Gerard Font Juvanteny

Leveraging the power of single-cell proteomics to understand the molecular heterogeneity of cell populations

MASTER'S THESIS

supervised by Dr. Joan T. Matamalas, Dr. Sasha A. Singh, Dr. Sarvesh Chelvanambi and Dr. Masanori Aikawa

Master's Degree in Biomedical Data Science

Tarragona, 2025

Dr. Joan T. Matamalas, certifies that the student Gerard Font Juvanteny has elaborated the work under his direction and he authorizes the presentation of this Master's Thesis for its evaluation.

Advisor signature:

Index

ABSTRACT

This master's thesis explores the evolving field of single-cell proteomics, with its challenges, opportunities, and potential for advancing our understanding of cellular biology and heterogeneity.

The project is centered on the study of macrophages, a cell type extensively researched for their critical roles in immune defense and inflammation, making them vital in research. In addition, this work supports a collaborative effort with a specialized research group at the Center for Interdisciplinary Cardiovascular Sciences (CICS) at Brigham and Women's Hospital, which is affiliated with Harvard Medical School.

Initially, it reviews the current state of single-cell proteomics, detailing each step from sample preparation to data analysis methods. A comprehensive meta-analysis is then performed using three bulk datasets and one single-cell dataset, with the focus on the biology of macrophages (human cell line - THP1 derived macrophages and mouse primary Bone marrow derived macrophages). The primary aim of the meta-analysis is to compare the findings from these studies, identifying and quantifying relevant proteins in both bulk and single-cell experiments.

Finally, the thesis explores the heterogeneity of single cells within populations of the same cell type to uncover the key proteins and pathways driving cellular variability. The analysis aims to provide information on unique protein expression profiles at the individual cell level to improve our understanding of cellular behavior.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the unconditional support and encouragement of numerous people.

First and foremost, I would like to express my sincere thanks to my family. To my wife, Sandra, your infinite patience, understanding and support made this journey possible. You were always there for me. To my daughters, Berta and Carla, I know my dedication to my studies meant I couldn't always be there for you as much as I wanted. Despite this, you recognized the importance of this journey for me and gave me your constant love and support. This achievement is as much yours as it is mine.

I am also very grateful to my fellow students in the master's program, whose camaraderie and support in difficult times (and laughs in the fun times) have enriched my academic experience. Their ideas, discussions, contributions and encouragement have been invaluable.

A very special thank you to Professors Joan T. Matamalas, Sasha A. Singh, Sarvesh Chelvanambi and Massanori Aikawa for the unique opportunity you offered me with the project for my thesis. Your guidance, expertise and willingness to collaborate have been essential in shaping this thesis. Working with you has been a privilege, and your support has made this challenging but rewarding journey possible.

And finally, I would also like to highlight the support of my work colleagues for their patience during periods of heavy workload during the program.

Thank you all for your continued support and make this possible.

LIST OF ACRONYMS/ABBREVIATIONS

Acronym/Abbreviation	Definition
BMDM	Bone Marrow Derived Macrophages
DDA	Data Dependent Acquisition
DEA	Differential Expression Analysis
DIA	Data Independent Acquisition
DNA	Deoxyribonucleic Acid
ES	Enrichment Score
FCS	Functional Class Scoring
FDR	False Discovery Rate
GO	Gene Ontology
GSEA	Gene Set Expression Analysis
HPG	Homopropargylglycine
HPLC	High Performance Liquid Chromatography
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid Chromatography Mass Spectrometry
LF-DIA	Label Free Data Independent Acquisition
mRNA	Messenger RNA
MS	Mass Spectrometry
MS/MS ***(MS2)***	Tandem Mass Spectrometry
NES	Normalized Enrichment Score
nPOP	Nano-ProteOmic sample Preparation
ORA	Over-Representation Analysis
PEA	Pathway Expression Analysis
plexDIA	Parallelized Single-Cell Proteomics by Data-Independent Acquisition
pSCoPE	Prioritized Single Cell Proteomics
PSMs	Peptide-Spectrum Matches
PTM	Post-Transactional Modifications
RNA	Ribonucleic Acid
SCoPE-MS	Single Cell ProtEomics by Mass Spectrometry
SCoPE-MS2	Single Cell ProtEomics by Mass Spectrometry (second version)
SCP	Single-Cell Proteomics
THP-1	Human monocytic cell line derived from an acute monocytic leukemia patient
TMT	Tandem Mass Tag
UMAP	Uniform Manifold Approximation and Projection

LIST OF FIGURES

LIST OF TABLES

BACKGROUND AND MOTIVATION (Introduction)

For decades, scientists and researchers have been studying cells, especially the molecules that not only make up the genetic material stored inside these cells but also those that play a fundamental role in the entire life cycle of organisms.

The first steps in the study of genetic material date back to the 19th century, when Johann Friedrich Miescher, a Swiss biologist, carried out a series of experiments that inspired other researchers who would later described the components of DNA (deoxyribonucleic acid), its shape, and its function in the transmission genetic information[1].

In 1957, Francis Crick published an article[2][3] in which he spoke for the first time about the Central Dogma of Molecular Biology. This was a very remarkable discovery as it explains the flow of genetic information from DNA to RNA (ribonucleic acid) and to proteins within a biological system. Since DNA contains the blueprint of life, it must be well preserved and protected by the cell. RNA is similar to DNA but differs in that it is a transcribed copy of DNA genetic information. This copy and its information are translated by the cell in the form of a protein, that executes this blueprint information.

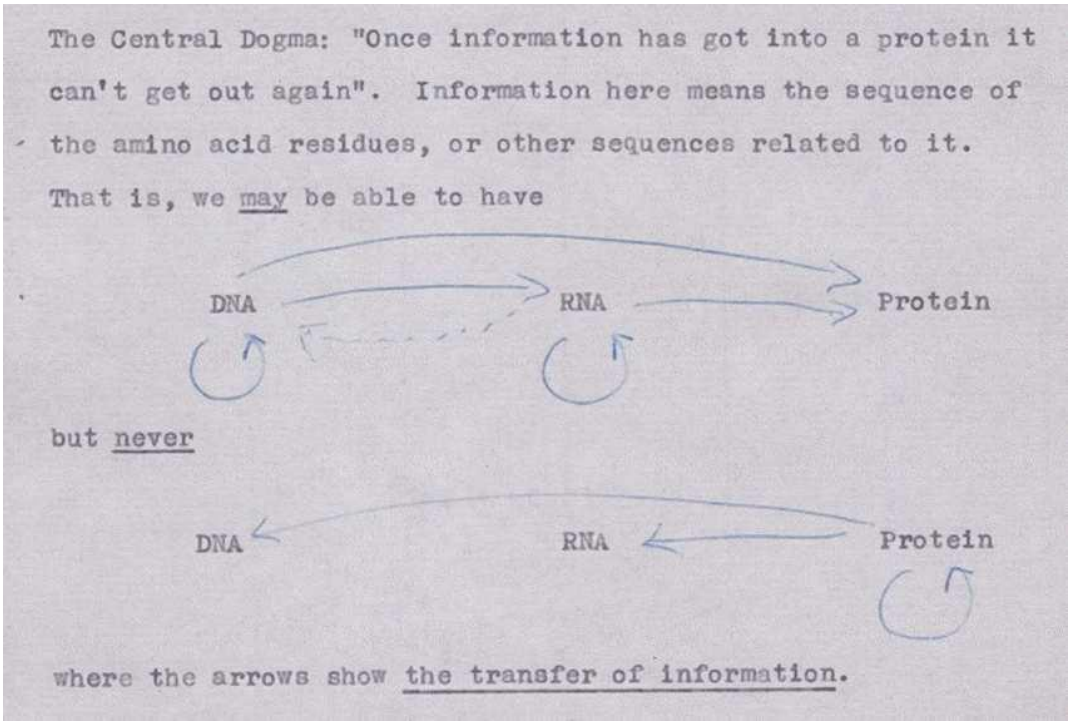


Figure 1 - Crick's first outline of the central dogma (Credit: Welcome Library, London)

The in-depth study of DNA and its genes, RNA, and proteins has led to the emergence of different specialized disciplines within biology. These disciplines explore the intricacies of these molecules and their interactions, tracing the central dogma of molecular biology from DNA to RNA to proteins. The so-called Omics[4] sciences encompass a number of disciplines, each of which focuses

on a specific set of biological molecules with the suffix "-omics" added to the name of the molecule. For example, genomics studies DNA, transcriptomics studies RNA, and proteomics studies proteins, but there are several others such as epigenomics, metabolomics and lipidomics.

By studying and analyzing each of these layers and their molecules, researchers can gain valuable insights into biological processes, health, and disease states and more.

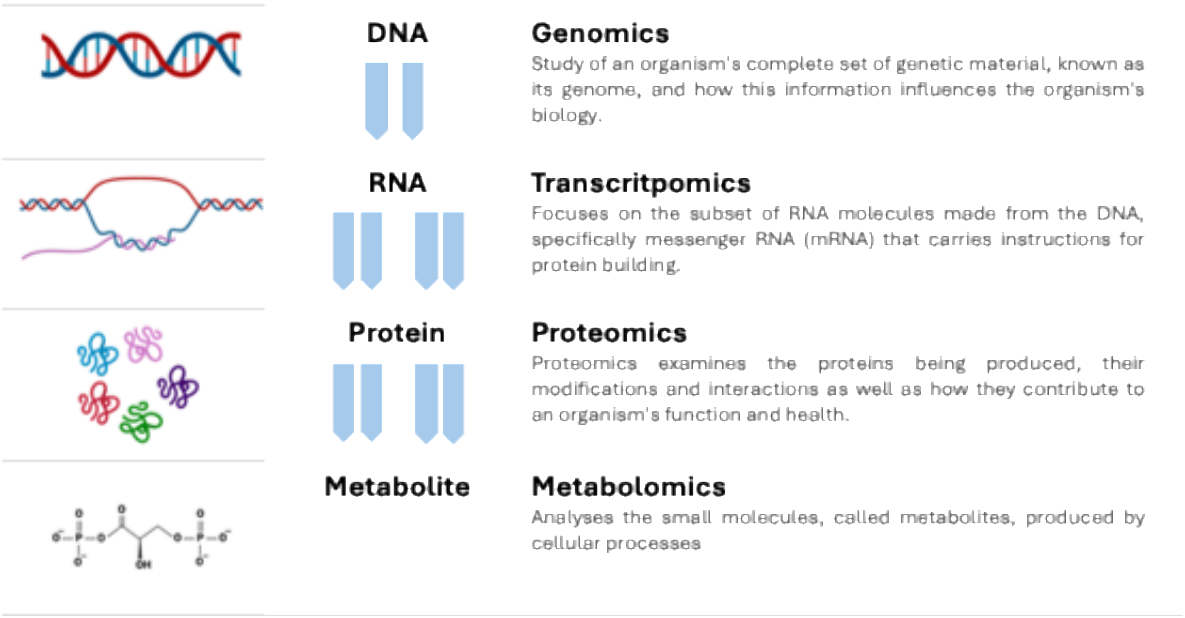


Figure 2 – Illustration of the hierarchy of biological information, from genes (genomics) to multiple mRNA transcripts (transcriptomics), resulting in multiple protein isoforms (proteomics), and small molecules (metabolomics). This mechanism underscores why the genome alone is not enough to account for biological diversity.

The study of molecules in each of these stages offers researchers the possibility of answering different questions. The study of DNA, which makes up the genetic material, allows us to understand, among other things, how variations and mutations in this genetic material affect the development of different diseases and pathologies. Although very interesting, genomics alone does not allow researchers to answer other biological questions that they face.

With genomics, researchers can answer some biological questions such as:

- What is an organism's genetic makeup (DNA sequence)?
- Are there genetic mutations associated with diseases?
- Are there genetic variations that influence individual responses to medications?

Genomics has a wide range of applications in biology and medicine. These are some examples:

- Genomic data can identify genetic predispositions to diseases, enabling early intervention and preventive measures.
- It can help tailor treatments based on an individual's genetic profile.
- Replacing or repairing defective genes to treat diseases.

To go further and to be able to answer additional biological questions, researchers also investigate how the different genes found in DNA are expressed to transcribe messenger RNA (mRNA) that in turn direct the cell to translate proteins. The science that studies RNA is therefore referred to as transcriptomics. Transcriptomics is based on the detection and study of mRNA found inside cells. As a result, researchers can answer questions such as:

- What genes are expressed in a particular cell or tissue?
- How does gene expression change in response to different stimuli?
- How do gene expression patterns differ between healthy and diseased cells?

Transcriptomics also has lots of potential applications in biology and medicine. For example, it can be used to:

- Identify new drug targets.
- Develop diagnostic tests for diseases.
- Monitor the response to treatment.
- Understand the molecular basis of disease.

By studying the transcriptome, researchers can gain a deeper understanding of how genes work and how they are regulated. This knowledge can be used to develop new treatments for diseases and to improve our understanding of human health and biology.

Like genomics, transcriptomics has limitations when it comes to answer biological questions. One reason is that there is no fixed relationship between the amount of mRNA found and the proteins that are ultimately generated. Another reason is that post-translational modifications (PTMs) can modify protein structure and function and these PTM are not captured with transcriptomics.

While genomics unveils the genetic blueprint and transcriptomics reveals gene expression, they have limitations in fully explaining cellular function. Proteomics studies bridge this gap by directly analyzing the proteome, offering insights into protein function, modifications, and interactions, providing a more comprehensive understanding of biological processes. These are some examples of the questions that proteomics studies aim to answer:

- What proteins are in a cell?
- What are the different protein isoforms (proteoforms) present in a cell?
- How much of each protein is present (abundance)?
- What are the specific functions of each protein?
- How do proteins interact with each other to form complexes?
- How do changes in protein abundance or function contribute to disease?

The questions that researchers can answer by studying each layer of the genome are distinct. The challenges, techniques, and unknowns associated with studying each layer of the genome are also distinct. The different molecules studied in each of these layers have different characteristics and environments. For example, DNA is a double-stranded (deoxyribonucleic acid) molecule that is contained within the nucleus of cells. RNA is a single-stranded (ribonucleic acid) molecule that is found in the nucleus and cytoplasm of cells. Proteins (polymers comprised of amino acids) are also

found in both the nucleus and the cytoplasm of cells. The different chemical structures and environments of these molecules make it challenging to study them using the same techniques.

Researchers approach these studies with two different methods, bulk and single-cell. With the bulk method, researchers analyze the average signals from thousands or millions of cells, providing a broad overview of the DNA, RNA, and proteins within a sample. This method is valuable for identifying general trends and patterns, but it masks the variability between individual cells. However, single-cell techniques allow researchers to examine each cell's unique characteristics. In single-cell epigenomics, DNA analysis reveals the regulatory states of individual cells, shedding light on gene accessibility and expression control. RNA transcriptomics at the single-cell level uncovers the specific gene expression profiles, showing how individual cells respond to different conditions. Finally, proteomics in single cells elucidates the functional state by detailing the abundance and modifications of proteins, capturing the heterogeneity within the same cell type.

Proteomics

Proteomics is the study of proteins, their varieties, quantities, functionalities, the relationship between them, and the relationship to certain diseases. Considering that proteins are one of the most important molecules that perform vital and biological functions in living cells, their study and understanding is of utmost importance.

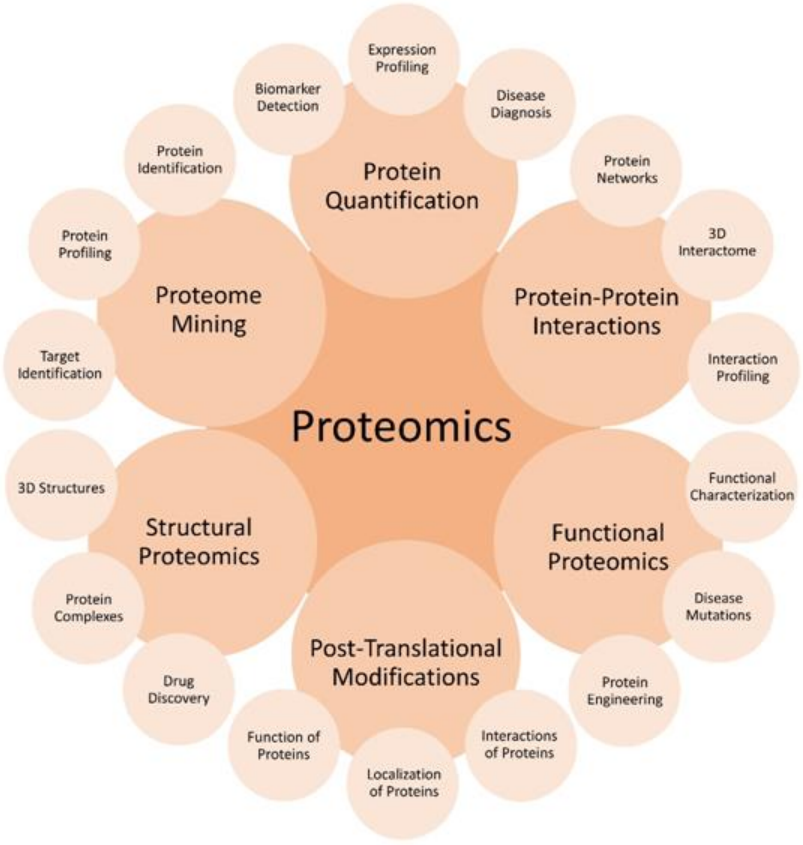


Figure 3 - Proteomics dissects protein function through various techniques. This figure explores key areas: identifying proteins (protein identification), measuring protein abundance (quantification), uncovering protein interactions (protein-protein interactions), analyzing protein structure (structural proteomics), etc… [5]

However, the study of the proteome involves some technical challenges that are more complex to solve than those experienced in the study of the genome or transcriptome. Some of reasons why proteomics requires to overcome some additional challenges are:

- Proteins have a more complex structure that needs to be maintained during the analysis.
- Proteins cannot be amplified like DNA.
- Proteomes are complex and display a wide dynamic range of proteins concentration [6]. More abundant proteins mask less abundant proteins, and the processes to remove them also deplete low abundant protein targets.
- Processing mass spectrometry-enabled proteomic data is a complex process [7].
- Proteins can be modified in a variety of ways, which can further complicate their study.
- Proteins have a specific function. They can interact independently or with other proteins by creating a protein complex. Proteins can be involved in multiple biological pathways.

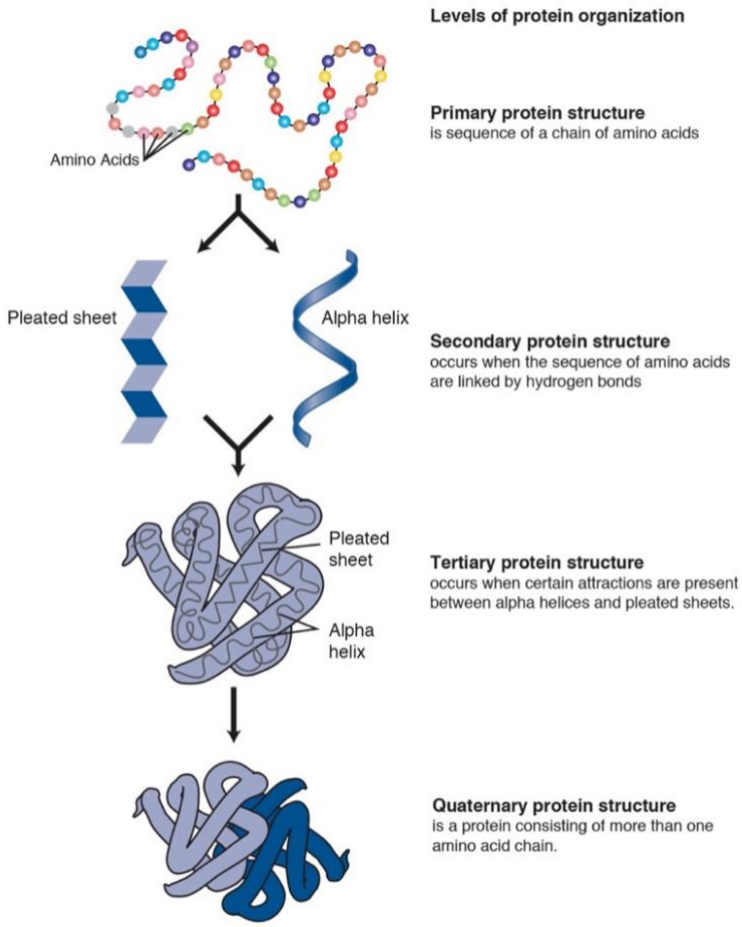


Figure 4 - Proteins fold into specific three-dimensional shapes, and these shapes are determined by four levels of protein structure (Protein structure by the National Human Genome Research Institute is in the public domain).

It is estimated that humans have between 30,000 and 40,000 genes potentially encoding different proteins, but considering alternative RNA splicing and PTMs, this number might be increased up to 2,000,000 proteins [8].

Proteomics is also important because it allows us to understand how PTMs and degradation affect the function and abundance of proteins, something that cannot be inferred from the study of the genome or transcriptome.

Bulk and Single-Cell Proteomics

Due to the inherent challenges of proteomics, scientists initially focused on studying the protein material that could be found in a group of cells. These materials have been studied using different techniques to generate proteomic profiles. However, the study of bulk cellular material, routinely used to investigate tissues, has some drawbacks since the resulting readout is an average of a likely heterogeneous population of cells and masks the unique contributions of different cell types. Understanding the differences between cell types, such as smooth muscle cells or macrophages, is crucial for identifying their specific roles in disease phenotypes.

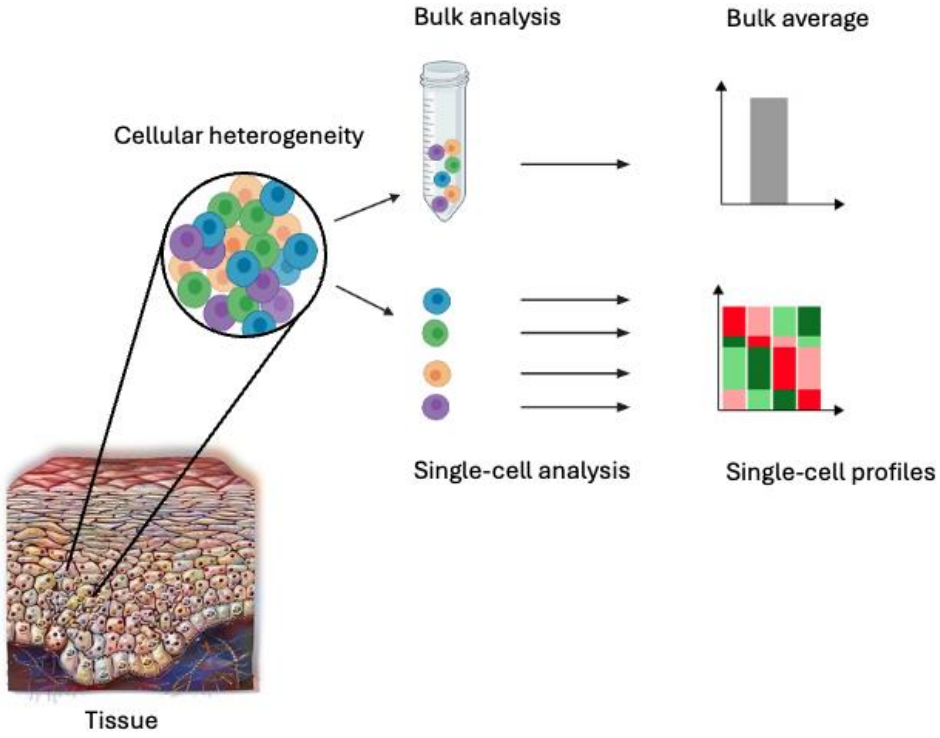


Figure 5 - Single-cell proteomics reveals the hidden diversity of protein abundance within individual cells, overcoming the limitations of bulk analysis that averages cellular heterogeneity (adapted from [9])

This limitation has led to the development of single-cell proteomics (SCP), which allows the study of the proteome of individual cells. This approach has the potential to provide much more detailed and specific information about the proteins that are present in different cell types.

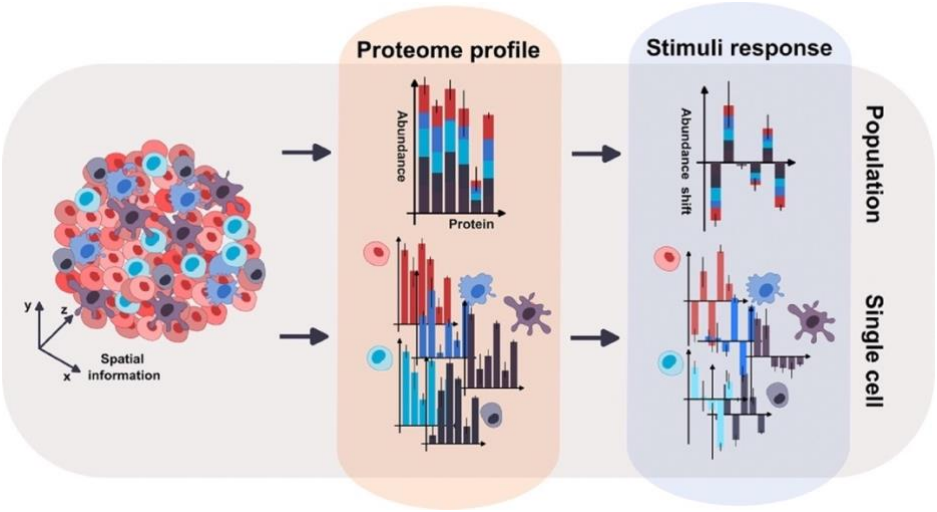


Figure 6 - Population (bulk) vs. single-cell tumor proteome resolution [10]

Single-cell proteomics provides precise answers to biological questions that could not be answered by bulk proteomics, for example, the proteins present in each cell type, the quantities, modifications, how these proteins are degraded, etc. [11]

Although the technologies used for single-cell epigenomics and single-cell transcriptomics provide are well established and provide valuable results, they have limitations for single-cell proteomic analysis. Similarly, the technologies traditionally employed for bulk proteomics also have numerous limitations, which makes single-cell analysis a challenge for researchers.

Spatial Proteomics

Spatial proteomics is a field that is advancing rapidly and that aims to enable the analysis of proteins within a tissue sample, down to the level of individual cells and even compartments within those cells.

- Tissue level: studying proteins present in tissues involve homogenizing them. Spatial proteomics allow to analyze how proteins are distributed across the tissue, revealing functional differences between regions. This is important for understanding tissue development, function, and disease.

- Cellular level: Within a cell, different proteins localize to specific compartments like the nucleus or membrane. Spatial proteomics enables the visualization of these locations, providing a much clearer picture of how proteins interact and carry out their functions.

- Subcellular level: Proteins can move around within compartments and spatial proteomics can track these movements, providing insights into cellular dynamics and processes like signaling and protein trafficking.

By combining microscopic imaging data with ultra-high-sensitivity proteomics, [12] it is now possible to analyze the spatial distribution of proteins at tissue, cell and subcellular levels [13].

Single-Cell Proteomics and Current Challenges

Most of the challenges that scientists and researchers are facing in the field of single-cell analysis are shared also with bulk sample analysis. However, some of them might be amplified in single-cell proteomics due to the limitation on the amount of available material. Although technologies and methods are evolving rapidly, there is still a long way to go before single-cell proteomics analysis can be developed at scale.

These are the most relevant challenges that proteomics in general, and single-cell proteomics in particular need to overcome:

Dynamic Range

Mass spectrometry has been used in proteomics since the 1970s [14]. Today, liquid chromatography mass spectrometry (LC-MS) is the most common approach for proteomics. In LC-MS, proteins or peptides (digested proteins) are separated using chromatography (usually "reverse-phase" chromatography) as they enter the mass spectrometer for subsequent analysis, or in other words, amino acid sequencing. However, the ability of a mass spectrometer to analyze a complex mixture relies in part on its dynamic range. This range refers to the widest spread of protein concentrations, from very low to very high, that the instrument can accurately detect at the same time. However, both the complexity of the sample and the type of mass spectrometer itself limit this dynamic range. This means that some proteins, especially those present in smaller amounts, might be entirely missed, or show up very weakly when surrounded by abundant proteins.

A single mammalian cell might contain 10,000 different types of proteins and the total number of individual protein molecules can be as high as few billions (10^9). However, the number of copies of each protein species varies greatly, ranging from just a hundred (10^2) to tens of millions (10^7) per cell [15].

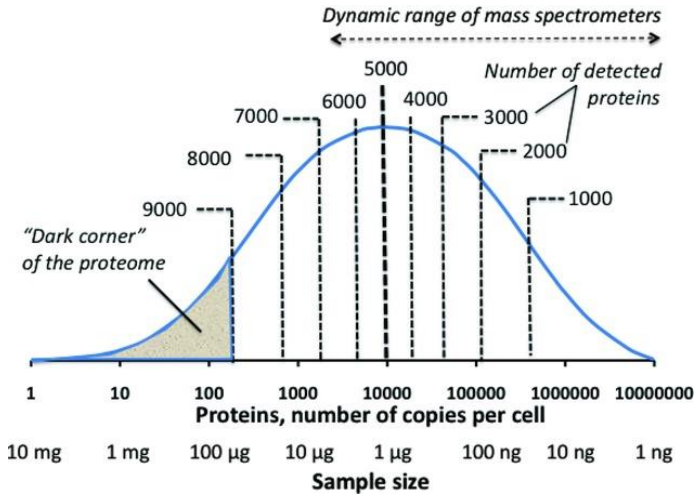


Figure 7 - Distribution of protein abundances is a bell-shape curve on a logarithmic copy number scale [13]

Some of the existing MS-based methods have quantified proteins with a copy number as low as 50,000 per cell. However, new methods such as SCoPE-MS and SCoPE2 have increased the sensitivity and are able quantify proteins that are present at only 1,000 copies per cell [17].

There are multiple approaches to overcome the challenge of dynamic range in proteomics, for example depletion strategies that remove the most abundant proteins from samples to let mass spectrometers detect less abundant proteins, but these methods limit the reproducibility and scalability [18]. Emerging methods usually offer a trade-off between quantifying low-abundance proteins or quantifying more proteins [17].

But venturing deeper into a proteome's dynamic range remains an important challenge in the field.

Sensitivity

As mentioned in the previous section, when working with very wide dynamic ranges, mass spectrometry has a bias towards more abundant proteins to the detriment of the less abundant ones. In other words, mass spectrometry applied to protein detection has a sensitivity problem, even more in the scope of single-cell proteomics.

This problem is accentuated by the fact that protein molecules cannot be amplified like nucleic acids, which implies the importance of sensitivity in single-cell proteomics methodologies. Furthermore, poor sensitivity contributes directly to the existence of a *dark proteome* [15].

Sensitivity may also be more challenging depending on the experiment type. Non-targeted methods have less sensitivity than targeted methods as sensitivity and coverage are interrelated concepts in proteomics.

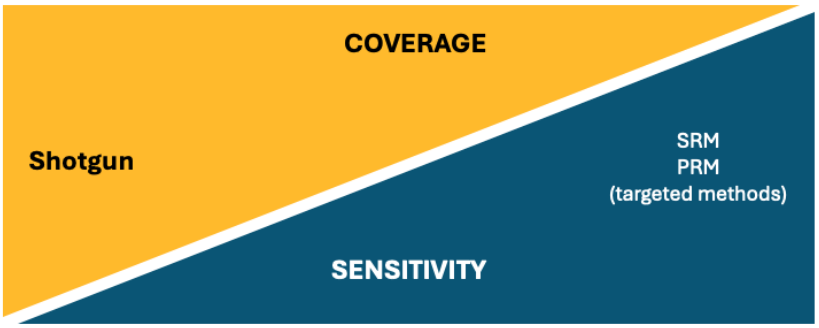


Figure 8 - Understanding the relationship between coverage and sensitivity allows choosing the most appropriate single-cell proteomics approach depending on the specific research objectives. Shotgun, unbiased proteome profiling; SRM (single reaction monitoring) and PRM (parallel reaction monitoring), targeted proteome profiling.

Throughput

Current methods for single-cell proteomics are often time-consuming and laborious, limiting the ability to analyze large numbers of cells efficiently [19].

From sample preparation, with the selection and processing of individual cells, and the analysis in the mass spectrometer up to the mass spectral data analysis itself, each action is time-consuming and makes single-cell proteomics a major challenge in terms of throughput [19].

The current single-cell proteomics methods and technologies are significantly more time-intensive than bulk proteomics. For example, analyzing 1,000 single cells requires approximately the same time as 500 bulk samples. Considering that a clinical study typically requires 100–200 bulk samples (200–400 hours), the single-cell equivalent involves 10,000–20,000 cells, requiring 10,000–20,000 hours.

In recent years there have been interesting initiatives to overcome the limitation of throughput in single-cell proteomics by using multiplexed analysis. As an example, we can find the recently created computational framework called *plexDIA*, that aims to increase the throughput of sensitive proteomics [20].

Material Loss

High-throughput mass spectrometric analysis with deep coverage relies on meticulous sample preparation. This is especially crucial when dealing with single-cell samples, where even minor losses can significantly impact the analysis [21].

Reproducibility

Another relevant challenge of single-cell proteomics is reproducibility, as it significantly influences the reliability and consistency of analytical results. Replication of single-cell proteomics data analysis is complex due to the limited amount of material available and because factors like sample preparation, technical noise and biological heterogeneity will introduce variability.

Moreover, current protein sequencing or identification instruments are still under development which makes reproducibility intra and inter-center a major challenge [22].

Peptide Sequencing and Protein Inference

One of the most used and powerful approaches for protein identification and quantification, mass spectrometry-enabled proteomics, relies on analyzing individual peptides, making it challenging to definitively identify proteins in complex organisms. This difficulty arises from sequence redundancy. Proteins within families, splice variants from the same gene, and differentially processed proteins can share highly similar sequences. While a single peptide might suggest a specific gene product is present, distinguishing between highly similar protein sequences or splice isoforms is challenging in most proteomic studies.

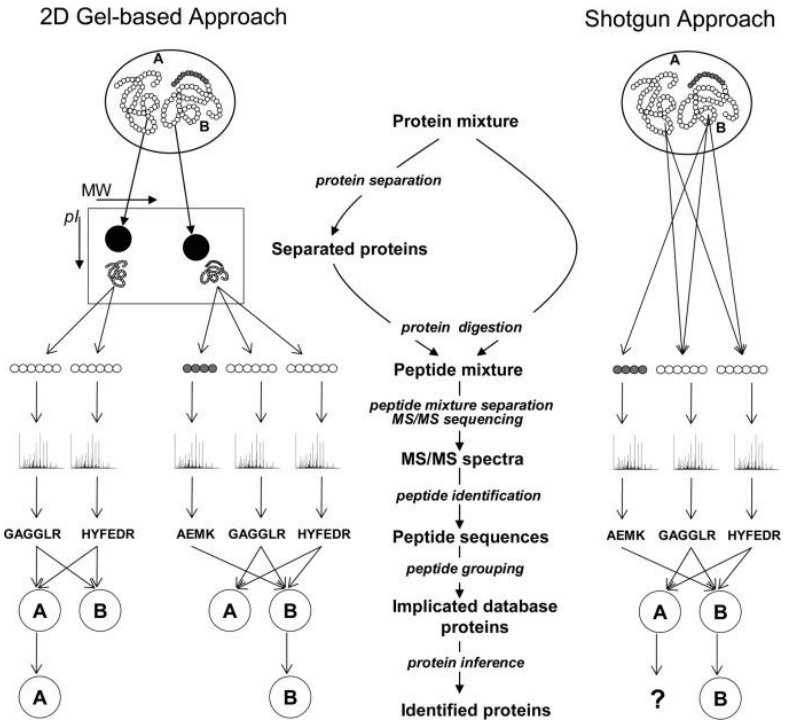


Figure 9 - Interpretation of Shotgun proteomic data. MS/MS referes to the tandem mass scans that are acquired during peptide sequencing. 2D gel separates proteins before digestion, aiding protein inference. Shotgun proteomics directly digests the sample, making it challenging to differentiate proteins with shared peptides [23].

The "Dark Proteome"

The dark proteome refers to the large number of proteins in a cell that remain uncharacterized due to their unique properties:

- Low abundance, thus isolation for functional studies is challenging.
- Lack of an ordered fold: Unlike many proteins with stable three-dimensional structures, dark proteins may be intrinsically disordered or have flexible regions, making them difficult to isolate without precipitating [24].
- Unknown function: Their primary amino acid sequence may have no homology to characterized protein domains. It is therefore difficult to predict their function(s).
- A large part of the proteome is in the dark [27]. The dark proteome constitutes a significant part of the total cellular protein content:
    - About 14% in archaea and bacteria.
    - Between 40-50% in eukaryotes and viruses.

The goal of single-cell proteomics is to analyze proteins within each cell, which allows a better understanding of cellular heterogeneity. However, the dark proteome still presents significant obstacles:

- Low abundance: Many obscure proteins are present in low abundances.

- Data interpretation: Low abundant proteins often yield low abundant MS2 sequencing data that in turn result in the lowest identification confidence scores.

Potential Benefits of Single-Cell Proteomics

Single-cell proteomics has the potential to revolutionize our understanding of cellular biology by offering a window into the intricate world of protein expression at the single-cell level. This approach transcends the limitations of traditional bulk proteomics, which averages information across a population of cells, masking the heterogeneity present within. These are some of the potential benefits of single-cell proteomics:

- **Bridging the Transcript-Protein Gap**: Unlike mRNA abundance, protein levels often provide a more accurate picture of cellular function. SCP allows us to directly correlate transcript and protein abundance at the single-cell level. Studies have shown only moderate correlation between the two, but the relationship between steady-state mRNA and protein levels remains unknown.
- **Dissecting Post-Translational Modifications (PTMs)**: Proteins undergo various modifications after synthesis, significantly impacting their function. SCP enables us to measure these PTMs within individual cells, providing deeper insights into protein activity and regulation.
- **Capturing Dynamic Proteome Changes**: Protein expression is not static; it fluctuates within cells over time. SCP can establish whether the dynamism measured at the bulk level is replicated across all cells or is influenced by a subpopulation of cells.
- **Identifying Disease Biomarkers**: Diseases often manifest as changes in protein expression at the single-cell level, or in disease-driving cell subpopulations[25][26]. SCP has the potential to identify these subpopulations that may express unique proteins which in turn may be developed into biomarkers or therapeutic targets. This could pave the way for earlier diagnosis and personalized treatment strategies.
- **Enhanced Multi-Omics Integration**: SCP data can be integrated with other single-cell omics data (e.g., transcriptomics), providing a more comprehensive understanding of cellular function.

Despite the many important challenges that SCP is facing, in recent years significant advances have been made in both technology and methodology that look promising for its development.

OBJECTIVES

The primary objective of this master thesis is to advance the understanding and application of single-cell proteomics, a rapidly evolving field that enables the study of protein expression at the resolution of individual cells. The thesis seeks to address several key goals.

Firstly, it aims to provide an in-depth overview of the current state-of-the-art in single-cell proteomics, covering all stages from sample preparation to data analysis.

Secondly, the thesis will perform a meta-analysis of existing datasets, comparing data from both bulk and single-cell proteomics experiments. This analysis will evaluate protein identification and quantification techniques, and will compare their results.

Finally, as the thesis is focused in single-cell proteomics, it will uncover significant differences in protein abundance levels through differential expression analysis, and assess the heterogeneity within the same cell types. The study will focus on macrophages cells, with the aim of uncovering unique cellular profiles and variability that provide deeper insights into their biology. Additionally, network science will be used to analyze the relationships between proteins and biological pathways. By constructing protein interaction networks, it will identify key proteins and pathways that influence cellular variability within cell types. The analysis will also extend to understanding the roles of critical proteins in driving heterogeneity across different cell types.

STATE-OF-THE-ART

In recent years, advances in the study of the proteome in single cells have progressed by tremendous speeds thanks to the evolution of technology and methodologies. This section aims to summarize the state of the art of single-cell proteomics by focusing on each of the three phases of a SCP study.

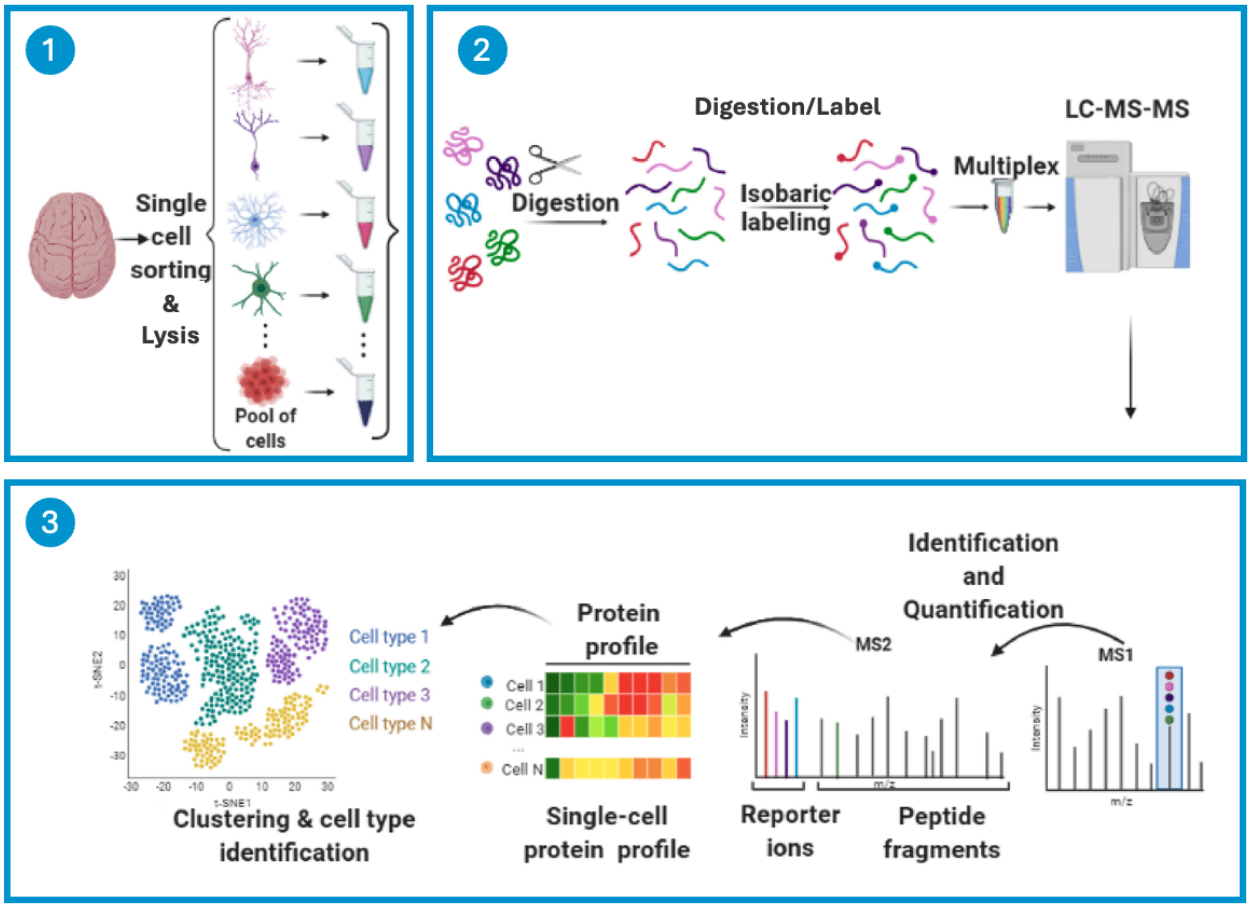


Figure 10 - Example workflow for Single-Cell Proteomics, with three different phases, Sample Preparation, Data Acquisition and Data Analysis (adapted from [28])

1. **Sample Preparation**: This stage focuses on isolating individual cells and preparing them for protein analysis, followed by cell lysis and protein extraction.

2. **Data Acquisition**: Proteins are hydrolyzed into smaller peptides using enzymes, a process referred to as proteolysis. These peptides might be stable isotopically labeled for enhanced detection to be analyzed by mass spectrometry. Mass spectrometers measure charged analytes, thus the outputs are the mass-to-charge. Peptide analytes are selected by the mass spectrometer for fragmentation (technically referred to as, dissociation) in a collision cell. The fragment spectra are then scanned again by the mass spectrometer. The peptide scan-dissociation-fragment scan method is called, tandem mass spectrometry (MS/MS).

3. **Data Analysis**: The mass spectrometry data is analyzed using bioinformatic tools to identify peptides based on the MS/MS data. The more abundant a protein is in a sample, the more likely its peptides will be sequenced. Protein quantification is therefore determined by the number and mass spectral signal intensity of its corresponding peptides. Statistical analysis helps identify differentially abundant proteins and uncover patterns in cellular protein abundance.

Each of these phases has its complexities, technologies, and specialists. In the case of SCP, we face important challenges in each of these phases, starting with the delicate phase of sample preparation, avoiding as much as possible the loss of material, passing through the phase of analysis of the material by MS and ending with the analysis of the data provided by MS.

Step 1: Sample preparation

Sample preparation is a crucial step in the case of single-cell proteomics. As we have seen previously, SCP has significant challenges in order to be a robust, reliable and successful tool for scientific researchers. Some of most prominent challenges are encountered during the sample preparation phase. Important factors such as minimization of material loss and sensitivity are closely linked to sample preparation.

In 2021, Leduc et all [29] developed a methodology called nPOP (nano-ProteOmic sample Preparation). nPOP offers a novel way to prepare single cells for proteomic analysis. It was designed to be widely accessible, inexpensive, robust, and automated sample preparation method.

nPOP utilizes droplets on a glass slide surface to prepare individual cells, in contrast of using well-plates or chips. This offers a highly flexible design, allowing researchers to customize the number of droplets per cluster to suit their specific multiplexing needs.

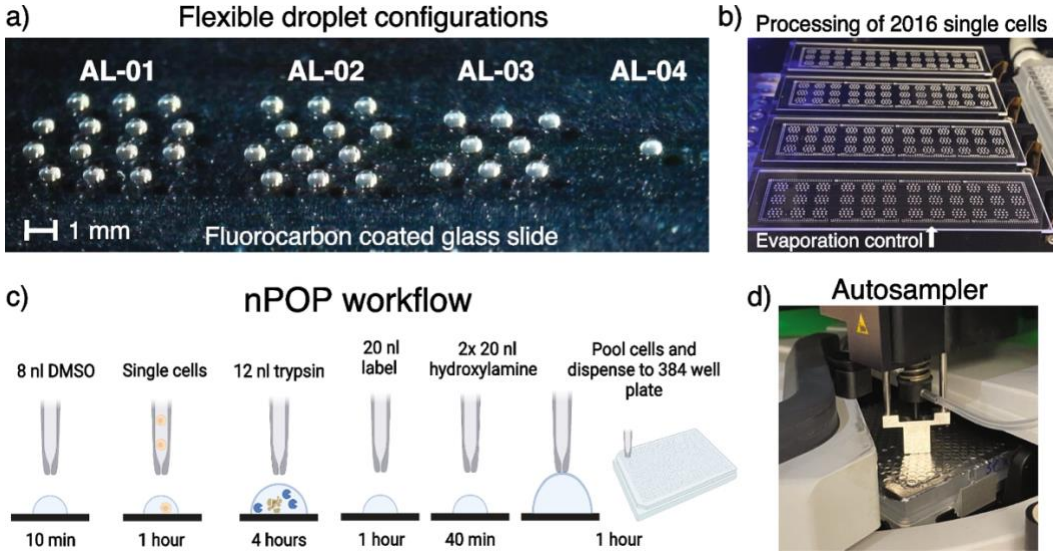


Figure 11 - The nPOP method enables the preparation of thousands of single cells on slides with various droplet layouts. Slides are arranged with 2016 single cells and are surrounded by water droplets to control humidity and placed on a cooling surface to prevent evaporation. The nPOP process involves cell lysis, protein digestion, peptide labeling with TMTpro, quenching the labeling reaction, and sample collection, all done in individual droplets. After labeling, the samples are pooled and transferred into a 384-well plate for automated LC-MS/MS injection. [30]

The use of this methodology allowed a significant leap in terms of throughput, at a relatively low cost, and allowing the study of hundreds of cells and cell subpopulations in parallel.

The nPOP method requires specialized equipment that was recently commercialized by Cellenion. The platform is called, the cellenONE®.

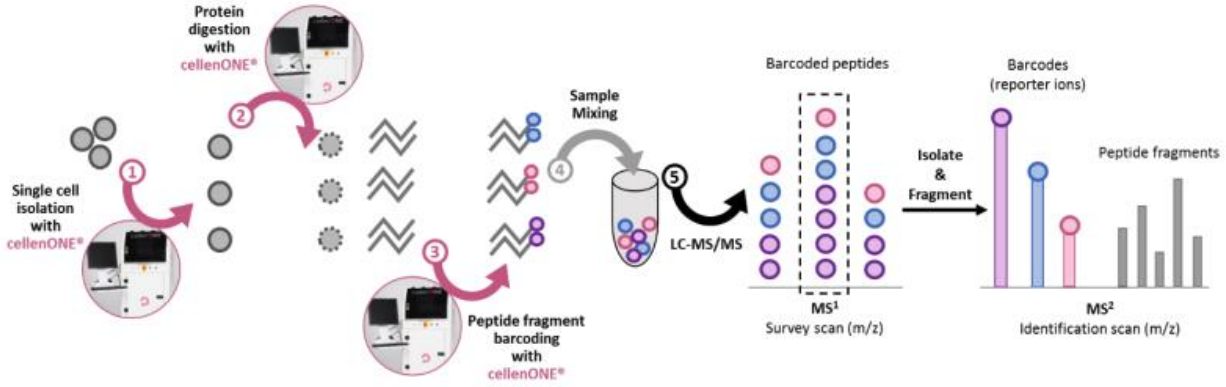


Figure 12 – Workflow of the analysis of single cell proteome using LC-MS/MS with SCoPE2 and cellenONE® [31]

By using the cellenONE®, researchers can combine single cell isolation and nanoliter dispensing, to perform the different steps of the sample preparation using a single device.

Figure 13 – The cellenONE® system isolates single cells using optical detection. Fluid with cells circulates through a capillary tip, divided into an ejection zone (green) and a sedimentation zone (pink). The ejection zone determines the content of the next droplet, while the sedimentation zone acts as a safety buffer. If the sedimentation zone is empty and the ejection zone contains one cell, a droplet is isolated, ensuring precise single-cell isolation [31]

In 2023, Leduc et al., already demonstrated accurate quantification of about 3,000 – 3,700 proteins per human cell by combining nPOP and plexDIA [30].

Although nPOP has proven to be a robust and accessible sample preparation method that provides promising results, other methods are also currently being developed that point to promising results, such as the combined use of cellenONE and proteoCHIP [32].

Step 2: Data acquisition

Once the proteins are extracted from the cell, they are digested into smaller peptides by enzymes, whose natural function is to break down proteins,. These peptides can then be labeled for enhanced detection and analyzed by mass spectrometers. Although there are alternatives to mass spectrometers, it is a dominant approach for single-cell proteomics due to its ability to identify and quantify thousands of proteins.

Researchers can employ one of three main proteomics experiments referred to as bottom-up, middle-down, and top-down proteomics. Bottom-up proteomics is the most commonly used method, since it is the easiest to execute.

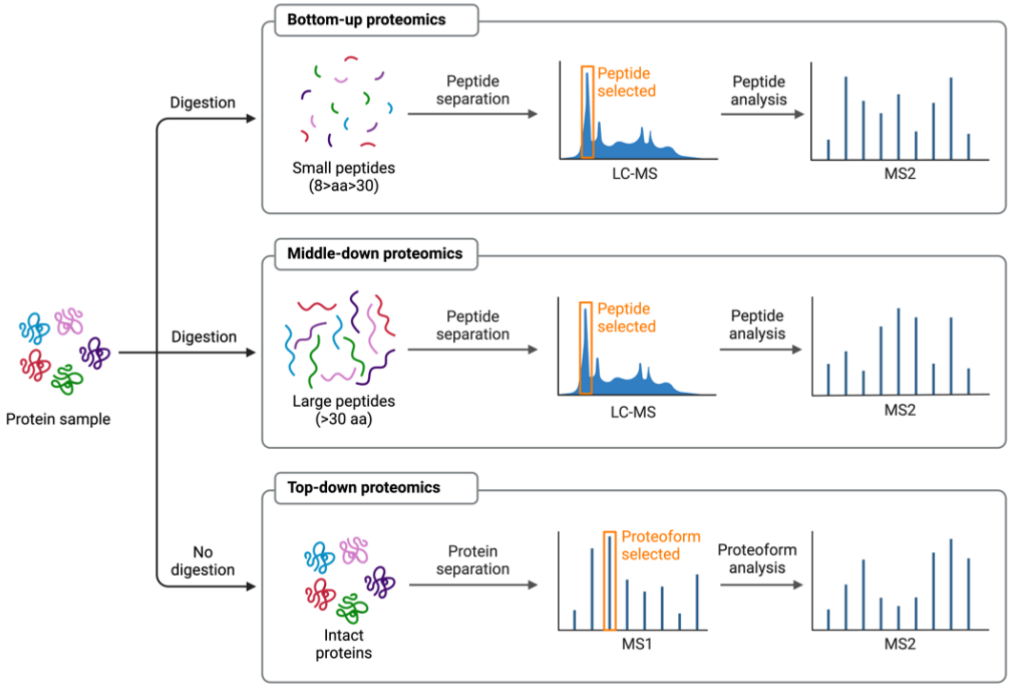


Figure 14 - Proteomic analysis strategies include bottom-up, where proteins are digested into peptides before mass spectrometry (MS) analysis, middle-down that involves partial digestion to produce larger fragments and Top-down, that analyzes intact proteins directly by MS.

Bottom-Up Proteomics

This technique, also known as shotgun proteomics, is based on the disassembling of the sample preparation. First, proteins are extracted and purified, then different enzymes (proteases) are used to hydrolyze proteins into smaller peptides to make them more conducive to chromatographic separation and mass spectrometric analysis (as compared to intact proteins). The peptides are separated based on their properties using chromatography, such as HPLC (high performance liquid chromatography), to disentangle complex mixtures. The HPLC instrument is coupled to the mass

spectrometer such that as they elute from the chromatographic column, they enter the mass spectrometer to be sequenced. Once the mass spectra are obtained, they are compared to databases to identify the peptides and reconstruct the original proteins.

One of the main advantages of this approach is the possibility to detect wider range of proteins, which makes it a good approach for protein identification and quantification.

Among the disadvantages, we encounter the problem that proteins with modifications or those poorly digested by proteases (enzymes that digest proteins into peptides) may be missing. Also, the identification can be difficult for proteins with very similar sequences.

Top-Down Proteomics

This approach tackles proteins in their entirety, offering insights into their structure and modifications.

Similar to bottom-up, proteins are extracted and purified to then be introduced into the mass spectrometer, but in this case, as intact proteins. In top-down proteomics, a second dissociation step is often introduced (MS/MS/MS) that in turn further fragments these larger peptides for improved amino acid sequencing.

The main advantage is that top-down provides information about intact proteins and their modifications, enabling us to study expressed proteoforms.

Although top-down proteomics has been improved in the last years, it still has many challenges to overcome, including poor sensitivity for detecting low-abundance proteins and limited dynamic range. Importantly, the mass spectra of intact proteins are far more complicated than those of peptides. Thus top-down proteomics is hampered by the complexity of analyzing, interpreting, and validating the results [33].

Middle-Down Proteomics

This technique aims to fill the gap between the two previous approaches. The lengths of the peptides are larger (20-200 amino acids) than those generated from bottom-up proteomics (8-25 amino acids), and smaller than the length of the average protein (400-500 amino acids). By doing so, the complexity of the resulting sample is smaller and the probability of detecting more unique peptides compared with bottom-up also increases [34], which means the possibility to identify more proteoforms.

Considering all pros and cons of the three approaches, the decision of using one or another method is usually made based on the research question:

- Protein identification and quantification: Bottom-up
- PTMs, isoforms and interactions: Top-down
- Protein details & low-abundance analysis: Middle-down

Bottom-Up strategies (DIA and DDA)

The most common approach for the identification and characterization of peptides and proteins is based on bottom-up (shotgun) proteomics, using data independent acquisition (DIA), data dependent acquisition (DDA), or a combination of both strategies.

DIA approach is based on the fragmentation of all peptide ions within a pre-defined mass range. This allows to capture a broader picture of the proteome but requires more complex data analysis and may result in failure to identify low abundance proteins due to spectral complexity. It requires the use of specialized software to identify and quantify proteins based on the full spectrum of fragment ions for each mass range.

DDA selectively fragments a limited number of the most intense precursor ions. This offers enhanced sequencing of chosen peptides but misses information on less abundant ones.

Both DIA and DDA data are annotated by use of a reference proteome (in silico digested and fragmented) or by use of previously acquired, confidently annotated mass spectral library.

Figure 15 - Discovery-based quantitative MS methods, like "shotgun" proteomics, aim to identify numerous proteins in diverse samples with minimal development. In data-dependent acquisition (DDA), the MS instrument scans peptide ions and fragments the most abundant ones. In data-independent acquisition (DIA), the instrument scans predetermined m/z ranges and fragments all ions within each range. (Advanced Analysis Center, University of Guelph)

The use of DIA or DDA highly depends on the type of study that needs to be conducted. For example, DIA can be a better approach for quantitative studies where the goal is to quantify a large number of proteins, or the samples are complex with a wide dynamic range of protein abundances. However, the complexity of the resulting spectra is higher and requires complex data analysis. DDA

may be the preferred option for targeted analysis of specific proteins and their characterization. The following table summarize the main differences between DDA and DIA[35]:

	DDA	DIA
Suitable for small proteomes	Yes	Yes
Suitable for large proteomes	No	Yes
Dynamic range	Moderate	High
Complexity of spectra	Low	High
Search requirements	Spectral library, Protein database	Spectral library, Protein database
Instrument method optimization	Low	High

Step 3: Data analysis

To extract meaningful insights from the raw data obtained from the mass spectrometer, it is required to go through a complex multi-step data analysis pipeline that can vary depending on the method.

As an example, the pipeline may include:

- **Peptide Identification**: Experimental tandem mass spectral linkage with proteins mainly involves two main approaches or a combination of both [36]:
    - **Database search:** MS data is analyzed using bioinformatics software to match peptide spectra to known protein sequences in databases [37]. This allows for the identification of proteins present in each single cell.

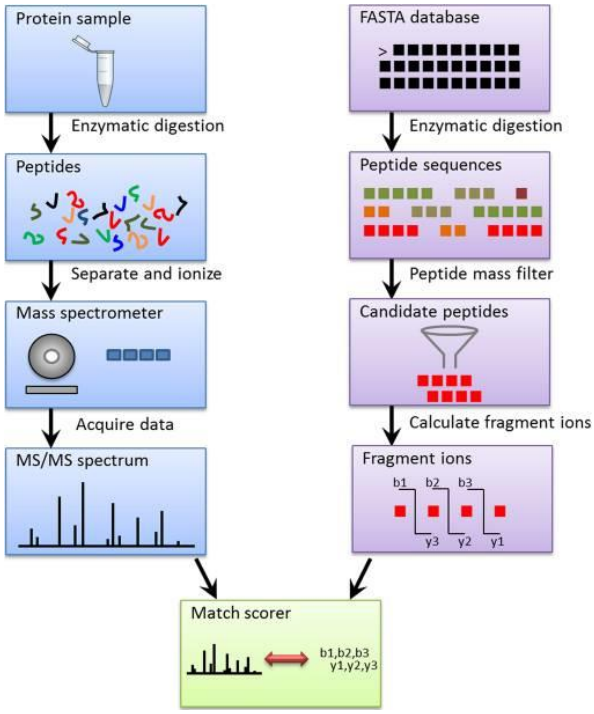
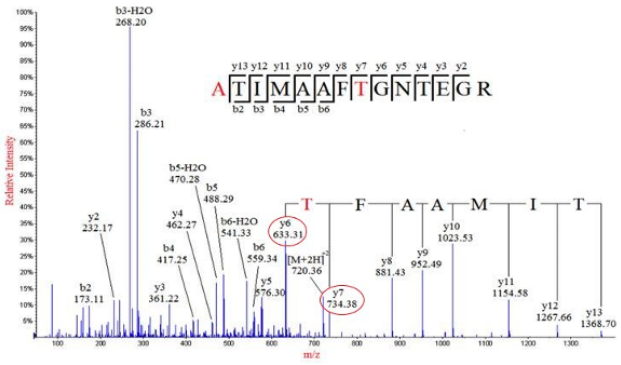


Figure 16 - Data acquisition and database search involve matching experimental MS/MS spectra with in silico digested protein sequences from a database. Theoretical fragment ions are calculated for candidate peptides and compared to the experimental spectrum to generate similarity scores. The best matching peptides and their scores are then reported [37]

- **De-novo sequencing**: Algorithms analyze the MS/MS spectra directly, attempting to deduce the amino acid sequence of the peptide based on the fragmentation patterns and mass information of the ions. This method is computationally complex and less accurate than database search, especially for longer peptides. It can be difficult to differentiate between isobaric amino acids with the same mass. This method is especially useful to identify novel proteins not present in any database or to analyze PTMs that alter peptide sequences.



Name	3-letter code	1-letter code	Residue Mass	Immonium ion	Related ions	Composition
Alanine	Ala	A	71.03711	44		C_3H_5NO
Arginine	Arg	R	156.10111	129	59,70,73,87,100,112	$C_6H_{12}N_4O$
Asparagine	Asn	N	114.04293	87	70	$C_4H_8N_2O_2$
Aspartic Acid	Asp	D	115.02694	88	70	$C_4H_5NO_3$
Cysteine	Cys	C	103.00919	76		C_3H_5NOS
Glutamic Acid	Glu	E	129.04259	102		$C_5H_7NO_3$
Glutamine	Gln	Q	128.05858	101	56,84,129	$C_5H_8N_2O_2$
Glycine	Gly	G	57.02146	30		C_2H_3NO
Histidine	His	H	137.05891	110	82,121,123,138,166	$C_6H_7N_3O$
Isoleucine	Ile	I	113.08406	86	44,72	$C_6H_{11}NO$
Leucine	Leu	L	113.08406	86	44,72	$C_6H_{11}NO$
Lysine	Lys	K	128.09496	101	70,84,112,129	$C_6H_{12}N_2O$
Methionine	Met	M	131.04049	104	61	C_5H_9NOS
Phenyalanine	Phe	F	147.06841	120	91	C_9H_9NO
Proline	Pro	P	97.05276	70		C_5H_7NO
Serine	Ser	S	87.03203	60		$C_3H_5NO_2$
Threonine	Thr	T	101.04768	74		$C_4H_7NO_2$
Tryptophan	Trp	W	186.07931	159	11,117,130,132,170,100	$C_{11}H_{10}N_2O$
Tyrosine	Tyr	Y	163.06333	136	91,107	$C_9H_9NO_2$
Valine	Val	V	99.06841	72	44,55,69	C_5H_9NO

Figure 17 - The core concept of de novo sequencing involves calculating the mass of an amino acid residue on the peptide backbone by measuring the mass difference between two fragment ions. The figure shows that the mass difference between the y7 and y6 ions is 101, corresponding to the mass of residue T. Therefore, by identifying either the y-ion or b-ion series in the spectrum, the peptide sequence can be deduced (image source: creative-proteomics.com)

- **Hybrid (database search + de-novo)**: The hybrid approach in peptide identification combines the advantages of both database search and de-novo sequencing to enhance protein identification, particularly when dealing with novel proteins or unknown modifications [38].

- **Protein Inference:** With the identified peptides, the next step is to reconstruct the original proteins by combining them. This step is critical as it is not an easy task and poses challenges that cannot always be solved. In the case of finding long peptide sequences, it is easier due to uniqueness, but if the peptides are short, it is much easier that they can be part of different proteins.
- **Protein Quantification**: The abundance of each identified protein is determined based on the intensity of its corresponding peptide signals in the MS data. This provides information on the relative protein abundance levels within individual cells.
- **Data Normalization**: Data usually needs to be normalized to account for technical variations and differences in cell size or protein extraction efficiency.
- **Imputation:** Protein data usually contains a lot of missing values and imputation methods are required to enable data analysis and downstream applications. For DIA experiments, spectra alignment is required to ensure data consistency between runs and to enable

accurate feature matching. It is an essential step in preparing proteomics data for downstream analyses.

- **Batch correction:** As single-cell proteomic experiments usually require to be split into batches, minor differences in protocols, reagents or instruments can introduce technical variability. This step aims to minimize these effects to avoid misleading results.
- **Statistical Analysis**: Statistical methods are used to analyze the proteomic data, identify differentially abundant proteins between different samples, and uncover protein co-expression patterns.
- **Enrichment Analysis:** Used to identify functionally significant groups of proteins within a dataset. It helps researchers understand the biological processes and pathways that the identified proteins might be involved in.
- **Protein-Protein Interaction Analysis:** Understanding protein-protein interactions allows us to disentangle complex biological pathways and mechanisms.
- **Pseudotime Analysis:** Additionally, by analyzing protein levels in many cells, scientists can order them based on how similar their proteins are. This creates a hypothetical timeline, which is known as ***pseudotime*** and it represents the cell development, even though they haven't tracked individual cells over time [39]. Pseudotime is a useful tool because it reveals how proteins change as cells differentiate into specialized types.
- **Visualization**: Data is visualized using various tools to create heatmaps, scatter plots, volcano plots and other graphical representations that aid in understanding cellular heterogeneity and protein abundance patterns.

Novel Methods for Data Acquisition and Data Analysis

In recent years there have been many advances in both techniques and methodologies to improve the sensitivity and resolution of single cell mass spectrometry analysis. One of the references worldwide in the research of single-cell proteomics is the Slavov Laboratory. They created a set of different mass spectrometry methods that label single cells to allow simultaneously analysis of multiple cells in parallel, increasing the throughput and reducing the cost per cell.

- **Shotgun Single-Cell Proteomics**
    - **SCoPE-MS**: It is an initial version of the “SCoPE” methods that quantify over a thousand proteins per single cell. The method is based in a multiplexing approach using tandem mass tags with a carrier proteome. It allows to improve the sensitivity. The method introduces carriers that act as reference points and help in quantifying the target proteins.

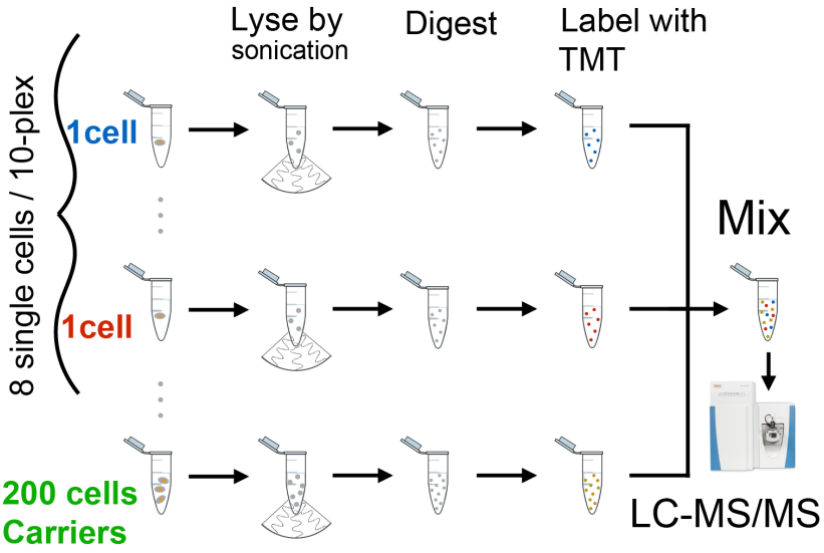


Figure 18 - The conceptual diagram and workflow of SCoPE-MS involves the following steps: live cells are individually lysed using sonication, their proteins are digested with trypsin, the resulting peptides are labeled with TMT (tandem mass tag) labels, combined, and then analyzed using LC-MS/MS [40]

- **SCoPE2**: Enhanced version that can quantify up to 3,000 proteins at lower price and that requires less time. By introducing the use of an isobaric carrier, it enhances peptide sequence identification [41][42].

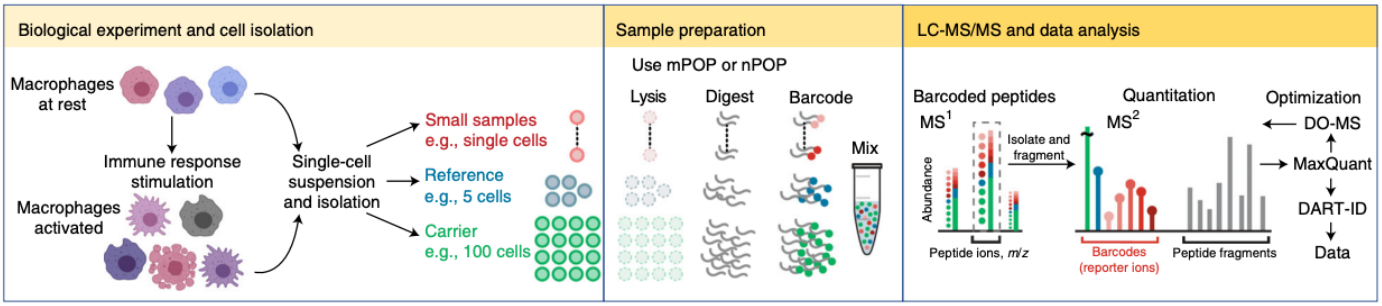


Figure 19 - The SCoPE2 workflow involves sorting cells into multiwell plates and lysing them with mPOP. Proteins are digested with trypsin, and the resulting peptides are labeled with TMT, combined, and analyzed by LC-MS/MS. Reference channels in SCoPE2 sets enable merging single cells from different sets into one dataset. LC-MS/MS analysis is optimized by DO-MS, and peptide identification is enhanced by DART-ID [42]

- **Prioritized Single-Cell Proteomics**

  - **pSCoPE:** it is an advanced method to guarantee the analysis of prioritized peptides while analyzing identifiable peptides at full cycle [43]. It allows to improve data consistency, sensitivity and depth of protein quantification, prioritizing proteins of interest.

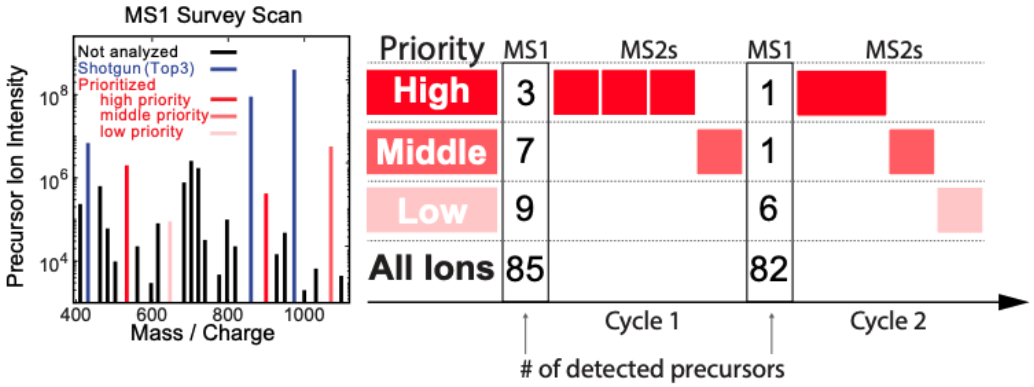


Figure 20 - Shotgun TopN analysis selects the N most abundant precursors for isolation and fragmentation (blue). Prioritized analysis first selects the highest priority precursors (solid red) and then those with lower priority (fading red tones). pSCoPE introduces prioritization to MaxQuant.Live to increase identification, consistency and protein coverage [43]

- **Parallel analysis of both single cells and peptides**

    - **plexDIA:** This is an experimental method that aims to provide solutions to increase the throughput of sensitive proteomics at cheaper cost than alternative methods like label-free DIA (LF-DIA). It acquires data on all peptides present in the sample to provide a comprehensive picture and works with both, bulk and single-cell samples.

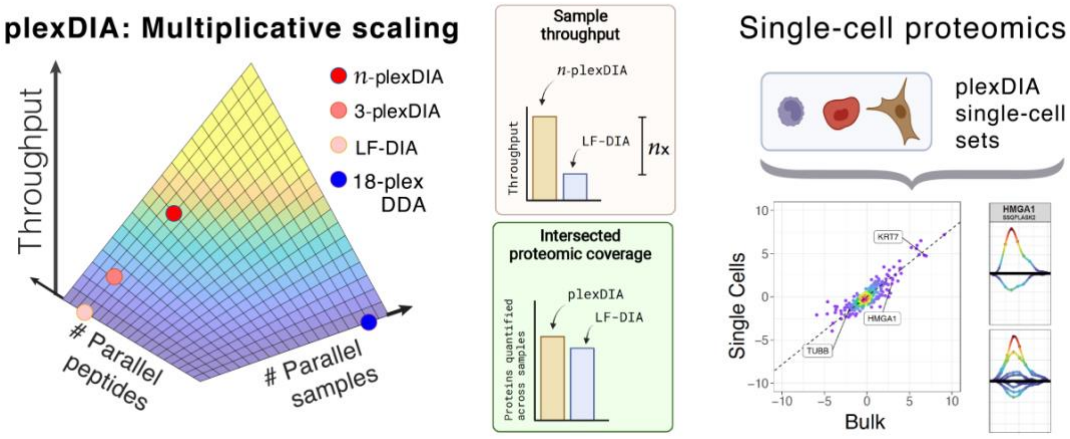


Figure 21 - plexDIA aims to increase the throughput of MS proteomics by combining the parallel analysis of multiple peptides with the parallel analysis of multiple samples [4444]

DESIGN AND DEVELOPMENT

Introduction

Among the primary objectives of this thesis was conducting a comprehensive meta-analysis. A Meta-analysis is a statistical method that combines data from multiple studies to derive a more comprehensive understanding of a particular research question or topic. By synthesizing results across diverse datasets, meta-analyses provide increased statistical power and more robust estimates of effect sizes. This approach is especially valuable when individual studies yield conflicting results or have limited sample sizes, as it aggregates evidence to draw more reliable and generalizable conclusions. Typically, meta-analysis involves systematically collecting data, assessing the quality of studies, and applying statistical techniques to integrate findings.

This meta-analysis integrates data from bulk proteomics datasets and one proteomics single-cell dataset, all focusing on the macrophages (derived from THP-1 and BMDMs). Specifically, the study includes cells of type M0, M1 and M2.

- M0 refers to the precursor, undifferentiated state, of macrophages. They have not yet been activated or polarized.
- M1 (Classically Activated Macrophages) are macrophages that have been induced by factors such as IFN-γ. They are pro-inflammatory and play a key role in defense against pathogens.
- M2 (Alternatively Activated Macrophages) are induced by factors like interleukin-4 (IL-4) and IL-13 and they are involved in tissue repair, wound healing and anti-inflammatory responses.

This analysis aims to determine if both approaches, bulk and single-cell proteomics, converge on similar conclusions regarding the biological characteristics and behavior of THP-1 cells. By comparing the outcomes from bulk and single-cell data, I can assess the consistency and reliability of findings across different methodologies.

Additionally, utilizing the single-cell dataset allows me to delve deeper into the heterogeneity of same type of macrophages, exploring the differences and unique subpopulations that might be masked in bulk analysis. This meta-analysis will help to better understand the biology of macrophages and the strengths and limitations of each proteomic technique.

The table below provides a summary with the number proteins identified in different experiments conducted by different researchers in the last years:

Study	Type	DIA/DDA	#Identified proteins	Protein FDR	Number of cells	Cell type
PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Iwata, H., et al. (2016) [45]	Bulk	DDA	9048	1%	*N/A (bulk)*	THP-1 derived macrophages
Proteomic characterization of four subtypes of M2 macrophages derived from human THP-1 cells. Li, P., et al. (2022) [46]	Bulk	DDA	6616	1%	*N/A (bulk)*	THP-1 derived macrophages
Comparative Proteomic Analysis of Polarized Human THP-1 and Mouse RAW264.7 Macrophages. Li, P., et al. 2021) [47]	Bulk	DDA	7349	1%	*N/A (bulk)*	THP-1 derived macrophages
Prioritized single-cell proteomics reveals molecular and functional polarization across primary macrophages. Huffman, R. G., et al. (2022) [43]	Single Cell	DIA	1123	1%	373	BMDMs derived macrophages

The datasets included in this research were derived from studies where the authors had already processed the raw mass spectrometry data. This approach was chosen to focus on the data analysis process rather than handling the vast raw data files generated by mass spectrometers, which was not part of the objectives of this thesis.

All datasets were downloaded from publicly available websites, with the exception of *Iwata et al. (2016)* [45], which was provided by the Center for Interdisciplinary Cardiovascular Sciences (CICS) at Brigham and Women's Hospital, an affiliate of Harvard Medical School (Boston, Massachusetts).

The analysis is structured into three different sections:

1. **Proteomic Profiling**: First, I analyze the proteomic profiles of the cells, identifying similarities and differences across the various studies. This comparative analysis provides insights into the consistency and variability of proteomic data.

   The aim is to answer questions like:
   - How many proteins were identified by the researchers in each study (using the same cell type)?
   - How many proteins were identified in each of the phenotypes?
   - How many proteins were identified using the two different methods (bulk and single-cell)?
   - Considering bulk versus single-cell, how many of these proteins were found exclusively in one of the subgroups (M0 and M1) and which were found in the two of them?
   - Which proteins were exclusively detected in single-cell analyses, which ones were uniquely identified in bulk samples, and which proteins were common to both types?
   - Etc…

2. **Protein Abundance Analysis**: In this section, I examine Protein Abundance levels accross different groups (M0, M1, and M2) whenever is possible. Note that the single-cell dataset does not contain M2 data.

   The goal is to analyse and compare bulk and single-cell data to answer questions like:
   - Are there relevant differences in the gene expression / proteina abundance for bulk versus single-cell data?
   - Are the results from the different experiments consistent?
   - What are the proteins found only in one of the cell types?

   However, one of the key objectives is to analyze the heterogeneity within the same cell types using the single-cell data. This allows to find answers to questions like:
   - Is it possible to identify clusters of cells in the same group?
   - What are the genes that are differently expressed in each of the clusters of the same group?

3. **Pathway Analysis:** Finally, I conduct a pathway analysis to examine the biological and functional significance of the identified proteins and genes within each group. This analysis aims to uncover the pathways and networks involved, providing deeper insights into macrophage biology and potential regulatory mechanisms.

Overview of Studies and Dataset Descriptions

According to the authors and the technical details provided in the different publications, the processed data for the three bulk datasets were generated using Thermo Proteome Discoverer (Thermo Scientific, Germany). This ensures that the formats are very similar, making the comparison of datasets much more straightforward and efficient.

Iwata, H., et al. (2016) [45]

In this publication, the authors compared the proteins present in THP-1 and RAW264.7 cells before and after stimulation with different cytokines.

Data was exported to a single excel file with 6 tabs (one for each of the M0, M1 and M2 states for THP-1 and RAW264.7). Each of the sheets contain the abundance of the identified proteins at 6 different time points (0, 8, 12, 24, 48 and 72 h), and the rows represents the different protein gene IDs.

Full description of the format available in the ANNEX I: Dataset format

Li, P., et al. (2022) [46]

The study analyzes and compares the protein expression profiles of four different M2 macrophage subtypes. By doing so, it seeks to identify unique and shared proteins among these subtypes to

better understand their functional roles. Ultimately, this research enhances the understanding of M2 macrophage heterogeneity and their potential implications in health and disease.

In this case, the processed data is stored in a single excel file, where the different columns represent the abundance, normalized abundance and relative abundance of different population groups (M0, M1, M2a, M2b, M2c, and M2d). To facilitate a comprehensive comparison of protein abundance across studies, a new column was added that averages the values of the M2 subgroups.

Full description of the format available in the ANNEX I: Dataset format

Li, P., et al. (2021) [47]

The authors aim to systematically compare the protein expression profiles of M1 and M2 macrophages derived from human THP-1 and mouse RAW264.7 cell lines. By analyzing over 5,000 proteins, the study identifies unique and common proteins across these macrophage types, highlighting significant differences in their polarization states.

Full description of the format available in the ANNEX I: Dataset format

Huffman, R. G., et al. (2022) [43]

The authors introduce a new prioritization algorithm called **pSCoPE** (Prioritized Single-Cell Proteomics) and compare the results to an exiting method, SCoPE2, to demonstrate the enhancement of the sensitivity, consistency, and throughput of single-cell proteomics by efficiently allocating mass spectrometry time to high-priority peptides.

The publicly available data consists in two different files, the first one with the intensity matrix where the rows represent the proteins and the columns the different cells. The second file is metadata that allows to identify the type of each of the cells, or the batch amongst others.

Full description of the format available in the ANNEX I: Dataset format

Data Availability

Processed data and metadata files are available at:

Iwata, H., et al. (2016)	-- *N/A* --
Li, P., et al. (2022)	http://proteomecentral.proteomexchange.org (identifier = PXD022320)
Li, P., et al. (2021)	http://proteomecentral.proteomexchange.org (identifier = PXD019800)
Huffman, R. G., et al. (2022)	https://scp.slavovlab.net/Huffman_et_al_2022#processed-single-cell-protein-data

Tools

The datasets were downloaded and analysed in a local Jupyter Notebook using python, public libraries and public API's.

Several libraries have been used to analyze the data and visualize the results (e.g. matplotlib, scanpy, GSEApy, etc). However, some others were implemented for the project. For example, due to the need of translating protein Gene ID's to GeneNames, I implemented a Python client to query UniProtKB REST API (https://www.uniprot.org/help/api). It works for Uniprot ID's, Uniparc ID's and Unisave (for old/deleted protein ID's). The implementation of Unisave translation methods was needed because of one of the datasets was using an old codification of protein Accessions.

Additionally, in order to generate the protein-protein interaction visualizations I implemented another API client, in this case for StringDB (https://string-db.org/help/api/).

The final code (notebook) and the API clients implemented as part of this thesis are published in a public repository in GitHub: https://github.com/powwowath/pyMSpro

EXPERIMENTS AND RESULTS

As described in the previous section, the meta-analysis is divided into three distinct sections. In each of these sections, I worked with the aim of answering the questions posed at the beginning of this project. These questions are based on the composition of macrophages according to their state, the study of gene expression differentiation based on the group, and how this composition can affect the organism through the biological functions of the various cell subtypes.

Exploratory data analysis (EDA) and data preparation

In data analytics projects, this initial step is crucial for gaining a better understanding of the data characteristics, including their distributions and relationships. It also involves detecting anomalies such as outliers, missing values, or any other irregularities in the data.

Understanding the format and designing new features

In this case, it was even more critical because the meta-analysis incorporated data from four different studies. Although three of these studies (bulk) used the same software to process the data, the formats and types of data varied depending on the specific study and its objectives. For instance, in one study (Iwata, H. et al. 2016), researchers published the protein abundance at six different time points post-cytokine stimulation. This was the only instance where temporal data was available at multiple points in time.

Another relevant variation in one of the datasets was the one previously explained for Li, P., et al. (2022) dataset, which contained the abundance data for four different M2 subtypes. To compare this data with other datasets, I created a new column with the averaged abundance for M2, using the data from the four different subtypes.

Data variability and consistency

Since the analysis included different datasets with equivalent or very similar data, it was important to analyze the variability of the data. The goal of this check was to ensure that the results from each dataset were consistent. To simplify the process I decided to check the relative protein abundance between the three bulk data sets.

I first focused on assessing the proteins identified in different experiments on the same cell types to ensure consistency and alignment of results between these experiments. The variability of the results obtained in the different bulk studies was significantly greater than anticipated. Initially, I identified the 25 genes with the highest number of isoforms in each of the three datasets. When I compared these results, I found that only two genes, **'AAK1'** and **'TMPO'**, were consistently present across all three lists. The consistency improved slightly when considering the two datasets generated by the same authors within a one-year interval, with six genes appearing in both lists. However, despite this improvement, the overall differences remained significantly higher than

expected, highlighting the substantial variability in protein abundance even within closely related studies.

To confirm the hypothesis, I checked with protein abundances for the same cell types. The correlation was checked using two approaches: first, by using the ratio of the normalized abudances of M1 over M0, and then by using the normalized abundances in each of the groups. The results revealed no overlap when interpolating the most abundant proteins in each group and the three datasets. This finding indicates that protein abundance values exhibit high variability between experiments, even for the same cell type and group. This variability might be caused by a number of factors, from being impacted severely by the experimental design to the challenges in achieving consistent measurements in proteomics experiments.

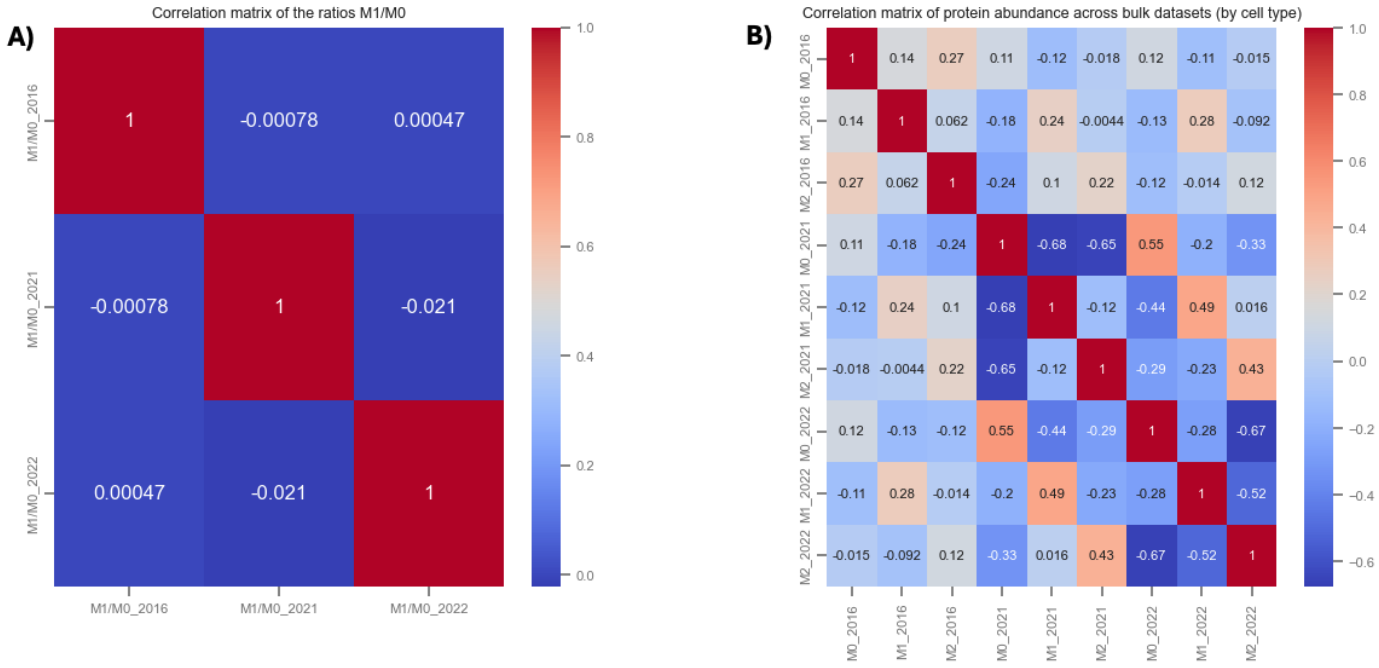


Figure 22 –Matrix (A) presents the correlation between the ratios M1/M0 for the three bulk datasets. The near-zero values indicate minimal to no correlation between the datasets. Matrix (B) offers a more detailed view, displaying correlations across multiple variables (M0, M1, M2) for the same datasets, also showing a general lack of correlation between most pairs. This suggests substantial variability in protein expression profiles across the different datasets.

Given this significant variability, aggregation of the data for meta-analysis was not possible. Consequently, some of the tasks planned in the project could not be completed, highlighting the challenge posed by the variability of the data.

Filtering considerations

During the exploratory data analysis phase, several data points were identified for filtering as they did not meet specific criteria, outlined below, and could potentially introduce errors in the analysis:

- **Values different from "1" in the column "126/126"** (Iwata et al. 2016)
  The processed data file includes a column labeled "126/126," utilized as a reference point by the authors of the study. This initial point (hour 0) required all proteins to have a value of 1. Any entries with values other than 1 were filtered out. In this case, the number of records not meeting the expected reference value was 44, which is a very small number compared to the total unique proteins in the dataset.
- **Score Sequest HT > 1.5**
  For all datasets, the criteria applied by Li P. et al. (2022) were used, requiring a Sequest HT Score greater than 1.5.
- **Keep proteins that has PSMs >= 5**
  As with the previous point, the same criteria used by the authors of Li P. et al. (2022) is applied here.

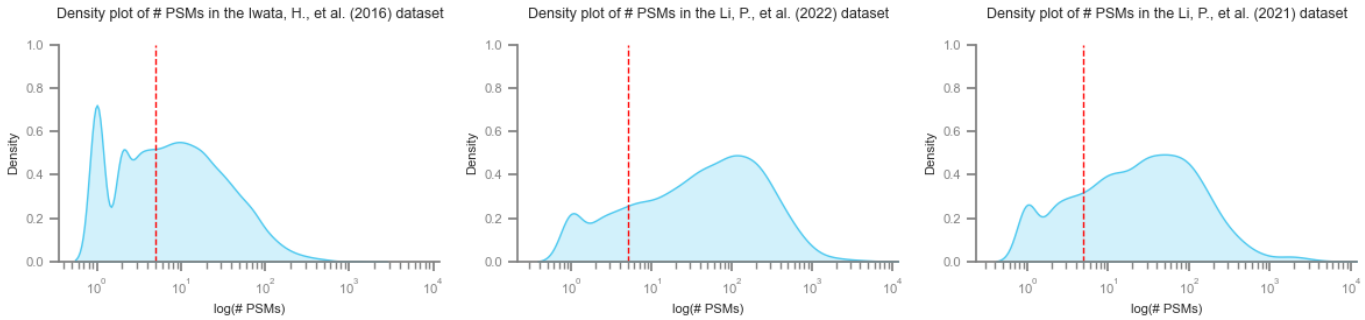


Figure 23 – Density plot of #PSMs in log scale. The red line is the Protein Spectrum Matches threshold, set to 5 as defined by Li P. et al. (2022)

- **Filter by the number of peptides**
  To avoid false positives, I employed the two-peptide rule, excluding proteins identified by the mass spectrometer with only a single peptide [48].
- **Removing *Null* values for Gene Name**
  The process of converting protein Gene IDs using the UniProtKB API successfully retrieved gene names for almost all the proteins. However, a very small number of API calls returned an http error code 400 or 404, and the resulting value in the data frame was Null. In this case and given the very small number (< 0,1% in just two of the datasets and 0% in the other two), these proteins were filtered out.
- **Removing Gene Names starting with "*None (*"**
  The client implemented to query the UniProtKB REST API returns *"None(protein_id)"* when the response it gets from the API is correct (*http response code = 200*) but does not contain a valid Gene Name. The code contains a filter to avoid unknown Gene Names. However, after checking the resulting data and executing the filter, no proteins were filtered as all of them were linked to a valid Gene Name.

Missing values

The exploration of missing values in the datasets after applying the described filters revealed the following percentages: Iwata, H., et al. (2016) with 0.48%, Li, P., et al. (2022) with 5.72%, Li, P., et al. (2021) with 6.36%, and Huffman, R. G., et al. (2022) with 0.00%. Due to the nature of the data, I interpreted the missing values in the Abundance columns as corresponding to proteins that were not found in the group of interest. This interpretation aligns with the expectation that certain proteins may not be detected in specific experimental conditions or biological samples, contributing to the observed variability in the datasets.

Only the proteins with missing values in the three groups (M0, M1 and M2) were filtered during the Differential Expression Analysis.

Regarding the single-cell dataset, I did not contain any missing values as the pipeline used by the authors imputed null values as they describe in the methodology [49].

Outliers

During the initial data exploration, I analyzed the relative abundance values to detect potential outliers. While some proteins exhibited values exceeding the interquartile range (IQR), I made the decision to retain all data points. This decision was based on the nature of the data, which probably suggests a high abundance of these proteins in the samples. Retaining these values ensures that biological variability and high expression levels of specific proteins are captured, which are critical for understanding the proteomic profile.

With respect to the differences in value scales across the various datasets, this discrepancy did not pose an issue. The analysis focused on the normalized abundance ratios between groups rather than the absolute abundance within each group.

Additional plots and all detailed information regarding the results obtained during the phase of Exploratory Data Analysis can be found in the Notebook (see Additional Materials).

Quantitative proteomic profiling

As the first step of the analysis, I summarized the number of proteins identified at the dataset level, both before and after applying the previously defined filters (A). To better understand the number of proteins identified for each cell type, I grouped the data by type (B).

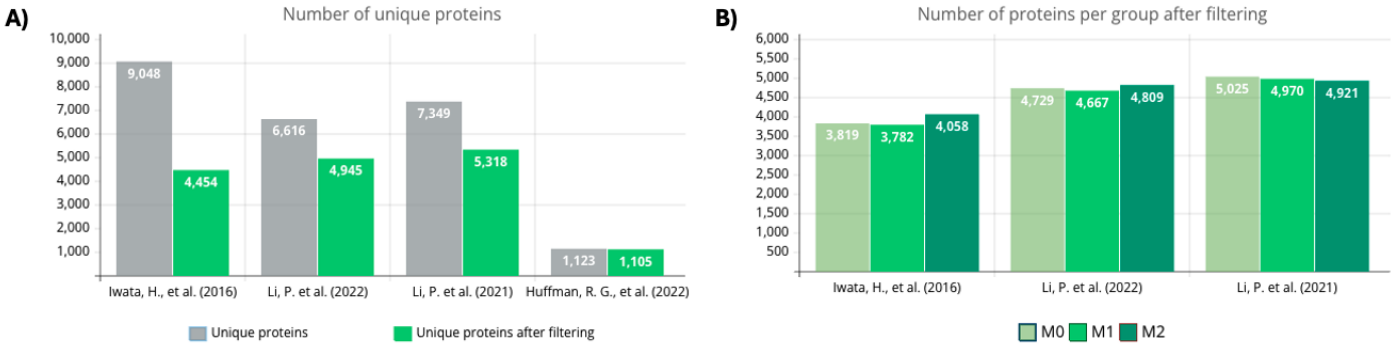


Figure 24 – (A) The charts display the number of unique proteins before and after applying the filter criteria. (B) Chart with the number of unique proteins identified across different datasets, grouped by type, after filtering.

In the case of Huffman et al. 2016, it was not possible to exactly determine which proteins were identified in each group due to the format used by the authors in their publication.

With the data already filtered and grouped by type, the next step was to analyze how these proteins were distributed across the different groups. The goal was to identify similarities among the results obtained in each study. In the case of Iwata H., et al. (2016), the number of proteins unique to each group was higher than in the other two bulk studies. For these studies, the criterion used to determine which proteins were present in each group was "Abundances (Grouped): Mx" > 0 (where x corresponds to 0, 1, or 2).

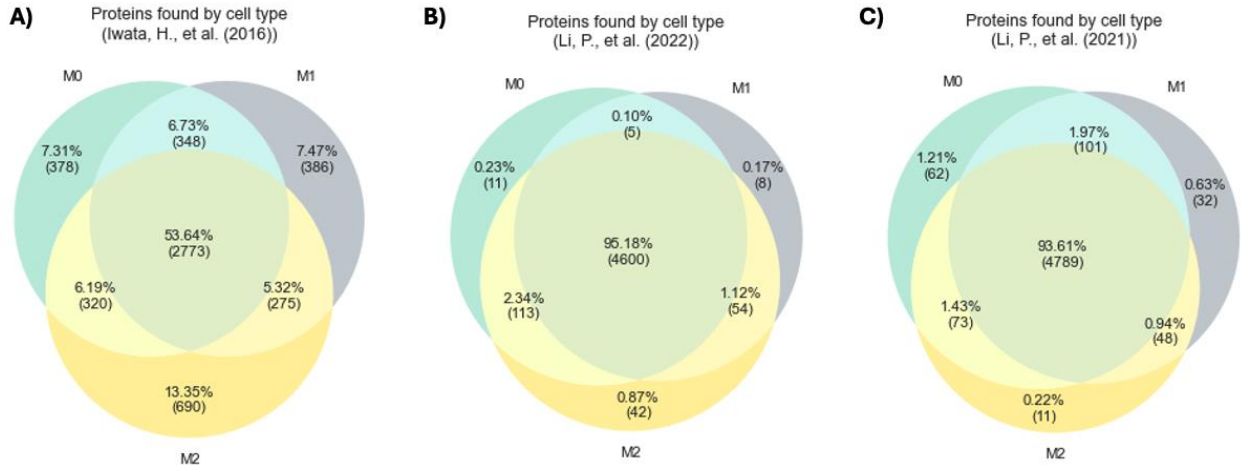


Figure 25 – Venn diagrams comparing the number of proteins identified in three cell types (M0, M1, M2) across the three different bulk datasets. The diagrams highlight the differences and similarities in protein identification across different studies and cell types, reflecting the variability and consistency in proteomic analyses. Note that for a better interpretability, the diagrams do not respect the overlap proportion.

At this stage, my focus was on the proteins present in all three datasets and how they were distributed across the different groups. To achieve this, I filtered only those proteins that were common to all. To study the overlapping proteins in the single-cell data for M0 and M1, I utilized the Gene Name. As shown in Figure 26.B, the majority of genes are expressed in both groups. However, a small number of genes are expressed exclusively in one of the two phenotypes.

Finally, when comparing the genes expressed in the bulk datasets with those expressed in the single-cell data, we observed that approximately 5% (57/1048) of the genes expressed in single-cell data were not expressed in the bulk datasets.

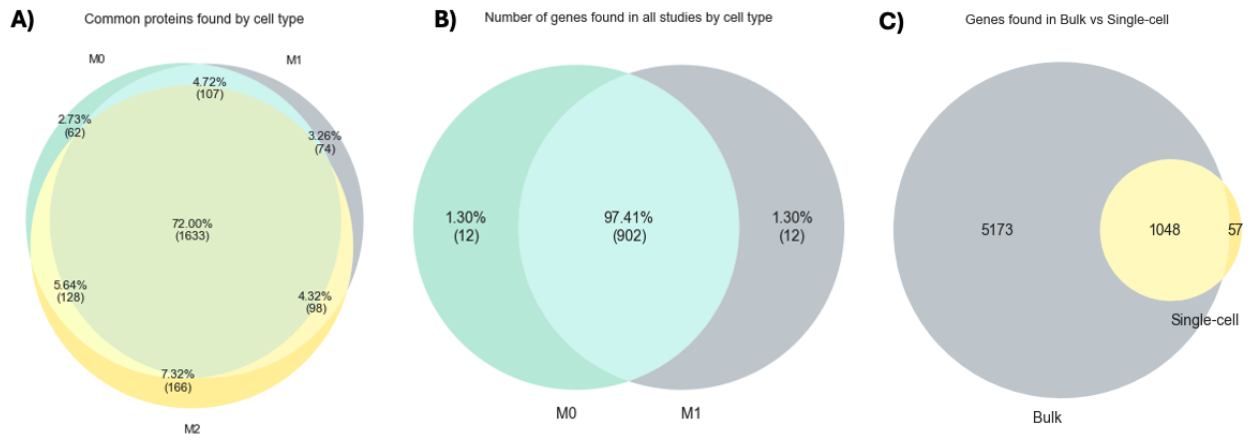


Figure 26 – (A) The diagram shows how the proteins identified in all bulk datasets are distributed among different groups. (B) The diagram represents the number of genes found in the four datasets and their distribution by group. (C) The diagram shows the number of genes found exclusively in one of the two methods and the number of genes expressed in both techniques.

As observed in the figure 26.C, most proteins identified in single-cell experiments were also present in bulk samples. However, some proteins were exclusively detected in either bulk or single-cell experiments. Potential reasons for these discrepancies could include:

- **Sensitivity:** Bulk proteomics often offers higher sensitivity, enabling the identification of low-abundance proteins that single-cell proteomics (SCP) might miss due to the smaller sample input. However, SCP provides higher resolution, allowing the detection of rare proteins.
- **Number of cells included in experiments:** Bulk experiments analyze a large number of cells together, while SCP examines individual cells. This difference in population size can lead to variations in the detected proteins.
- **Heterogeneity:** SCP can identify proteins that are present only in specific cells within a population. In contrast, bulk experiments average the signal from many cells, potentially masking these proteins.
- **Post-Translational Modifications (PTMs):** SCP can capture PTMs that may be missed in bulk experiments, as these modifications might not be uniformly distributed across the entire cell population.
- **Method and Technical variability:** The three bulk datasets were acquired using Data-Dependent Acquisition (DDA), while the single-cell dataset was obtained through Data-Independent Acquisition (DIA). When working with proteomics data, the method of data acquisition can significantly impact the proteins detected in the experiment.
  The proteins found only in the single-cell dataset might be due to DIA's full scanning approach, which captures a wider spectrum of proteins compared to DDA's selective targeting. The field of proteomics is rapidly evolving, with continuous improvements in

methods and technologies. Variations in sample preparation and processing can significantly impact the final results, contributing to differences in protein identification between bulk and single-cell experiments.

The data obtained during this phase was used as the base to later compare the protein levels in the different groups and also between the different techniques.

Differential Expression Analysis (DEA)

Given the aforementioned lack of consistency among the data sets and the significant and widespread variations in protein abundance, I decided not to perform differential expression analysis on the bulk data. These inconsistencies made the results unreliable and made it clear that the data could not provide a valid basis for meaningful comparisons.

DEA (Bulk datasets)

The lack of correlation among the bulk datasets led me to decide against combining them for PCA and UMAP calculations. Instead, I decided to further investigate the clusterization tendency of the dataset that has relative abundance rates for six different time points (Iwata, H., et al. 2016). The goal was to try to identify patterns of evolving relative abundances for all the proteins and classes.

The dataset contains the relative abundance of the proteins at the initial time and then after 8 hours, 12 hours, 24 hours, 48 hours, and 72 hours, with all values normalized to the initial time point.

Time series clustering is a technique used to group time-dependent data into clusters, where each cluster contains time series that exhibit similar patterns or behaviors over time. To carry out this part, I opted for TimeSeriesKMeans, which is a specific algorithm designed for time series clustering. It extends the traditional KMeans algorithm by incorporating distance metrics like Dynamic Time Warping (DTW), which accounts for temporal distortions between time series.

As the data contains the relative abundance of all the proteins identified in M0, M1, and M2, I decided to call this algorithm using 3 as the number of clusters to identify.

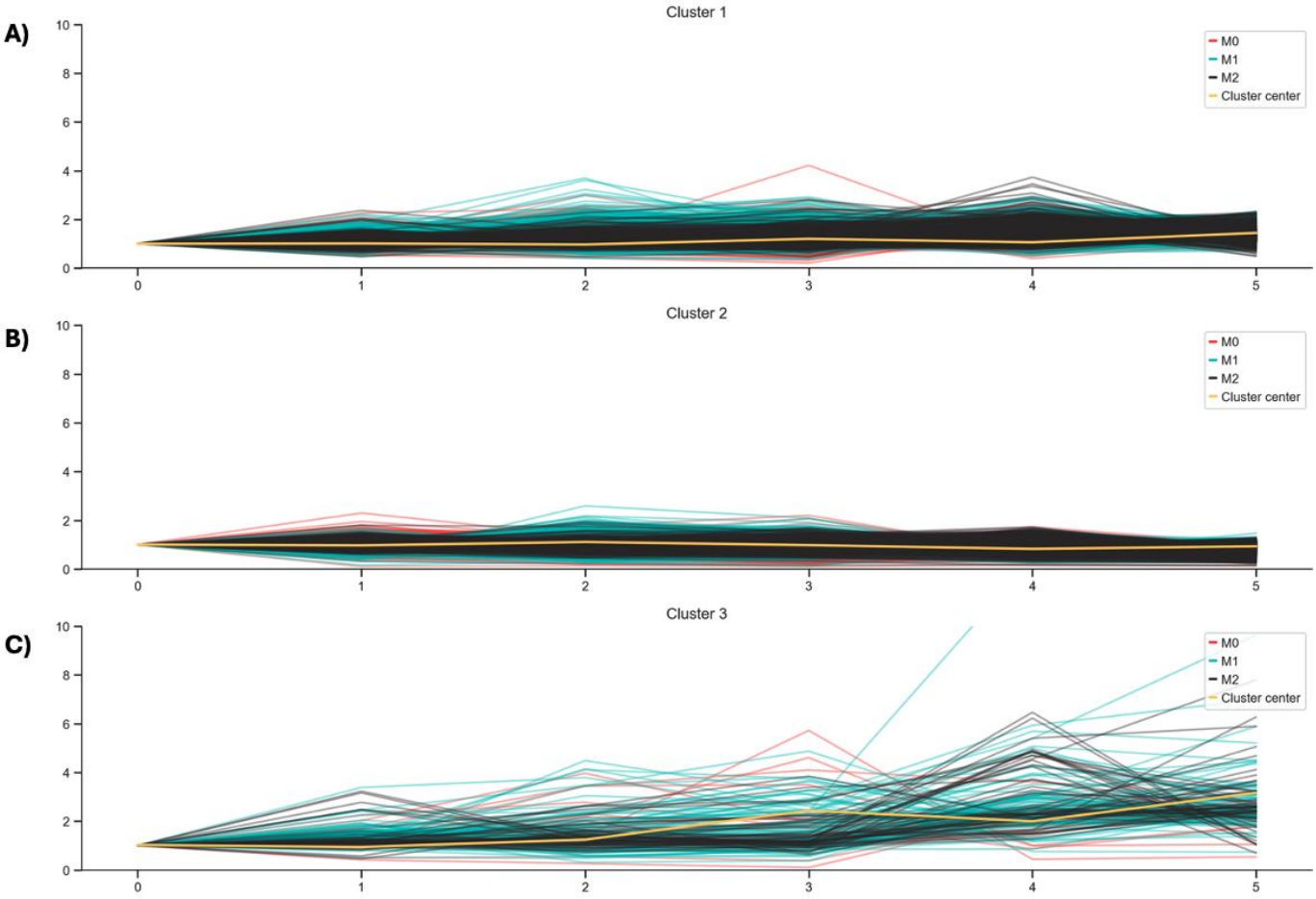


Figure 27 - Time series clustering with three clusters and corresponding lines for each of the proteins represented in three different colors according to their class. The yellow line represents the cluster center.

After performing time series clustering on the proteomics dataset, the results are summarized in the table below.

	Number of proteins	Percentage (%)
Cluster 1	**2926**	**25,10%**
M0	*632*	5,42%
M1	*1224*	10,50%
M2	*1070*	9,18%
Cluster 2	**8585**	**73,63%**
M0	*3171*	27,20%
M1	*2476*	21,24%
M2	*2938*	25,20%
Cluster 3	**148**	**1,27%**
M0	*16*	0,14%
M1	*82*	0,70%
M2	*50*	0,43%

Table 1 - Table with the number of proteins and their corresponding percentages in each of the three identified clusters. Additionally, for each cluster, the number of proteins and their percentages are further broken down by the subcategories M0, M1 and M2.

These results provide an overview of how proteins are distributed across the different clusters and subcategories. Further investigation of these clusters and the genes and proteins that form them could help to understand complex biological phenomena.

DEA (Single-cell dataset)

Due to the variability of bulk data and the interest on single-cell data, I then focused in the single-cell dataset. I performed a principal component analysis (PCA) to simplify and visualize the data. PCA reduces high-dimensional data to key components that capture most of the variance, making it easier to interpret. I was able to identify two distinct groups in the 373 cells, corresponding to the untreated cells and the LPS stimulated cells.

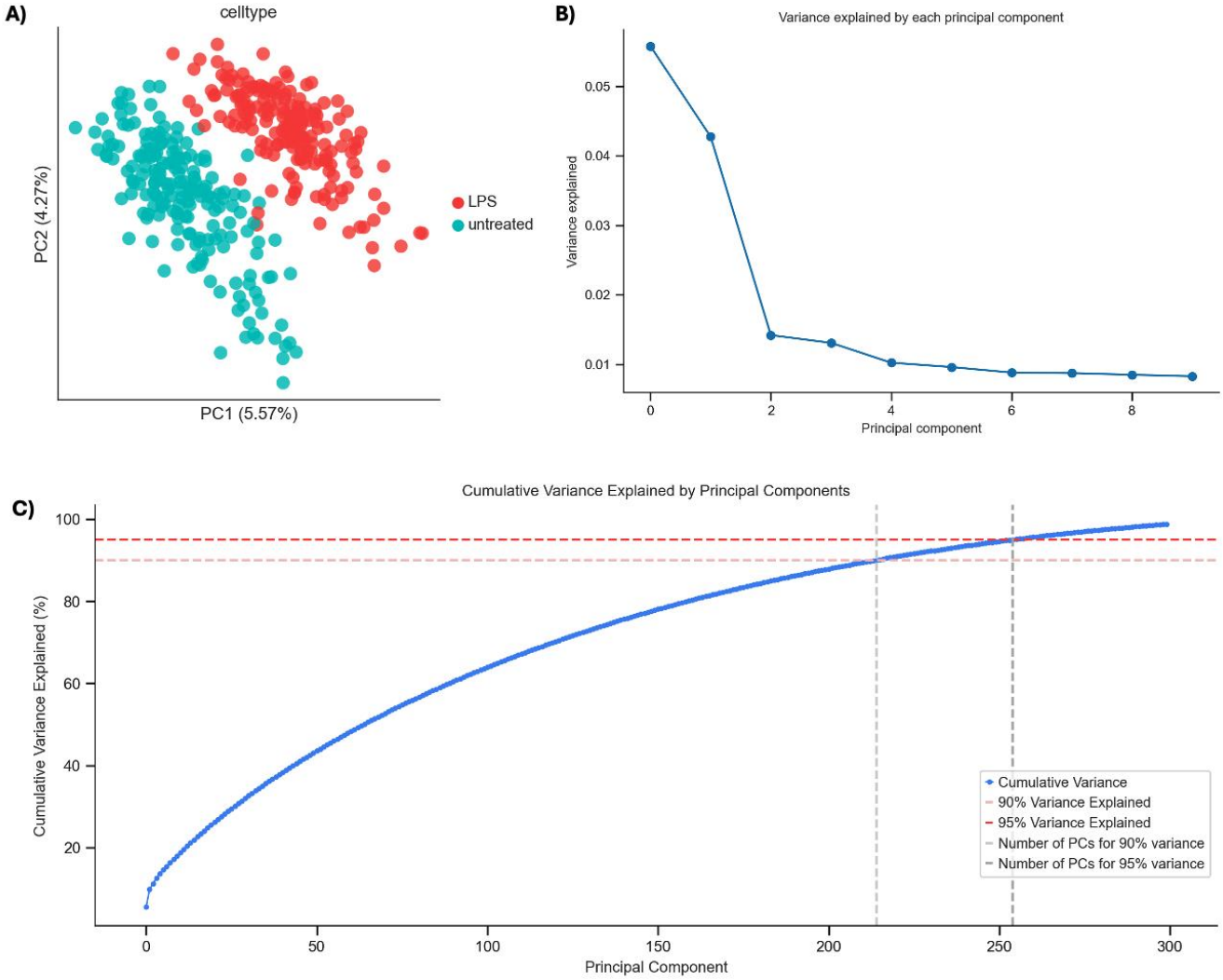


Figure 28 - (A) Principal Component Analysis (PCA) Plot of M0 and M1 Macrophages: The distinct clustering of the cell types suggests significant differences in their proteomic profiles, with PC1 and PC2 capturing the main sources of variance between the cell types. (B) Scree Plot: Variance explained by each principal component. The plot confirms that the first two principal components account for the majority of the variance. (C) Cummulative variance plot with the 90% and 95% thresholds as red dashed lines and their corresponding number of PC's as dashed grey lines.

The key genes driving the separation of these two clusters are known to play a critical role in macrophage polarization and immune responses.

The gene that contributes the most is Interferon **Induced Protein With Tetratricopeptide Repeats 1** (IFIT1), which is involved in the body's defense against viruses, helping to block viral replication[50].

The second one **Superoxide Dismutase 2** (SOD2) protects cells from damage caused by harmful molecules called free radicals [51].

And ISG15 and ISG20 are part of the immune response, helping to tag and eliminate viral proteins and damaged cellular components.

Gene / Description	Loading
IFIT1 Interferon Induced Protein With Tetratricopeptide Repeats 1	0.144957
SOD2 Superoxide Dismutase 2	0.139648
ISG15 ISG15 Ubiquitin Like Modifier	0.129619
ISG20 Interferon Stimulated Exonuclease Gene 20	0.126352
GBP2 Guanylate Binding Protein 2	0.125942
FCER1G Fc Epsilon Receptor Ig	0.122075
UBXN4 UBX Domain Protein 4	0.114359
PRDX1 Peroxiredoxin 1	0.109036
IFIT3 Interferon Induced Protein With Tetratricopeptide Repeats 3	0.093034
SAMHD1 SAM And HD Domain Containing Deoxynucleoside Triphosphate Triphosphohydrolase 1	0.091062

Table 2 - Loadings for the genes that contribute more with the separation of the two groups (untreated and LPS)

The next step was to use Uniform Manifold Approximation and Projection (UMAP) to further explore the single-cell data. UMAP is a powerful dimensionality reduction technique that preserves local and global structures in the data, making it a powerful tool for visualizing and understanding complex biological data sets. By applying UMAP, I aimed to uncover patterns and relationships within the single-cell macrophage proteomic data. This allowed me to identify **two well separated clusters within the untreated group.**

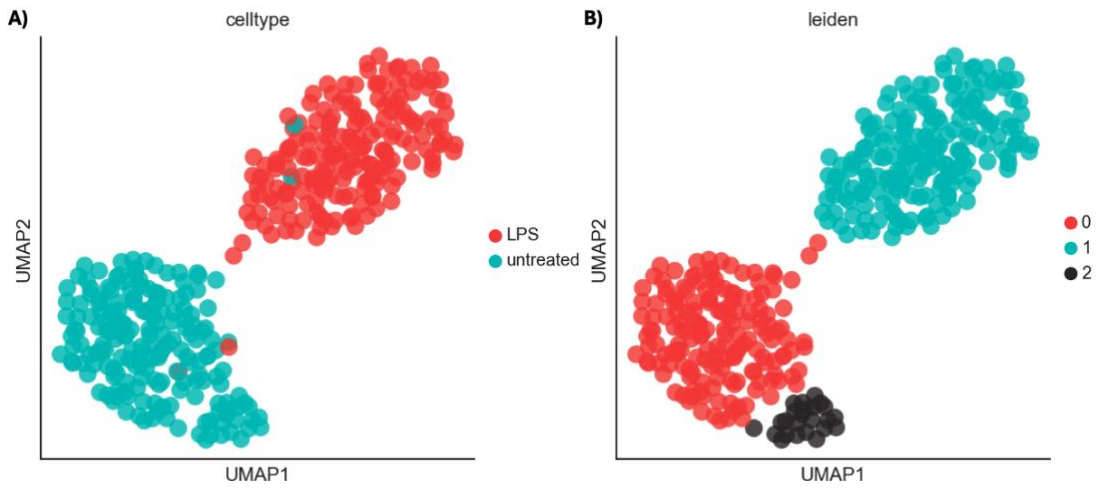


Figure 29 - (A) UMAP plot by celltType: The plot highlights the separation and clustering of cells based on their treatment status (untreated/LPS). (B) UMAP plot by Leiden clustering: Three different clusters were identified using Leiden algorithm for the only two different cell subgroups.

I performed a t-test to compare the differential protein abundance between these two groups (untreated and LPS). A t-test is a statistical method used to determine if there is a significant difference in the means of two groups, which, in this context, helps to identify proteins that are differentially expressed due to LPS stimulation. The aim was to identify specific proteins that show significant changes in expression levels, thus highlighting the impact of LPS on macrophage function.

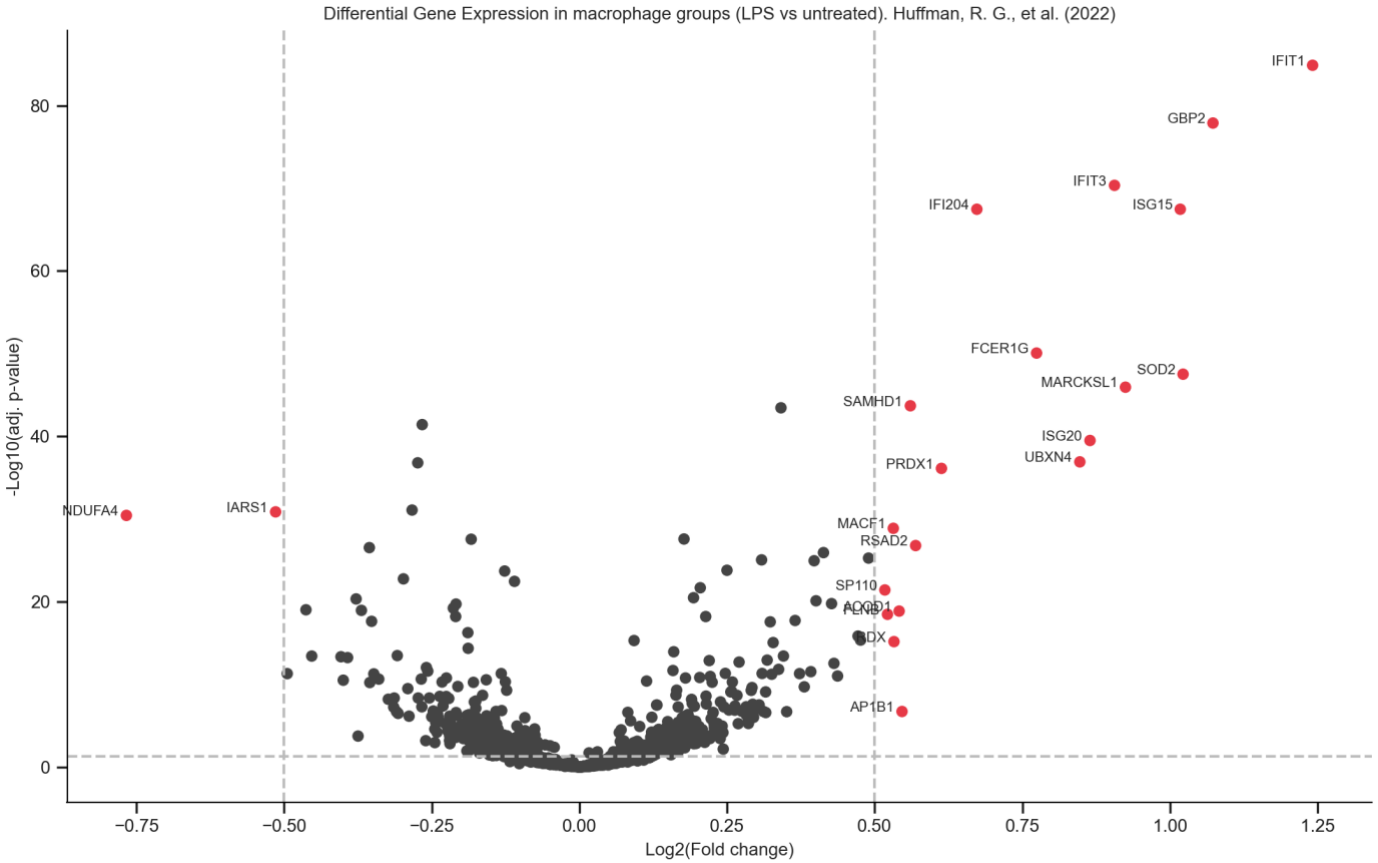


Figure 30 - Volcano plot for the Huffman, R. G., et al. (2022) dataset. The plot illustrates the relationship between the fold change (Log2(Fold Change)) and statistical significance (-Log10(adj. p-value)) of differentially expressed proteins in BMDMs derived macrophages under LPS versus untreated conditions. Each point represents a protein, with black points indicating proteins that are not significantly differentially expressed and red points indicating proteins that are significantly differentially expressed (adj. p-value < 0.05). Highlighted proteins, such as IFIT1, GBP2, ISG15, and SOD2, demonstrate significant differential expression and are key drivers in the observed immune response.

Heterogeneity within M0 cells

Having the single-cell dataset provided me with a unique opportunity to study **heterogeneity** within the same cell type, focusing on the M0 group (since I had previously identified two subgroups).

For this, I went through PCA and UMAP again, but this time only with the data from the untreated cells. The result was similar and evident in the two methods I used, a clear division between the two subgroups of cells, one containing 150 and the other one 37.

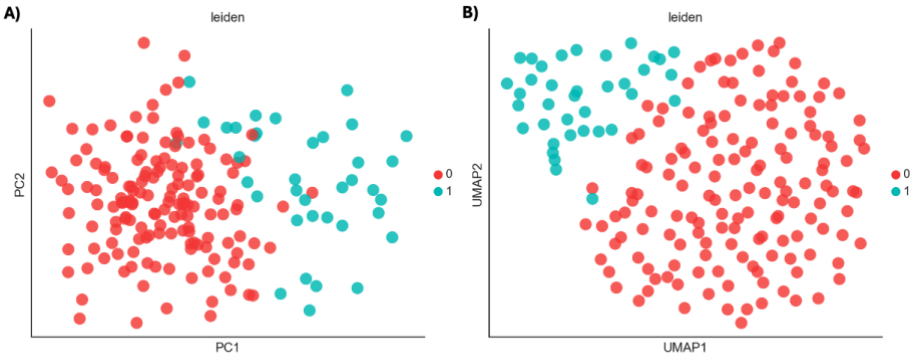


Figure 31 - Separation between the two cell subgroups as determined by PCA and UMAP methods. Both visualizations reveal a relatively clear distinction between the subgroups.

By analyzing the PCA loadings, I could identifying the most significant genes contributing to the first principal component (PC1).

The two genes with higher loadings were S100A11 and RBM39.

The gene **Calcium Binding Protein A11** (S100A11) has biological functions such as regulating cell growth, enzyme activity, and the inflammatory response. It is involved in the regulatory process of cancers, metabolic diseases, neurological diseases, vascular calcification, and inflammatory diseases [52].

The second most significant gene, **RNA Binding Motif Protein 39** (RBM39), is an RNA-binding protein involved in transcriptional co-regulation and alternative RNA splicing.

Gene / Description	Loading
S100A11 Calcium Binding Protein A11	0.221746
RBM39 RNA Binding Motif Protein 39	0.197669
GPNMB Glycoprotein Nmb	0.146500
EVL Enah/Vasp-Like	0.123098
ATP6V1G1 ATPase H+ Transporting V1 Subunit G1	0.116201
NIBAN2 Niban Apoptosis Regulator 2	0.109679
ATP6V1E1 ATPase H+ Transporting V1 Subunit E1	0.107943
ESD Esterase D	0.105199
MUG1 Murinoglobulin 1	0.086481
ATP6V1B2 ATPase H+ Transporting V1 Subunit B2	0.086423

Table 3 - Loadings for the genes that contribute more with the separation of the two subgroups identified in the M0 cells.

I then performed a t-test to compare the differential protein abundance between these two subgroups or clusters. The analysis revealed that the RBM39 and S100A11 genes exhibit significant up and down regulation, each surpassing the threshold of +/-1 in log2 fold change, respectively. Additionally, nine other genes exceeded the +/-0.5 threshold, indicating notable differential expression.

It is important to note that heteregeneity within the same group may be due to several factors, such as differences in cell state or environment, among others.

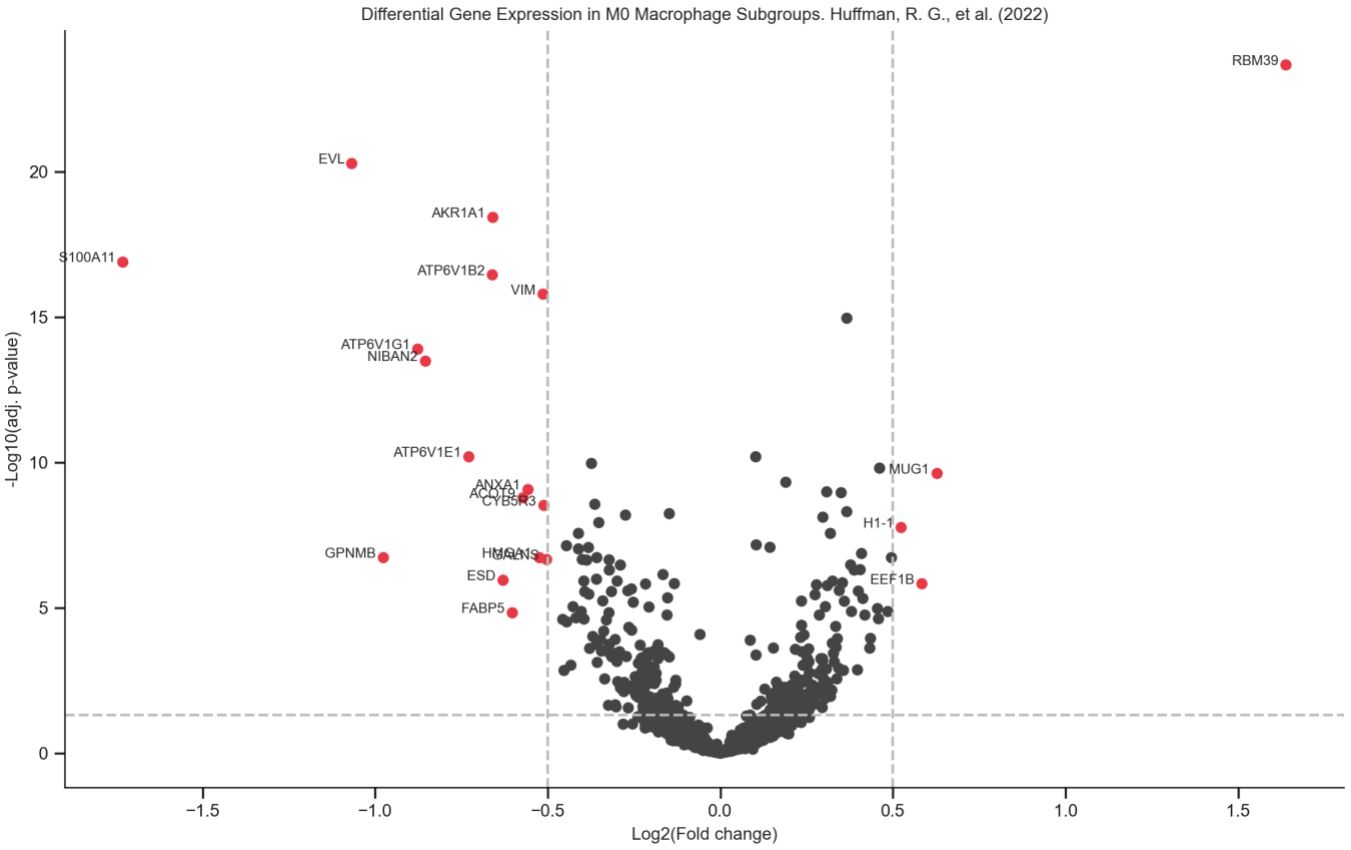


Figure 32 - This volcano plot illustrates the differential gene expression analysis between two clusters of M0 cells. Three of the up/down-regulated genes exceed the fold change thresholds of ±1, indicating significant changes in expression. Additionally, several genes fall below the -0.5 fold change threshold, showing substantial downregulation. These findings highlight the differences in gene expression between the two M0 cell clusters, making evident the cellular heterogeneity within the same cell type.

Heterogeneity within M1 cells

After analyzing the M0 cells, I proceeded to examine the M1 cells with the same steps performed before. While clusters were not identified for the M1 group when conducting the analysis with both untreated and LPS-stimulated cells, the PCA and UMAP results still revealed a tendency towards clustering, albeit with some partial overlap. This indicates that, similar to the untreated group, the LPS cells also exhibit a degree of heterogeneity.

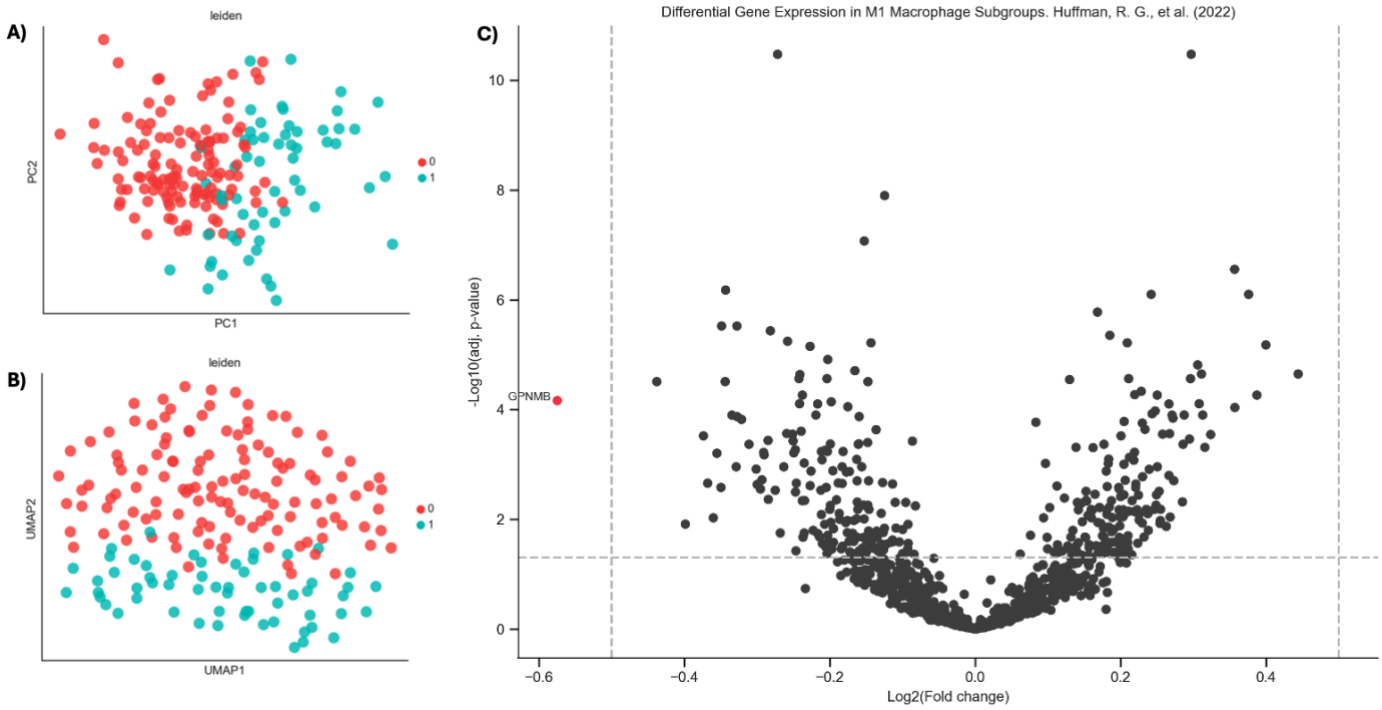


Figure 33 – (A, B) Both PCA and UMAP illustrate the clustering patterns of M1 macrophages. The UMAP plot, in particular, highlights more distinct and well-separated clusters compared to PCA. (C) Volcano Plot with the Differential Gene Expression Analysis (DEA) for M1 Cell Type. The volcano plot shows only a gene exciding the fold change thresholds of ±0.5, indicating significant upregulation or downregulation in M1 cells beyond these thresholds. In contrast, comparisons with M0 cells reveal some genes surpassing the ±1 fold change threshold, highlighting the differences between these cell types.

These findings highlight the intrinsic diversity also within the M1 macrophage population, suggesting that even under inflammatory conditions induced by LPS, the cells display a range of responses and characteristics.

Pathway Analysis

To gain deeper insights into the biological significance and gene functions contributing to the heterogeneity within the groups and the process of macrophage polarization, I performed Enrichment Analysis using both ORA and FCS methods[53]. GSEApy[54], a Python/Rust implementation of GSEA (FCS) and wrapper for Enrichr (ORA), was used to perform Gene Set Enrichment Analysis[55][56].

The main difference between overrepresentation analysis (ORA) and functional class scoring (FCS) lies in its approach to gene set analysis. ORA focuses on identifying specific pathways or sets of genes that are statistically overrepresented in a given list of genes (usually differentially expressed genes) by comparing the observed frequency of these genes to what would be expected by chance. This method highlights pathways that are significantly enriched.

On the other hand, FCS evaluates the global expression patterns of predefined sets of genes, assessing the collective behavior of groups of genes based on a ranking, rather than individual genes. It scores these gene sets based on their association with the biological condition studied, highlighting functional classes that show coordinated expression changes relevant to the experimental data.

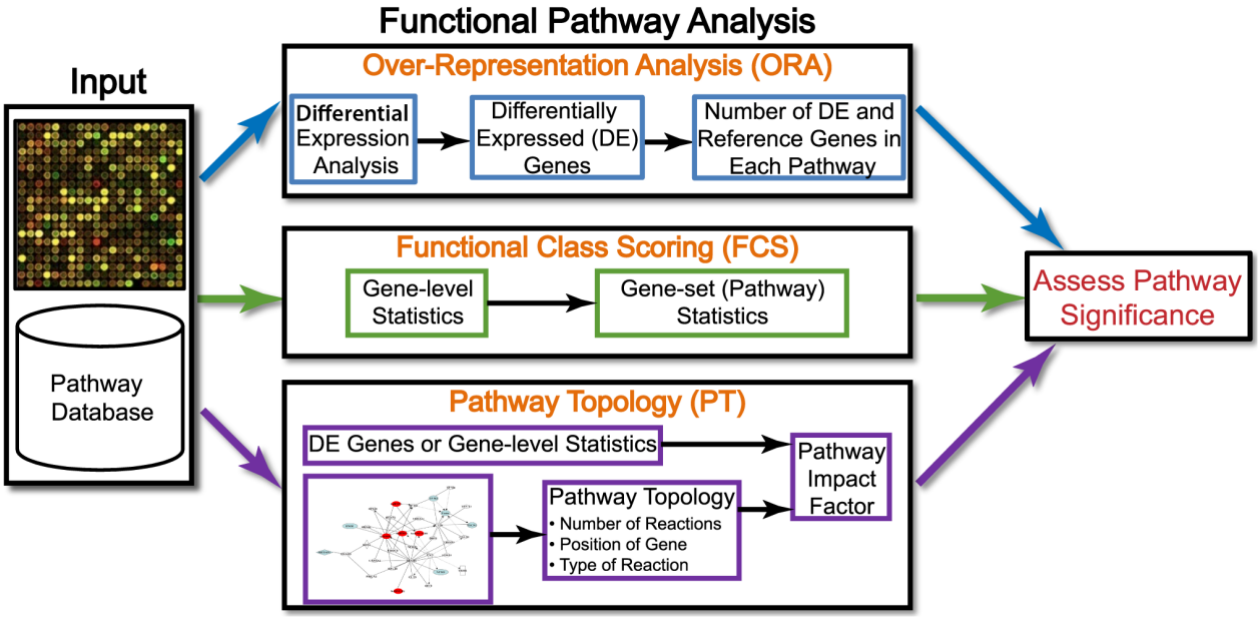


Figure 34 - The figure outlines the three main analysis methods: Over-Representation Analysis (ORA), Functional Class Scoring (FCS), and Pathway Topology (PT)[53]. Each method employs different statistical approaches to assess pathway significance and ultimately evaluate pathway impact factors.

The following two functions were used taking advantadge of the capabilities offered by GSEApy:

- ***Enrichr (ORA)***: It only requires a set of differentially expressed genes to identify significantly overrepresented gene sets. A threshold of 0.05 for the adjusted p-value was used to select the differentially expressed genes.
- ***Prerank (FCS)***: This method requires a pre-ranked list of genes as input to perform Gene Set Enrichment Analysis (GSEA). It is ideal for cases where you have data from two different groups with differential expression scores. For this purpose, I computed the "RANK" value as follows:

$$s_i\ (Rank) = Significance(FoldChange\ gene_i) * \log_{10}(p - value_i)$$

The result is a sorted list of genes with their rank. At the top of the list there are the most significant and upregulated genes, while at the bottom of the list the most significant and downregulated genes. In the middle of the list, there are those that are not significant.

	Gene	RANK
0	IFIT1	105.362426
1	GBP2	83.550830
2	ISG15	68.632580
3	IFIT3	63.735229
4	SOD2	48.560299
..	...	...
556	ATP1A3	-9.437372
557	CNPY4	-10.072146
558	MTHFD1	-11.025824
559	IARS1	-15.873025
560	NDUFA4	-23.339302

Table 4 - Ranked list of genes for untreated vs LPS cells. As can be observed, IFIT1 gene is the most significant and up-regulated gene, while NDUFA4 is the most significant and down-regulated gene.

In this rank metric, genes that are up-regulated with relatively low p-values are positioned at the top of the list, while down-regulated genes with low p-values are placed at the bottom.

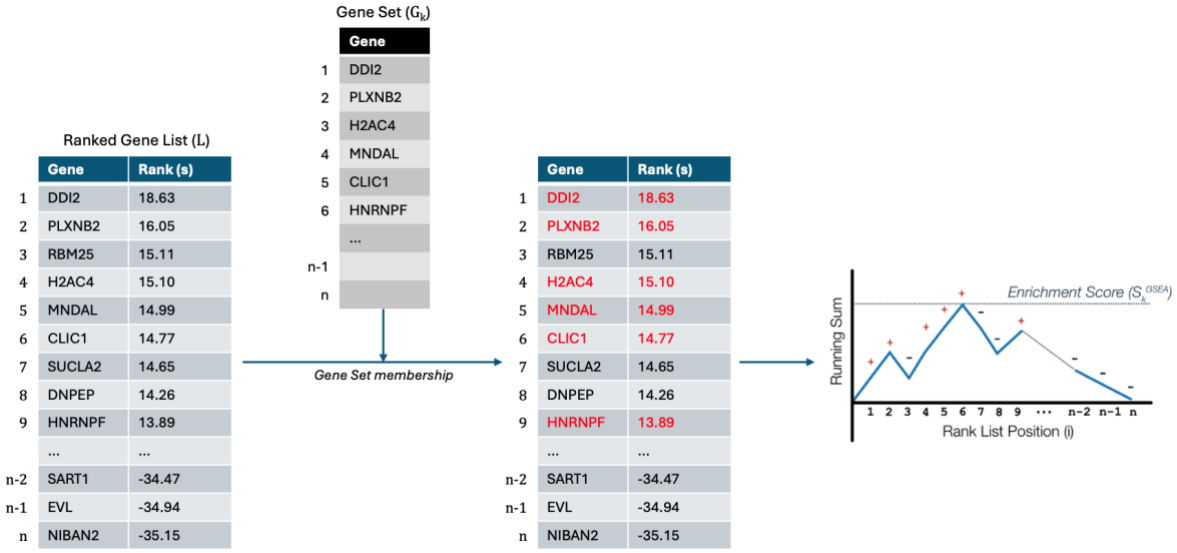


Figure 35 – To compute the GSEA Enrichment Score (S), a ranked gene list (L) is needed as well as a candidate gene set (G). The process involves computing a running sum by iterating through each gene in the ranked list. If the gene is part of the candidate gene set (indicated with a red +), the sum is incremented; otherwise is decremented (indicated with a black -). The Enrichment Score (S) is defined as the highest value reached by this running sum at any point in the list. It can also move in the negative direction. The Enrichment Score (S) is actually the maximum absolute value of the running sum.

ORA methods have some well-known limitations that FCS methods try to overcome. Some of the most relevant limitations are:

- Arbitrary thresholds: ORA relies on setting thresholds such as the fold change or p-value to define the get a list of differentially expressed genes. Small changes in these thresholds can lead to different results, which introduces subjectivity and potential bias.
- Unused quantitative data: ORA methods analyze only differentially expressed genes and ignores the magnitude of the changes like expression level or fold change.

- It ignores gene correlation: ORA assumes genes act independently within pathways, which is not true in most of the cases. By ignoring gene correlation, it can miss a interesting and complex biological pathways.
- False positives: ORA often produces false positives due to multiple testing, especially when many gene sets are analyzed simultaneously.

To capture a more complete picture, and especially considering the limitations of ORA methods, I employed both the GSEApy *prerank* (FCS) and *enrichr* methods (ORA).

The following databases of gene sets from GO[57], KEGG[58], REACTOME[59] and MSigDB[56] were used:

- GO_Biological_Process_2023
- GO_Cellular_Component_2023
- GO_Molecular_Function_2023
- KEGG_2019_Mouse
- Reactome_2022
- MSigDB_Hallmark_2020

Pathways activated or downregulated in both untreated and LPS stimulated cells

The results of analyzing the most up/down-regulated pathways with ORA include "Infectious Disease," "Cellular Responses to Stress," "Innate Immune System," and "Oxidative Metabolism," among others. These results confirm some of the significant biological processes and pathways active that were expected of macrophages.

From a quantitative point of view, 948 pathways have been identified as significant using a threshold of 0.05 for the Adjusted P-value. Using a less stringent threshold (0.25), this figure rises to 1,984 pathways identified as overrepresented.

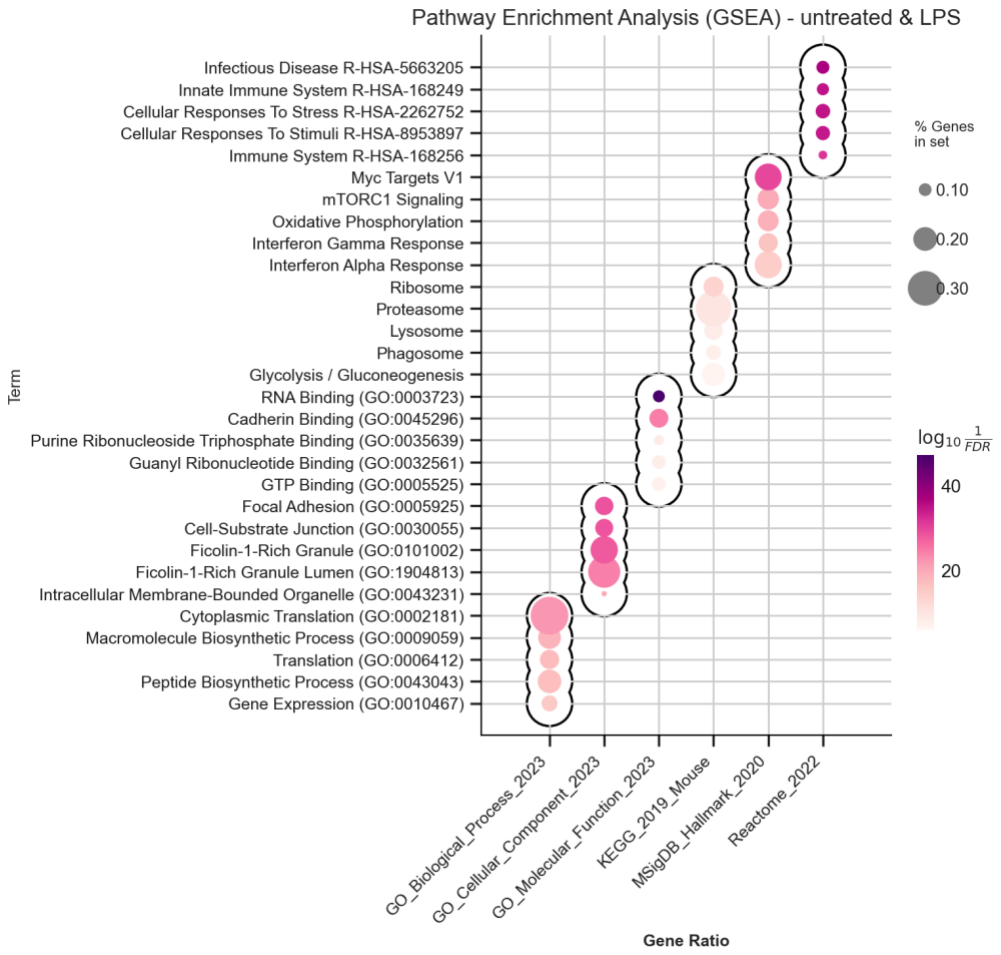


Figure 36 – This figure displays the top 5 enriched terms for each database, illustrating the effects of LPS treatment compared to untreated cells. The color gradient indicates the $\log_{10}(1/FDR)$, reflecting the statistical significance of each term. The size of each marker represents the percentage of genes in the dataset that overlap with the corresponding gene sets, highlighting the pathways most significantly impacted by the treatment.

As highlighted before, GSEA Prerank (FCS) was calculated to compare the results with Enrichr (ORA) results. As a prerequisite, I first calculated the rank for each of the genes using the formula mentioned earlier. Then I computed GSEA using Prerank function and plotted the results to visualize the Enrichment Score (ES) of the overrepresented pathways.

The following criteria was used to highlight pathways that are not only statistically significant but also show substantial enrichment:

- False Discovery Rate (FDR) q-value less than 0.25
- Absolute value of the Normalized Enrichment Score (NES) greater than 1

As can be seen in the figure below, only three pathways were identified, all of them due to up-regulation.

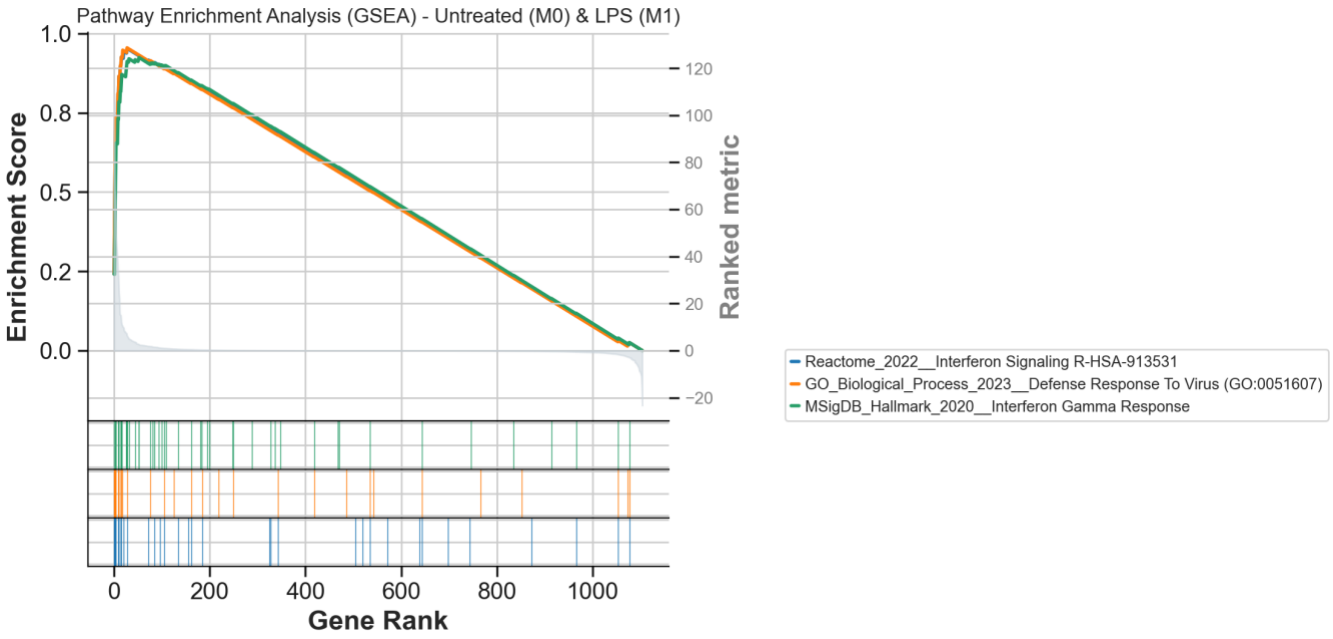


Figure 37 – Gene Set Enrichment Analysis (GSEA) illustrating the variation in the running enrichment score as we progress through the ranked list of genes. The plot provides a visual representation of the enrichment score (ES) across the ranked list of genes.

These pathways are also strongly related to the defense system of the organisms, and therefore to the cell type studied.

However, from a quantitative point of view, the results with ORA (1,984 pathways with a threshold of 0.25) are drastically different from those obtained with FCS (3 pathways with the same threshold).

Pathways activated or downregulated in untreated cells (M0)

After the initial pathway enrichment analysis in which all cells in the dataset (untreated and LPS) were included, another analysis was performed but focused exclusively on untreated cells. This analysis was intended to explore the heterogeneity within this group, identifying potential variations and distinct gene expression patterns that might be masked by comparing different treatments.

Again, both methods were used to have the complete picture for untreated cells.

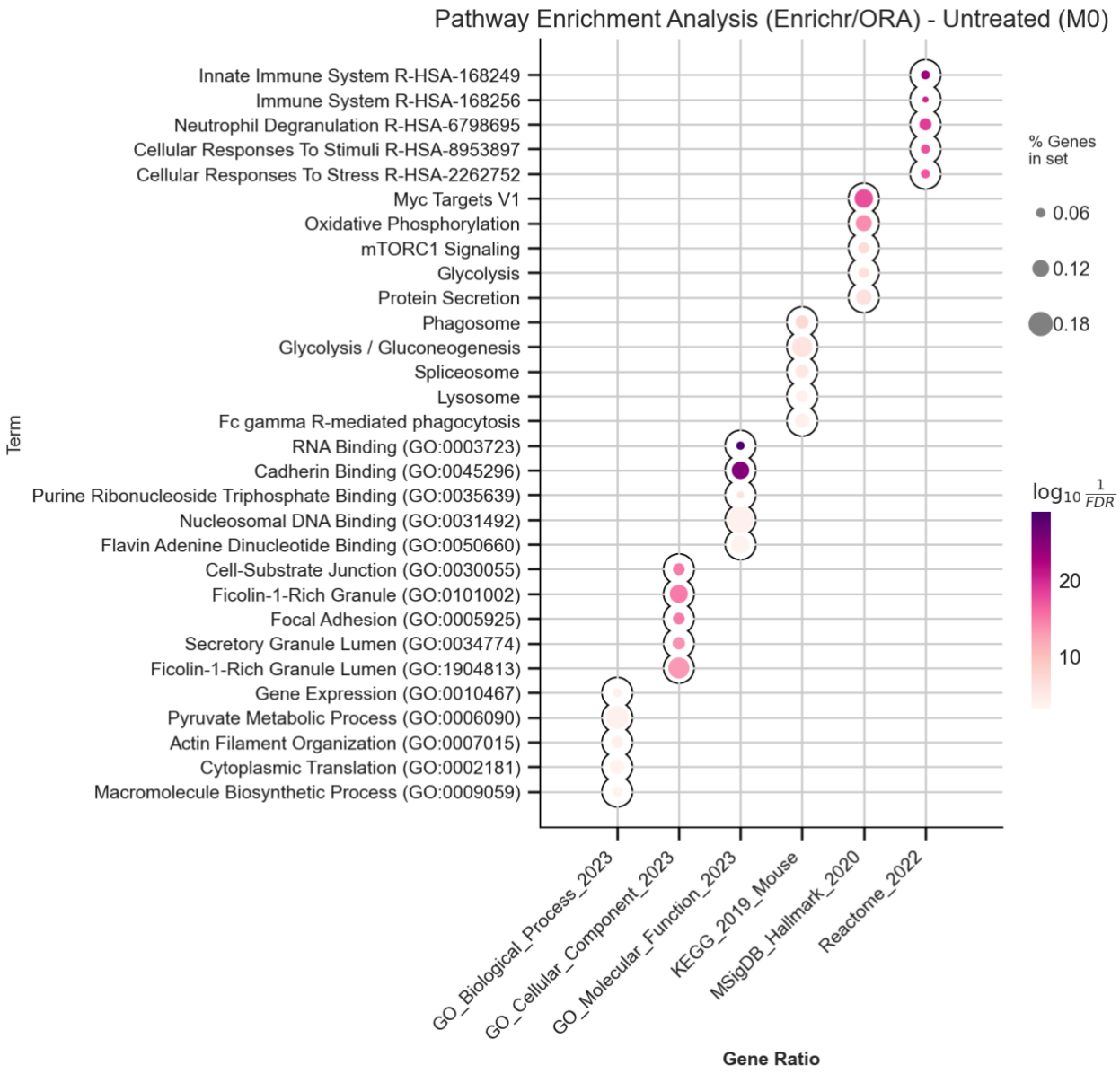


Figure 38 – Top 5 enriched terms for each database, illustrating the differences between the two identified clusters for untreated cells.

In this case, 584 pathways were identified with 0.05 threshold and 1,286 with 0.25.

The Prerank was also calculated for the untreated cells for the two identified clusters. However, none of the pathways was identified as significatnly overrepresented. As can be seen in the following table, sorted by ascending FDR q-value, none of the pathways meet the criteria of having a value below 0.25:

Term	ES	NES	NOM p-val	FDR q-val	FWER p-val	Tag %	Gene %	Lead_genes
Reactome_2022__Ion Channel Transport R-HSA-983712	-0.951278	-1.597795	0.001047	0.265182	0.3322	5/17	2.08%	ATP6V1G1;ATP6V1B2;ATP6V1E1;ATP6V1C1;ATP6V1A
GO_Cellular_Component_2023__Secretory Granule ...	-0.796372	-1.600125	0.001693	0.368576	0.3136	20/90	9.59%	S100A11;CYB5R3;GALNS;ALDOA;FABP5;ACTR2;HK3;GLB...
GO_Cellular_Component_2023__Lysosome (GO:0005764)	-0.795367	-1.550953	0.005172	0.413138	0.756	21/70	11.40%	ATP6V1G1;ATP6V1B2;ATP6V1E1;GALNS;ATP6V1C1;ATP6...
Reactome_2022__RHO GTPases Activate Formins R-...	-0.912279	-1.52893	0.009463	0.415702	0.8884	2/19	1.36%	EVL;DYNC1LI1
GO_Biological_Process_2023__Monoatomic Cation ...	-0.902288	-1.485378	0.021931	0.425335	0.9874	5/16	5.70%	ATP6V1B2;ATP6V1E1;ATP6V1C1;ATP6V1A;ATP1A1
...	...	...	...	...	...	...	...	...
GO_Biological_Process_2023__Negative Regulatio...	-0.444917	-0.746035	0.79611	1.0	1.0	1/17	5.07%	RTN4
Reactome_2022__Beta-catenin Independent WNT Si...	-0.500817	-0.924725	0.578223	1.0	1.0	9/39	20.45%	GNB1;PSMB6;PSMA7;GNB4;PPP3CB;PSMB4;PSMD13;PSMA...
Reactome_2022__FBXL7 Down-Regulates AURKA Duri...	-0.415299	-0.741557	0.814706	1.0	1.0	7/29	20.45%	PSMB6;PSMA7;SKP1;PSMB4;PSMD13;PSMA6;PSMA2
Reactome_2022__GSK3B And BTRC:CUL1-mediated-de...	-0.415299	-0.741557	0.814706	1.0	1.0	7/29	20.45%	PSMB6;PSMA7;SKP1;PSMB4;PSMD13;PSMA6;PSMA2
GO_Biological_Process_2023__RNA Metabolic Proc...	0.949994	1.740786	0.000472	1.0	0.4518	1/18	0.09%	RBM39

Table 5 - Results of GSEA (FCS) method for untreated cells sorted by ascending FDR q-val

Pathways activated or downregulated in LPS stimulated cells (M1)

Finally, Pathway Enrichment Analysis was also computed for LPS-treated cells using both methods.

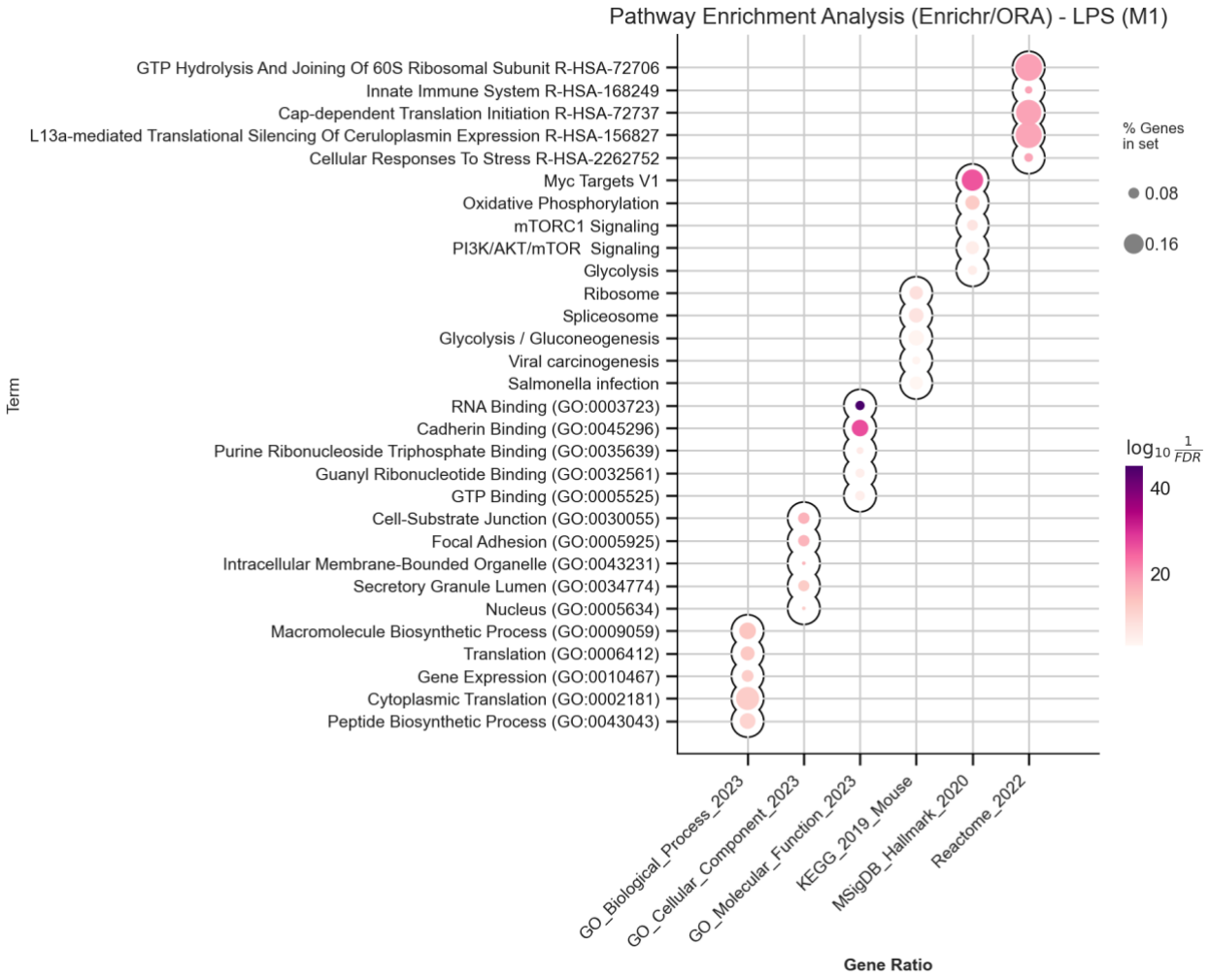


Figure 39 - Top 5 enriched terms for each database, illustrating the differences between the two identified clusters for the LPS cell group.

For LPS-treated cells, 516 pathways were identified as overrepresented using the 0.05 threshold. Using the less stringent threshold of 0.25, this number raised to 1,378.

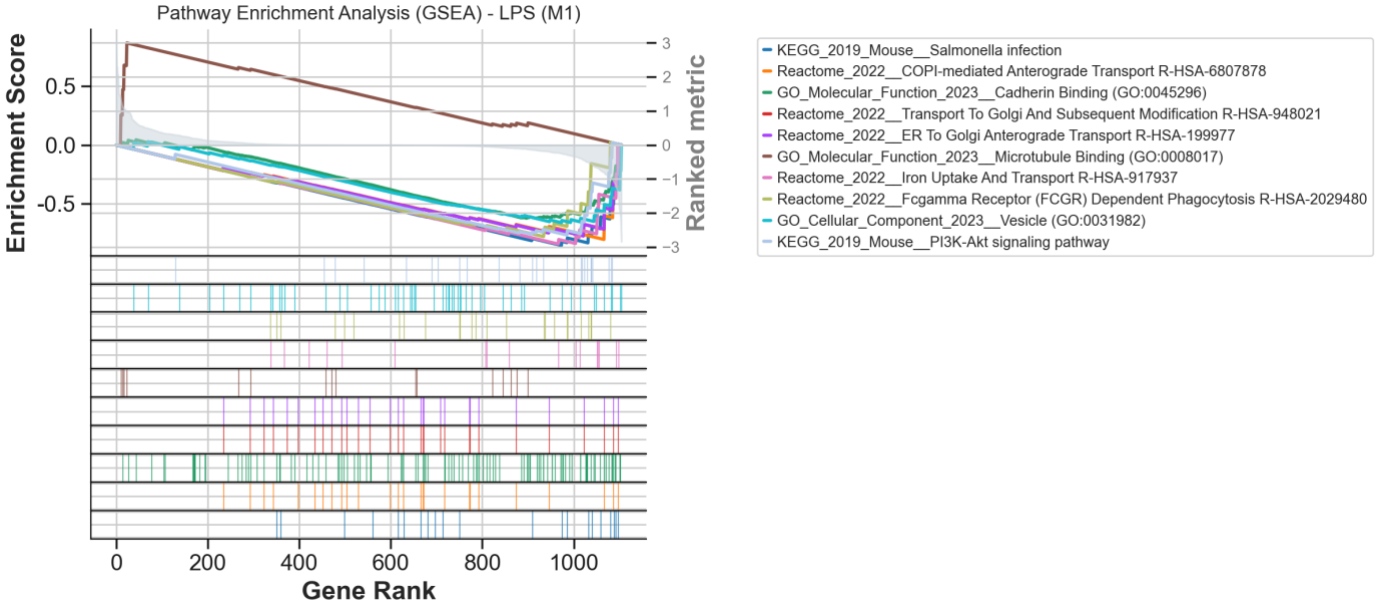


Figure 40 - This figure illustrates the results of a Gene Set Enrichment Analysis (GSEA) performed on LPS-treated cells.

Using the FCS method, no pathways were identified when a stringent threshold of 0.05 was used. However, when the threshold was relaxed to 0.25, 15 pathways emerged as overrepresented, with the majority being associated with down-regulation.

To further investigate the genes that play a role in the variability observed between individual cells, a network representation was prepared to visualize the relationship between the genes and the overrepresented pathways identified with the Prerank FCS method.

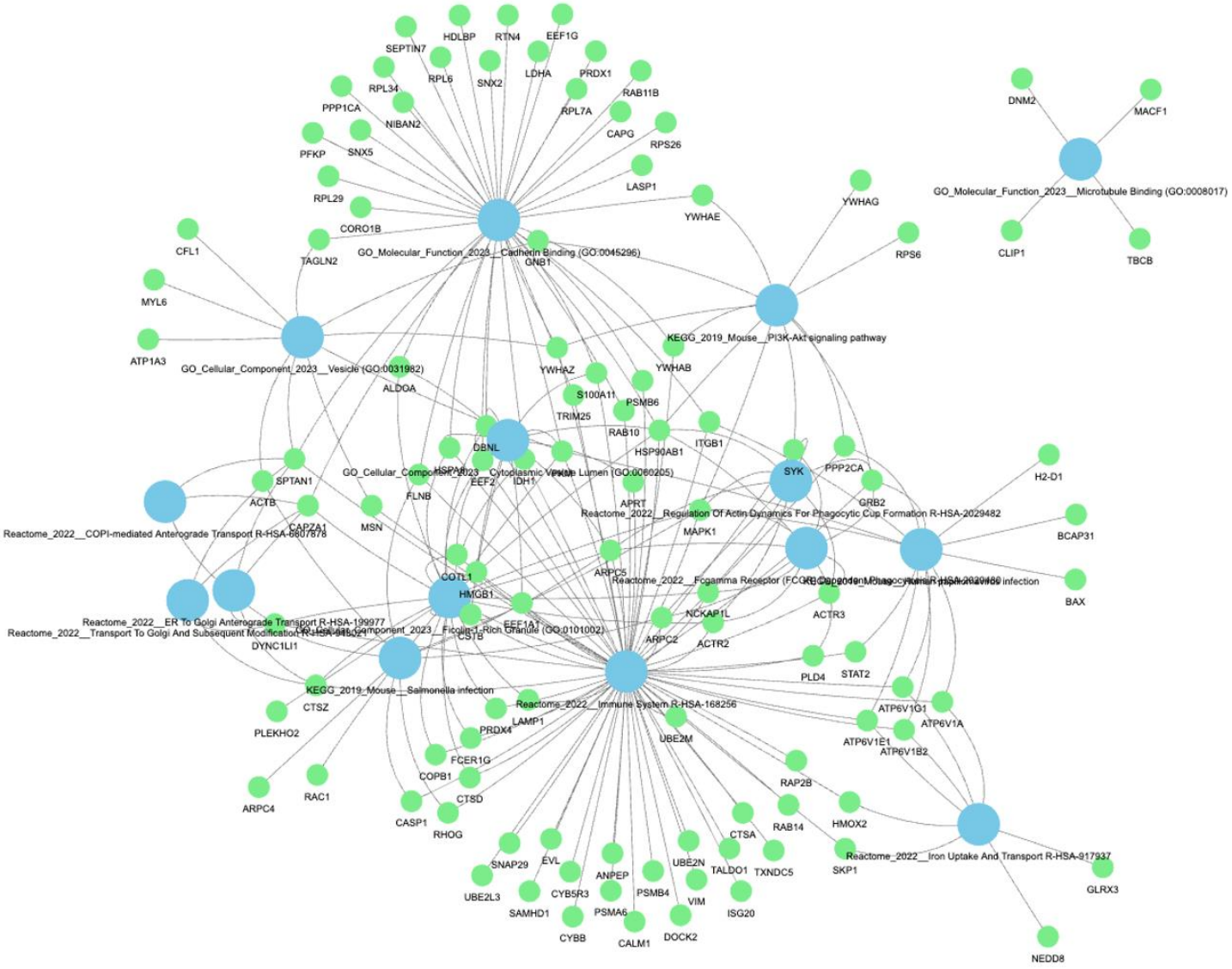


Figure 41 - This figure illustrates the gene-pathway interaction network, depicting the intricate relationships between various genes and their associated pathways. Genes are represented by green nodes, while pathways are shown as blue nodes.

For an easier interpretation of the results, especially for the M1 cell group that contains 15 pathways and several genes, an HTML versión of this network was implemented with *Pyvis* (https://github.com/WestHealth/pyvis) library.

This library provides users with the ability to filter by nodes and move nodes in the network for better visualization and understanding. The HTML files can be opened and networks visualized using a standard web browser.

Links to the HTML files can be found in this document, in the section Additional Materials.

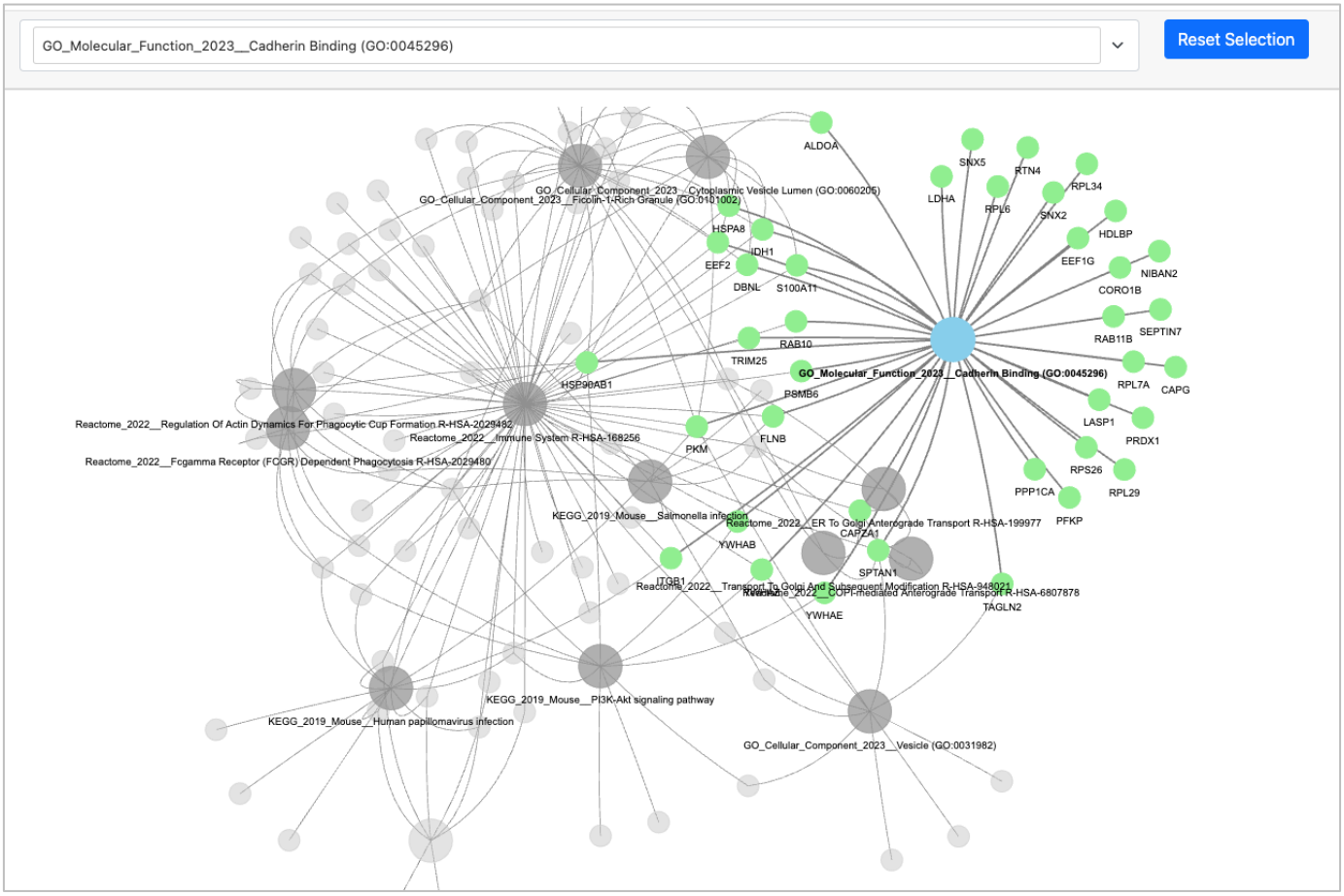


Figure 42 - The gene-pathway interaction network for M1 (LPS-treated) cells features 15 overrepresented pathways identified using the FCS method. These pathways are associated with related genes, facilitating a comprehensive understanding of their roles. In the figure, one of the pathways is highlighted to enhance interpretation and provide clearer insights into the network's biological significance.

An enrichment map was also elaborated to visualize the network with the gene-set enrichment analysis results. Each of the nodes represent a pathway and the edges represent the overlap between these pathways.

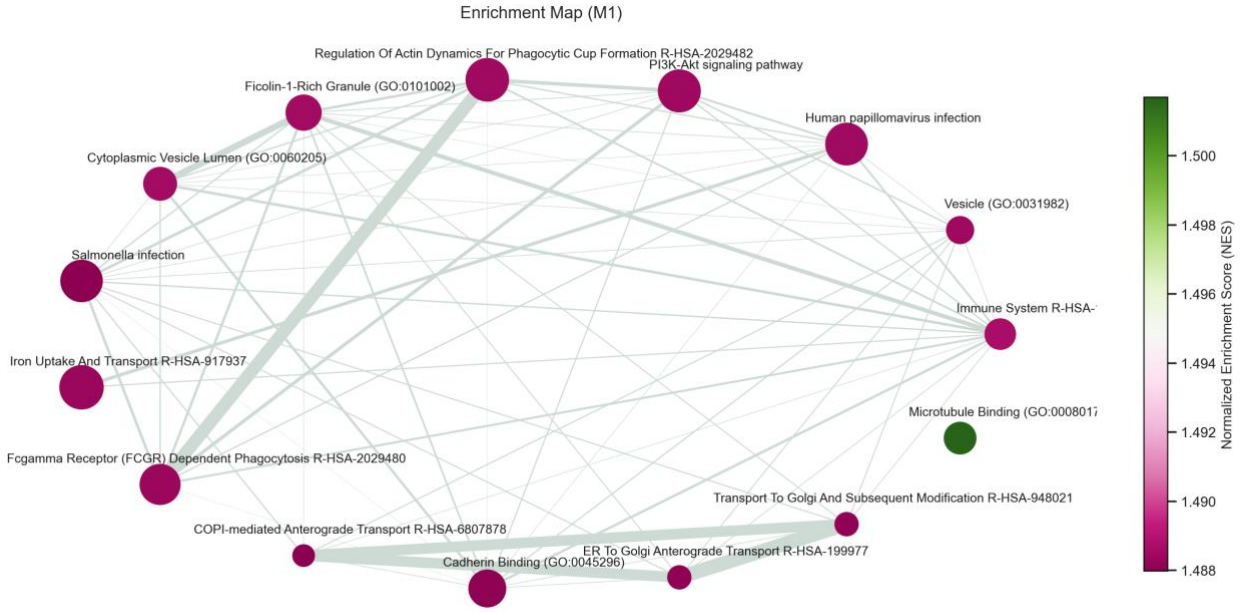


Figure 43 - Enrichment Map of M1 (LPS-treated) cells. Nodes represent enriched biological pathways, with node size proportional to the significance of each pathway. Edges indicate shared genes between pathways. The color gradient, ranging from green (lower NES) to magenta (higher NES), represents the Normalized Enrichment Scores. Highlighted pathway enhances ease of interpretation and emphasizes key interactions within the network.

All the results from both ORA and FCS methods were saved into CSV files with a convenient format to be later used with String-DB online tools.

Protein to Protein Interaction (PPI)

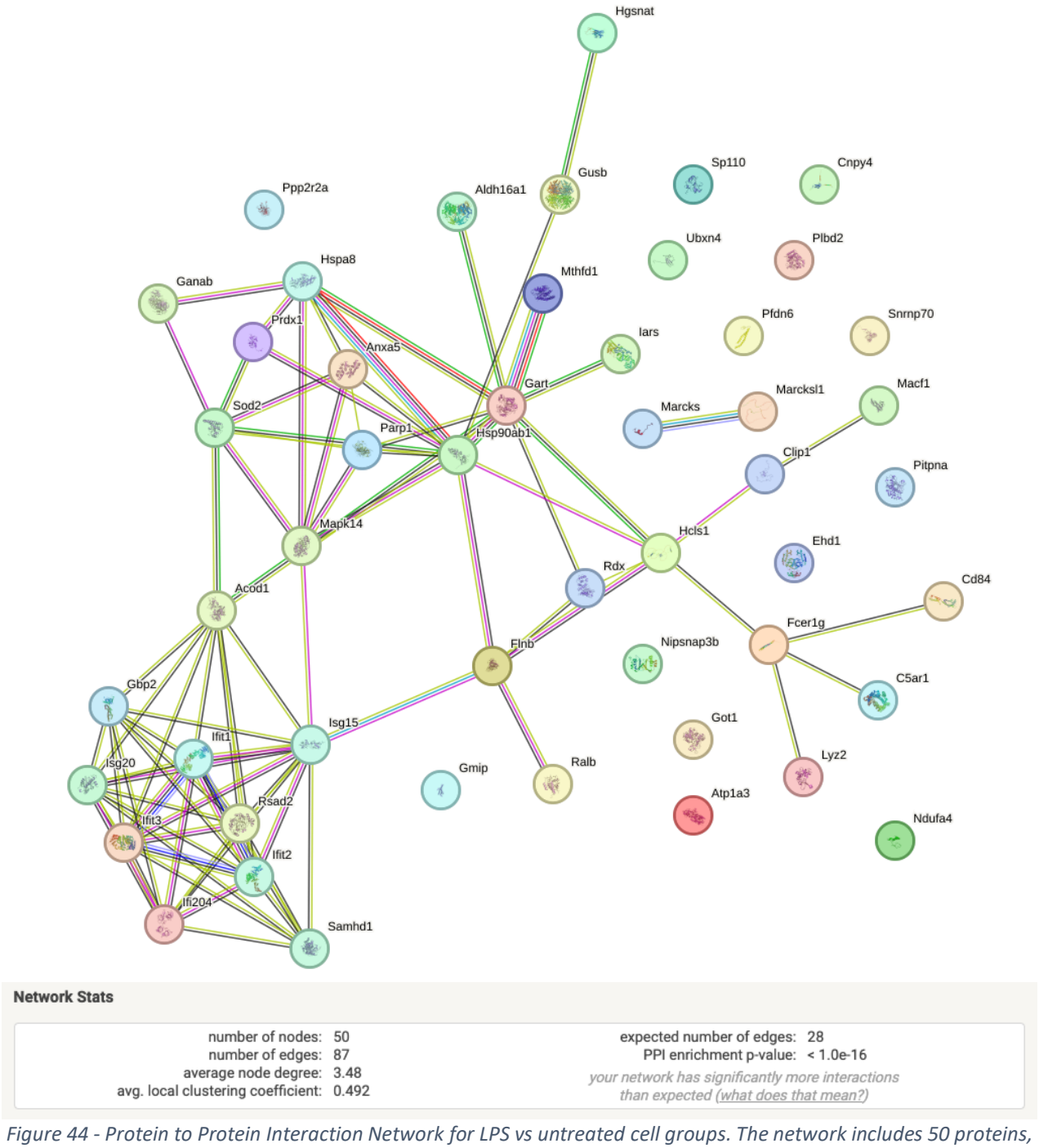
The analysis of protein-protein interactions (PPIs) is crucial because it provides deep insight into the intricate network of biological processes and molecular mechanisms within cells. Proteins rarely function in isolation. Their interactions form the basis of cellular functions and regulatory networks. Understanding these interactions helps to understand the role of proteins in various pathways, and can help identifying potential therapeutic targets and uncover disease mechanisms.

These networks can help predicting the functional consequences of protein interactions and how alterations can lead to pathological states. In addition, analysis of PPIs can also contribute with valuable information for drug discovery, as targeting specific protein interactions can modulate key biological processes and offer new treatment strategies.

For the three scenarios it was analyzed the interactions of:

- Most up-regulated proteins: To highlight the interactions of proteins that are actively promoted in response to the experimental condition.
- Most down-regulated proteins: To highlight the proteins that are being suppressed in the experimental condition.
- All together: To capture a full spectrum of the biological changes.

The study included the 25 most up-regulated and the 25 most down-regulated proteins.



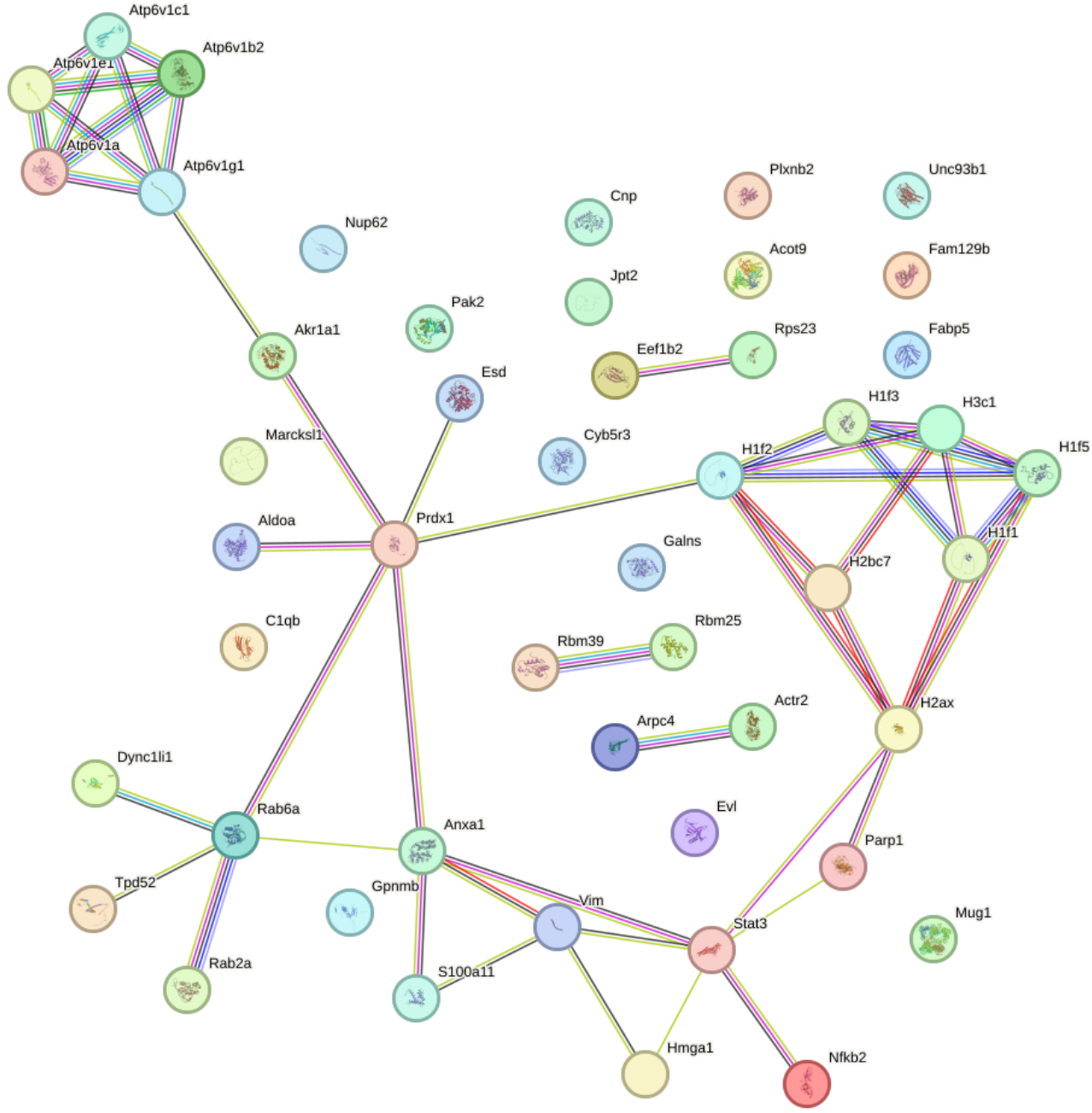
Network Stats

number of nodes:	50	expected number of edges:	28
number of edges:	87	PPI enrichment p-value:	< 1.0e-16
average node degree:	3.48		
avg. local clustering coefficient:	0.492		

your network has significantly more interactions than expected (what does that mean?)

Figure 44 - Protein to Protein Interaction Network for LPS vs untreated cell groups. The network includes 50 proteins, represented as nodes, and 87 edges indicating their interactions. The average node degree is 3.48, and the average local clustering coeficient 0.492. The PPI enrichment p-value is less than 1.0e-16, which indicates that the proteins are, at least, partially biologically connected as a group.

The same analysis was performed for the two clusters identified in the untreated cell group.



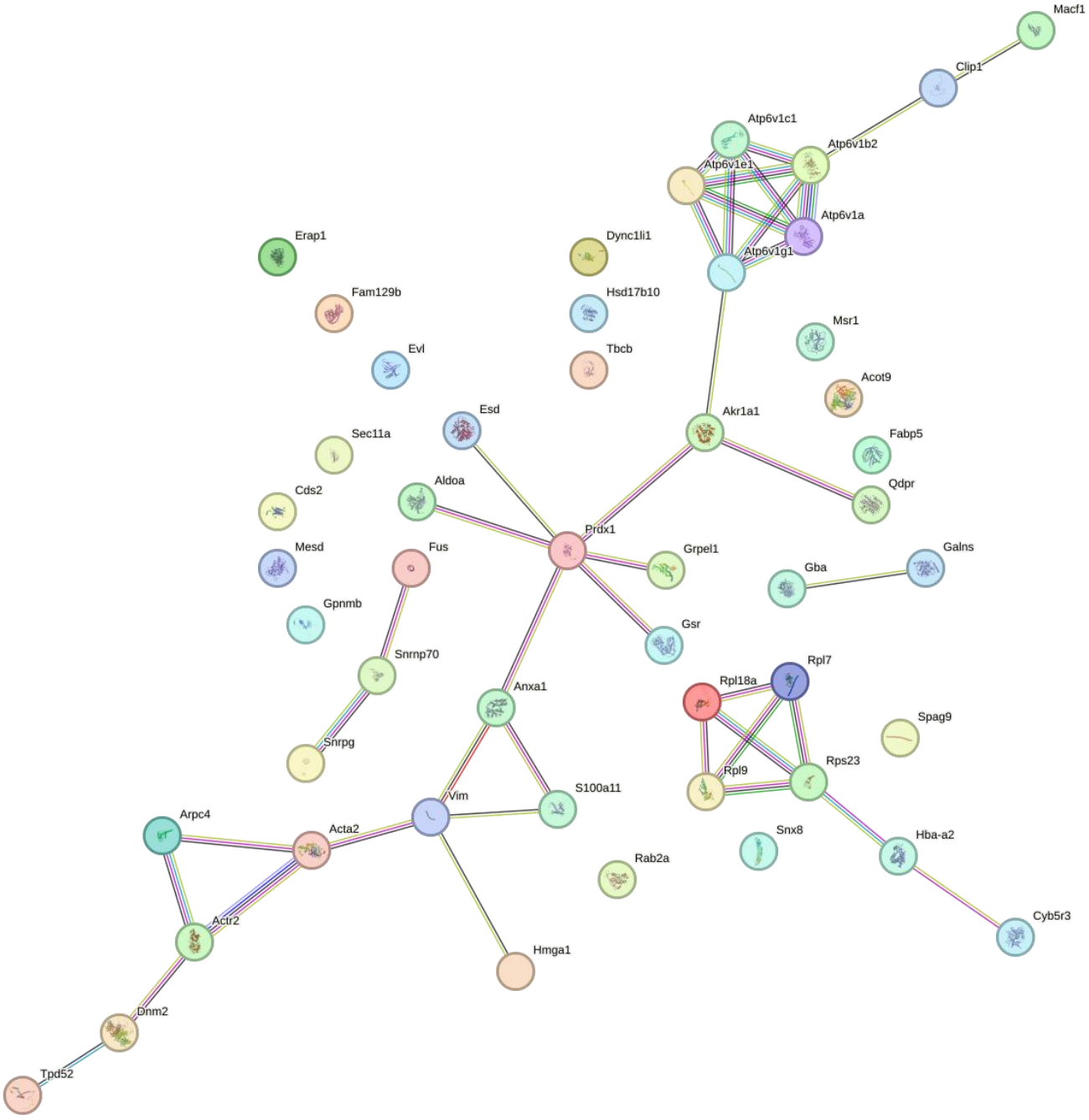
Network Stats

number of nodes:	49	expected number of edges:	18
number of edges:	50	PPI enrichment p-value:	2.32e-10
average node degree:	2.04		
avg. local clustering coefficient:	0.508		

your network has significantly more interactions than expected (what does that mean?)

Figure 45 - Protein to Protein Interaction Network for the two clusters identified in the untreated cell group. The network includes 49 proteins (nodes), and 50 interactions (edges). The average node degree is 2.04, and the average local clustering coeficient 0.508. The PPI enrichment p-value is less than 2.32e-10, which also indicates that the proteins are, at least, partially biologically connected as a group.

Finally, the analysis was also done for the two clusters identified in the LPS-treated cells.



Network Stats

number of nodes:	50	expected number of edges:	19
number of edges:	41	PPI enrichment p-value:	5.59e-06
average node degree:	1.64	*your network has significantly more interactions than expected (what does that mean?)*	
avg. local clustering coefficient:	0.477		

Figure 46 - Protein to Protein Interaction Network for the two clusters identified in LPS-treated cells. The network includes 50 proteins (nodes), and 41 interactions (edges). The average node degree is 1.64, and the average local clustering coeficient 0.477. The PPI enrichment p-value is less than 5.59e-06, which also indicates that the proteins are, at least, partially biologically connected as a group.

Quality Control

To assess and ensure the quality of the results, several parts of the data analysis were replicated using multiple approaches. Specifically, some parts were re-executed in R to validate the consistency of the findings. Additionally, complementary analyses for the Functional Enrichment Analysis were performed using the String-DB online tools, with the prepared data to cross-verify and confirm the robustness of the results.

All the files with the code and the results can be found in the ANNEX III: Additional Materials section.

Additional Materials

During the project, several files and materials have been generated. This is a compilation of all the links to these materials.

Source code: https://github.com/powwowath/pyMSpro/

These are the most relevant files and folders within this repository:

- Notebook with all the analysis - 00_workflow_META-ANALYSIS.ipynb
- Output folder with the results - ./results/
- Source code of String-DB and UniprotKB clients and other tools - ./src/
- Quality Control R scripts - ./QC_*

Gene-pathway networks can be found at:

- Network visualization (M1)
- Network visualization (M0 - M1)

Additionally, a public web page has been published will all the links to these materials (https://www.athzone.com/mbds/).

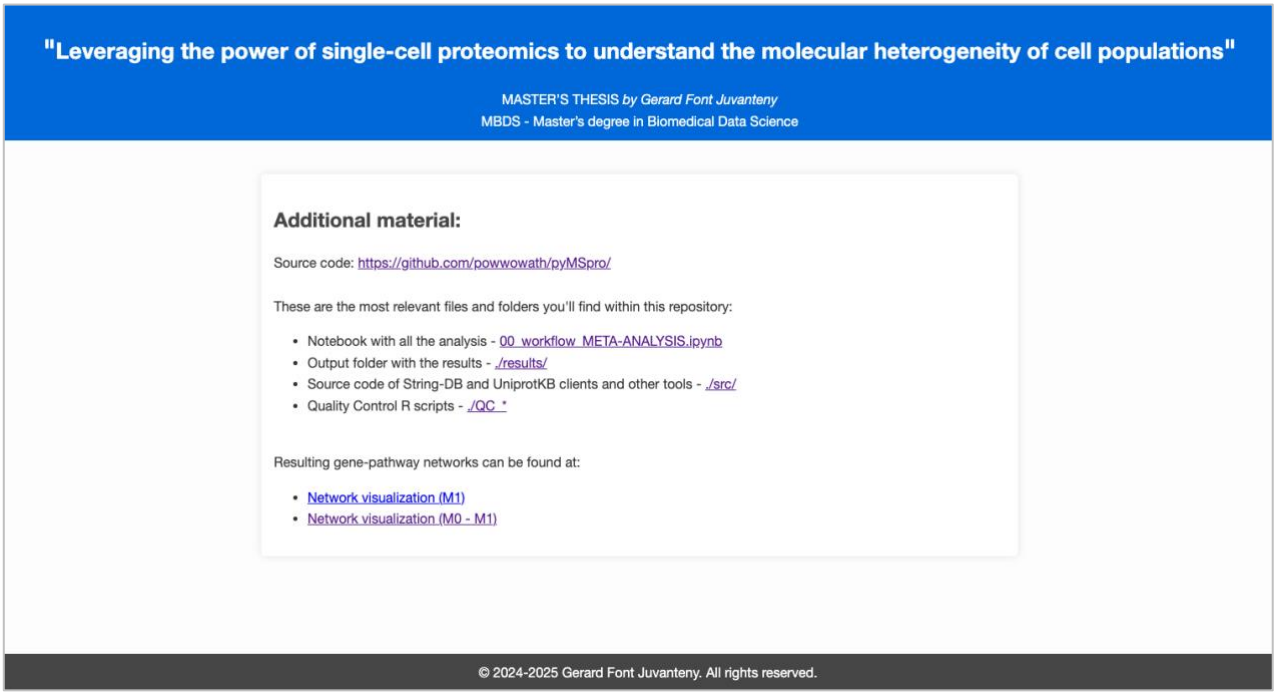


Figure 47 - Screenshot of the public website with all the links to the Additional materials.

CONCLUSIONS

Discussion

One of the objectives of this thesis was to conduct a meta-analysis of publicly available datasets to compare quantitative results between bulk proteomics and single-cell proteomics (SCP). Throughout the project, I compared the results obtained from each experiment to identify both similarities and differences. At the level of protein identification, the percentage of proteins found in the different bulk experiments was relatively low, ranging from 40% to 50%. It is important to note that all three studies utilized the Data-Dependent Acquisition (DDA) method, which may account for the observed differences due to the variability in ion selection criteria.

Furthermore, when incorporating the single-cell dataset (acquired with DIA) into the comparison, only 936 proteins were consistently identified across all datasets. However, this number represents 85% of the genes present in the single-cell datasets. This high percentage strongly suggests that the methodology and techniques employed in the single-cell experiment were highly effective, yielding reliable and robust results.

The scenario changed considerably when I analyzed the abundances of the identified proteins. The correlation between the three bulk datasets was nearly non-existent, indicating that different parameters were likely used during the studies, or that the results were not fully reliable due to the inherent challenges of proteomics. This variability significantly limited the scope of the meta-analysis, but nevertheless was still an important observation in the comparative analysis.

Another objective of this project was to analyze the differential expression in the data obtained from single cells. Using PCA and UMAP, I confirmed the clear separation of the classes (untreated and LPS), with genes such as IFIT1, SOD2, ISG15, and ISG20 being major contributors to this differentiation. However, to reach a high level of accuracy it required the use of a high number of features, suggesting that there are not strongly correlated, and that the data is complex and highly multidimensional.

One of the most significant findings was the identification of two well-differentiated clusters within the untreated cells from the same cell line. This discovery highlights the inherent heterogeneity present even within seemingly homogeneous cell populations. To further explore this heterogeneity, I performed a differential expression analysis in the three studied scenarios that include all the cells, only untreated cells and only LPS-treated cells.

The analysis of the M0 group (untreated cells) revealed two distinct groups, containing 150 and 37 cells, respectively. Some of the genes that contributed significantly to this separation were S100A11, RBM39, GPNMB and EVL.

The results of the Functional Enrichment analysis for the M0 cells were not particularly revealing. The analysis was unable to identify any pathways that were significantly overrepresented when comparing the two clusters using the FCS results. This lack of significant findings suggests that cells in the two clusters identified in M0 may not possess distinct functional pathways that differentiate them from each other.

For the LPS stimulated cells, a clustering tendency was also observed using both PCA and UMAP, though the separation was not as pronounced, and the contribution of individual genes to cluster generation was lower. The data visualization through a volcano plot provided further insights, detailing which genes were upregulated and downregulated, along with their respective levels of significance for this division. Despite the less distinct clustering, heterogeneity remains apparent in the LPS cells, underscoring the complexity and variation within these cellular responses.

During the enrichment analysis of this group of cells with GSEA (FCS), 15 overrepresented pathways were identified that are likely to be crucial to the biological differences between the two groups of cells. These pathways are related to immune activation, pathogen response, inflammatory mediator secretion, etc., which align with the classical pro-inflammatory role of M1 macrophages in innate immunity. The differences between the two clusters might reflect functional specialization within the M1 population, variations in cell states, differential exposure to microenvironmental signals, temporal dynamics of activation, or a combination of these and other factors.

The Protein-to-Protein Interaction (PPI) study revealed that the interactions observed between proteins for the three scenarios were highly unlikely to have occurred by chance. These highly significant interactions provide a strong basis for further research. They can help guide more detailed investigations, such as functional assays or therapeutic target validation.

Conclusions

This project allowed me to work with and understand different proteomic data formats for both, bulk and single-cell datasets, discern how various macrophage types differ in terms of protein abundance, and evaluate each group's biological characteristics. Furthermore, having single-cell data provided me with the opportunity to analyze and verify intrinsic heterogeneity at the cellular level.

However, given the current state of proteomics research, and in the still emerging field of single-cell proteomics, I encountered, and I had to face several of the known challenges of proteomics and especially single-cell proteomics. These obstacles highlight the complexities inherent in this field of study and underscore the critical advances that are still needed to unlock its full potential.

What becomes evident from the study of the available data is that the proteomic profile is not only dynamic but also highly diverse, even within the same cell type. Delving deeper into the evolution and dynamism of the proteome at single-cell level promises to unlock a more profound understanding of its biological mechanisms and functions.

Limitations

During the execution of this master thesis, I encountered several relevant limitations that are important to mention to understand and contextualize the results and conclusions.

One of the primary motivations of this project was to explore heterogeneity within the same cell type using a new dataset prepared with a novel method. Unfortunately, the resulting data failed to meet the quality standards and had to be discarded. This setback prompted me to seek alternative datasets to continue the project. But the number of high-quality datasets available in public repositories is not yet as high as for other types of analysis.

Among the few datasets available, it was very difficult to find datasets that study the same cell type (in my case, THP-1 human monocytic cell line). And those datasets that are based on the same cell type often have different formats, which means having to transform data or even to generate new features in some cases. These data preparation activities might introduce potential changes to the data that have an impact on the results. In cases where data are available in public repositories, most of them are published only raw data (the files generated by the mass spectrometer). Since these are proteomic data, the files often exceed one TB in size, requiring substantial storage and computational capacity.

As previously mentioned, I had to exclude the datasets with raw data only from my analysis, as big data management was beyond the scope of this thesis. Instead, I focused on datasets where the preprocessed files were available. For the three bulk datasets the published files were generated using all of them with the same software Proteome Discoverer (Thermo Scientific) but different versions. As the data had not undergone extensive filtering, I could implement my own filtering criteria during the analysis, ensuring the data were relevant and manageable for my study.

Another significant limitation of the project was the unavailability of THP-1 single-cell data, as I could not find any datasets in public repositories. Consequently, I had to perform the macrophage cell comparison using murine bone marrow-derived macrophages (BMDMs) instead. While this choice was necessary due to data limitations, the conclusions and the findings need to be contextualized.

Focusing on the single-cell dataset, the intensity matrix did not contain any missing values as they were imputed during processing, as described in the workflow [49]. This means that the dataset may include proteins initially detected in only one of the studied groups but now represented with imputed values for the other group. While this ensures a complete dataset for statistical analyses, it introduces a limitation: imputed values can obscure the distinction between proteins genuinely shared between groups and those uniquely present in one group. Consequently, the analysis might overlook proteins exclusive to a particular group, potentially masking critical biological differences or unique characteristics of that group. This imputation strategy could lead to challenges in accurately interpreting distinct biological mechanisms of each group, potentially affecting the robustness and completeness of the results.

A final limitation of my study lies in the technique used to identify and quantify the proteins. For the three bulk datasets, the data were acquired using Data-Dependent Acquisition (DDA), while the single-cell data were obtained using Data-Independent Acquisition (DIA). The use of different methodologies might introduce inconsistencies, as DDA can be biased towards more abundant peptides, potentially missing low-abundance signals, whereas DIA provides a more unbiased proteomic profile.

Assumptions

Throughout the course of this project, I have made several assumptions that are indispensable to mention for the sake of transparency. These assumptions are intended to provide clarity and facilitate a more accurate understanding and interpretation of the results presented.

- Due to the lack of single cell datasets for THP-1 derived macrophages, I assumed that comparing macrophages derived from THP-1 cells with macrophages derived from murine bone marrow (BMDMs), the proteomic profiles and cellular behaviors of the two types would be sufficiently comparable.

- A second important assumption is that the processed data resulting from DDA (Data Dependent Acquisition) experiments are, to some extent, comparable with those processed using the DIA (Data Independent Acquisition) method. Throughout the analysis, I took this factor into account when formulating conclusions and explaining the results.

- The dataset Iwata, H., et al. (2016) has a column named "Score", while the other two bulk datasets have a column named "Score Sequest HT: Sequest HT". According to the Proteome Discoverer Manual, I assumed that the column "Score" in the Iwata, H., et al. (2016) dataset is equivalent to the Sequest Score in the other datasets. For both columns, the formula used to compute the Protein Score is:

$$(sum\_of\_all\_cross\text{-}correlation\_factors\_of\_0.8\_or\_above) + (peptide\_charge \times peptide\_relevance\_factor)$$

- To assess the presence of a protein in the different groups (M0, M1 and M2) in Li, P., et al. (2021) and Li, P., et al. (2022) datasets, I did the following assumption:

```
"Abundances (Grouped): M0" > 0  # Protein found in M0 group
"Abundances (Grouped): M1" > 0  # Protein found in M1 group
"Abundances (Grouped): M2" > 0  # Protein found in M2 group
```

Ethical-Social Impact, Sustainability, and Diversity

The study of omics in general, and single cell proteomics in particular, represents a very important tool in the research of new biomarkers[60] and treatments, especially in personalized medicine. To ensure that the benefits reach everyone it is imperative to consider the ethical-social aspect, sustainability and diversity in the research.

The ability to examine cellular heterogeneity can lead to more personalized and effective treatments for diseases like cancer, which directly aligns with the ethical imperative to provide the best possible care for patients. This power also raises concerns about **data privacy**, especially when dealing with human samples. Ensuring that patient consent is fully informed, and that data is **securely stored** and used responsibly is also paramount. In 2022, for example, Ivo-Fierro Monti et

all [61] concluded that identification of an individual using proteomic data alone was very difficult and unlikely. However, during the last few years many research initiatives have been initiated and the technology and methods to re-identify individuals using proteomic data alone have continued to advance, opening up new possibilities.

From a societal standpoint, single-cell proteomics has the potential to **democratize healthcare** by enabling precise diagnostics and treatments tailored to each patient. This can reduce health disparities by making advanced diagnostics accessible to underserved populations, improving the overall equity of healthcare. In addition, the development of **cost-effective** single-cell technologies could expand the availability of these tools and further reduce health disparities.

Sustainability in single-cell proteomics involves both environmental and economic dimensions. Economically, while the initial costs of single-cell technologies are high, their potential to reduce the need for extensive and repeated testing could lower long-term healthcare costs. Additionally, the development of more efficient and sustainable technologies is crucial, emphasizing the need for green chemistry practices and recyclable materials. As an example, with the development of nPOP method, Leduc et all [29] reduced the cost of sample preparation materials to few cents per cell.

Environmentally, the reduction in sample sizes inherent in single-cell analysis can decrease the volume of reagents and materials used, which minimizes waste and lowers the **ecological footprint** of research labs.

And finally, efforts to increase the **diversity of populations** in genomics research are essential [62] and must be proactive, ensuring that underrepresented groups are included in studies to capture a full spectrum of biological variability. This can be achieved through community engagement, transparent communication, and partnerships with institutions serving diverse populations. Additionally, supporting diverse talent within the scientific community is very important. This includes providing opportunities for underrepresented minorities in science, technology, engineering, and mathematics (STEM) fields, which can enrich the research landscape and enhance the relevance and application of single-cell proteomics.

FUTURE WORK

As the field of single-cell proteomics continues to evolve, there remain several promising avenues for further research and development. Building upon the findings and methodologies established in this thesis, future work will aim to address current limitations, enhance analytical techniques, and explore novel applications.

The complexity and heterogeneity of single-cell data requires the continuous refinement of computational tools and the integration of cutting-edge technologies to improve accuracy, scalability, and interpretability. Furthermore, expanding the scope of single-cell proteomics to new biological contexts and leveraging interdisciplinary approaches will undoubtedly contribute to a deeper understanding of cellular functions and interactions.

The following initiatives outline the key areas where future efforts will be directed to advance the field and maximize the impact of single-cell proteomics research.

- **Development of standardized protocols and methodologies**: Currently, there is a considerable dependency on technical variables, such as the manufacturers and machines employed, leading to inconsistencies and variability in results. Establishing standardized sample preparation protocols is essential to ensure uniformity and reproducibility across different laboratories and studies. Furthermore, clear criteria for identifying and quantifying proteins must be established to accurately interpret and compare data. This includes setting benchmarks for the confidence in protein identification, the accuracy of quantification, and the methodologies for data analysis

- **Improving accuracy and repeatability**. Enhancing technologies to improve the quantitative accuracy and sensitivity of single-cell proteomics measurements is crucial. Innovations in mass spectrometry and protein labeling techniques can contribute to more precise quantification of protein abundance. Despite recent advancements, there remains considerable room for improvement in repeatability between experimental runs, underscoring the need for ongoing development and refinement of these technologies to achieve more consistent and reliable results in single-cell proteomics.

- **Expand dataset availability**. Accessing high-quality single-cell datasets it's still a challenge. This scarcity of reliable datasets hinders the validation and comparison of new findings, limiting the broader application and development of single-cell proteomics methodologies and data analysis. In most of the public datasets, only RAW data or excessively pre-processed data (sources for figures in publications) are provided. While having access to the RAW data ensures that one can use one's own criteria in processing the data, these files require enormous storage space and have very high computational requirements. For this reason, it would be interesting to additionally publish the preprocessed data without excessive manipulation and filtering.

- **Enhancing Data Integration**. The datasets explored offered already processed data with a highly customized format according to the authors. To reduce the complexity of integrating

data from different studies, it is imperative to standardize data formats as well as metadata.

- **Longitudinal single-cell proteomics**. By tracking protein expression profiles in individual cells over time, researchers can gain unique insights into cellular dynamics, disease progression, and treatment responses. However, to achieve this goal it is very important to improve, as mentioned above, the accuracy and repeatability between runs.

- **Include single-cell proteomics in multi-omics studies**. By combining information from different molecular levels, researchers can gain a more comprehensive understanding of cellular function, dysfunction, and response to stimuli [63].

- **Functional Enrichment Analysis is still challenging**. The community must keep improving pathway databases to include new discoveries, but also to reduce pathway redundancy. The development of new advanced statistical methods could also contribute to more robust results, with less false negatives and without inflating false positives. These are just a few of the various actions that are required to enhance the outcomes of this type of analysis.

- **Development of an open, robust, and parametrizable software package or data platform for Single-Cell.** The goal is to create a standardized toolset to prepare and analyze single-cell data, significantly reducing the effort required for such tasks. It could include the implementation and publication of API's, as well as the possibility to work with multiple cloud providers. The current available tools rely heavily on specific versions of different packages, which are often not well-maintained or documented. Additionally, these tools are spread across different programming languages, primarily R and occasionally Python. These issues greatly hinder the reproducibility of experiments and often compel researchers to spend considerable effort developing their own tools.

REFERENCES

1. Avery, O. T., Macleod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of Experimental Medicine*, *79*(2), 137. https://doi.org/10.1084/JEM.79.2.137

2. CRICK, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, *12*, 138–163.

3. Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, *15*(9). https://doi.org/10.1371/JOURNAL.PBIO.2003243

4. Perakakis, N., Yazdani, A., Karniadakis, G. E., & Mantzoros, C. (2018). Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism - Clinical and Experimental*, *87*, A1–A9. https://doi.org/10.1016/J.METABOL.2018.08.002

5. Zengin, T., Kılıç, S., Yenigün, U., Erdoğmuş, S., & Baysal, Ö. (2017). A smart way for talking with proteins; proteomics. *MOJ Proteomics & Bioinformatics*, *Volume 5*(Issue 3). https://doi.org/10.15406/MOJPB.2017.05.00160

6. Chandramouli, K., & Qian, P.-Y. (2009). Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. *Human Genomics and Proteomics : HGP*, *2009*(1). https://doi.org/10.4061/2009/239204

7. Baldwin, M. A. (2004). Protein Identification by Mass Spectrometry. *Molecular & Cellular Proteomics*, *3*(1), 1–9. https://doi.org/10.1074/mcp.r300012-mcp200

8. Kosak, S. T., & Groudine, M. (2004). Gene order and dynamic domains. *Science (New York, N.Y.)*, *306*(5696), 644–647. https://doi.org/10.1126/SCIENCE.1103864

9. Lee, S., Vu, H. M., Lee, J. H., Lim, H., & Kim, M. S. (2023). Advances in Mass Spectrometry-Based Single Cell Analysis. *Biology 2023*, *Vol. 12*, *Page 395*, *12*(3), 395. https://doi.org/10.3390/BIOLOGY12030395

10. Petrosius, V., & Schoof, E. M. (2023). Recent advances in the field of single-cell proteomics. *Translational Oncology*, *27*, 101556. https://doi.org/10.1016/J.TRANON.2022.101556

11. Ahmad, R., & Budnik, B. (2023). A review of the current state of single-cell proteomics and future perspective. *Analytical and Bioanalytical Chemistry*, *415*(28), 6889–6899. https://doi.org/10.1007/S00216-023-04759-8/TABLES/1

12. Rosenberger, F. A., Thielert, M., Strauss, M. T., Schweizer, L., Ammar, C., Mädler, S. C., Metousis, A., Skowronek, P., Wahle, M., Madden, K., Gote-Schniering, J., Semenova, A., Schiller, H. B., Rodriguez, E., Nordmann, T. M., Mund, A., & Mann, M. (2023). Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nature Methods 2023 20:10*, *20*(10), 1530–1536. https://doi.org/10.1038/s41592-023-02007-6

13. Lundberg, E., & Borner, G. H. H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology 2018 20:5*, *20*(5), 285–302. https://doi.org/10.1038/s41580-018-0094-y

14. Hass, G. M., Nau, H., Biemann, K., Grahn, D. T., Ericsson, L. H., & Neurath, H. (1975). The amino acid sequence of a carboxypeptidase inhibitor from potatoes. *Biochemistry*, *14*(6), 1334–1342. https://doi.org/10.1021/BI00677A036

15. Kim, S., Kamarulzaman, L., & Taniguchi, Y. (2023). Recent methodological advances towards single-cell proteomics. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *99*(8), 306. https://doi.org/10.2183/PJAB.99.021

16. Zubarev, R. A. (2013). The challenge of the proteome dynamic range and its implications for in-depth proteomics. *PROTEOMICS*, *13*(5), 723–726. https://doi.org/10.1002/PMIC.201200451

17. Slavov, N. (2020). Unpicking the proteome in single cells:: Single-cell mass spectrometry will help reveal mechanisms that underpin health and disease. *Science (New York, N.Y.)*, *367*(6477), 512. https://doi.org/10.1126/SCIENCE.AAZ6695

18. Newman, S. S., Wilson, B. D., Mamerow, D., Wollant, B. C., Nyein, H., Rosenberg-Hasson, Y., Maecker, H. T., Eisenstein, M., & Soh, H. T. (2023). Extending the dynamic range of biomarker quantification through molecular equalization. *Nature Communications 2023 14:1*, *14*(1), 1–10. https://doi.org/10.1038/s41467-023-39772-z

19. Mali, S. B. (2023). Single cell proteomics. Potential applications in Head and Neck oncology. *Oral Oncology*, *146*, 106586. https://doi.org/10.1016/J.ORALONCOLOGY.2023.106586

20. Derks, J., Leduc, A., Wallmann, G., Huffman, R. G., Willetts, M., Khan, S., Specht, H., Ralser, M., Demichev, V., & Slavov, N. (2022). Increasing the throughput of sensitive proteomics by plexDIA. *Nature Biotechnology 2022 41:1*, *41*(1), 50–59. https://doi.org/10.1038/s41587-022-01389-w

21. Kassem, S., van der Pan, K., de Jager, A. L., Naber, B. A. E., de Laat, I. F., Louis, A., van Dongen, J. J. M., Teodosio, C., & Díez, P. (2021). Proteomics for Low Cell Numbers: How to Optimize the Sample Preparation Workflow for Mass Spectrometry Analysis. *Journal of Proteome Research*, *20*(9), 4217–4230. https://doi.org/10.1021/ACS.JPROTEOME.1C00321/ASSET/IMAGES/LARGE/PR1C00321_0002.JPEG

22. Single-cell proteomics: challenges and prospects. (2023). *Nature Methods 2023 20:3*, *20*(3), 317–318. https://doi.org/10.1038/s41592-023-01828-9

23. Nesvizhskii, A. I., & Aebersold, R. (2005). Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics*, *4*(10), 1419–1440. https://doi.org/10.1074/MCP.R500012-MCP200

24. Perdigão, N., Rosa, A. C., & O'Donoghue, S. I. (2017). The Dark Proteome Database. *BioData Mining*, *10*(1). https://doi.org/10.1186/S13040-017-0144-6

25. Decano, J. L., Iwamoto, Y., Goto, S., Lee, J. Y., Matamalas, J. T., Halu, A., Blaser, M., Lee, L. H., Pieper, B., Chelvanambi, S., Silva-Nicolau, J., Bartoli-Leonard, F., Higashi, H., Shibata, H., Vyas, P., Wang, J., Gostjeva, E., Body, S. C., Singh, S. A., … Aikawa, E. (2022). A disease-driver population within interstitial cells of human calcific aortic valves identified via single-cell and proteomic profiling. *Cell Reports*, *39*(2), 110685. https://doi.org/10.1016/j.celrep.2022.110685

26. Decano, J. L., Maiorino, E., Matamalas, J. T., Chelvanambi, S., Tiemeijer, B. M., Yanagihara, Y., Mukai, S., Jha, P. K., Pestana, D. V. S., D'Souza, E., Whelan, M., Ge, R., Asano, T., Sharma, A., Libby, P., Singh, S. A., Aikawa, E., & Aikawa, M. (2023). Cellular Heterogeneity of Activated Primary Human Macrophages and Associated Drug-Gene Networks: From Biology to Precision Therapeutics. *Circulation*, *148*(19), 1459–1478. https://doi.org/10.1161/CIRCULATIONAHA.123.064794

27. Perdigão, N., & Rosa, A. (2019). Dark Proteome Database: Studies on Dark Proteins. *High-Throughput*, *8*(2). https://doi.org/10.3390/HT8020008

28. Goto-Silva, L., & Junqueira, M. (2021). Single-cell proteomics: A treasure trove in neurobiology. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1869*(7), 140658. https://doi.org/10.1016/J.BBAPAP.2021.140658

29. Leduc, A., Huffman, R. G., Cantlon, J., Khan, S., & Slavov, N. (2022). Exploring functional protein covariation across single cells using nPOP. *BioRxiv*, 2021.04.24.441211. https://doi.org/10.1101/2021.04.24.441211

30. Leduc, A., Koury, L., Cantlon, J., & Slavov, N. (2023). Massively parallel sample preparation for multiplexed single-cell proteomics using nPOP. *BioRxiv*, 2023.11.27.568927. https://doi.org/10.1101/2023.11.27.568927

31. Hartlmayr, D. N., Ctortecka, C., Mechtler, K., Eickhoff, H., Tourniaire, G., & Seth, A. (n.d.). *cellenONE ® : all-in-one solution for single cell proteomics using LC-MS/MS sample preparation*.

32. Hartlmayr, D., Ctortecka, C., Seth, A., Mendjan, S., Tourniaire, G., Mechtler, K., & Biocenter, V. (n.d.). *An automated workflow for label-free and multiplexed single cell*

proteomics sample preparation at unprecedented sensitivity. https://doi.org/10.1101/2021.04.14.439828

33. Wu, S., Ma, H., Prytkova, I., Stenoien, D., & Paša-Tolić, L. (2017). Proteomics, Top-Down. *Encyclopedia of Spectroscopy and Spectrometry*, 774–778. https://doi.org/10.1016/B978-0-12-409547-2.12138-9

34. Pandeswari, P. B., & Sabareesh, V. (2018). Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Advances*, *9*(1), 313. https://doi.org/10.1039/C8RA07200K

35. *Discovery-based Quantitative MS methods | Advanced Analysis Centre*. (n.d.). Retrieved April 21, 2024, from https://www.uoguelph.ca/aac/facilities/mass-spectrometry-facility/quantitative-proteomics/discovery-based-quantitative-ms
36. Chen, C., Hou, J., Tanner, J. J., & Cheng, J. (2020). Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. International Journal of Molecular Sciences, 21(8). https://doi.org/10.3390/IJMS21082873

37. Eng, J. K., Searle, B. C., Clauser, K. R., & Tabb, D. L. (2011). A Face in the Crowd: Recognizing Peptides Through Database Search. *Molecular & Cellular Proteomics : MCP*, *10*(11). https://doi.org/10.1074/MCP.R111.009522

38. Wang, P., & Wilson, S. R. (2013). Mass spectrometry-based protein identification by integrating de novo sequencing with database searching. *BMC Bioinformatics*, *14 Suppl 2*(2), 1–9. https://doi.org/10.1186/1471-2105-14-S2-S24/FIGURES/5

39. Wei, J., Zhou, T., Zhang, X., & Tian, T. (2021). DTFLOW: Inference and Visualization of Single-cell Pseudotime Trajectory Using Diffusion Propagation. *Genomics, Proteomics & Bioinformatics*, *19*(2), 306–318. https://doi.org/10.1016/J.GPB.2020.08.003

40. Budnik, B., Levy, E., Harmange, G., & Slavov, N. (2018). SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, *19*(1). https://doi.org/10.1186/S13059-018-1547-5

41. Specht, H., Emmott, E., Perlman, D. H., Koller, A., & Slavov, N. (2019). High-throughput single-cell proteomics quantifies the emergence of macrophage heterogeneity. *BioRxiv*, 665307. https://doi.org/10.1101/665307

42. Petelski, A. A., Emmott, E., Leduc, A., Huffman, R. G., Specht, H., Perlman, D. H., & Slavov, N. (n.d.). *Multiplexed single-cell proteomics using SCoPE2*. https://doi.org/10.1038/s41596-021-00616-z

43. Huffman, R. G., Leduc, A., Wichmann, C., di Gioia, M., Borriello, F., Specht, H., Derks, J., Khan, S., Khoury, L., Emmott, E., Petelski, A. A., Perlman, D. H., Urgen Cox, J. ¨, Zanoni, I., & Slavov, N. (2022). Prioritized single-cell proteomics reveals molecular and functional polarization across primary macrophages. *BioRxiv*, 2022.03.16.484655. https://doi.org/10.1101/2022.03.16.484655

44. Derks, J., Leduc, A., Wallmann, G., Huffman, R. G., Willetts, M., Khan, S., Specht, H., Ralser, M., Demichev, V., & Slavov, N. (2022). Increasing the throughput of sensitive proteomics by plexDIA. *BioRxiv*, 2021.11.03.467007. https://doi.org/10.1101/2021.11.03.467007

45. Iwata, H., Goettsch, C., Sharma, A., Ricchiuto, P., Goh, W. W. bin, Halu, A., Yamada, I., Yoshida, H., Hara, T., Wei, M., Inoue, N., Fukuda, D., Mojcher, A., Mattson, P. C., Barabási, A. L., Boothby, M., Aikawa, E., Singh, S. A., & Aikawa, M. (2016). PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. *Nature Communications 2016 7:1*, *7*(1), 1–19. https://doi.org/10.1038/ncomms12849

46. Li, P., Ma, C., Li, J., You, S., Dang, L., Wu, J., Hao, Z., Li, J., Zhi, Y., Chen, L., & Sun, S. (2022). Proteomic characterization of four subtypes of M2 macrophages derived from human THP-1 cells. *Journal of Zhejiang University. Science. B*, *23*(5), 407–422. https://doi.org/10.1631/jzus.B2100930

47. Li, P., Hao, Z., Wu, J., Ma, C., Xu, Y., Li, J., Lan, R., Zhu, B., Ren, P., Fan, D., & Sun, S. (2021). Comparative Proteomic Analysis of Polarized Human THP-1 and Mouse RAW264.7 Macrophages. *Frontiers in Immunology*, *12*, 700009. https://doi.org/10.3389/FIMMU.2021.700009/FULL

48. Barnouin, K. (2011). Guidelines for experimental design and data analysis of proteomic mass spectrometry-based experiments. *Amino Acids*, *40*(2), 259–260. https://doi.org/10.1007/S00726-010-0750-9

49. Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., Kharchenko, P., Koller, A., & Slavov, N. (2021). Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, *22*(1), 50. https://doi.org/10.1186/s13059-021-02267-5

50. *IFIT1 Gene - GeneCards | IFIT1 Protein | IFIT1 Antibody*. (n.d.). Retrieved January 15, 2025, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=IFIT1&keywords=IFIT1

51. *SOD2 Gene - GeneCards | SODM Protein | SODM Antibody*. (n.d.). Retrieved January 15, 2025, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=SOD2

52. Zhang, L., Zhu, T., Miao, H., & Liang, B. (2021). The Calcium Binding Protein S100A11 and Its Roles in Diseases. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/FCELL.2021.693262

53. Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology*, *8*(2), e1002375. https://doi.org/10.1371/JOURNAL.PCBI.1002375

54. Fang, Z., Liu, X., & Peltz, G. (2023). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*, *39*(1). https://doi.org/10.1093/BIOINFORMATICS/BTAC757

55. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., … Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, *34*(3), 267–273. https://doi.org/10.1038/NG1180

56. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102

57. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 2000 25:1*, *25*(1), 25–29. https://doi.org/10.1038/75556

58. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34. https://doi.org/10.1093/NAR/27.1.29

59. Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., & Hermjakob, H. (2017). Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*, *18*(1), 142. https://doi.org/10.1186/s12859-017-1559-2

60. Bader, J. M., Albrecht, V., & Mann, M. (2023). MS-Based Proteomics of Body Fluids: The End of the Beginning. *Molecular & Cellular Proteomics*, *22*(7), 100577. https://doi.org/10.1016/j.mcpro.2023.100577

61. Fierro-Monti, I., Wright, J. C., Choudhary, J. S., & Vizcaíno, J. A. (2022). Identifying individuals using proteomics: are we there yet? *Frontiers in Molecular Biosciences*, *9*. https://doi.org/10.3389/FMOLB.2022.1062031

62. Martinez-Martin, N., & Magnus, D. (2019). Privacy and ethical challenges in next-generation sequencing. *Expert Review of Precision Medicine and Drug Development*, *4*(2), 95. https://doi.org/10.1080/23808993.2019.1599685

63. Jiang, Y. R., Zhu, L., Cao, L. R., Wu, Q., Chen, J. B., Wang, Y., Wu, J., Zhang, T. Y., Wang, Z. L., Guan, Z. Y., Xu, Q. Q., Fan, Q. X., Shi, S. W., Wang, H. F., Pan, J. Z., Fu, X. D., Wang, Y., & Fang, Q. (2023). Simultaneous deep transcriptome and proteome profiling in a single mouse oocyte. *Cell Reports*, *42*(11). https://doi.org/10.1016/J.CELREP.2023.113455/ATTACHMENT/35B5BDCB-255E-4F38-8E2D-A6CB4A52CDD6/MMC7.PDF

ANNEX I: Dataset format

Dataset 1: (file: SI_THP_RAW_proteomics_data_ss.xlsx)

- **Accession**: This is a unique identifier assigned to a protein, often from databases like UniProt or RefSeq.
- **Description**: A brief description of the protein's function or characteristics.
- **Score**: A numerical value representing the confidence level of the protein identification. Higher scores indicate more reliable identifications.
- **Coverage**: The percentage of the protein's amino acid sequence that has been covered by identified peptides.
- **#Proteins**: The number of proteins associated with the identified peptides.
- **#Unique Peptides**: The number of distinct peptide sequences identified.
- **#Peptides**: The total number of peptides identified, including duplicates.
- **#PSMs**: The number of Peptide Spectrum Matches, which represent the number of times a peptide sequence has been identified in the mass spectra.
- **126/126 Count**: Counts at moment 0hr
- **127/126 Count**: Counts at moment 8hr
- **128/126 Count**: Counts at moment 12hr
- **129/126 Count**: Counts at moment 24hr
- **130/126 Count**: Counts at moment 48hr
- **131/126 Count**: Counts at moment 72hr
- **126/126**: Relative abundance at moment 0hr
- **127/126**: Relative abundance at moment 8hr
- **128/126**: Relative abundance at moment 12hr
- **129/126**: Relative abundance at moment 24hr
- **130/126**: Relative abundance at moment 48hr
- **131/126**: Relative abundance at moment 72hr

Dataset 2: PD_Globalprotomics_2024.xlsx

- **Protein FDR Confidence: Combined**: Reflects the level of confidence in the identification of each protein.
- **Master**: A unique identifier assigned to each protein group.
- **Accession**: A database accession number (e.g., UniProt) for the protein.
- **Description**: A textual description of the protein's function and characteristics.
- **Sum PEP Score**: The sum of peptide-spectrum match (PSM) scores, indicating overall confidence in protein identification.
- **# Peptides**: The number of peptides identified for the protein.
- **# PSMs**: The total number of PSMs for the protein.
- **# Protein Unique Peptides**: The number of peptides uniquely assigned to this protein.
- **# Unique Peptides**: The total number of unique peptides identified.
- **# AAs**: The number of amino acids in the protein sequence.
- **MW [kDa]**: The molecular weight of the protein in kilodaltons.
- **Score Sequest HT: Sequest HT**: The score assigned by the Sequest HT search engine for the protein identification.

- **# Razor Peptides**: The number of peptides uniquely assigned to this protein group, helping to distinguish it from other similar proteins.
- Quantitative Proteomics Data:
    - **Abundance Ratio**: These columns represent the relative abundance of a protein in different experimental conditions (e.g., M0, M1, M2a, M2b, M2c, M2d). The ratios are calculated by dividing the abundance of a protein in one condition by the abundance in another condition.
    - **Abundances (Grouped)**: These columns represent the absolute or normalized abundance of a protein in each experimental group.
    - **Abundances (Normalized)**: These columns represent the normalized abundance of a protein in each individual sample (F1-F18). Normalization is a process to account for experimental variability and ensure accurate comparison of protein abundance across samples.
- **Modifications**: This column indicates any post-translational modifications (PTMs) identified on the protein, such as phosphorylation or acetylation.

Dataset 3: THP1_global_M0_M1_M2_proteins.xlsx

- **Checked**: Indicates whether the protein identification was manually checked.
- **Protein FDR Confidence**: False Discovery Rate (FDR) confidence level for the protein identification.
- **Combined**: Indicates whether the protein identification is based on a combination of different search engines or data sources.
- **Master Unique Sequence ID**: A unique identifier assigned to the master protein.
- **Protein Group IDs**: Identifiers for protein groups, which can include isoforms or related proteins.
- **Accession**: Accession number from a protein database (e.g., UniProt).
- **Description**: Brief description of the protein function.
- **FASTA Title Lines**: The FASTA header line corresponding to the protein sequence.
- **Exp. q-value: Combined**: The q-value (FDR-adjusted p-value) for the protein identification.
- **Sum PEP Score**: The summed peptide expectation score for the protein.
- **# Decoy Protein: Combined**: The number of decoy proteins identified in the analysis.
- **Coverage [%]**: The percentage of the protein sequence covered by identified peptides.
- **# Peptides**: The number of peptides identified for the protein.
- **# PSMs**: The number of peptide spectrum matches (PSMs) identified.
- **# Protein Unique Peptides**: The number of peptides unique to the protein.
- **# Unique Peptides**: The number of unique peptide sequences identified.
- **# AAs**: The number of amino acids in the protein sequence.
- **MW [kDa]**: The molecular weight of the protein in kilodaltons.
- **calc. pI**: The calculated isoelectric point of the protein.
- **Score Sequest HT: Sequest HT**: The score assigned by the Sequest HT search engine.
- **Coverage [%] (by Search Engine): Sequest HT**: The coverage by Sequest HT.
- **# PSMs (by Search Engine): Sequest HT**: The number of PSMs identified by Sequest HT.
- **# Peptides (by Search Engine): Sequest HT**: The number of peptides identified by Sequest HT.

- **# Razor Peptides**: The number of razor peptides, which are uniquely assigned to a single protein group.
- Quantitative Data:
  - **Abundance Ratio: (M1) / (M0)**: The ratio of protein abundance between two conditions (M1 and M0).
  - **Abundance Ratio (log2): (M1) / (M0)**: The log2-transformed abundance ratio.
  - **Abundance Ratio P-Value**: The p-value for the difference in abundance between the two conditions.
  - **Abundance Ratio Adj. P-Value**: The adjusted p-value for the abundance ratio.
  - **Abundance Ratio Variability [%]**: The variability in the abundance ratio.
  - **Abundance Ratio Weight**: The weight assigned to the abundance ratio.
  - **Abundances (Grouped): M0, M1, M2**: The grouped abundance values for each condition.
  - **Abundances (Grouped) CV [%]**: The coefficient of variation for the grouped abundance values.
  - **Abundances (Grouped) Count**: The number of replicates for each condition.
  - **Abundances (Scaled)**: The scaled abundance values for each sample.
  - **Abundances (Normalized)**: The normalized abundance values for each sample.
  - **Abundance**: The absolute abundance value for each sample.
  - **Abundances Count**: The number of replicates for each sample.
- Additional Sample columns:
  - **Found in File**
  - **Found in Sample**
  - **Found in Sample Group**
- **# Protein Groups**: The number of protein groups identified in the sample.
- **Modifications**: Any identified post-translational modifications on the protein.

Datasets 1, 2 and 3 were generated using Proteome Discoverer (Thermo Scientific). Detailed methodologies, parameterization, and additional information can be found in the ***Proteome Discoverer User Manual:*** *https://assets.thermofisher.com/TFS-Assets/CMD/manuals/Man-XCALI-97808-Proteome-Discoverer-User-ManXCALI97808-EN.pdf*

Dataset 4: limmaCorrected_normed_prePCA_Priori_mrri02_PIF50_DART_1pFDR.csv and its metadata, BMDM_pSCoPE_SampleGuide.csv

- Source-code: https://github.com/SlavovLab/pSCoPE
- **Processed data (proteins)**: Intensity matrix where each row represents an identified protein and each column represents a single cell. The final intensity matrix contains log2-transformed and normalized quantitation values for peptides or proteins across different single cell runs. Each row represents a peptide or protein, and each column represents a single cell or experimental run.
- **Metadata**: The relevant fields for this meta-analysis are:
  - **id**: Cell id

- **celltype**: Cell type ("LPS" for M1 macrophage, "untreated" for M0)

Interpretation of the values

According to the methodology section of the publication as well as the source-code of pSCoPE and SCoPE2 implementations (both available at https://github.com/SlavovLab/), the values of the matrix correspond to the intensities of the identified proteins in each cell after filtering, normalization, imputation, log2 transformation and batch correction. A log2-transformed value of 0 indicates that the quantitation level is equal to the reference level (channel 2 in this experiment). A positive value indicates that the quantion level of a protein is higher than the reference level, and a negative value indicates that is lower than the reference level.

ANNEX II: Other references (publications, data repositories, code repositories, videos, software and others)

- https://cics.bwh.harvard.edu/
- https://cen.acs.org/biological-chemistry/proteomics/proteins-remain-hidden-dark-proteome/100/i3
- https://slavovlab.net/Slavov-Lab-Publications/2023_Proteogenomics.pdf
- https://www.youtube.com/watch?v=JPpKL1uzE0I&ab_channel=TheAhmedLab%3ANorthwesternNeurosurgery
- Brunner, A., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O. B., Bache, N., Apalategui, A., Lubeck, M., Richter, S., Fischer, D. S., Raether, O., Park, M. A., Meier, F., Theis, F. J., & Mann, M. (2022). Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Molecular Systems Biology*, *18*(3), 10798. https://doi.org/10.15252/MSB.202110798/SUPPL_FILE/MSB202110798-SUP-0005-DATASETEV4.XLSX
- https://frontlinegenomics.com/an-overview-of-single-cell-and-spatial-proteomics/
- Xu, Y., Wang, X., Li, Y., Mao, Y., Su, Y., Yang, Y., Gao, W., Fu, C., Chen, W., Ye, X., Liang, F., Bai, P., Sun, Y., Xu, R., & Tian, R. (2023). Multimodal single cell-resolved spatial proteomics reveals pancreatic tumor heterogeneity. *BioRxiv*, 2023.11.04.565590. https://doi.org/10.1101/2023.11.04.565590
- Mund, A., Brunner, A. D., & Mann, M. (2022). Unbiased spatial proteomics with single-cell resolution in tissues. *Molecular Cell*, *82*(12), 2335–2349. https://doi.org/10.1016/J.MOLCEL.2022.05.022
- https://www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/protein-targeting-and-traffic
- https://www.youtube.com/watch?v=rvfvRgk0MfA&ab_channel=ndsuvirtualcell
- https://www.youtube.com/watch?v=4sWnK7OqK-k&ab_channel=Med-Ace
- https://ab604.github.io/docs/bspr_workshop_2018/index.html
- https://github.com/statway/DTFLOW
- https://www.youtube.com/watch?v=U0KaHSL-Yn8&ab_channel=UCIGenPALS

- Ding, J., Sharon, N., & Bar-Joseph, Z. (2022). Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics*, *23*(6), 355–368. https://doi.org/10.1038/S41576-021-00444-7
- https://scproteomicsdb.com/
- https://github.com/SlavovLab/SCoPE2/
- https://pnnl-comp-mass-spec.github.io/proteomics-data-analysis-tutorial/
- https://uclouvain-cbio.github.io/SCP.replication/articles/SCoPE2.html
- https://orangedatamining.com/
- https://www.youtube.com/playlist?list=PL74zbKNRyduROO3uH2HgYAnBjUJ44v3si
- https://www.youtube.com/watch?v=VcbbG7Y5qIs
- https://www.youtube.com/watch?v=IQzTHJbiRK8
- https://www.youtube.com/watch?v=VhimYFFgQ98&ab_channel=EnvironmentalMolecularSciencesLaboratory%28EMSL%29
- https://www.youtube.com/watch?v=C1YaDPP0GYQ&list=PL6ecvCJPoOtnAPDdvNndBSqFKc9FyqpBZ&index=14&ab_channel=EnvironmentalMolecularSciencesLaboratory%28EMSL%29
- https://www.youtube.com/watch?v=Lo_J5FNjIDk&ab_channel=ShortChemistry
- https://www.nature.com/articles/s41467-023-44323-7
- https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00074
- https://www.denbi.de/online-training-media-library/843-introduction-to-computational-proteomics
- https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00441
- https://www.youtube.com/watch?v=KhRAyUNYFyE&ab_channel=LarsJuhlJensen
- https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/introduction/slides-plain.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7216093/
- https://www.sciencedirect.com/science/article/pii/S0168165617302869

- https://anndata.readthedocs.io/en/latest/tutorials/notebooks/getting-started.html
- https://www.technologynetworks.com/proteomics/lists/data-dependent-vs-data-independent-proteomic-analysis-331712
- https://www.creative-proteomics.com/subcell/itraqtmt-label-free-dia-dda-in-proteomic.htm
- Chen, X., Sun, Y., Zhang, T., Shu, L., Roepstorff, P., & Yang, F. (2021). Quantitative Proteomics Using Isobaric Labeling: A Practical Guide. *Genomics, Proteomics & Bioinformatics*, *19*(5), 689. https://doi.org/10.1016/J.GPB.2021.08.012
- Matzinger, M., Mayer, R. L., & Mechtler, K. (2023). Label-free single cell proteomics utilizing ultrafast LC and MS instrumentation: A valuable complementary technique to multiplexing. *PROTEOMICS*, *23*(13–14), 2200162. https://doi.org/10.1002/PMIC.202200162
- https://statomics.github.io/SGA2020/assets/backgroundProteomicsDataAnalysis.pdf
- https://www.biotech.iastate.edu/wp_biotech/wp-content/uploads/2023/07/ProteomeDiscovererInstructions.pdf
- https://bioinfomagician.wordpress.com/2013/10/14/p-vs-q-vs-pep-values-in-mass-spec-database-search/
- https://www.sc-best-practices.org/preamble.html
- https://pmc.ncbi.nlm.nih.gov/articles/PMC10683783/
- https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/introduction/slides-plain.html
- https://www.ometalabs.net/resources/ms1vsms2