VENKATASUBRAMANI, A. V., KOCHANOVA, N. Y., AFTAB, W., CHOUDHURY, R., FORNE, I. & IMHOF, A. 2021. The Integrity of the HMR complex is necessary for centromeric binding and reproductive isolation in Drosophila. *PLoS genetics,* 17, e1009744.
MACLEAN, B., TOMAZELA, D. M., SHULMAN, N., CHAMBERS, M., FINNEY, G. L., FREWEN, B., KERN, R., TABB, D. L., LIEBLER, D. C. & MACCOSS, M. J. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* 26, 966-968.
MARCOTTE, E. M., PELLEGRINI, M., NG, H.-L., RICE, D. W., YEATES, T. O. & EISENBERG, D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science,* 285, 751-753.
MATRIX-SCIENCE. 2021. *Mascot* [Online]. Available: http://www.matrixscience.com/ [Accessed 11th Aug. 2021].
MAYER, C., ADAM, M., GLASHAUSER, L., DIETRICH, K., SCHWARZER, J. U., KÖHN, F. M., STRAUSS, L., WELTER, H., POUTANEN, M. & MAYERHOFER, A. 2016. Sterile inflammation as a factor in human male infertility: Involvement of Toll like receptor 2, biglycan and peritubular cells. *Scientific reports,* 6, 1-10.
MEISTERMANN, H., NORRIS, J. L., AERNI, H.-R., CORNETT, D. S., FRIEDLEIN, A., ERSKINE, A. R., AUGUSTIN, A., MUDRY, M. C. D. V., RUEPP, S. & SUTER, L. 2006. Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat. *Molecular & Cellular Proteomics,* 5, 1876-1886.
MORTENSEN, P., GOUW, J. W., OLSEN, J. V., ONG, S.-E., RIGBOLT, K. T., BUNKENBORG, J., COX, J. R., FOSTER, L. J., HECK, A. J. & BLAGOEV, B. 2010. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *Journal of proteome research,* 9, 393-403.
NYGAARD, V., RØDLAND, E. A. & HOVIG, E. 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics,* 17, 29-39.
OPENJS-FOUNDATION. 2021. *Electron apps in different domains* [Online]. Available: https://www.electronjs.org/apps [Accessed 24th Apr. 2021].
OUGHTRED, R., STARK, C., BREITKREUTZ, B.-J., RUST, J., BOUCHER, L., CHANG, C., KOLAS, N., O'DONNELL, L., LEUNG, G. & MCADAM, R. 2018. The BioGRID interaction database: 2019 update. *Nucleic acids research,* 47, D529-D541.
PAGEL, P., KOVAC, S., OESTERHELD, M., BRAUNER, B., DUNGER-KALTENBACH, I., FRISHMAN, G., MONTRONE, C., MARK, P., STÜMPFLEN, V., MEWES, H.-W., RUEPP, A. & FRISHMAN, D. 2004. The MIPS mammalian protein–protein interaction database. *Bioinformatics,* 21, 832-834.
PALMER, A., PHAPALE, P., CHERNYAVSKY, I., LAVIGNE, R., FAY, D., TARASOV, A., KOVALEV, V., FUCHSER, J., NIKOLENKO, S. & PINEAU, C. 2017. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature methods,* 14, 57-60.
PAUKER, V. I., BERTZBACH, L. D., HOHMANN, A., KHEIMAR, A., TEIFKE, J. P., METTENLEITER, T. C., KARGER, A. & KAUFER, B. B. 2019. Imaging Mass Spectrometry and Proteome Analysis of Marek's Disease Virus-Induced Tumors. *Msphere,* 4.
PAZOS, F. & VALENCIA, A. 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering,* 14, 609-614.
PHANSE, S., WAN, C., BORGESON, B., TU, F., DREW, K., CLARK, G., XIONG, X., KAGAN, O., KWAN, J. & BEZGINOV, A. 2016. Proteome-wide dataset supporting the study of ancient metazoan macromolecular complexes. *Data in brief,* 6, 715-721.
PU, S., WONG, J., TURNER, B., CHO, E. & WODAK, S. J. 2008. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research,* 37, 825-831.

RAUSER, S., MARQUARDT, C., BALLUFF, B., DEININGER, S.-O., ALBERS, C., BELAU, E., HARTMER, R., SUCKAU, D., SPECHT, K. & EBERT, M. P. 2010. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of proteome research,* 9**,** 1854-1863.

RIGBOLT, K. T., VANSELOW, J. T. & BLAGOEV, B. 2011. GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Molecular & Cellular Proteomics,* 10.

RITCHIE ME, P. B., WU D, HU Y, LAW CW, SHI W, SMYTH GK removeBatchEffect: Remove Batch Effect https://rdrr.io/bioc/limma/man/removeBatchEffect.html [Accessed on 27th June 2021]

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* 43**,** e47-e47.

ROLLAND, T., TAŞAN, M., CHARLOTEAUX, B., PEVZNER, S. J., ZHONG, Q., SAHNI, N., YI, S., LEMMENS, I., FONTANILLO, C. & MOSCA, R. 2014. A proteome-scale map of the human interactome network. *Cell,* 159**,** 1212-1226.

RÖST, H. L., AEBERSOLD, R. & SCHUBERT, O. T. 2017. Automated SWATH data analysis using targeted extraction of ion chromatograms. *Proteomics.* Springer.

RÖST, H. L., SACHSENBERG, T., AICHE, S., BIELOW, C., WEISSER, H., AICHELER, F., ANDREOTTI, S., EHRLICH, H.-C., GUTENBRUNNER, P. & KENAR, E. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods,* 13**,** 741-748.

ROUX, K. J., KIM, D. I. & BURKE, B. 2013. BioID: a screen for protein-protein interactions. *Current protocols in protein science,* 74**,** 19.23.-1-19.23. 14.

ROUX, K. J., KIM, D. I., RAIDA, M. & BURKE, B. 2012. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of cell biology,* 196**,** 801-810.

RSTUDIO-INC. 2021. *Debugging Shiny applications* [Online]. Available: https://shiny.rstudio.com/articles/debugging.html [Accessed 2021/07/05].

SALVATORI, R., AFTAB, W., FORNE, I., IMHOF, A., OTT, M. & SINGH, A. P. 2020a. Mapping protein networks in yeast mitochondria using proximity-dependent biotin identification coupled to proteomics. *STAR protocols***,** 100219.

SALVATORI, R., KEHREIN, K., SINGH, A. P., AFTAB, W., MÖLLER-HERGT, B. V., FORNE, I., IMHOF, A. & OTT, M. 2020b. Molecular wiring of a mitochondrial translational feedback loop. *Molecular cell,* 77**,** 887-900. e5.

SCHOBER, Y., GUENTHER, S., SPENGLER, B. & RÖMPP, A. 2012. High-resolution matrix-assisted laser desorption/ionization imaging of tryptic peptides from tissue. *Rapid Communications in Mass Spectrometry,* 26**,** 1141-1146.

SCHOPP, I. M., RAMIREZ, C. C. A., DEBELJAK, J., KREIBICH, E., SKRIBBE, M., WILD, K. & BÉTHUNE, J. 2017. Split-BioID a conditional proteomics approach to monitor the composition of spatiotemporally defined protein complexes. *Nature communications,* 8**,** 1-14.

SCHWAMMLE, V., LEÓN, I. R. & JENSEN, O. N. 2013. Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. *Journal of proteome research,* 12**,** 3874-3883.

SCHWANHÄUSSER, B., BUSSE, D., LI, N., DITTMAR, G., SCHUCHHARDT, J., WOLF, J., CHEN, W. & SELBACH, M. 2011. Global quantification of mammalian gene expression control. *Nature,* 473**,** 337.

SCIEX-PEAKVIEW Fast and Accurate Mass Spec Data Interrogation Software https://sciex.com/products/software/peakview-software [Accessed on 18/06/2021]

SCIEX-PROTEINPILOT Identify and Quantify Proteins Faster, With More Confidence https://sciex.com/products/software/proteinpilot-software [Accessed on 17/04/2021]

SCILS SCiLS Lab for imaging mass spectrometry https://scils.de/download/ [Accessed on 20/06/2021]

SEARS, R. M., MAY, D. G. & ROUX, K. J. 2019. BioID as a tool for protein-proximity labeling in living cells. *Enzyme-Mediated Ligation Methods.* Springer.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research,* 13**,** 2498-2504.

SINGH, A. P., SALVATORI, R., AFTAB, W., AUFSCHNAITER, A., CARLSTRÖM, A., FORNE, I., IMHOF, A. & OTT, M. 2020. Molecular connectivity of mitochondrial gene expression and OXPHOS biogenesis. *Molecular Cell,* 79**,** 1051-1065. e10.

SKINNIDER M, F. L. PrInCE: Predicting Interactomes from Co-Elution. R package version 1.8.0. https://www.bioconductor.org/packages/release/bioc/html/PrInCE.html [Accessed on 18/06/2021]

SMYTH, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology,* 3.

SRIHARI, S., YONG, C. H. & WONG, L. 2017. *Computational prediction of protein complexes from protein interaction networks*, Morgan & Claypool.

SUGIURA, Y. & SETOU, M. 2010. Matrix-assisted laser desorption/ionization and nanoparticle-based imaging mass spectrometry for small metabolites: a practical protocol. *Mass Spectrometry Imaging.* Springer.

SUN, T., ZHOU, B., LAI, L. & PEI, J. 2017. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics,* 18**,** 277.

SZKLARCZYK, D., GABLE, A. L., LYON, D., JUNGE, A., WYDER, S., HUERTA-CEPAS, J., SIMONOVIC, M., DONCHEVA, N. T., MORRIS, J. H., BORK, P., JENSEN, L. J. & MERING, CHRISTIAN V. 2018. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research,* 47**,** D607-D613.

TIAN, K., SHAO, M., WANG, Y., GUAN, J. & ZHOU, S. 2016. Boosting compound-protein interaction prediction by deep learning. *Methods,* 110**,** 64-72.

TILLIER, E. R. & CHARLEBOIS, R. L. 2009. The human protein coevolution network. *Genome research,* 19**,** 1861-1871.

TING, L., COWLEY, M. J., HOON, S. L., GUILHAUS, M., RAFTERY, M. J. & CAVICCHIOLI, R. 2009. Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & Cellular Proteomics,* 8**,** 2227-2242.

TISSUE-ATLAS. 2021. *The human proteome - The Human Protein Atlas* [Online]. Available: https://www.proteinatlas.org/humanproteome/tissue [Accessed 2021/04/24/13:43:37].

TIWARY, S., LEVY, R., GUTENBRUNNER, P., SOTO, F. S., PALANIAPPAN, K. K., DEMING, L., BERNDL, M., BRANT, A., CIMERMANCIC, P. & COX, J. 2019. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature methods***,** 1.

TSOU, C.-C., AVTONOMOV, D., LARSEN, B., TUCHOLSKA, M., CHOI, H., GINGRAS, A.-C. & NESVIZHSKII, A. I. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods,* 12**,** 258.

TYANOVA, S., TEMU, T. & COX, J. 2016a. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols,* 11**,** 2301-2319.

TYANOVA, S., TEMU, T., SINITCYN, P., CARLSON, A., HEIN, M. Y., GEIGER, T., MANN, M. & COX, J. 2016b. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods,* 13**,** 731.

VAN OOIJEN, M. P., JONG, V. L., EIJKEMANS, M. J., HECK, A. J., ANDEWEG, A. C., BINAI, N. A. & VAN DEN HAM, H.-J. 2018. Identification of differentially expressed peptides in high-throughput proteomics data. *Briefings in bioinformatics,* 19**,** 971-981.

VARNAITĖ, R. & MACNEILL, S. A. 2016. Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics,* 16**,** 2503-2518.

WANG, Z., ZHANG, X.-C., LE, M. H., XU, D., STACEY, G. & CHENG, J. 2011. A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. *PloS one,* 6**,** e17906.

WASIM AFTAB, A. I. ComplexMiner: Software to explore protein complexes in datasets generated by native LC followed by MS. *Under preperation.*

WEB-PAGE-HTMLWIDGETS. 2021. *htmlwidgets for R* [Online]. Available: https://www.htmlwidgets.org/ [Accessed 29th April 2021 2021].

WEB-PAGE-SHINY. 2021. *R Shiny* [Online]. Available: https://shiny.rstudio.com/ [Accessed 29th April 2021].

WEBB-ROBERTSON, B.-J. M., WIBERG, H. K., MATZKE, M. M., BROWN, J. N., WANG, J., MCDERMOTT, J. E., SMITH, R. D., RODLAND, K. D., METZ, T. O. & POUNDS, J. G. 2015. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research,* 14**,** 1993-2001.

WEINER, A. K., SIDOLI, S., DISKIN, S. J. & GARCIA, B. A. 2018. Graphical interpretation and analysis of proteins and their ontologies (GiaPronto): a one-click graph visualization software for proteomics data sets. *Molecular & Cellular Proteomics,* 17**,** 1426-1431.

YANG, H., NEPUSZ, T. & PACCANARO, A. 2012. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics,* 28**,** 1383-1389.

YIN, W. & SCHÜTZE, H. Convolutional neural network for paraphrase identification. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015. 901-911.

YU, C. & HUANG, L. 2018. Cross-linking mass spectrometry (XL-MS): An emerging technology for interactomics and structural biology. *Analytical chemistry,* 90**,** 144.

YU, G. & HE, Q.-Y. 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems,* 12**,** 477-479.

YU, H., TARDIVO, L., TAM, S., WEINER, E., GEBREAB, F., FAN, C., SVRZIKAPA, N., HIROZANE-KISHIKAWA, T., RIETMAN, E. & YANG, X. 2011. Next-generation sequencing to generate interactome datasets. *Nature methods,* 8**,** 478.

ZIESKE, L. R. 2006. A perspective on the use of iTRAQ™ reagent technology for protein complex and profiling studies. *Journal of experimental botany,* 57**,** 1501-1508.

Appendix A

ELSEVIER LICENSE TERMS AND CONDITIONS

Apr 28, 2021

This Agreement between Ludwig Maximilian University of Munich -- Wasim Aftab ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5057610757079
License date	Apr 28, 2021
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Proteomics
Licensed Content Title	From classical to new generation approaches: An excursus of -omics methods for investigation of protein-protein interaction networks
Licensed Content Author	Ilaria Iacobucci,Vittoria Monaco,Flora Cozzolino,Maria Monti
Licensed Content Date	Jan 6, 2021
Licensed Content Volume	230
Licensed Content Issue	n/a
Licensed Content Pages	1
Start Page	103990
End Page	0
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Computational methods for exploratory analysis of proteomics data
Institution name	Biomedical Center Munich - LMU Munich
Expected presentation date	Nov 2021
Portions	Fig. 1. Practical flowchart for the choice of the more suitable protein complexes isolation strategy based on researcher requirements.
Requestor Location	Biomedical Center Munich - LMU Munich Grosshaderner Str. 9 Planegg-Martinsried Munich, 82152 Germany Attn: Ludwig Maximilian University of Munich
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must

be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions:**

v1.10

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B

Consent to Publish

Series Title:

Proteomics, Metabolomics, Interactomics and Systems Biology

Published under the imprint

Springer

Title of Book/Volume/Conference: Separation techniques Applied to Omics Sciences - From Principles to Relevant Applications

Editor(s) name(s): Ana Valeria Colnaghi Simionato

Title of Contribution:

Author(s) full name(s):

Corresponding Author's name, address, affiliation and e-mail:

When Author is more than one person the expression "Author" as used in this agreement will apply collectively unless otherwise indicated.

§ 1 Rights Granted

Author hereby grants and assigns to **Springer International Publishing AG** (hereinafter called Publisher) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, anthologies, advance printing, reprints or print to order, microfilm editions, audiograms and videograms), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and data networks (e.g. the Internet) for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or search engines, and posting the Contribution on social media accounts closely related to the Work), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. Publisher especially has the right to permit others to use individual illustrations, tables or text quotations and may use the Contribution for advertising purposes. For the purposes of use in electronic forms, Publisher may adjust the Contribution to the respective form of use and include links (e.g. frames or inline-links) or otherwise combine it with other works and/or remove links or combinations with other works provided in the Contribution. For the avoidance of doubt, all provisions of this contract apply regardless of whether the Contribution and/or the Work itself constitutes a database under applicable copyright laws or not.

The copyright in the Contribution shall be vested in the name of Publisher. Author has asserted his/her right(s) to be identified as the originator of this Contribution in all editions and versions of the Work and parts thereof, published in all forms and media. Publisher may take, either in its own name or in that of Author, any necessary steps to protect the rights granted under this Agreement against infringement by third parties. It will have a copyright notice inserted into all editions of the Work according to the provisions of the Universal Copyright Convention (UCC).

The parties acknowledge that there may be no basis for claim of copyright in the United States to a Contribution prepared by an officer or employee of the United States government as part of that person's official duties. If the Contribution was performed under a United States government contract, but Author is not a United States government employee, Publisher grants the United States government royalty-free permission to reproduce all or part of the Contribution and to authorise others to do so for United States government purposes. If the Contribution was prepared or published by or under the direction or control of Her Majesty (i.e., the constitutional monarch of the Commonwealth realm) or any Crown government department, the copyright in the Contribution shall, subject to any agreement with Author, belong to Her Majesty. If Author is an officer or employee of the United States government or of the Crown, reference will be made to this status on the signature page.

§ 2 Rights retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other research colleagues, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the original source of publication is cited according to current citation standards.

§ 3 Warranties

Author agrees, at the request of Publisher, to execute all documents and do all things reasonably required by Publisher in order to confer to Publisher all rights intended to be granted under this Agreement. Author warrants that the Contribution is original except for such excerpts from copyrighted works (including illustrations, tables, animations and text quotations) as may be included with the permission of the copyright holder thereof, in which case(s) Author is required to obtain written permission to the extent necessary and to indicate the precise sources of the excerpts in the manuscript. Author is also requested to store the signed permission forms and to make them available to Publisher if required.

Author warrants that Author is entitled to grant the rights in accordance with Clause 1 "Rights Granted", that Author has not assigned such rights to third parties, that the Contribution has not heretofore been published in whole or in part, that the Contribution contains no libellous or defamatory statements and does not infringe on any copyright, trademark, patent, statutory right or proprietary right of others, including rights obtained through licences; and that Author will indemnify Publisher against any costs, expenses or damages for which Publisher may become liable as a result of any claim which, if true, would constitute a breach by Author of any of Author's representations or warranties in this Agreement.

Author agrees to amend the Contribution to remove any potential obscenity, defamation, libel, malicious falsehood or otherwise unlawful part(s) identified at any time. Any such removal or alteration shall not affect the warranty and indemnity given by Author in this Agreement.

§ 4 Delivery of Contribution and Publication

Author shall deliver the Contribution to the responsible Editor on a date to be agreed upon, electronically in Microsoft Word format or in such form as may be agreed in writing with Publisher. The Contribution shall be in a form acceptable to the Publisher (acting reasonably) and in line with the instructions contained in the guidelines and Author shall provide at the same time, or earlier if the Publisher reasonably requests, any editorial, publicity or other form required by the Publisher.

Publisher will undertake the publication and distribution of the Work in print and electronic form at its own expense and risk within a reasonable time after it has given notice of its acceptance of the Work to Author in writing.

§ 5 Author's Discount for Books and Electronic Access

Author is entitled to purchase for his/her personal use (if ordered directly from Publisher) the Work or other books published by Publisher at a discount of 40% off the list price for as long as there is a contractual arrangement between Author and Publisher and subject to applicable book price regulation.

Resale of such copies or of free copies is not permitted.

Publisher shall provide electronic access to the electronic final published version of the Work on Publisher's Internet portal, currently known as SpringerLink, to Author. Furthermore, Author has the right to download and disseminate single chapters from the electronic final published version of the Work for his/her private and professional non-commercial research and classroom use (e.g. sharing the chapter by mail or in hard copy form with research colleagues for their professional non-commercial research and classroom use, or to use it for presentations or handouts for students). Author is also entitled to use single chapters for the further development of his/her scientific career (e.g. by copying and attaching chapters to an electronic or hard copy job or grant application).

When Author is more than one person each of the co-authors may share single chapters of the Work with other scientists or research colleagues as described above. In each case, Publisher grants the rights to Author under this clause provided that Author has obtained the prior consent of any co-author(s) of the respective chapter.

§ 6 Termination

Either party shall be entitled to terminate this Agreement forthwith by notice in writing to the other party if the other party commits a material breach of the terms of the Agreement which cannot be remedied or, if such breach can be remedied, fails to remedy such breach within 28 days of being given written notice to do so.

On termination of this Agreement in accordance with its terms, all rights and obligations of Publisher and Author under this Agreement will cease immediately, except that any terms of this Agreement that expressly or by implication survive termination of this Agreement shall remain in full force and effect.

§ 7 Governing Law and Jurisdiction

If any difference shall arise between Author and Publisher concerning the meaning of this Agreement or the rights and liabilities of the parties, the parties shall engage in good faith discussions to attempt to seek a mutually satisfactory resolution of the dispute. This agreement shall be governed by, and shall be construed in accordance with, the laws of Switzerland. The courts of Zug, Switzerland shall have the exclusive jurisdiction.

Corresponding Author signs for and accepts responsibility for releasing this material on behalf of any and all Co-Authors.

Signature of Corresponding Author: **Date:**

.....

.....

☐ I'm an employee of the US Government and transfer the rights to the extent transferable (Title 17 §105 U.S.C. applies)
☐ I'm an employee of the Crown and copyright on the Contribution belongs to Her Majesty

For internal use only:

Order Number: 87011426

GPU/PD/PS: 2/28/643

Legal Entity Number: 1128 Springer International Publishing AG

Springer-C-CTP-05/2016

Appendix C

Author rights

The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): 1. Extend an article to book length 2. Include an article in a subsequent compilation of their own work 3. Re-use portions, excerpts, and their own figures or tables in other works.	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): 1. In their own classroom teaching. Electronic and physical distribution of copies is permitted 2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees 3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use 4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration 5. Include in a thesis or dissertation (provided this is not published commercially) 6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

Acknowledgements

In my life, I got the opportunity to interact with a diverse group of fascinating people who have enriched me with their kindness, sacrifices, views, and experiences. They have all been instrumental in bringing me here. The most challenging part of writing my thesis has been thanking them in few lines. I am blessed to have outstanding professional mentors and a beautiful family, friends, and wife. I would like to take this opportunity to express my gratitude to all of them:

To begin, I want to express my profound gratitude to Prof. Dr. Axel Imhof for enabling me to work on my PhD in his group at BMC, LMU. Without his meticulous scrutiny, scientific direction, and assistance, I couldn't have completed this scientific endeavor.

I am grateful to Dr. Shibojyoti Lahiri for inspiring discussions and helpful ideas that aided in the completion of my PhD studies. I appreciate your time and effort for carefully reading my thesis.

I am indebted to Edith Müller and Caroline Brieger for their constant support with administrative issues, which enabled me to concentrate on my research.

I am also thankful to Dr. Teresa Barth for correcting the summary (in German) of my thesis, Dr. Abeer Singh, Dr. Ignasi Forne, Prof. Dr. Martin Ott, Dr. Roger Salvatory for the fruitful collaborations and discussions, Dr. Andreas Schmidt for interesting discussions and having helped me learn about proteomics, Andrea Lukacs, Dr. Andreas Schmidt, Marc Wirth, Dr. Ignasi Forne for performing experiments for the complexomics project, the members of the Imhof group and the Department of Molecular Biology at BMC, Munich for providing a great research environment.

I am obliged to QBM graduate school and its faculty and staff for their support and funding.

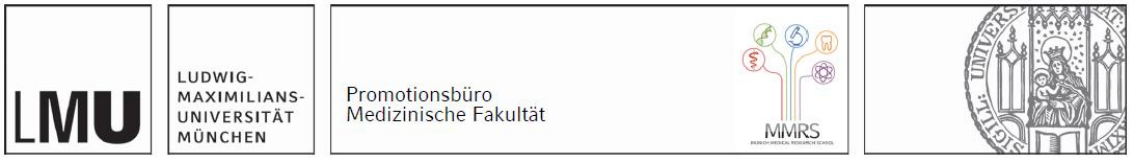
Most importantly I would like to thank my parents for encouraging me to pursue higher education. And I am sure that you would have been immensely proud of me. I am grateful for your encouragement to explore things. Thank you for instilling in me the ability to reason rationally. Thank you for all the sacrifices you have made for me. Both of your lives have taught me the distinction between love and attachment, which has aided me in resolving several issues throughout my life.

I appreciate my siblings for their support, help and efforts. And I am grateful to you for believing in me.

I want to express my gratitude to my wife for understanding me, for never failing to encourage me, and for seeing the positive in everything and I would like to thank my daughter, who joined us during a tough period of my PhD, for giving me unlimited happiness and pleasure. I could never have hoped for more.

Finally, I am indebted to my friends for their encouragement and stress-busting conversations during difficult phases of my life.

Affidavit



Affidavit

Aftab, Wasim
_____

Surname, first name

_____

Street

_____

Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Computational methods for exploratory analysis of proteomics data

is my own work. I have only used the sources indicated and have not made unauthorized use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

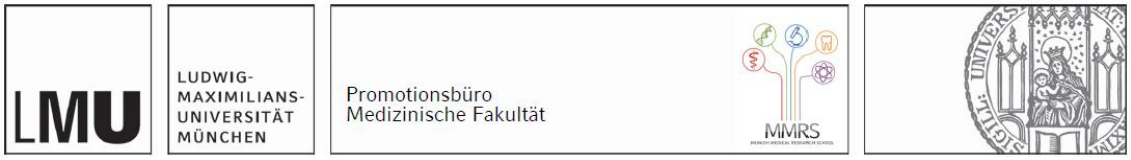
München, 19th July 2022
_____

place, date

Wasim Aftab
_____

Signature doctoral candidate

Confirmation of congruency



Confirmation of congruency between printed and electronic version of the doctoral thesis

Aftab, Wasim

_____

Surname, first name

_____

Street

_____

Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Computational methods for exploratory analysis of proteomics data

is congruent with the printed version both in content and format.

München, 19th July 2021 Wasim Aftab

_____ _____

place, date Signature doctoral candidate

Curriculum vitae

Name: Wasim Aftab

Professional Summary

I'm passionate about solving real world problems using computer. I've software engineering and data analysis experience in a variety of fields ranging from quantum chemistry to astrophysics, electrical engineering, and proteomics. During my PhD, I focused on computational methods to facilitate proteomic data analysis, and the results were published in premier journals in the bioscience domain.

Work History

- Wissenschaftlicher mitarbeiter at Biomedical Center Munich, Germany (Sep. 2018-Present)
- Senior software engineer at iCelero Technologies Pvt. Ltd., Bangalore, India (Apr. 2012 – Dec. 2012)
- Research Assistant at JNCASR (Aug 2009-Mar 2012)

LinkedIn Page

https://www.linkedin.com/in/wasim-aftab-78a38016/

GitHub Page

https://github.com/wasimaftab

Education

- **PhD** in Bioinformatics, Biomedical Center Munich, LMU of Munich, Germany (Sep 2015 – Present)
- **M.Sc.** in Computer Engineering, Department of ECE, KAU, Saudi Arabia (Jan 2013 – Jun 2015)
- **B.E.** in Computer Science and Engineering, VTU, Belgaum, Karnataka, India (Sep 2005 – Jul 2009)

Selected Publications

- **AFTAB, W.**, LAHIRI, S. & IMHOF, A. 2022. ImShot: An open-source software for probabilistic identification of proteins in situ and visualization of proteomics data. Molecular & Cellular Proteomics, 100242.
- **AFTAB, W.** & IMHOF, A. 2021. Discovery of Native Protein Complexes by Liquid Chromatography Followed by Quantitative Mass Spectrometry. Separation Techniques Applied to Omics Sciences. Springer.
- SALVATORI, R., **AFTAB, W.**, FORNE, I., IMHOF, A., OTT, M. & SINGH, A. P. 2020. Mapping protein networks in yeast mitochondria using proximity-dependent biotin identification coupled to proteomics. STAR protocols, 100219.

Conferences & Workshops:

- Presented poster in *FEBS Workshop on Chromatin Proteomics, Crete, Greece 3-8 Oct. 2016*
- Presented poster in Quantitative Proteomics: Strategies and Tools to Probe Biology EMBL Heidelberg, Germany, 5 - 10 May 2019

Honors and Awards

- Graduate studies scholarship at KAU, Jeddah, Saudi Arabia (Jan 2013-Jun 2015)
- Fellowship from DFG to pursue Dr. rer. nat. in LMU, Munich (Sep. 2015-Aug. 2018)

List of publications

1. **AFTAB, W.**, LAHIRI, S. & IMHOF, A. 2022. ImShot: An open-source software for probabilistic identification of proteins in situ and visualization of proteomics data. Molecular & Cellular Proteomics, 100242.

2. **AFTAB, W.** & IMHOF, A. 2021. Discovery of Native Protein Complexes by Liquid Chromatography Followed by Quantitative Mass Spectrometry. Separation Techniques Applied to Omics Sciences. Springer.

3. SALVATORI, R., **AFTAB, W.**, FORNE, I., IMHOF, A., OTT, M. & SINGH, A. P. 2020. Mapping protein networks in yeast mitochondria using proximity-dependent biotin identification coupled to proteomics. STAR protocols, 100219.

4. LAHIRI, S., **AFTAB, W.**, WALENTA, L., STRAUSS, L., POUTANEN, M., MAYERHOFER, A. & IMHOF, A. 2021. MALDI-IMS combined with shotgun proteomics identify and localize new factors in male infertility. Life Science Alliance, 4, e202000672.

5. SINGH, A. P., SALVATORI, R., **AFTAB, W.**, AUFSCHNAITER, A., CARLSTRÖM, A., FORNE, I., IMHOF, A. & OTT, M. 2020. Molecular connectivity of mitochondrial gene expression and OXPHOS biogenesis. Molecular Cell, 79, 1051-1065. e10.

6. SALVATORI, R., KEHREIN, K., SINGH, A. P., **AFTAB, W.**, MÖLLER-HERGT, B. V., FORNE, I., IMHOF, A. & OTT, M. 2020. Molecular wiring of a mitochondrial translational feedback loop. Molecular cell, 77, 887-900. e5.

7. LUKACS, A., THOMAE, A. W., KRUEGER, P., SCHAUER, T., VENKATASUBRAMANI, A. V., KOCHANOVA, N. Y., **AFTAB, W.**, CHOUDHURY, R., FORNE, I. & IMHOF, A. 2021. The Integrity of the HMR complex is necessary for centromeric binding and reproductive isolation in Drosophila. PLoS genetics, 17, e1009744.

8. BASCH, M., WAGNER, M., ROLLAND, S., CARBONELL, A., ZENG, R., KHOSRAVI, S., SCHMIDT, A., **AFTAB, W.**, IMHOF, A. & WAGENER, J. 2020. Msp1 cooperates with the proteasome for extraction of arrested mitochondrial import intermediates. Molecular biology of the cell, 31, 753-767.