POLITECNICO DI TORINO

Corso di Laurea
in Ingegneria Matematica

Tesi di Laurea

Statistical methods for multi-omics data integration: a study on Ehlers-Danlos syndrome

Supervisors
Prof. Enrico Bibbona
Ing. Eugenio Del Prete

Candidate
Alice Zorzan

Anno Accademico 2024-2025

Contents

List of Tables

List of Figures

Abstract

Rare diseases represent a significant challenge for biomedical research due to limited data availability and complex molecular interactions. Multi-omics integration emerges as a promising strategy to overcome individual omics limitations and provide comprehensive biological insights. This work presents a comparative analysis of multi-omics integration approaches applied to Ehlers-Danlos Syndrome (EDS), a group of hereditary rare diseases characterized by collagen production defects. The research operates on three analytical levels: first, single omics analysis, performed on transcriptomics and proteomic data, followed by multiomics integration through statistical techniques. All the levels of analysis are performed on bulk (2D) and spheroid cell (3D) cultures, to capture shared information and compare possible differences. Data were collected from fibroblast cultures of 14 patients (10 patients with disease and 4 healthy controls) under both 2D and 3D conditions. Transcriptomics analysis was mainly focused on the use of DESeq2 framework, based on negative binomial generalized linear models, while proteomics analysis compared three statistical approaches: classical ANOVA, linear models for micro-array data (limma) and non-parametric Wilcoxon test. After this, multi-omics integration was performed, by using three complementary methodologies: Multi-Omics Factor Analysis (MOFA), for identifying shared latent factors across the omics layers; iClusterPlus, that aims to integrate and cluster samples through Bayesian latent variables; and finally Similarity Network Fusion (SNF), for patient similarity network fusion. Each approach was evaluated for its ability to merge multiple omics layers together by using different approaches and for identifying biologically relevant genes. Results show the potential of integration techniques to capture molecular patterns, providing biological insights regarding the most significant genes. The analysis demonstrates how multi-omics integration can reveal further biological insights, by obtaining improved clustering and differentiation along the fused components. While transcriptomic and proteomic analyses alone resulted in partial and inconsistent clustering, patient and control groups are well separated across the latent fused space after applying the integration techniques (especially with SNF). In particular, they produced a clear distinction between patients and controls, with all controls grouped together and the number of misclassified patients reduced to only 3. Moreover, the fused similarity space highlights two main components of separation, driven by a restricted set of genes, which represent promising candidates for investigating Ehlers-Danlos syndrome mechanisms. The comparative and systematic evaluation performed on the different methods emphasizes both the strengths and limitations of each integrative approach, contributing to a deeper understanding of their applicability when applied to rare diseases.

Chapter 1

Introduction

Rare diseases represent a challenge for biomedical research due to the limited availability of data and the complexity of the molecular interactions involved. These two aspects require methodologies and tools to extract the most of the limited amount of data available, but at the same time to simplify the overview of these complex diseases. Omics data integration represents a promising strategy in this sense, since it combines different types of data from omics disciplines (e.g. genomics, transcriptomics and proteomics) to obtain a wider and more complete view of biological processes involved.
This integration helps to overcome the limitations of each individual technology, by merging together the different information extracted from each omics layer and obtaining more meaningful results about the disease.
Focusing on rare disease, integrative approaches open the opportunity to uncover hidden patterns across molecular and biological dimensions, supporting the identification of biomarkers and possible therapeutic targets.

1.1 Aim of the work

The aim of this thesis is to apply and evaluate different integration strategies to study multi-omics data. In particular, starting from a individual analysis conducted separately on transcriptomics and proteomics data, we moved to the study of integrative approaches based on different principles. In this way, the comparison was performed across single omics and multi omics methods, with the ultimate goal of identifying key molecular features, such as genes, proteins or pathways involved in the disease.
In particular, transcriptomics and proteomics focus on two different expressions of the same sample: the first corresponding to the gene expressions in the mRNA, while the latter representing the proteins abundances [11]. Thus, the two data offer complementary insights into the genes expression in a patient, providing a wider understanding.

This study focuses on the Ehlers-Danlos syndrome (EDS), a group of hereditary rare diseases characterized by a defect in the production of collagen, an essential component

of connective tissue [12]. The spectrum of clinical manifestations ranges from joint hypermobility and skin fragility, to abnormal wound healing, due to the collagen function of connect, protect and support the body's tissues [12].
Data are obtained from fibroblast cultures in both bulk (2 dimensional) and spheroid (3 dimensional) conditions, extracting them on the same samples but using different techniques.
By applying statistical and computational tools, it was possible to assess differential expressions, clustering patients and genes behaviors and pathway insights.

The goal is to perform the analysis at three different levels:

- Investigate the common points and differences between transcriptomics and proteomics data, by performing a single omics analysis on each;
- Explore potential patterns or significant signatures between 2D and 3D cell culture conditions, within and across the omics layers;
- Analyze the data as unified framework using the integration techniques, in order to have a comprehensive comparison among all the samples and expressions, identifying possible shared features across conditions and data types.

1.2 State of the art

In the last decades, the raising and development of new technologies has allowed a better understanding and study of rare genetic diseases, that require special care due to their limited availability of patient data and phenotypic heterogeneity.
Among these technologies, transcriptomics and proteomics represent good tools for investigating gene and protein expressions. Omics technologies have a high range of applications and have been used to capture several factors and insights, such as static genomic alterations, proteomics dynamics and temporal perturbations [11].
In the context of rare diseases, experimental data are often scarce, since the number of patients and available samples is relatively low, but the use of advanced omics methods can help find correlations and patterns through statistical and mathematical analysis [9].

1.2.1 Transcriptomics analysis

So far, transcriptomics has stood out as the best tool to describe and quantify the difference between physiological and pathological condition or between before and after treatment [10]. The innovation introduced by this technique is the unified analysis of all the RNA molecules, taking into account the whole pool of cell RNA.
We can say that the transcriptome, i.e. the set of all RNA transcripts, constitutes a peculiarity of the individual cell at a given time or condition. The expression of transcripts, in fact, changes depending on the conditions of the extra and intracellular environment. The goal and power of transcriptomics analysis is not only to investigate the cellular transcriptome, but also its variations from cell to cell or tissue to tissue, as a result of changes

in the cellular or tissue conditions [10].

Transcriptome is studied through two main technologies: microarray, developed in 1995, and RNA Sequencing, invented in the 2000s and that has revolutionized transcriptomics by offering higher resolution and unbiased identification of genes [7]. Microarray allowed the assay of thousands of transcripts simultaneously at a greatly reduced cost per gene [26]. These arrays consist of a solid substrate onto which DNA probes are attached, which encode a gene and assess its expression. The biggest limitations of this tool are the inability to detect de novo transcripts and the necessity of large sample size [11], even if it still maintains a wide use in the scientific community, thanks to its maturity and relative low cost.

On the other hand, RNA Sequencing has improved the transcriptome analysis, allowing for unbiased and high-throughput identification and quantification of genes, including novel ones [7]. It utilizes Next-Generation Sequencing (NGS) technologies to analyze the cellular transcriptome by sequencing RNA molecules. The method is based on a quantitative nature: starting from the sequences of bases, they are then mapped to a reference genome or transcriptome to identify expressed genes. These reads are used to count the genes in the transcript and serve as a measure of gene's expression [10].
After extracting the genes counts, it is essential to proceed with computational, statistical and machine learning techniques to analyze the huge amount of data obtained. These methods are used for tasks such as differential gene expression analysis or for identifying targets for treatments or further studies [7].
Deep learning methods may also be used for their ability to integrate heterogeneous datasets and uncover complex relationships within transcriptomic data [5].

After the raw data is obtained as integer number, representing the genes counts, some computational steps are required to prepare the data for following analysis.
The preprocessing pipeline usually includes prefiltering cut, normalization (i.e. using rlog function) and balancing distributions among the genes expressions. Through heat maps and PCA plot it is possible to assess the overall similarity and dissimilarity among the samples and see the main genes that are able to separate the data.
At this point, the most common downstream application is the differential gene expression analysis, which aims to find the genes with expression levels that differ significantly between two or more conditions. Statistical tools such as DESeq2, edgeR and Limma are among the most widely adopted, relying on negative binomial or linear models adapted for RNA-Seq data [2]. These methods produce a list of differentially expressed genes, typically ranked by adjusted p-values and fold changes.
Finally, to find a bond between the statistical output and the biological context, functional enrichment analysis is usually performed. This part helps identify and assess if some specific pathways, molecular functions or biological processes need extra attention and further analysis. These identified genes are over or under expressed in the samples, compared to a reference condition or control group.
The main tools used for this step are ClusterProfiler and Enrichr to perform both Gene Ontology or pathway enrichment analyses. [2].

1.2.2 Proteomics analysis

From the translated mRNA, we obtain proteins, which represent the final step of the dogma on the omics scale. The study of all the expressed proteins in a cell or tissue is called proteomics, involving at the same time the study of their functions and structures. Cellular proteins can have either a structural or a functional role. It is these last that are of particular interest for all the analysis, since they are mainly responsible for the activities of the cell. Some examples of this case are intra and inter cellular messengers, receptors and enzymes [9].
Proteomics investigates the functional relevance of all expressed proteins in a cell or tissue by interrogating the information flow through protein signaling. The study of proteome serves as a reliable tool to measure cellular alterations during cellular state transitions, as for example in the context of carcinogenesis, since most biological functionalities are activated and controlled by proteins. The fact that a protein is expressed only by a certain cell type or only under certain pathological conditions allows it to be used as a target for therapies, along with the benefit of reducing side effects [9].
In contrast to genomics and transcriptomics, where the plan and directions for cellular processes are being investigated, proteomics examines the actual molecular machinery that carries out these processes. The power of this field lies in its ability to image the dynamic patterns of protein expression, modification and interaction that reflect the true functional cellular and tissue state. The proteome, being the complete set of proteins expressed from a genome at a given time, is the most direct expression of cellular phenotype and function. Protein expression levels, post-translational modifications, and protein-protein interactions vary vastly among different cell types, developmental stages, disease states and environmental conditions [6].

Proteomics technologies have evolved through several generations, but mass spectrometry has remained the primary analytical platform since the 1990s. Traditional approaches like two-dimensional gel electrophoresis, pioneered in the 1970s, provided resolution and visualization of hundreds of proteins in parallel but were limited in their dynamic range and reproducibility [39]. Nowadays, proteomics analysis relies strongly on mass spectrometry, that enables detailed characterization of protein sequences and post-translational modifications.

Mass spectrometry is a analytical technique used to measure the mass-to-charge ratio (m/z) of ions, providing information about the composition and isotopic signature of a sample. In proteomics, MS is primarily used to identify and quantify proteins and their constituent peptides. The general process involves ionizing the sample, separating these ions based on their m/z and detecting them [43].
In particular, proteomics data represents the area under the intensity curve of the mass spectrometry for each peptide. Each value reflects the signal intensity based on the peptide ion counts and abundance. Recent advances have introduced data-independent acquisition methods, which offer greater reproducibility and quantitative accuracy [6]. These methods have enhanced protein quantification reliability and made possible more robust comparative analyses under different biological conditions.

The main advantages of this technique are:

- High Sensitivity and Specificity, being able to identify substances at very low concentrations;
- Versatility, for analyzing a wide range of molecules, from small chemicals to large proteins;
- Identification of changes, since it is good at identifying post-translational modifications on proteins, which are a crucial point to understand proteins function and regulation;
- Quantitative analysis, that is provided with different labeling strategies or label-free approaches.

The most relevant limitations of the Mass spectrometry method are related to the complexity of the data analysis required after generating the data, using sophisticated computational tools, and the high cost and accessibility that the instruments need [43].

Following data collection, proteomics datasets are subjected to computational processing to go from raw spectral information to biological insights. The preprocessing pipeline typically includes shrinkage of the values range by applying a function transformation, imputation of missing values and filtering in order to correct for technical variability.
After performing the preprocessing steps, statistical and machine learning techniques are essential for analyzing the data generated by proteomics. Machine learning algorithms are employed for tasks such as predicting protein-protein interactions, identifying disease markers and classifying protein functions [1].
Several machine learning methods play a crucial role in different phases of the data analysis, such as handling missing values by performing data imputation and clustering the data to uncover hidden relationships. Some examples are the use of *Random Forest* or *Bayesian Principal Component Analysis* for missing values imputation, or *K-nearest neighbors* to cluster proteins and samples.
The most significant downstream application in proteomics is differential protein expression analysis, which identifies proteins whose abundance levels between experimental conditions differ significantly. Advanced statistical methods, often taken from transcriptomics but adapted to suit the characteristics of proteomics data, are used to normalize for missing values, technical noise and batch effects [43]. An example is the Limma package, that allows to fit a linear model for each protein, while shrinking the variances using empirical Bayes approach and performing multiple test correction with Benjamini-Hochberg. After preprocessing on cleaning proteomics data, statistical approaches are still the most used: this includes the application of ANOVA tests, t-tests, linear regression and non parametric tests, as the Wilcoxon Test. Depending on the type of data and the type of analysis one wants to perform, the best tool may vary.

1.2.3 Multi-omics integration

Integrative multi-omics approaches have emerged as a powerful strategy to combine heterogeneous biological data into one analytical framework. These methods aim to capture

the complexity that is present between different molecular layers and biological processes. This innovative approach has arisen because of the complexity, high dimensionality and heterogeneity of multi-omics data. It is a challenge for researchers to extract valuable information from this data [1].
This approach allows for a more global analysis of the data than any individual form of omics technology, since it allows the investigation of data in a complete way, taking into consideration the different steps involving the genes (DNA - mRNA - proteins) [11]. The overall assumption that forms the foundation of multi-omics integration is that the cell systems operate following complex network behaviors, in which the genes information is shared among genomics, transcriptomics and proteomics levels. By combining different layers of data, researchers are able to analyze a greater amount of data extracted from different sites of the cell, enabling multi-dimensional insights that would not be accessible through single omics approaches alone.
Transcriptomics and proteomics are most frequently used in combination, followed by the combination of transcriptomics with epigenomics and proteomics with metabolomics [5]. In fact, putting together the data from mRNA and proteins, it is possible to document both the regulatory instructions hidden in mRNA expression patterns and their functional consequences in protein abundance and activity.

There are two potential approaches for multi-omics data analysis [33]. The first approach focuses on various analysis across the different omics layers in the context of pathways and mechanisms. The key point is that it might use information from different databases to put together the different components of the same disease. The objectives are mainly to gain disease insights and identify significant molecular players involved in the disease. A second and more demanding approach is the integration of multi-omics datasets collected from the same set of patient samples. in this case, the analysis looks for correlations to discover patterns in the features in order to understand the mechanisms of the disease and of that sample set [5].

Strategies for multi-omics integration can be classified broadly into three approaches: early integration, late integration, and intermediate integration [49]. In early integration, the different omics datasets are combined into one table or a single matrix which is then used as input to apply analytical methods. In late integration, models are applied to each dataset independently, and after that a second model combines their predictions. Thus, the results are combined on the interpretation level afterwards, for instance, combining differently expressed genes and proteins into pathway analysis. Lastly, in intermediate integration, a model learns a joint representation of the datasets, performing integration as part of the analytical process itself. These methods are the most sophisticated ones and they typically work by dimensionality reduction or network-based methods [5].

Among all the types of integration techniques, the following ones stand out for their distinctive methodological principles:

- Matrix factorization techniques: very popular in multi-omics integration due to their ability to reveal underlying factors to characterize shared patterns across different

molecular layers [4]. Among the most widely used techniques we have the Non-negative Matrix Factorization (NMF), which breaks down the integrated omics data matrix into two low-dimension matrices, both with the property of no negative elements. This non-negativity makes the resulting matrices easier to inspect and remains coherent to the data being considered (since it always represents counts or areas). For example, when applied to transcriptomic and proteomic information from cancer tissue, NMF can identify molecular subtypes with coordinated changes in protein and mRNA expression, that reveal subtypes that could be overlooked if each omics layer is considered individually.
Another example is the Multi-omics Factor Analysis (MOFA), which is a more sophisticated variant that was specifically designed for multi-omics data. The main idea of this approach is to identify the factors that explain variation within and between different omics layers. MOFA has been largely used for integration of omics data, especially for its ability to handle bulk data and samples scarcity. [4].

- Principal Component Analysis (PCA) and its variants like sparse PCA and kernel PCA are some of the other dimension reduction methods available for multi-omics integration. PCA is primarily used for the exploration and identification of the largest sources of variation within omics datasets. The aim of the method is to reduce the dimensionality of the input data, while at the same time retaining as much variance (hence information) as possible [28]. The procedure consists of constructing axes along which the data is projected. These new axes are the combinations of the original features and are selected based on their explained variance. So, the first ax will always capture the majority of variance from the original data, with each subsequent Principal Component capturing less than the one before it.
  Variation of the PCA method, such as Sparse and Kernel PCA, are still based on the idea of projecting high-dimensional data onto lower-dimensional spaces with maximal variance preservation, but using more sophisticated approach to handle the data and its projection. The main advantage of PCA-based methods is its ability to visualize and discover the major sources of variation between samples or genes [17].

- Network-based integration strategies exploit the inherent networked character of biological systems by constructing and analyzing integrated networks from multiple omics layers [30]. The strategies rely on the concept that proteins and genes do not function in isolation but rather in complex interactions and regulatory networks. One popular strategy is to construct heterogeneous networks with the nodes representing different molecular entities (transcripts, proteins, genes) and the edges representing different types of relations (protein-protein interactions, co-expression, regulation) [14]. These networks can map central hub proteins that serve as key nodes connecting transcriptional and proteomics perturbations, providing insight into potential therapeutic targets.
  An example is the Similarity Network Fusion (SNF), a network-based strategy that constructs patient similarity networks for each omics layer separately, before performing their integration. Then, the method iterate by fusing together the layers into an integrated patient similarity network [30]. These types of approach works well when the data is incomplete or the dataset quality is not uniform. Moreover,

networks can be built in complex ways, taking into account additional, annotated information, allowing the creation of a global unified knowledge.

The selection of suitable integration methods depends on many factors, including the biological questions being addressed, the composition and type of the datasets, and the level of interpretability needed. Each method has its own advantages and limitations, and it is the researcher's duty to understand which type could work better for the specific task. For instance, matrix factorization methods are well suited for identifying global patterns and molecular subtypes, but network methods are well suited to identify crucial regulatory nodes. Unfortunately, finding the perfect tool for the analysis may be difficult, especially for complex data or small samples. This happens frequently when handling omics data related to rare genetic diseases: for this reason it is often reasonable to apply different techniques on the data and compare them and their results.
Throughout this work, many techniques are applied to perform a specific task and then they are compared to find the one that works better and extracts more information. In the next section a resume by chapters of the work is provided.

1.3 Resume by chapters

The aim of this thesis is to provide a deep analysis of transcriptomics and proteomics data, followed by the application of multi-omics data integration approaches taken on Ehlers-Danlos syndrome samples. The study goes from first principles and theoretical explanation, through methods' application and then to empirical results and interpretation.

Chapter 2: Materials and Methods. In this section, the theoretical and practical framework of the research is presented. The Mathematical Background section provides the statistical concepts required for understanding the analytical approaches, focusing on the explanation of the mathematical and statistical knowledge required for the analysis. Following this preliminary section, the core of the work is presented in the Advanced Methods part. Statistical concepts of differential expression analysis are presented, referring to the transcriptomics study, including negative binomial models used in DESeq2 for RNA-seq data, while for proteomics techniques we present analysis as linear models through limma,one-way ANOVA and non-parametric alternatives such as Wilcoxon tests. Multiple testing correction methods, missing value imputation techniques (BPCA, PPCA, SVD, KNN) and filtering approaches are explained alongside data preprocessing.
Chapter 3: Results. The empirical results are organized in three levels of analysis. First, we start with intensive examination of transcriptomics findings, both for the 2D data and for the 3D data. These results comprehend graphical visualizations, like PCA plots and Heatmaps, but also numerical outcomes from the analysis performed.
The Proteomics section has the same structure, with the only difference of the application of tools designed for the proteome data. In this case, statistical analysis includes normality (Shapiro-Wilk test) and homoscedasticity (Levene's test) testing, followed by differential protein expression using three approaches: ANOVA, limma linear model and non-parametric Wilcoxon tests. By comparing the results obtained with the transcriptomics and proteomics analysis, it is interesting to notice possible common points or key

genes more involved in the disease.
The integration part constitutes the core contribution of this work, evaluating and comparing different multi-omics integration methods. MOFA (Multi-Omics Factor Analysis) is used for the identification of hidden factors that capture coordinated transcriptome and proteome variations, illustrated with factor plots. iClusterPlus performs simultaneous clustering across omics layers using Gaussian models of different numbers of latent variables and provides visualizations in latent space and cluster ellipses. Similarity Network Fusion (SNF) builds patient similarity networks for each omics layer, merges them into a large network and applies spectral clustering with feature ranking by Normalized Mutual Information (NMI) scores. Each integration strategy is evaluated based on the strength of its ability to stratify patient groups and identify biologically relevant molecular signatures.

Chapter 4: Conclusions. The last section presents the key findings and their implications of the research carried out over multi-omics data. First, it summarizes the results obtained with a single omics analysis, performed separately on transcriptomics and proteomics data. Some recurrent patterns are highlighted, such as the presence of a outlier sample (RR) and the higher significance in the 3D culture analysis, that can preserve more biologically relevant pattern than the 2D data. The section continues with the presentation of the main findings obtained with the multi-omics integration techniques and a comparison of these results. In particular, the SNF method is the one that performed a more reliable and efficient methodology, especially for the case of rare genetic diseases, in which the sample sizes are limited.
Moreover, the section presents the main limitations of this work and suggests further studies that might be followed to improve the quality of the analysis.

Chapter 2

Mathematical background

In order to investigate the mechanisms and patterns of Ehlers-Danlos Syndrome through multi-omics data, a series of mathematical and biological methods were applied. The following sections describe in detail the tools and techniques used throughout the study. The aim is to describe the mathematical, biological and computational approaches used during all the phases of the study, trying to present them in a multidisciplinary way, but putting more emphasis on the statistical and machine learning tools used.
We start by introducing and describing the methods utilized to preprocess, visualize, impute, check statistical assumptions and perform statistical analysis.

2.1 Data description and notation

This section describes the data nature of the omics datasets analyzed in this study, establishing the mathematical notation used throughout the following section.
Both the transcriptomics and proteomics datasets consist of measurements from 14 patients, including 10 patients with the pathological condition and 4 healthy controls.
The transcriptomic matrix $\mathbf{T}$ contains RNA sequencing data, that are expressed as integer raw counts, represented as a matrix of dimension $G_T \times N$, with:

- G_T corresponding to the number of genes detected in the transcriptomic analysis;
- $N = 14$ being the total number of samples;
- Each element $T_{ij} \in \mathbb{N}$ represents the raw count for gene i in sample j.

In this way, the rows correspond to the genes, identified by the unique gene symbols (Ensembl annotation), while the columns correspond to the samples ID code: $j = 1, \ldots, 10$ for patients (EXP) and $j = 11, \ldots, 14$ for controls (CTR).
The transcriptomic values are discrete integer values, since they represent the direct measurements of the number of sequencing reads mapped to each gene. No normalization or transformation is performed before the preprocessing, that is then explained in the

Section 2.2.

The proteomics dataset $\mathbf{P}$ is organized as a matrix of dimension $G_P \times N$ and positive real values, with:

- G_P corresponding to the number of proteins quantified in the laboratory;
- $N = 14$ being the total number of samples (as for transcriptomic data);
- Each element $P_{ij} \in \mathbb{R}_{\geq 0}$ representing the measure quantity (called abundance) for protein i in sample j.

As said for the transcriptomics matrix, also for $\mathbf{P}$ the rows correspond to proteins and the columns to the patients sample.
The measurements indicate the protein abundance with continuous values, obtained through the mass spectrometry method and label-free quantification. Specifically, proteomics values can be expressed as *Quantity* or *iBAQ* (intensity-based absolute quantification). The first metric reflects the total signal intensity for each protein, maintaining raw and unprocessed the data collected from the laboratory. The *iBAQ* values account instead for protein-specific theoretical peptide numbers, making it more suitable for absolute and complex comparisons across different types of proteins.
In this study, the *Quantity* metric is preferred rather than the *iBAQ* , since the analysis performed does not required additional changes and we wanted to preserve the original data as much as possible.

Due to the different sensitivities and detection capabilities of RNA sequencing tool and mass spectrometry technology, the gene sets collected in the transcriptomic matrix $\mathbf{T}$ and proteomics matrix $\mathbf{P}$ differ. In addition, within the same omics layer, the number of genes in 2D and 3D data are different. This means that $G_{T,2D} \neq G_{T,3D} \neq G_{P,2D} \neq G_{P,3D}$, thus when merging the datasets together for the multiomics integration analysis, only the genes present in both datasets are retained.

Specifically, the datasets analyzed in this work contain the following number of genes:

	2D	3D
Transcriptomics	12 046	12 578
Proteomics	5 544	5 369

Table 2.1. Number of genes detected in each omics layer (2D and 3D).

After merging the samples from the two different cultures, the transcriptomics layer contains 11 418 genes, whereas the proteomics data is composed of 5 050 genes.

2.2 Preprocessing for transcriptomics data

In the data analysis process, the first step is always characterized by data cleaning and filtering, since it is common to have noise, outliers, redundancy and any type of obstacle that make raw data difficult to manage. It is then essential to first detect the issues present in the data and then to solve them by performing accurate and specific approaches.
In this work, the most important techniques of preprocessing applied are filtering, normalization and log transformation.

For transcriptomics data, the genic expression is represented as an integer number, corresponding to the count for each gene. The main features of mRNA counts are:

- a long right-skewed tail in the distribution, due to few highly expressed genes;
- a wide dynamic range, that covers several orders of magnitude;
- a high concentration of values near or at 0, reflecting low or undetected expression for many genes;
- heteroschedasticity, with the variance greater than the mean, a phenomenon called over-dispersion.

These characteristics motivate specific preprocessing strategies, such as filtering low-count genes, normalization and transformation, to improve the reliability and interpretability of the analyses.

2.2.1 Low expression filtering

Raw RNA expression data are represented by an integer, that corresponds to the expression count of that specific gene in the sample. These counts range in a wide interval, with low number of counts associated with large proportion of genes, joint with a lack of upper limit for expression. These factors allow for a long right tail to grow, that is very noticeable when visualizing the data in a density plot. To avoid this behavior, that does not allow to approximate the data with known distribution, we performed a low expression cut.
The idea is that if a gene is low expressed in all the samples analyzed, including both control and treatment, then its significancy for the study is very likely to be inconsistent. In fact, when studying mRNA data, differential expression is the common tool to be applied, meaning that we look for changes in the gene expression between two or more groups, such as comparing case vs control or treatment1 vs treatment2. More in general, the goal is to identify the source of variation such that we can separate the interesting from the uninteresting part. So, if the variation of a gene expression has very low level (close to 0), then that gene is not relevant for the study performed.

The low expression filtering, also known as independent or gene-level filtering, erases the rows with very low counts, since these genes are not likely to see significant difference due

to high dispersion. This step is performed during the preprocessing part, so the genes that are cut out from the dataset are no longer involved in any type of further analysis, and neither they are tested through the statistical method after. This reflects a reduce in the number of genes that are considered for multiple testing in the differential analysis, and thus in a increase of the fraction of significant genes.
To sum up, this tool has the goal of cutting out those tests referred to genes that have no or little chance of showing significant evidence, without even looking at their statistic.

There are different types of gene-levels filtering:

1. Cut the genes that have all counts equal to 0. In this case, if these genes were tested, they would obtain $basemean = 0,\ p - value = NA,\ p.adj = NA$.

2. Cut the genes that have extreme count outliers, using the method of Cook's distance [36]. In this case, the tests would result in $p - value = NA,\ p.adj = NA$.

3. Cut the genes with low mean normalized count, that would obtain a $p.adj = NA$.

This last type is the one used in this work.
The filtering criterion was established as:

$$\sum_{j=1}^{N} T_{ij} \geq \theta$$

where T_{ij} represents the raw count for gene i in sample j, N is the total number of samples ($N = 14$), and $\theta = 50$ is the minimum sum threshold. The value of the thresold θ is set empirically, starting from low value and increasing it until the filtering shows a good cut in the density and box plots. This approach eliminates genes with consistently low counts that contribute primarily to noise rather than to biological signal.

2.2.2 Regularized log transformation

Following filtering, the count data were analyzed with the DESeq2 framework. Here, the analysis is performed on the raw data counts through the use of GLM, that are explained in the Section 3.2. To enable sample visualization and clustering, transcriptomic counts were subsequently transformed using the regularized logarithm (rlog) transformation. This transformation is crucial to stabilize variance across the range of expression values and to provide a more reliable representation for downstream analyses.

For transcriptomics data, the regularized log transformation allows the count data to pass to the log2 scale in a way which minimizes differences between samples for rows with small counts and which normalizes with respect to library size [24].
The rlog transformation employs the same negative binomial generalized linear model used in the differential analysis in DESeq2 (see Section 3.2), but with a modified design matrix where each sample is treated as an independent factor. Specifically, the transformation

is built upon the negative binomial generalized linear model, where for each gene i and sample j, the raw counts T_{ij} are modeled as:

$$T_{ij} \sim \text{NegBin}(\mu_{ij}, \alpha_i).$$

μ_{ij} represents the mean parameter and α_i the gene-specific dispersion parameter (more details in Sec. 3.2).
In this context, q_{ij} represents the fitted values from this GLM, corresponding to the expected counts for gene i in sample j after accounting for size factors and applying empirical Bayes shrinkage to sample-specific coefficients.

The rlog transformation addresses the mean-variance relationship inherent in count data through:

$$rlog(T_{ij}) = \log_2(q_{ij}).$$

This transformation stabilizes variance across the dynamic range of expression values, making the data suitable for downstream analysis that assume homoschedastic errors. As nearby count values for low counts genes are almost as likely as the observed count, the rlog shrinkage is greater for low counts. For high counts, the rlog shrinkage has a much weaker effect.
To sum up the concept, the transformed values, are equal to

$$rlog(T_{ij}) = \log_2(q_{ij}) = \beta_{i0} + \beta_{ij},$$

with β_{i0} being the baseline expression level for gene i and β_{ij} representing the sample-specific coefficient for gene i in sample j, with prior distributions applied for regularization [24].
This means the transformation applies shrinkage to the sample-specific effects β_{ij}, particularly for genes with low counts, which is what provides the variance stabilization. This regularization is what makes rlog transformation particularly effective for variance stabilization compared to a simple log transformation.

2.3 Preprocessing for proteomics data

Regarding proteomics data, the registrations represent the area under the intensity curve of the mass spectrometry for each peptide. Each value reflects the signal intensity based on the peptide ion counts and abundance. The main issues that need to be solved are the huge values range of the proteome expressions and the high number of missing values. To cope with them, it is usual to apply log transformation to shrink the values range and to use imputation of missing values techniques to fill or erase the features with no value available.

Log transformation

For proteomics data, the logarithm transformation is applied as first step of the preprocessing phase. This is essential to pass from the raw data to normalized values that are easier to study and deal with, especially for the shrinkage of the values range. The application of this transformation allows to separate and erase the relation between mean and variance and to enhance the high variances in respect to the low variances.

As opposed to the transformation applied on transcriptome data, it is not required to perform a rlog tranformation for the proteomics values, but it is sufficient to apply a standard log transformation in basis 2 to shrink the values range. The initial raw data are characterized by a minimum equal to 0 and maximum greater than 3 million.

2.4 Visualizations and plots

Another important step to include in the data analysis is to visualize it through different representations. Hidden patterns, relevant information and important features of the values are often noticeable by looking at the plots and extracting them just by observing the whole context. In this study, the most used methods for plotting data are boxplot, pca plot, density plot and heatmap representation. All of them are useful for explorative data analysis and for highlighting variability, detecting potential outliers and assessing data distribution.

2.4.1 Boxplot and density plot

Starting from the most simple, the boxplot is a method for demonstrating graphically the locality, spread and skewness groups of numerical data through their quartiles. Its power lays on the capacity of detecting outliers and also on the fact that it is non-parametric: it displays variation in samples of a statistical population without making any assumptions of the underlying statistical distribution.
The boxplot tool is a way of representing the dataset based on five values summary: the minimum, the maximum, the sample median, the first and third quartiles. Specifically:

- The minimum is the lowest data point in the data set excluding any outliers;
- The maximum is the highest data point in the data set excluding any outliers;
- The median is the middle value in the data set;
- First quartile is the value under which 25 % of data points are found when they are in increasing order.
- Third quartile is the value under which 75 % of data points are found when they are in increasing order [45].

Boxplot's advantages are the detection of outlier and the visualization of possible simmetries. On the other side, this tool is not relevant for detailed analysis of the data as it deals with a summary of the data distribution.

Density plot has some similarities with the boxplot tool, but instead of focusing only on the five numbers summary, it deals with the whole distribution of the data. In fact, density plots use kernel density estimation to create a smoothed, continuous curve that approximates the underlying distribution, over a continuous interval or time period. In this way it is possible not only to provide a general idea of the range of values of the data, but also to visualize the empirical distribution extracted from it. Density curves can have all shapes and sizes and they allow us to gain a fast visual understanding of the trend of values in the dataset. In particular, this tool is common used for its ability to visualize:

- Skewness of the distribution, since it is fast to notice if the curve is left skewed, right skewed or has no skew.
- The number of peaks, that can tell us if the distribution is unimodal or multimodal, meaning that has two or more peaks.
- Similarity with known distributions, to understand whether is better to use specific statistical methods or others.
- Important characteristics of data, that need to be handle before proceeding with further analysis [20].

2.4.2 PCA plot

Principal Component Analysis is a technique widely used to reduce the dimensionality of input data while retaining the most significant variations. It involves the following steps:

1. Calculate the mean of the feature vector: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
2. Compute the covariance matrix as

$$\mathrm{Cov}(x_i, x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_j - \bar{x})$$

   where $\bar{x}$ is the mean of the feature vector;
3. Calculate the eigenvalues and eigenvectors of the Covariance matrix. The eigenvectors are then sorted in descending order based on their corresponding eigenvalues, which represent the variance explained by each principal component;
4. By fixing the percentage of explained variation that we want to retain or the number of dimensions that we want to keep, the directions (principal components) that capture the most variance in the data are maintain accordingly [19].

PCA is not only strictly used for dimensionality reduction, but also for visualizing the data distribution along the first two or three principal components, i.e. the directions that explain the majority of the variation in the data. It helps to visualize high-dimensional data by projecting it into a lower-dimensional space, such as a 2D or 3D plot. This simplifies data interpretation and exploration. The power of this type of plot lays in the ability of expressing how the samples are positioned according to the features that originate most variance. Applying PCA can help to preprocess or extract the most informative features from datasets with many variables, while preserving relevant information.

A PCA plot is a scatter plot created by using the first two principal components as axes. The plot shows the relationships between observations and the new variables (the principal components). The position of each point shows the values of PC1 and PC2 for that observation. The direction and length of the plot arrows indicate the loadings of the variables, that is, how each variable contributes to the principal components. If a variable has a high loading for a particular component, it is strongly correlated with that component [18]. This can highlight which variables have a significant impact on data variations.

An example is shown in Figure 2.1, that illustrates the geometric interpretation of Principal Component Analysis. The data points (in yellow) are distributed in a three-dimensional space defined by the first three features X_1, X_2, X_3. PCA identifies a new coordinate system (represented by the dotted line) such that the variance of the projected data points along this new axis is maximized. This axis corresponds to the first principal component, which captures the direction of maximum variability in the dataset.

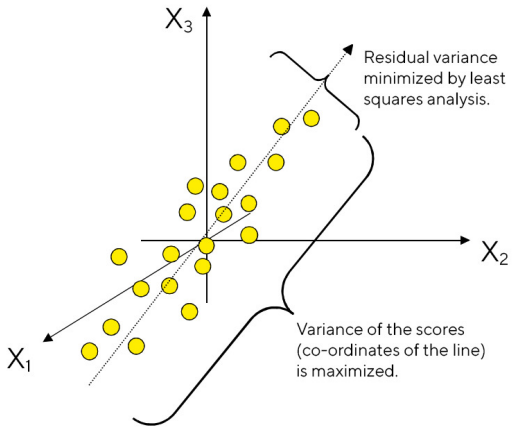

Figure 2.1. Geometric interpretation of PCA [41].

The projection of the original points onto this axis reduces the dimensionality of the data while preserving the most informative structure. At the same time, PCA minimizes the residual variance, meaning the distance between the original data points and their projections, using a least squares minimization. This dual goal, of maximizing variance and at

the same time minimizing reconstruction error, makes PCA a powerful method for both dimensionality reduction and visualization.

2.4.3 Heatmap plot

Heatmap is a visualization tool widely used in genomics and bioinformatics to display high-throughput gene expression data in a compact and understandable form. The method represents numerical values of gene expressions as a color-encoded matrix whose cell color intensity corresponds to the degree of gene expression. In this way the tool enables a rapid identification of expression patterns in genes and samples [32].
The heatmap plot is obtained from the mathematical representation of gene expression data as a matrix $\mathbf{T} \in \mathbb{R}^{G_T \times N}$ for the transcriptomics data, as presented in the Section 2.1. To improve pattern recognition and comparability between genes whose expressions range differently, it is commonly used to apply a standardization first. An example of standardization is the z-score normalization that takes the original expression value, subtracts to it the mean expression of gene i across all samples, and divides the result by the standard deviation.

To find the most informative features, genes $g_1, ..., g_{G_T}$ are ranked in descending order of their variance within samples according to the following formula:

$$Var(g_i) = \frac{1}{N-1} \sum_j (T_{ij} - \mu_i)^2,$$

with $\mu_i = \frac{1}{N} \sum_j T_{ij}$.
The top 50 most variable genes are selected because they typically capture the most biologically informative features that induce sample differentiation and explain the primary sources of variation in the dataset.

In addition, the plot contains hierarchical clustering to reveal the underlying data structure, even if often it is noticeable by just looking at the heatmap. The distances are computed pairwise by the clustering algorithm through Euclidean distance and are used to construct the dendrogram through complete linkage criteria, which are also displayed along with the heatmap for marking sample and gene relationships [37].
The key advantages of this tool are:

- Easy pattern recognition, thanks to the color-coding system that facilitates fast visual identification of gene clusters and sample groupings, making it simpler to identify biological modules and pathways.

- Dimensionality reduction, reducing high-dimensional data to understandable two-dimensional representations with significant information. The focus is concentrated on the 50 most variant genes and the application of clustering performed with the top genes information.

- Scalability, since it is capable of displaying a huge number of genes and samples within a single figure, making them ideal for exploratory data analysis and result presentation.

2.5 Statistical assumptions check for proteomics data

Before performing differential expression analysis using Analysis of Variance (ANOVA), it is required to ensure that the data satisfies the fundamental assumptions of this statistical test. ANOVA is a parametric test and relies on some distributional and variance assumptions in order to have valid results and to maintain proper Type I and Type II error rates.
There are two main assumptions that must be ensured, which are:

1. Normality: Residuals (or equivalently, the observations within each group) should be normally distributed;
2. Homoscedasticity: The variance of observations should be the same for all the groups in an experiment (homogeneity of variances).

Violations of such assumptions can create inflated Type I error rates, reduced statistical power and incorrect conclusions about differential protein expressions. Therefore, specific statistical tests are employed to investigate the validity of such assumptions before applying ANOVA-based methods.

2.5.1 Shapiro-Wilk test

The Shapiro-Wilk test is widely used to check the normality assumption, particularly for small to moderate sample sizes ($N < 50$). This test evaluates the null hypothesis that the data comes from a normal distribution [27].

Given the gene i, the Shapiro-Wilk test statistic is defined as:

$$W = \frac{\left(\sum_{j=1}^{N} a_j \cdot \tilde{T}_{ij}\right)^2}{\sum_{j=1}^{N}(\tilde{T}_{ij} - \bar{T}_i)^2}$$

where:

- $\tilde{T}_{ij}$ are the ordered sample values of the statistics, meaning that are the jth-smallest number in the sample, in respect to the gene i;
- $\bar{T}_i$ is the sample mean of the values of gene i;
- a_j are coefficients derived from the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and the covariance matrix of those normal order statistics. The coefficients a_j are calculated such that the numerator represents the best linear unbiased estimator of the standard deviation under the assumption of normality.

The test statistic W ranges from 0 to 1, with values closer to 1 indicating greater evidence for normality. There is no name for the distribution of W. The cutoff values for the statistics are calculated through Monte Carlo simulations.

The null hypothesis of normality is rejected when W is significantly small, corresponding to a p-value less than the chosen significance level (typically $\alpha = 0.05$). In this case there is sufficient evidence that the data tested are not normally distributed. If W reaches small values, it means that the linear combination of order statistics deviates substantially from what would be expected under normality.

2.5.2 Levene's test

Levene's test is a robust technique for testing equality of variances across groups, i.e. homoscedasticity of the data, and it is less sensitive to non-normality samples in respect to other tests. Therefore, even if data does not apply with the normality assumption, this test is still valid for testing the variances. For this reason, the Levene's approach is a good choice for proteomics data that could be fairly non-normal.

Levene's test is based on the ANOVA of the absolute deviations from group medians (or means). Given the gene i, we denote with P_{kj} the observation for the gene i, group k and sample j. The test statistic is:

$$L = \frac{(N-K)}{(K-1)} \cdot \frac{\sum_{k=1}^{K} n_k (\bar{Z}_{k\cdot} - \bar{Z}_{\cdot\cdot})^2}{\sum_{k=1}^{K} \sum_{j=1}^{n_k} (Z_{kj} - \bar{Z}_{k\cdot})^2} \tag{2.1}$$

where:

- K is the number of groups ($K = 2$);
- n_k is the number of observations in group k, thus $n_1 = 10$ for the patients and $n_2 = 4$ for the controls;
- $N = \sum_{k=1}^{K} n_k$ is the total sample size, $N = 14$;
- $Z_{kj} = |P_{kj} - \tilde{P}_{k\cdot}|$ where $\tilde{P}_{k\cdot}$ is the median of group k
- $\bar{Z}_{k\cdot} = \frac{1}{n_k} \sum_{j=1}^{n_k} Z_{kj}$ is the mean of Z_{kj} for group k
- $\bar{Z}_{\cdot\cdot} = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Z_{kj}$ is the overall mean of Z_{kj}

Under the null hypothesis of equal variances, the statistic L follows an F-distribution with $(K-1, N-K)$ degrees of freedom [31].

The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ is rejected when L exceeds the critical value from the F-distribution at the chosen significance level. Large values of L indicate substantial differences in the spread of observations across groups, suggesting heteroscedasticity.

2.6 Statistical methods for proteomics

The statistical analysis of proteomic data requires robust methods capable of handling high-dimensional datasets while controlling for multiple testing and addressing the specific challenges inherent to mass spectrometry-based quantitative proteomics. Three complementary approaches were implemented: classical ANOVA for parametric testing under normality assumptions, limma for improved statistical power through empirical Bayes moderation, and Wilcoxon rank-sum test for non-parametric analysis when distributional assumptions are violated.

2.6.1 Analysis of variance (ANOVA)

The one-way ANOVA model follows the linear regression model, in which qualitative variables X_i, such as the presence or absence of a disease or the use of a treatment, function as predictive independent variables related to the dependent variable Y_i. In proteomics data this is widely use since it is often interesting and useful to study the relation between protein with high abundance or differential expression and the presence of specific biological conditions, such as disease states, treatment responses, or other phenotypic traits. In this case, given the gene i, the ANOVA model is written as:

$$P_{kj} = \mu + \alpha_k + \epsilon_{kj} \tag{2.2}$$

where:

- P_{kj} represents the log-transformed intensity of gene i in group k and sample j;
- μ is the overall mean intensity for gene i;
- α_k is the effect of group k, as for example the effect of the treatment applied to the samples in that group;
- $\epsilon_{kj} \sim \mathcal{N}(0, \sigma^2)$ are the independent error terms, that represent the noise in the data.

The null hypothesis tests the equality among the group means:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{at least one } \mu_i \neq \mu_j$$

The statistic is computed as shown previously in the Equation 2.1.
Under the null hypothesis, the statistic follows a F-distribution with degrees of freedom of $F \sim F_{k-1,N-k}$.

The ANOVA method is able to work properly when the following assumptions hold true:

1. Normality: the protein intensities must follow normal distributions. This property can be checked with the Shapiro-Wilk test;

2. Homoscedasticity: variances across groups must be equal: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$. This is tested using Levene test;

3. Independence: All the observations are independent within and between groups.

This technique stands out for its wide use across different field and application. Its strengths lay in the well-established statistical theory under assumptions, the computational efficiency for large datasets and the optimal power when assumptions are met.
On the other hand, the approach remains very sensitive to violations of normality and homoscedasticity assumptions, susceptible to outliers affecting variance estimations and limited power with small sample sizes. In addition, proteomics data is characterized by high number of proteins recorded, while rare diseases are scarce in samples, leading to dataset with few patients and lots of protein intensities. The ANOVA model does not take this into account, and so does not consider the information shared across proteins, but instead look only at each protein separately.

2.6.2 Linear models for microarray data (limma)

The limma approach extends classical linear modeling through empirical Bayes moderation of variance estimates for proteomics data analysis. The package enables the consistent application of linear models to normally distributed omics data in general, with a specific focus on microarray data. The power of this tool is that includes in the linear model an empirical Bayes method that borrows information across features to estimate the standard error and calculate the t-statistics according to it. This approach is demonstrably more powerful than a standard t-tests or ANOVA approach when the number of samples is low [13].

For gene i, the linear model is:

$$p_i = X\beta_i + \epsilon_i, \tag{2.3}$$

where p_i is the intensity vector, X is the design matrix, β_i contains the coefficients and the error is $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 I)$.

The key innovation of this tool is the empirical Bayes moderation integrated in it, with the estimation of gene-specific variance moderated. The linear model for gene i has residual variance σ_i^2 with sample value s_i^2 and degrees of freedom d_i [44]. The empirical Bayes method assumes a scaled chisquare prior distribution for $1/\sigma_i^2$ with mean $1/s_0^2$ and degrees of freedom d_0:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{s_0^2} \cdot \frac{\chi_{d_0}^2}{d_0}. \tag{2.4}$$

The moderated variance combines gene-specific and pooled variance estimates, obtaining the posterior values for the residual variances as:

$$\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i} \tag{2.5}$$

where:

- s_0^2 is the pooled variance estimate across all proteins;
- d_0 and d_i are the prior and residual degrees of freedom for each gene i.

Finally, for testing $H_0 : \beta_{ik} = 0$ the moderated t-statistic that takes into account the posterior variances is:

$$t_{ik} = \frac{\hat{\beta}_{ik}}{\tilde{s}_i \sqrt{c_{kk}}} \tag{2.6}$$

where c_{kk} is the k-th diagonal element of $(X^T X)^{-1}$.
Under the null hypothesis, t_{ik} follows a t-distribution with $d_0 + d_i$ degrees of freedom [44].
Hyperparameters are estimated using method of moments.

Among the main advantages of this approach we find the higher statistical power, since borrowing information across proteins reduces variance estimates, that is beneficial for small sample sizes; and the improved stability, preventing inflation of significance for proteins with artificially low sample variances. With proteomics data, the limma tool is more appropriate because of its ability to handle good high-dimensional data and small sample size, since it has been designed specifically for omics datasets with thousands of features. In fact, the empirical Bayes framework is particularly advantageous in proteomic studies where sample sizes are often limited due to cost constraints, while the number of measured proteins is large.

However, the method still presents some limitations: it requires normality assumption for protein intensities, it may be overly conservative when true variances vary substantially across proteins, and the computational complexity increases with dataset size.

2.6.3 Wilcoxon Rank-Sum test

The Wilcoxon rank-sum test provides a non parametric alternative, meaning that it does not require strong assumptions on data distribution or characteristics. In particular, this test is used when normality assumptions are violated.
The Wilcoxon method tests if two populations distributions are identical or not, therefore the hypothesis is:

$$H_0 : \text{identical distributions}$$
$$H_0 : \text{distributions differ in location (median).}$$

Given the gene i, for two independent samples $X_1, \ldots, X_{n_1}$ being the first group and $Y_1, \ldots, Y_{n_2}$ being the second, the test statistic is based on ranks rather than raw intensities. Specifically, it takes the joint ranking of the observation from the two samples, considering them as extracted from the same group and ordered.

Let R_j denote the rank of observation j in the combined sample of size $N = n_1 + n_2$. The Wilcoxon statistic is:

$$U = \sum_{j=1}^{n_1} R_j \tag{2.7}$$

For large samples, i.e. when $n_k > 5$, the test statistic follows:

$$Z = \frac{U - \mu_U}{\sigma_U} \sim \mathcal{N}(0,1) \tag{2.8}$$

where:

- $\mu_U = \frac{n_1 n_2}{2}$;
- $\sigma_U^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12}$.

For large samples, the distribution of the Wilcoxon statistic U can be justified through the asymptotic theory of U-statistics. In this context, the classical Central Limit Theorem cannot be directly applied, since the ranks R_j are not independent. Instead, one relies on the framework of U-statistics. The Wilcoxon statistic can in fact be expressed as a U-statistic of degree two, as it is based on pairwise comparisons between elements of two independent samples.
More precisely, given two independent samples of sizes n_1 and n_2, the Wilcoxon statistic U can be written in the form

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j), \tag{2.9}$$

with kernel $h(x, y) = \mathbf{1}\{x < y\}$. This representation shows that U is a U-statistic of degree two, as it coincides with its definition.
A fundamental result, shown in [46], proves that, if the kernel has finite second moments, any U-statistic is asymptotically normal:

$$\sqrt{n}\,(U - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{2.10}$$

where $\theta = \mathbb{E}[h(X_1, \ldots, X_r)]$ and σ^2. This result does not rely on independence of the ranks themselves, but rather on the structure of U-statistics and the projection method used in their asymptotic analysis.
This justifies the use of the standardized normal approximation in the large-sample case.

In general, the test evaluates:

$$H_0 : \mathbb{P}(X > Y) = 0.5$$
$$H_1 : \mathbb{P}(X > Y) \neq 0.5$$

This is equivalent to testing the equality of population medians under symmetric distributions, that means testing if the distributions are identical in location or not.

The main advantages of this technique lay in the absence of assumptions about underlying data distribution and in the robustness to outliers, since rank-based statistics are less sensitive to extreme values. Another important characteristic of this approach is its power even for small sample size data, maintaining exact p-values available also when the samples are scarce. This is a factor that allows more robustness in the analysis of proteomics data, especially with rare diseases, due to the low availability of patients.

However, the Wilcoxon Rank-Sum test is limited by the absence of interpretable measures of biological significance and by a lower statistical power, especially when normality holds it is less powerful than the parametric techniques.

The three approaches provide complementary perspectives on differential protein expression and for this reason they have been all implemented in this study. By comparing the results, it is then possible to notice common points or divergences between the methods outputs.
To sum up:

- ANOVA: Optimal under normality with equal variances, provides interpretable F-statistics and detailed effect size estimation. Best suited for well-controlled experimental conditions with adequate sample sizes.
- limma: Superior power than ANOVA for small sample sizes through variance moderation and robust to moderate variations from normality while maintaining parametric efficiency. Ideal for typical proteomic studies with limited biological replicates.
- Wilcoxon Rank-Sum test: the only non parametric, effective for detecting median differences regardless of distributional assumptions. Essential when data transformation fails to achieve normality or when robust inference is prioritized.

This multi-method approach provides comprehensive coverage of potential differential expression patterns, ensuring the identification of biologically relevant proteins using different techniques.

Chapter 3

Advanced methods

3.1 Imputation of missing values

Missing values represent one of the main issues that need to be addressed in proteomics data. The absence of registrations recorded in the samples might reaches high level that interferes with the mathematical analysis. These missing values, in fact, threaten the integrity of subsequent statistical analyses by reduction of statistical power, introduction of bias and failure to represent the true sample [23]. Over the years, several categories of missing value imputation methods have been developed and adapted for proteomics data. By using the mass spectrometry technique to identify proteins in a cell or tissue, thousands of different proteins can be quantified in a single MS injection. The major issue is that the power of statistical inference and downstream functional analysis is greatly impacted by the presence of missing values in the protein abundance data. Multiple factors contribute to the presence of missing data in proteome, including biological factors, such as non existing proteins and protein abundances below the instrument detection limit, as well as analytical factors, such as sample loss in preparation, mis-cleavage of peptides during digestion and poor ionization efficiency [21].
Missing values might also come from other various factors such as scratches on the slide, spotting problems, dusts, experimental errors and so on. In practice, every experiment contains missing entries and in some extreme cases the level of genes affected by missing values can be up to 90 % [48]. Moreover, most of the classic multivariate analysis methods for proteomics data cannot be used when the data have missing values. Therefore, we need to treat missing values appropriately.

In general, we can distinguish two categories of missing values for proteomics data:

- Missing at random, known as MAR missing values, mostly result from technical limitations and stochastic fluctuations in an abundance-independent manner;
- Missing not at random, called MNAR missing values, are more abundance-dependent that can be explained by the measurability of the corresponding peptides [21].

Missing values in proteomic data are a mixture of MAR and MNAR.

First thing that we implemented in the preprocessing is the filtering for high level of missing values. In fact, imputation methods for missing values are able to fill with good accuracy the missing entries, in a efficient and precise way. However, this is not possible when the registrations for a specific protein are missing in almost all the samples: the methods can still be applied but the result obtained can not be considered satisfactory and accurate. Therefore, before applying the imputation technique, it is essential to filter those genes that have a percentage of missing entries greater than a threshold. Usually the limit is set at 70 %, meaning that we erase all the rows with less than 70 % of observations recorder, i.e different from NaN, for all groups.
This condition must be verified for at least one group. If, for example, the dataset is composed of two groups, i.e. "control" and "treatment", we ask to each protein to reach a threshold of at least 70 % of recorded observations for the control group or 70 % of recorded observations for the treatment group. The only case in which that protein is erased from the dataset is when for both groups the threshold is not reached.
Mathematically, for a given gene i, the following condition must hold for at least one group k:

$$\frac{\#\mathrm{NA}_k(i)}{n_k} < 1 - \tau \tag{3.1}$$

where:

- $\#\mathrm{NA}_k(i)$ is the number of missing values for gene i in group k,
- n_k is the total number of samples in group k,
- τ is the required threshold, for example $\tau = 0.7$ for 70% completeness.

Only the proteins that satisfy this filtering criterion proceed to the imputation phase to ensure that the subsequent missing value prediction is performed on data with sufficient information to provide valid results.

Following this filtering process, imputation of the missing values that are still present in the dataset must be carried out with appropriate statistical procedures. The imputation approach is critical as it can impact downstream analyses, such as differential expression analysis, pathway enrichment and biomarker identification. Different imputation methods have been developed to address the different characteristics that can be found in proteomics data.
Because of the mixed type of missing values in proteomics data (both MAR and MNAR), we can not know a priori which will be the best method for the data imputation. Here, we employed and compared different techniques to determine the best imputation for our data, testing and evaluating them by performing a masking on the dataset. With the use of the cross validation masking test, it was possible to 'hide' some values, perform different methods and evaluate them with the RMSE. The applied technique consists of random masking, which better reflects the behavior of randomly occurring missing values. This approach was considered sufficient for the purposes of the study, although it does not fully account for the presence of MNAR data. The one that reached the lowest squared

error was then used to perform the actual imputation over the missing values.

The methods performed and tested are:

- Probabilistic PCA is a improved version of PCA that assumes that the observed data are generated by a linear model with Gaussian noise and is based on maximum likelihood estimation. PCA is one of the most reliable techniques for dimensionality reduction as it minimizes the reconstruction loss on variance during the data compression. This property can be utilized for imputing the missing data points by first estimating the distribution of the compressed information based on the available data and then reconstructing the missing data from the compressed information by projecting data points [16].
  Different versions of PCA algorithms to handle missing data are now present and they mainly differ in the assumption on the relationship between the original data points and the latent data points. In general, this kind of methods is powerful for imputation of missing values in proteomics data since the basic idea on which they are founded is that the observed variables (proteins) live in a space with low latent dimensionality. This means that their variability can be explained by few principal components, instead of needing high amount of features to explain it. What we can do is to initially replace missing values with crude and simple estimates (for example row or column mean) and thus obtain a fictitious “complete” matrix. At this point, we apply PCA or any type of versions of it to this "complete" matrix, reducing the dimensionality. Finally as last step we rebuild the matrix in the initial space, so the values of the missing entries are overwritten with the rebuilt ones. The iterative procedure is repeated until convergence, i.e. until the imputed values no longer change significantly.
  Specifically, Probabilistic PCA is based on the union of two concepts: the dimensionality reduction performed by PCA and the estimation of the missing data through the maximum likelihood technique [16].

- Random Forest method is applied to impute missing values particularly in the case of mixed-type data. It can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations [38]. It uses a series of decision trees to impute missing values. Each tree is trained on a subset of the data and provides a prediction for the missing values. The final prediction is an average of the predictions made by the individual trees. The main disadvantage of this technique is that the training of the whole forest is very time consuming and with big amount of data the algorithm easily becomes slow.

- Bayesian Principal Component Analysis, which is a improved version of PCA method was developed especially for missing value estimation. Scores and loadings obtained with Bayesian PCA slightly differ from those obtained with conventional PCA. The algorithm does not force orthogonality between factor loadings, as a result factor loadings are not necessarily orthogonal. However, it was found that including an orthogonality criterion made the predictions worse. Bayesian PCA works iteratively

and the complexity is growing with $O(n^3)$ because several matrix inversions are required [34]. The size of the matrices to invert depends on the number of components used for re-estimation. The relatively high complexity of the method is a result of several matrix inversions required in each step. Considering the case in which the maximum number of iteration steps is needed, the approximate complexity is given by the term

$$maxsteps \cdot row_{missing} \cdot O(n^3),$$

where $row_{missing}$ is the number of rows containing missing values and $O(n^3)$ is the complexity for inverting a matrix of size n [34].
The model assumes that the observed data are generated by a linear model with Gaussian noise using Bayesian statistics: a priori is performed on μ, σ and latent variables. Posterior distributions of all parameters are calculated and the final estimates are obtained as averages over these posteriors, often via Bayesian EM algorithms or approximations [35]. This reduces overfitting, better handles uncertainty and imputation when data are sparse or noisy.

- K-Nearest Neighbors, that for each missing value looks for the "k" nearest neighbours and uses the mean (or median) of the values of these neighbours to impute the missing entry. It works by finding the "nearest neighbors" (rows) that have similar patterns to the row with missing data and then to calculate the missing values [22]. The main steps are:

  1. Each row is treated as a coordinate in a multi-dimensional space (each column represents a dimension).
  2. The algorithm calculates the distance between rows to identify the ones that are the most similar.
  3. The missing value is then estimated based on the values of the closest rows.

  The distance between samples is generally measured using a Euclidean distance. The metric used to compute the distances between samples highly affects the results: it might happen that using different metrics leads to very different outputs, both when the method is used for imputation and for clustering.

3.2 Differential analysis for transcriptomics

Differential expression analysis of RNA data presents unique statistical challenges, that are not present in other types of omics data, such as the continuous intensity measurements from proteomics. First of all, the RNA data are characterized by:

- Low number of counts associated with large portion of genes;
- Lack of upper limit for expression, that leads to long right tail;
- Large dynamic range.

These features, together with the discrete nature of read counts make the data never normally distributed. Also the Poisson distribution, that can look good to approximate this kind of data, does not satisfy some features of trascriptomics, as for example the presence of heteroschedasticity and the difference between mean and variance values.
Figure 3.1 shows the overdispersion observed in the transcriptomics dataset T.

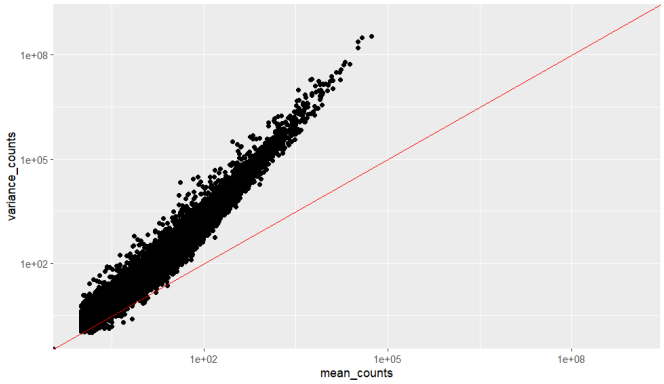

Figure 3.1. Plot of Mean - Variance for transcriptomics data.

The most used method to analyze mRNA data is implemented in the "DESeq" package and consists of a robust approach based on negative binomial generalized linear models (GLMs) with empirical Bayes shrinkage for dispersion estimation, specifically designed to handle the characteristics of RNA sequencing count data.

The main steps of the analysis can be summarized as:

1. Modeling the raw counts, using normalization size factors s_j;
2. Estimating gene-wise dispersion and then shrinking them using the empirical Bayes technique;
3. Fitting the Negative Binomial GLM for each gene, shrink the log fold changes and perform hypothesis testing using Wald test or Likelihood ratio test;
4. Applying the rlog transformation (see Sec. 2.2.2) to visualize and cluster the samples.

These steps are summarized in the following map (Fig. 3.2).
The raw count data T_{ij} for gene i in sample j shows overdispersion, therefore using a negative binomial distribution we can express this phenomenon:

$$T_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i), \tag{3.2}$$

where μ_{ij} is the expected count, computed as the fitted values from the GLM q_{ij} multiplied by a size factor s_j: $\mu_{ij} = s_j \cdot q_{ij}$, and α_i is the gene-specific dispersion parameter [15].
The negative binomial distribution accounts for overdispersion through the relationship:

$$\mathrm{Var}(T_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2 \tag{3.3}$$

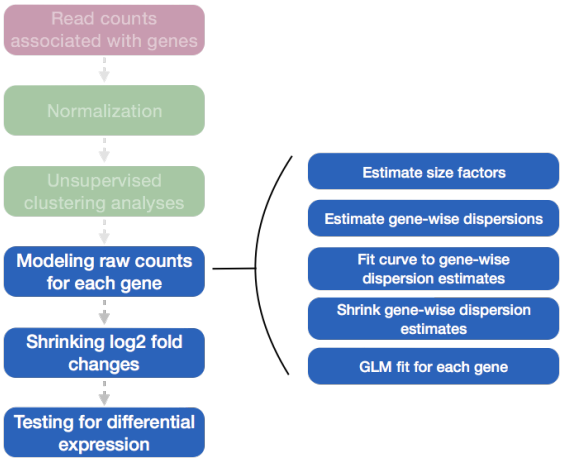

Figure 3.2. DESeq2 workflow map.

This formulation captures the relationship that was seen empirically in transcriptomics data, where variance typically exceeds the mean due to biological and technical variability.

For a comparison between only two conditions, the expected count is modeled through a log-link GLM. The generalized linear model for each gene i can be written as:

$$\log(\mu_{ij}) = \beta_{i0} + \beta_{i1} x_j, \tag{3.4}$$

where x_j is the condition indicator for sample j, β_{i0} the intercept and β_{i1} represents the $\log_2$ fold change estimate between conditions.

Initial gene-wise dispersion estimates $\hat{\alpha}_i$ are obtained through maximum likelihood estimation. However, these estimates are unreliable for genes with low counts or few replicates, leading to either false positives (underestimated dispersion) or reduced power (overestimated dispersion).

To cope with this issue, the approach uses a complex shrinkage approach to improve the dispersion estimation. This empirical Bayes shrinkage is performed by applying the following formula:

$$\alpha_i^{\text{shrunk}} = \frac{\alpha_i^{\text{prior}} \cdot w_i^{\text{prior}} + \hat{\alpha}_i \cdot w_i^{\text{obs}}}{w_i^{\text{prior}} + w_i^{\text{obs}}}, \tag{3.5}$$

where:

- $\hat{\alpha}_i$ is the raw gene-wise estimate from maximum likelihood,
- α_i^{prior} is the value predicted by the fitted dispersion curve, that follows Eq. 3.6;
- w_i^{prior} and w_i^{obs} are precision weights, respectively assigned to the prior and observed from the data.

Specifically, the prior dispersion α_i^{prior} follows a parametric relationship, meaning that is estimated by fitting a parametric curve to the scatterplot of $\hat{\alpha}_i$ versus the mean expression μ_i. This curve typically follows the form:

$$\alpha_i^{\text{prior}} = \frac{a_1}{\mu_i} + a_0, \tag{3.6}$$

where a_0 and a_1 are fitted coefficients that capture the global mean-dispersion trend across all genes.

In this way, it captures the empirical observation that dispersion decreases with increasing mean expression, with an asymptotic minimum value given by the parametric curve of α_i^{prior} [15].Genes with large counts will rely more on $\hat{\alpha}_i$, while low-information genes are pulled towards the prior curve. This balances variance and bias in dispersion estimation.

For testing differential expression, the hypothesis is based on the concept of Log Fold Change (LFC), that represents the logarithmic ratio of expression levels between conditions. Given a gene i, the Log fold change between two group conditions is:

$$\text{LFC}_i = \log_2\left(\frac{\mu_{i,\text{group 1}}}{\mu_{i,\text{group 2}}}\right) = \log_2(\mu_{i,\text{group 1}}) - \log_2(\mu_{i,\text{group 2}}), \tag{3.7}$$

where $\mu_{i,\text{group 1}}$ and $\mu_{i,\text{group 2}}$ are the mean counts for gene i in the two groups.
After defining the LFC, the hypothesis is expressed as:

$$\begin{aligned} &H_0 : LFC = 0 \quad \Longrightarrow \text{no differential expression across the groups} \\ &H_1 : LFC \neq 0. \end{aligned}$$

The approach utilizes as test:

- Wald Test, applied when the data has two group conditions and is based on the asymptotic normality of maximum likelihood estimators. The Wald test statistic is computed as:

$$W_i = \frac{\hat{\beta}_{i1}}{\text{SE}(\hat{\beta}_{i1})}, \tag{3.8}$$

  where $\hat{\beta}_{i1}$ is the estimated LFC and $\text{SE}(\hat{\beta}_{i1})$ its standard error, derived from the covariance matrix of the maximum likelihood estimates.
  Then, the statistic is compared to a normal distribution computing the corresponding p-value. Under H_0, W_i follows asymptotically a standard normal distribution:

$$W_i \sim \mathcal{N}(0,1). \tag{3.9}$$

  This test allows to assess whether the observed log fold change is significantly different from zero, accounting for the estimated variability in the model.

- Likelihood Ratio Test, when the data has more than two groups, that is a comparison between the fit of two different models [25]. The test is expressed as:

$$LR = -2\log\left(\frac{L_{model1}}{L_{model2}}\right)$$

To address the inflation of log fold change estimates for low-count genes, DESeq2 implements adaptive shrinkage:

$$\beta_i^{\text{shrunk}} = \frac{\hat{\beta}_i}{\hat{\beta}_i^2/s_i^2 + 1/\tau^2} \tag{3.10}$$

where τ^2 is estimated adaptively based on the distribution of effect sizes across genes.

3.2.1 Multiple test correction

Given the high-dimensional nature of omics data, multiple testing correction is essential for both transcriptomics and proteomics in this study. Since each p-value, associated to each gene, is the result of a single test, the more genes we test, the more we inflate the False Positive Rate.
The Benjamini-Hochberg procedure controls the False Positive Rate at level α.
For m hypothesis tests with p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$:

1. Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$
2. Reject hypotheses $H_{(1)}, \ldots, H_{(k)}$

The adjusted p-values are computed as:

$$p_{\text{adj},(i)} = \min\left(1, \min_{j \geq i} \frac{m \cdot p_{(j)}}{j}\right) \tag{3.11}$$

3.2.2 Analysis results

After performing the analysis with DeSeq, we obtain the Log fold change, the p-values and the p-adjusted with the multiple test correction for each gene.
Usually, for further investigation, two sequential filters are applied to filter out the genes that do not show difference in the expressions across groups:

1. Significance filter: $p_{\text{adj}} \leq \alpha$, where α can be regulated based on the experiment and the genes involved. It is commonly set equal to 0.05 or 0.1.
2. Effect size filter: $|\log_2(\text{FC})| \geq 1$, that ensures that identified genes exhibit:

$$\frac{\mu_{\text{treatment}}}{\mu_{\text{control}}} \geq 2 \quad \text{or} \quad \frac{\mu_{\text{treatment}}}{\mu_{\text{control}}} \leq 0.5 \tag{3.12}$$

   This dual filtering approach balances statistical significance with biological relevance. In particular, this filter corresponds to ask to the means within groups to be "sufficiently different", by explicitly asking to be one the double of the other [25].

Among the strengths of this technique, the choice of the Negative binomial distribution, that handles well overdispersion, and the use of shrinkage for gene-wise dispersion, that prevents unreliable estimates from low-count genes, stand out for their power to manage trascriptomics data. In addition, the GLM framework and the multiple test corrections

accommodate complex designs with multiple factors, ensuring a higher performance and power.

However, the biggest issues show up for small sample sizes and low count genes. In the first case, the Wald test performance may be low when $n < 3$ per group, while in the second case the power of the method may be limited due to the conservative approach, that may miss true positives in low expressed genes.

To summarize, the statistical assumptions and validation are:

- Negative binomial distribution, for the count data;
- Independence of samples;
- Homogeneity of dispersion parameters, that need to be consistent within gene across conditions;
- Logarithm of expected counts is linear in model parameters.

The approach addresses some relevant challenges in the mRNA data analysis, such as the discrete count values, the quadratic mean-variance relationship and the gene-wise testing that requires extra care to be handled with the multiple test correction.
The integration of robust statistical methodology and specific considerations for the genomic data characteristics makes this approach particularly suited for differential expression studies.

3.3 Multi-omics integration: MOFA

Multi-Omics Factor Analysis (MOFA) represents an unsupervised statistical method designed for the integrative analysis of multi-omics datasets. MOFA can be viewed as a versatile but also rigorous generalization of the principal component analysis (PCA) concept, in order to use it for multi-omics data and their factorial integration.
The framework faces the challenges of integrating two layers of omics data together: extracting biologically meaningful signals from the datasets while accounting for the different properties and noise characteristics of each omics type. Given several data matrices with measurements of multiple omics data types on the same or on overlapping sets of samples, MOFA infers an interpretable low-dimensional representation in terms of a few latent factors that aims to capture relevant signals in the input data.
Practically, MOFA disentangles the sources of variation in the data, identifying the factors that are shared across multiple data modalities from the factors that drive variability in a single data modality [3].

Mathematically, MOFA starts from $M = 2$ data matrices (one transcriptomics matrix $\mathbf{T}$ and one proteomics matrix $\mathbf{P}$) of dimension $N \times G_T$ and $N \times G_P$ respectively. In respect to the notation introduced in the Section 2.1, MOFA needs the two matrices transpose. As mentioned before, N is the number of samples, while G_T and G_P the number of genes for

the transcriptomics and proteomics data. MOFA implements a multi-view probabilistic factor model where each omics dataset is modeled as:

$$\mathbf{T} = \mathbf{Z} \cdot \mathbf{W}^{(1)T} + \boldsymbol{\epsilon}^{(1)} \tag{3.13}$$

$$\mathbf{P} = \mathbf{Z} \cdot \mathbf{W}^{(2)T} + \boldsymbol{\epsilon}^{(2)} \tag{3.14}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times K}$ is the shared latent factor matrix with $k = 1, ..., K$ factors, $\mathbf{W}^{(1)} \in \mathbb{R}^{G_T \times K}$ and $\mathbf{W}^{(2)} \in \mathbb{R}^{G_P \times K}$ contain the view-specific factor loadings and $\boldsymbol{\epsilon}^{(m)}$ represents the noise terms for each omics layer m ($m = 1$ for transcriptomics and $m = 2$ for proteomics). In addition, we denote as $g = 1, ..., G_T$ for transcriptomics and $g = 1, ..., G_P$ for proteomics the genes in the matrices.
In Figure 3.3, a graphical overview of the MOFA methodology is presented.

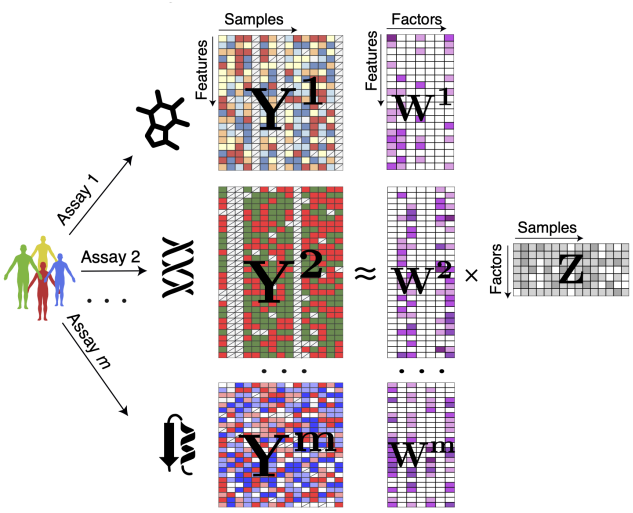

Figure 3.3. Graphical overview of the MOFA methodology.

The model uses a Bayesian approach, where we place prior distributions on all the unknown variables of the model, applying the following prior functions.

- For the factors: $\mathbf{Z} \sim \mathcal{N}(0, 1)$
- For the loadings, the weights are parameterized as a product of a Bernoulli distributed random variable and a normally distributed random variable: $\mathbf{W} = S \cdot \hat{\mathbf{W}}$, where $\hat{w}_{gk}^{(m)} \sim \mathcal{N}(0, 1/\alpha_k^{(m)})$ and $s_{gk}^m \sim Ber(\theta_k^m)$
- For the parameter $\alpha_k^{(m)}$, which controls the strength of factor k in view m, a uninformative conjugate prior is defined as, $\alpha_k^{(m)} \sim \text{Gamma}(a_0, b_0)$
- For the parameter θ_k^m, which determines the feature-wise sparsity level of factor k in view m, a uninformative conjugate prior is applied as well, as $\theta_k^m \sim Beta(a_0^\theta, b_0^\theta)$ [4].

This hierarchical structure the model to automatically determine factor relevance and induce sparsity in the loading matrices.

One of the main issues that need to be faced when dealing with multiomics integration is the presence of heterogeneous data across the layers. In the case of trascriptomics and proteomics samples, the first type contains discrete and integer data, while the second is composed of continuous data. MOFA accommodates different likelihood functions for each data type.
Taking as example the transcriptomics matrix $\mathbf{T}$ for simplicity (the same holds true for the matrix $\mathbf{P}$), we have:

- For continuous data, a Gaussian likelihood is used:

  $$T_{ng} \sim \mathcal{N}(0, 1/\tau_g),$$

  where τ_g is defined as the precision parameter.

- For count data, a Poisson likelihood is more appropriate instead, and it is expressed as

  $$T_{ng} \sim \text{Poisson}(\lambda(Z_n, w_g k^T)),$$

  where $\mu_{ng} = \sum_{k=1}^{K} z_{nk} w_{gk}$, where λ denotes the function $\lambda(x) = \log(1 + e^x)$.

The first step, after a model has been trained, is to disentangle the variation explained by each factor in each view. The proportion of variance explained by factor k in view m is computed as:

$$\text{Transcriptomics matrix:} \quad R^2_{k,1} = \frac{\text{Var}(\mathbf{Z}_k \mathbf{W}_k^{(1)T})}{\text{Var}(\mathbf{T})} \tag{3.15}$$

$$\text{Proteomics matrix:} \quad R^2_{k,2} = \frac{\text{Var}(\mathbf{Z}_k \mathbf{W}_k^{(2)T})}{\text{Var}(\mathbf{P})} \tag{3.16}$$

where $\mathbf{Z}_k$ represents the k-th factor and $\mathbf{W}_k^{(m)}$ the corresponding loadings for view m. The total variance explained across all factors for view m is: $R^2_{total,m} = \sum_{k=1}^{K} R^2_{k,m}$ [4].

The general concept on which the MOFA approach is constructed is to identify the factors that are manifested in multiple modalities, by revealing the shared axes of variation between the different omics layers. To give a biological interpretation to the prevalent factors, we identify the features with high absolute loadings, investigate their biological pathways through pathway enrichment and highlight sample associations to correlate factor values with sample metadata.

The main advantages of MOFA method include its unsupervised nature, which identifies shared patterns without requiring prior knowledge of sample relationships, and its flexibility in handling different data types through appropriate likelihood functions. The method

is able to manage the presence of missing values through its probabilistic formulation and works well with large datasets. The variance decomposition and loading analysis allows to achieve biological interpretation thanks to the factor interpretability and the analysis of the genes and proteins that retain the majority of variance.
However, MOFA has several limitations. The method assumes linear relationships between factors and observed variables and assumes that factors follow a normal distributions. Moreover, the variational optimization may converge to local rather than global minimum, while the performance may be sensitive to the choices of the hyperparameters. Lastly, the iterative nature of the optimization can be computationally intensive for very large datasets.

Summing up the key statistical assumptions:

- Observed variables are linear combinations of latent factors;
- Latent factors are assumed to be uncorrelated;
- Observations are conditionally independent given the factors.

The MOFA method tries to address the critical challenges in multi-omics integration through the use of dimensionality reduction of the high-dimensional multi-omics, different likelihood and techniques to handle heterogeneous data and Bayesian inference to maintain the statistical robustness.
In the next section, the iClusterPlus method for integrating multiomics data is presented as an alternative to matrix factorization approaches.

3.4 Multi-omics integration: IClusterPlus

iClusterPlus (extension of ICluster method) represents a joint latent variable model designed for integrative clustering of multiomics data, measured on the same set of samples. The method is based on a Bayesian approach, applied to the latent variable model. This model is composed of latent variables that are capable of spanning a low dimensional subspace without losing too much information: the subspace can still capture the general structure of multi-omics data and thus can be used for clustering the samples and genes. An ideal integrative clustering approach would allow joint inference from the multiomics data and generate a single integrated cluster through simultaneously capturing patterns of genetic alterations that are:

- consistent across multiple data types;
- specific to individual data types;
- weak to individual data type but consistent across datasets, that would emerge only as a result of combining levels of evidence [42].

Therefore, the goal of the ICluster approach is to develop such an integrative framework, by facing the two main challenges of this structure: first, to capture both concordant and

unique alterations across data types, second, to highlight covariance between data types but also the variance and covariance within data types.
While traditional dimension reduction techniques such as Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) work well for single data types, when it comes to dimension reduction of multiple correlated datasets they fail. Instead, iCluster addresses this limitation connecting the PCA approach with latent variable models, where the principal components can be computed using maximum-likelihood estimation under a Gaussian latent variable framework [29].
The mathematical implementation of iCluster follows these key steps:

- Joint Latent Variable Modeling: samples are modeled as unobserved latent variables, estimated in a substantially lower dimensional space.
- Distributions: The model assumes a Gaussian latent variable structure where residual variance is captured by an additional error term. With the extension of IClusterPlus, it is allowed for omics types to arise from other distributions than the Gaussian, such as multinomial, multivariate Gaussian and Poisson.
- Bayesian inference algorithm: Parameter optimization is performed through the Expectation-Maximization algorithm (EM) for ICluster, maximizing a penalized log-likelihood function and, through Markov Chain Monte Carlo (MCMC) for IClusterPlus, creating a sequence of simulations that converges to a stationary distribution.
- Lasso Regularization: A sparse solution is obtained with the Lasso regularization (L1 penalty), that shrinks to zero the coefficients corresponding to non-informative features. A sparse result is preferred since the variance is reduced and the clustering performance improved.
- Statistical inference: at this point, the model can be simultaneous inferred on the different omics datasets [42].

Mathematically, the iCluster model assumes that M omics data matrices $\mathbf{X}^{(m)}$ for $m = 1, \ldots, M$ are related to a set of K latent variables $\mathbf{Z} = (z_1, \ldots, z_K)^M$, each of dimension N. Each data matrix $\mathbf{X}^{(m)}$ has dimensions $G_m \times N$, where G_m represents the number of features in the m-th omics layer and N is the number of samples. The decomposition of data matrix $\mathbf{X}^{(m)}$ into the product of omics-specific weight matrix $\mathbf{W}^{(m)}$ and the shared factor matrix can be written as:

$$\mathbf{X}^{(m)} = \mathbf{W}^{(m)}\mathbf{Z} + \mathbf{E}^{(m)}, \tag{3.17}$$

where $\mathbf{W}^{(m)}$ is the $G_m \times K$ weight matrix for the m-th data type, $\mathbf{Z}$ is the $K \times N$ matrix of latent variables common across all data types and $\mathbf{E}^{(m)}$ represents the random error matrix of dimension $G_m \times N$ [8].
The latent variables $\mathbf{Z}$ represent the underlying biological processes that produce coordinated patterns of variation across multiple omics layers. The weight matrices $\mathbf{W}^{(m)}$ represent instead the contribution of each latent factor to the observed features in each omics dataset.

The method can integrate four different data types including continuous, count, binary and multi-categorical data. Considering M omics datasets to be merged, we denote x_{ijt} an omics variable for the jth ($j = 1, 2, \ldots, G_m$) omics feature of the ith ($i = 1, 2, \ldots, N$) sample in the mth ($m = 1, 2, \ldots, M$) dataset.
For the distributional assumptions, given the sample i, the latent variables corresponding are modeled as multivariate normal:

$$\mathbf{z}_i \sim \mathcal{N}_K(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z). \tag{3.18}$$

The error terms are assumed to follow:

$$\mathbf{e}_{ij}^{(m)} \sim \mathcal{N}(0, \sigma_{jt}^2), \tag{3.19}$$

where $\mathbf{e}_{ij}^{(m)}$ is the error for feature j in data type m for sample i.
When x_{ijt} is a continuous variable, x_{ijt} and z_i are related through a standard linear regression:

$$x_{ijt} = \alpha_{jt} + \beta_{jt} z_i + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim \mathcal{N}(0, \sigma_{jt}^2),$$

where α_{jt} is the intercept and $\beta_{jt} = (\beta_{1jt}, \ldots, \beta_{kjt})$ are the slope coefficients. When the data is not continuous, the relationship between x_{ijt} and z_i changes according to the corresponding regression (for example, with count data the regression is Poisson).

The model incorporates regularization to ensure sparsity in the weight matrices, using the Lasso penalty function, with the following form:

$$P(\mathbf{W}^{(m)}) = \lambda_m \sum_{j=1}^{G_m} \sum_{k=1}^{K} |w_{jk}^{(m)}| \tag{3.20}$$

where λ_m is the regularization parameter for data type m and $w_{jk}^{(m)}$ represents the (j, k)-th element of the weight matrix $\mathbf{W}^{(m)}$ [8].

After the parameter estimation phase, performed with the MCMC algorithm, the clustering assignment for each sample is determined based on the posterior mean of the latent variables. The sample clusters are determined by using K-means clustering technique, which separates the N samples into $k+1$ clusters. In order to achieve an optimal solution for the iClusterPlus model, a small number of k are usually tested, starting from $k = 1$ and reaching usually $k = 3$ or $k = 4$.

To summarize the method's idea, in Figure 3.4 the general workflow of the iCluster method is illustrated.
The key advantage of the IClusterPlus method is its combination of different tools that enables a strong and good handle of multiomics data and their integration. The dimensionality reduction facilitates interpretation, while the Bayesian approach ensures computational efficiency and biological and clinical interpretability is enhanced thanks to the sample clustering the identification of biologically meaningful genes.

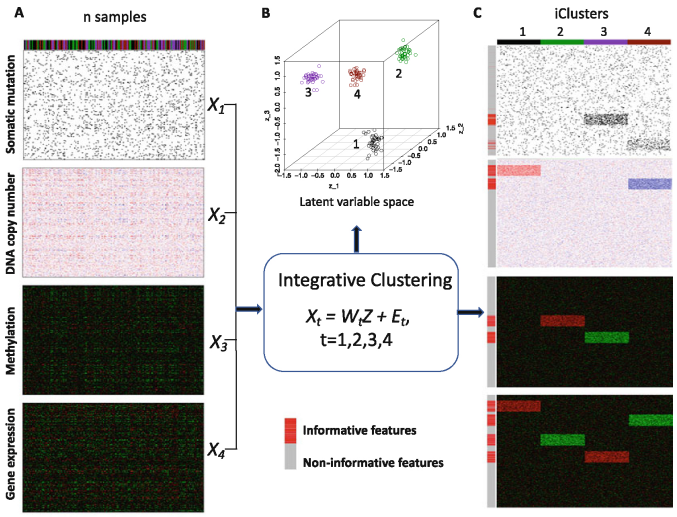

Figure 3.4. Graphical overview of the iCluster methodology.

On the other hand, the method presents several limitations that must be considered. The method shows sensitivity to scaling, as different omics data types may range in different scales and distributions, but also to missing values in any omics layer. For these reasons, a good preprocessing that imputes the missing data and aligns the value intervals of the different omics is essential to be implemented. Regularization parameter selection can be challenging, since the choice of λ_t values has a strong impact on the results. Finally, the complexity grows as the number of latent factors increases, making biological interpretation more difficult for higher-dimensional latent spaces [42].

In the next section, the last integration approach performed in this study is described.

3.5 Multi-omics integration: SNF

Similarity Network Fusion (SNF) represents a network-based computational framework designed for integrating multiomics data by constructing and merging similarity networks. SNF solves the multiomics integration problem by building networks of samples for each data type (transcriptomics and proteomics in this case) and then fusing these into one network that represents the full totality of data. The method addresses two challenges simultaneously: leveraging complementarity across different omics layers while preserving the unique information of each data.
The construction of patient similarity networks starts from the nodes, representing the samples analyzed, and then passes to the creation of the edges, which encode similarities between patient profiles. A patient similarity network is thus represented as a graph $G = (V, E)$. The vertices V correspond to the patients $\{x_1, x_2, \ldots, x_N\}$ and the edges E are weighted by how similar the patients are. Edge weights are represented by an $N \times N$

similarity matrix $\mathbf{W}$ with $W(i,j)$ indicating the similarity between patients x_i and x_j.The key step of SNF is to iteratively and simultaneously update the global patient similarity matrix of each layer using a local K-nearest neighbors (KNN) method, combined with the global similarity matrices of the other layers [30]. This iterative fusion process enables the method to capture both the information within each omics layer and the information across all data.

Mathematically, the SNF algorithm operates on $M = 2$ data matrices $\mathbf{T} \in \mathbb{R}^{N \times G_T}$ and $\mathbf{P} \in \mathbb{R}^{N \times G_P}$, where N represents the number of samples, G_T and G_P the number of genes respectively.
We denote $\rho(x_i, x_j)$ as the distance between patients x_i and x_j, that corresponds to the Euclidean metric when the data is continuous. The similarity matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ is then determined with a scaled exponential similarity kernel as:

$$W(i,j) = \exp\left(-\frac{\rho(x_i, x_j)^2}{\mu\, \epsilon_{i,j}}\right), \tag{3.21}$$

where μ is a hyperparameter that can be empirically set and $\epsilon_{i,j}$ is used to eliminate the scaling problem [47].
The scaling factor $\epsilon_{i,j}$ is defined as:

$$\epsilon_{i,j} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j))}{2}, \tag{3.22}$$

where $\text{mean}(\rho(x_i, N_i))$ represents the average distance between sample x_i and its K nearest neighbors N_i. This local scaling ensures that the similarity measure adapts to the local density of data points in the feature space.

At this point, the following step is to merge together the information from the multiple similarity matrices of the omics layers. To compute the fused matrix, we define a full and sparse kernel on the vertex set V. The full kernel is a normalized weight matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ defined as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix whose entries are $D(i,i) = \sum_j W(i,j)$, so that $\sum_j P(i,j) = 1$. However, this normalization may suffer from numerical instability since it involves self-similarities on the diagonal entries of $\mathbf{W}$. One way to perform a better normalization is as follows:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k \neq i} W(i,k)}, & \text{if } j \neq i \\ 1/2 & \text{otherwise} \end{cases} \tag{3.23}$$

This normalization will be free of the scale of self-similarity in the diagonal entries and $\sum_j P(i,j) = 1$ still holds.

The core SNF fusion process starts, updating each similarity matrix by incorporating information from the others. To do so, the algorithm of the K nearest neighbors (KNN) is applied to retain only the information and influence from the closest neighbors for each node (i.e., each patient). Given the graph of a omics layer G, let n_i represent the set of

neighbors of x_i (including x_i) in G. We use then the KNN technique to measure local affinity as:

$$S(i,j) = \begin{cases} 2 \cdot P(i,j), & \text{if } x_j \in N_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.24}$$

This operation sets the similarities between non-neighboring points to zero [47]. The sparse matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ preserves only the K strongest connections for each sample, enforcing the assumption that local similarities are more reliable than remote ones.
Note that $\mathbf{P}$ carries the full information about the similarity of each patient to all others, whereas $\mathbf{S}$ only encodes the similarity to the K most similar patients for each patient.

Considering the case with two omics layers, i.e. $m = 2$, we calculate the similarity matrices $\mathbf{W}^{(1)} \in \mathbb{R}^{N \times N}$ and $\mathbf{W}^{(2)} \in \mathbb{R}^{N \times N}$ as in Equation 3.21. Then the kernel matrices $\mathbf{S}^{(1)} \in \mathbb{R}^{N \times N}$ and $\mathbf{S}^{(2)} \in \mathbb{R}^{N \times N}$ are obtained as in Equation 3.24.

Let $\mathbf{P}_0^{(1)} \in \mathbb{R}^{N \times N}$ and $\mathbf{P}_0^{(2)} \in \mathbb{R}^{N \times N}$ represent the initial two status matrices at $t = 0$. The key step of SNF is to iteratively update the similarity matrix corresponding to each of the data types as follows:

$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \cdot \mathbf{P}_t^{(2)} \cdot \left(\mathbf{S}^{(1)}\right)^T, \tag{3.25}$$

$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \cdot \mathbf{P}_t^{(1)} \cdot \left(\mathbf{S}^{(2)}\right)^T, \tag{3.26}$$

where $\mathbf{P}_t^{(1)}$ refers to the first data type after t iterations and $\mathbf{P}_t^{(2)}$ to the second data type. The algorithm alternates between updating the status matrices and maintaining the affinity matrices until convergence. This procedure updates the matrices at each step, generating two parallel diffusion processes that fuse together progressively [40].

The convergence criterion is typically defined as:

$$\max\left(\|\mathbf{P}_{t+1}^{(1)} - \mathbf{P}_t^{(1)}\|_F, \|\mathbf{P}_{t+1}^{(2)} - \mathbf{P}_t^{(2)}\|_F\right) < \tau, \tag{3.27}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and τ is a predefined tolerance threshold (typically $\tau = 10^{-6}$).
After t steps, the final fused network is obtained by averaging the converged status matrices as:

$$\mathbf{P}_{fused} = \frac{1}{2}\left(\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)}\right), \tag{3.28}$$

where $\mathbf{P}_{fused} \in \mathbb{R}^{N \times N}$ represents the integrated similarity matrix containing information from both omics layers.

The following illustration (Fig. 3.5) gives a general overview on SNF methodology.

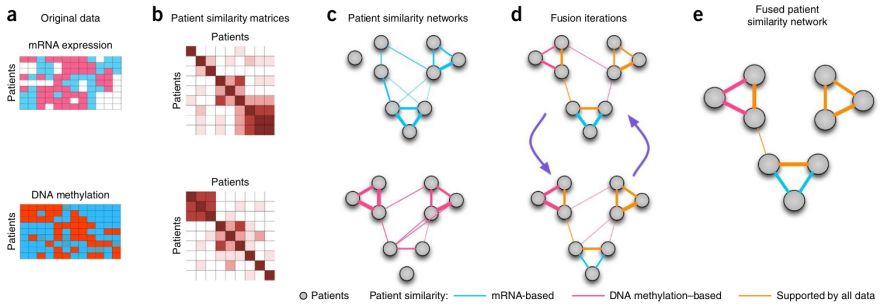

Figure 3.5. Graphical overview of the SNF methodology.

Sample clustering

Following fusion, clustering is usually performed to highlight affinities among samples, using spectral clustering on P_{fused}. In multivariate statistics, spectral clustering techniques make use of the eigenvalues of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions.
This algorithm decomposes the fused matrix using its eigenvalues. As first step, the normalized Laplacian matrix is constructed as:

$$\mathbf{L} = \mathbf{D}_{fused}^{-1/2} \mathbf{P}_{fused} \mathbf{D}_{fused}^{-1/2}, \tag{3.29}$$

where $\mathbf{D}_{fused}$ is the degree matrix with entries $D_{fused}(i,i) = \sum_j P_{fused}(i,j)$.

The second step involves computing the first k eigenvectors corresponding to the largest eigenvalues of $\mathbf{L}$, forming the matrix $\mathbf{U} \in \mathbb{R}^{N \times k}$, that has each eigenvector as column.
Each row of $\mathbf{U}$ represents a sample in the spectral embedding space. K-means clustering is then applied to the rows of $\mathbf{U}$ to identify k clusters of samples.

Feature importance and selection

To identify the most informative molecular signatures that contribute to the integrated sample structure, feature importance is assessed using the Normalized Mutual Information (NMI) metric. For each gene g in the original omics datasets, the NMI score is computed as:

$$\mathrm{NMI}(g, C) = \frac{2 \cdot I(g, C)}{H(g) + H(C)}, \tag{3.30}$$

where g represents the gene values across all samples, C represents the cluster assignments obtained from spectral clustering, $I(g, C)$ is the mutual information between feature g and cluster labels C, and $H(\cdot)$ denotes entropy.
The mutual information $I(g, C)$ quantifies the amount of information shared between gene g and cluster structure C:

$$I(g, C) = \sum_{g \in \mathcal{G}} \sum_{c \in \mathcal{C}} p(g, c) \log \frac{p(g, c)}{p(g)p(c)}, \tag{3.31}$$

where $\mathcal{G}$ and $\mathcal{C}$ represent the discrete sets of gene values and cluster labels, respectively, and $p(\cdot)$ denotes probability distributions estimated from the data.
The entropy terms are defined as:

$$H(f) = -\sum_{g \in \mathcal{G}} p(g) \log p(g), \quad H(C) = -\sum_{c \in \mathcal{C}} p(c) \log p(c). \tag{3.32}$$

For continuous features, discretization is performed using equal-frequency binning or k-means clustering to compute the NMI scores.
NMI ranges from 0 to 1, where values close to 0 indicate little relationship between the gene and cluster structure, while values close to 1 suggest high correlation. Features are ranked according to their NMI scores in descending order, in order to select the top-k most informative features (for example, $k = 20$ for the top 20 genes).
The feature selection process operates as follows:

1. For each gene g_i in the transcriptomics data matrix $\mathbf{T}$ (where $i = 1, \ldots, G_T$), compute $\text{NMI}(g_i, C)$.

2. For each feature g_i in the proteomics data matrix $\mathbf{P}$ (where $i = 1, \ldots, G_P$), compute $\text{NMI}(g_i, C)$.

3. Combine all NMI scores into a single ranking list and select the top-k genes with highest NMI values.

4. The selected genes are the ones most strongly associated with the clustering structure of the fused sample.

To summarize the procedure of this method, the main steps are:

1. Pairwise distance matrices are computed for each omics layer, using a metric appropriate to the nature of the data (continuous, discrete, or boolean).

2. Similarity matrices $\mathbf{W}$ are constructed from the distance matrices using a Gaussian kernel. The matrix captures local similarity relationships, with higher values indicating greater similarity between samples.

3. Normalized weight matrices $\mathbf{P}$ are obtained from the similarity ones by applying the normalization step.

4. The core SNF fusion process iteratively updates each affinity matrix by incorporating information from all other modalities until convergence.

5. The final fused similarity matrix $\mathbf{P}_{fused} \in \mathbb{R}^{N \times N}$ is obtained by averaging the converged status matrices.

6. Spectral clustering is applied to identify sample clusters, and feature importance is computed using NMI scores to select the most informative molecular signatures.

The choice of hyperparameters significantly impacts SNF performance. The number of nearest neighbors K controls the connectivity of the similarity networks, with small values leading to fragmented networks and large values causing loss of local structure. The scaling parameter μ in Equation 3.21 controls the emphasis on distance differences between samples, with higher values creating more aggressive contrast between similar and dissimilar samples.

SNF offers several advantages for multi-omics integration. Network-based representation naturally captures complex, non-linear relationships between samples that may not be apparent in individual omics analyses. In addition, the iterative fusion process enables a better communication between different omics layers, potentially revealing hidden relationships that emerge only through integration. The computational efficiency is maintained through the use of sparse matrices and convergent iterative algorithms [47].

However, SNF has several limitations that must be considered. Parameter sensitivity requires careful tuning of K and μ parameters, that potentially may need extensive cross-validation. At the same time, the choice of initialization parameters may interfere with the convergence, since the iterative algorithm can converge to local optima instead to the global one. Moreover, the computational scalability becomes problematic for very large datasets, due to the quadratic complexity of similarity matrices ($O(N^2)$ space complexity and $O(N^3)$ time complexity per iteration).
The method's strength lies in its ability to create a unified similarity structure, providing a more comprehensive view of sample relationships than single-omics approaches. The network-based nature of the approach captures complex and non-linear biological relationships, while keeping computational costs manageable for datasets of moderate size.

Chapter 4

Results

4.1 Transcriptomics

Boxplots

Boxplots of the raw data, both for 2D and 3D datasets, show a distribution strongly compressed near zero: the boxes are almost flattened at the lower part of the axis, indicating the presence of a large fraction of features with very low or null counts (Figure 4.1). After applying low-filtering, the boxplots show a more standard distribution, with visible medians and interquartile ranges and comparable trends across samples. For both 2D and 3D samples no major differences emerge between patients and controls, except for one control sample (labeled *C3*) in the 2D case, which shows a lower-shifted distribution both before and after filtering. The presence of a limited number of genes with very high counts is noticeable across all samples, as indicated by the upper outliers in the boxplots (Figure 4.2). These outliers correspond to highly expressed genes, which are typical in RNA-seq datasets and generally represent a small subset of genes with strong biological activity.

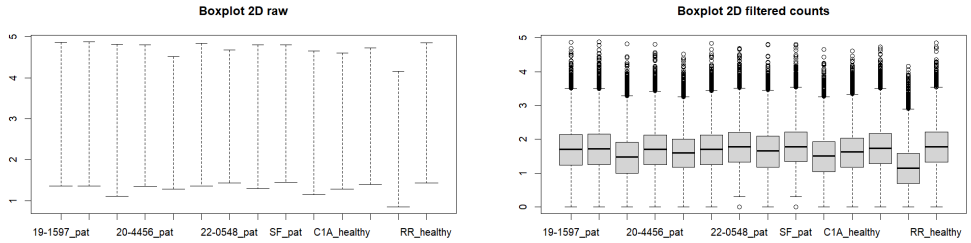

Figure 4.1. Boxplots of read counts per sample for 2D data.

We can notice that the spheroid data (3D) displays the same overall behavior as the bulk data (2D), maintaining similar behavior.

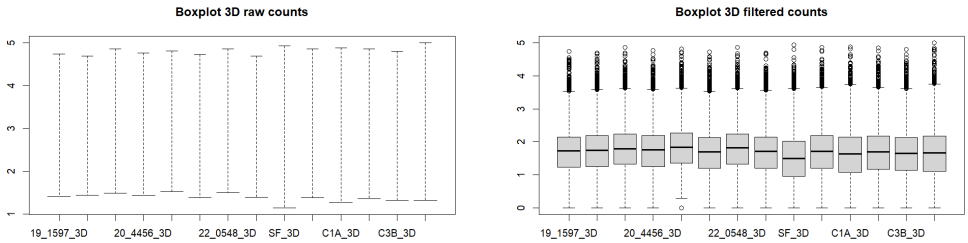

Figure 4.2. Boxplots of read counts per sample for 3D data.

Density plots

Looking at the density plots (Figure 4.3 and 4.4), the raw data exhibit a characteristic initial dip: a high density of values close to zero, followed by a decrease and then the typical right-skewed profile. Right tails (highly expressed genes) are consistent across samples and reflect what shown in the boxplots: few genes with high counts are present across all samples. After low-filtering, this initial bump almost disappears, confirming that most low-expressed features were removed. However, a residual anomaly remains noticeable in sample *C3* (Figure 4.3), which already displayed deviations in the boxplots. Overall, filtering reduces apparent sparsity and improves cross-sample comparability.

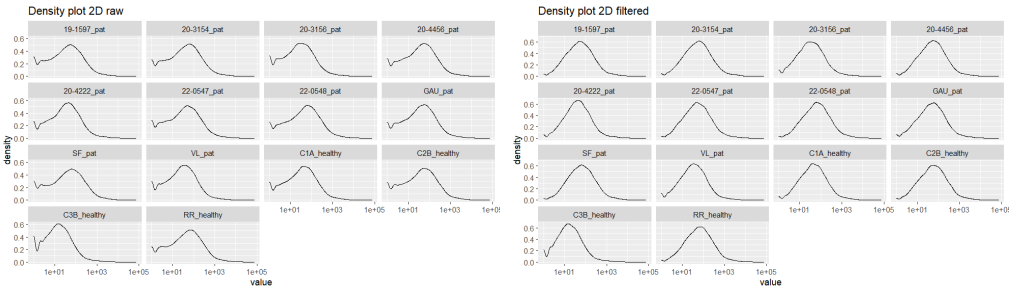

Figure 4.3. Density plots of read counts per sample for 2D data.

Heatmap

The heatmap highlights groups of genes with coherent expression patterns, showing blocks of over-expressed and under-expressed genes across subsets of samples.
Observing the hierarchical clustering of the 2D data (Figure 4.5), three control samples cluster distinctly apart from the patient group, while one control (*RR*, not *C3*) fails to group with the other controls. In the second half of the matrix, controls generally display

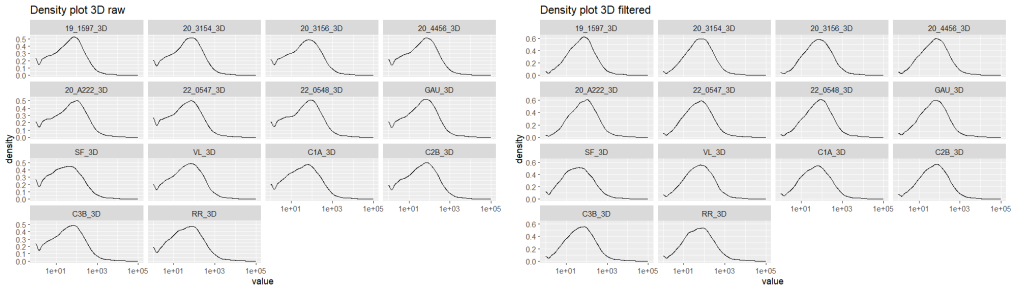

Figure 4.4. Density plots of read counts per sample for 3D data.

lower expression values (blue), while patients show higher expression levels (yellow to orange), highlighting a clear differential expression pattern. However, the separated control sample (*RR*) exhibits the opposite trend, showing a profile more similar to the patients ones than those of the other controls.
This indicates that while biological signals distinguish patients from controls, variability across samples remains. Such variability may reflect biological heterogeneity or technical factors, such as the presence of batch effect.

Similarly, the heatmap of the 3D data (Figure 4.6) also reveals blocks of coherently expressed genes, with distinct patterns emerging between patients and controls. As in the 2D condition, three controls cluster together, separating from most of the patients. However, they are not completely separated from all the patients, but on the contrary they are clustered with three patients. Once again the *RR* sample does not follow this trend and groups apart, confirming its divergent behavior across both culture systems.
Moreover, while patients generally show higher expression levels (yellow to orange) compared to controls (blue), the distinction appears less sharp than in the 2D data, suggesting that spheroid cultures (3D) may introduce additional variability or reduce the strength of the separation. This result highlights a possible outlier sample (RR) and a softer contrast between conditions compared to the bulk data (2D).
The presence of clusters of differentially expressed genes supports the potential for downstream differential expression analysis.

PCA

The Principal Component Analysis reveals patterns in the transcriptomic data structure, with some differences between 2D and 3D culture conditions. For the 2D data (Figure 4.7), the PCA plot of PC1 vs PC2 shows a moderate separation between patients (blue points) and controls (red points). The first two principal components capture the major sources of variation, with patients generally clustering toward the right side of PC1 and controls showing more scattered distribution. Notably, one control sample appears as an outlier in the upper region of the plot (*RR* sample). The PC3 vs PC4 plot reveals additional structure, with samples distributed across multiple quadrants, capturing high

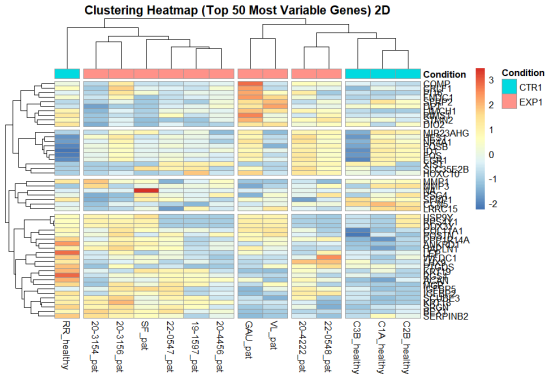

Figure 4.5. Heatmap of the top 50 most variable genes for 2D data.

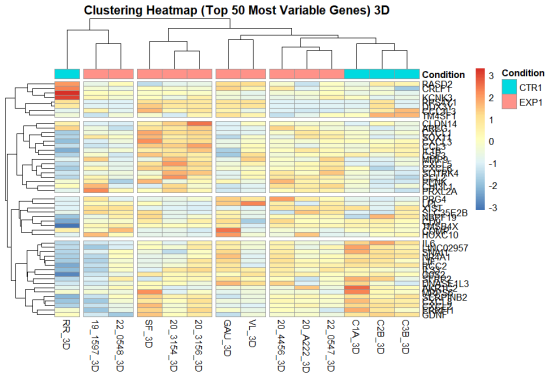

Figure 4.6. Heatmap of the top 50 most variable genes for 3D data.

variance of two of the controls with the rest of the samples (the one in the right side correspond to the sample *RR*).
The 3D spheroid data (Figure 4.8) exhibits markedly different clustering patterns. The PC1 vs PC2 plot shows a more evident separation between conditions, with most patient samples forming a distinct cluster in the upper portion of the plot and control samples positioned in the lower region. Still, the major proportion of variance is retained in the outlier control sample (*RR*), positioned in the down left side. The PC3 vs PC4 analysis shows more dispersed clustering, indicating that spheroid culture may introduce additional biological variability across all samples.
Overall, the first two components tend to better distinguish the two clusters between patients and controls, even if the highest portion of variance is retained in the outlier control (*RR*).

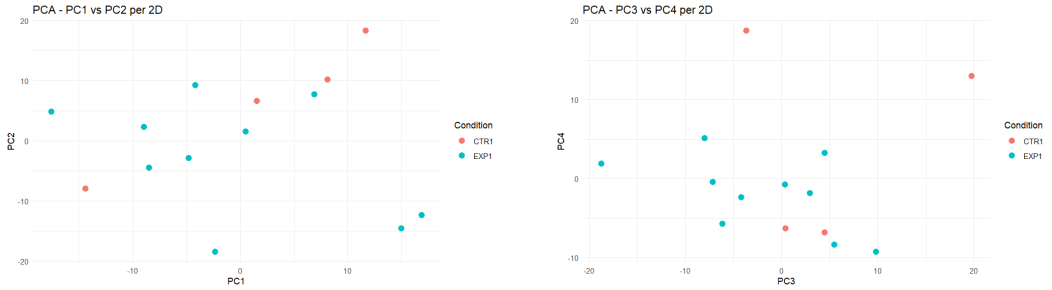

Figure 4.7. PCA Plots - PC1 vs PC2 and PC3 vs PC4 for 2D data.

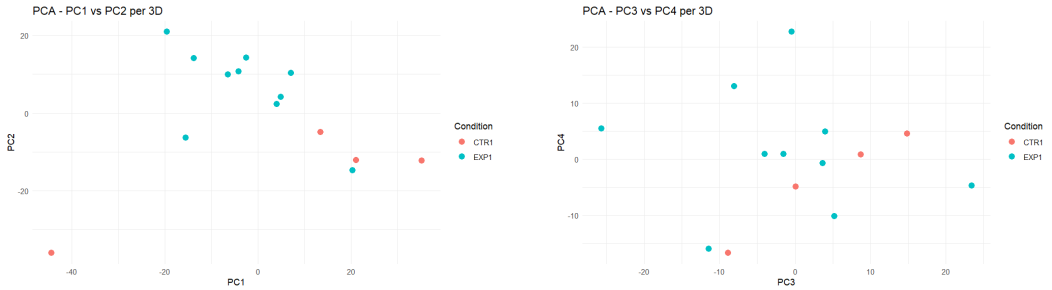

Figure 4.8. PCA Plots - PC1 vs PC2 and PC3 vs PC4 for 3D data.

Differential expression analysis

The differential expression analysis yielded substantially different results between culture conditions, as shown in Table 4.1. The number of significant genes identified in the spheroid culture is consistently higher. Among the significant genes with log-fold change values above the threshold, the 2D data displays a balanced proportion of up- and down-regulated genes, whereas in the 3D condition the majority of genes is up-regulated.

Condition	2D data	3D data
Total genes analyzed	12047	12578
Genes with p.adj < 0.1	26	296
Genes with p.adj < 0.1 and $\lvert\log_2 FC\rvert \geq 1$	23	218
$\rightarrow$ Down-regulated	13	72
$\rightarrow$ Up-regulated	10	146

Table 4.1. Summary of differential expression analysis results

The volcano plots provide complementary visualization of the differential expression results (Figure 4.9). The 2D volcano plot reveals a limited number of significantly differentially expressed genes, with only 23 genes meeting both statistical significance (p.adj < 0.1) and biological relevance ($\lvert \log_2 FC\rvert > 1$) criteria, marked with red colour. As

summed up in the Table 4.1, the distribution shows a relatively balanced pattern of up- and down-regulation. Most genes cluster around a fold change close to zero, indicating low transcriptional differences between patients and controls under 2D culture conditions. Oppositely, the 3D volcano plot demonstrates a different scenario with 218 genes meeting significance criteria. The plot reveals also an asymmetry, with the majority of genes being up-regulated, suggesting that Ehlers-Danlos syndrome mainly involves activation rather than repression of transcriptional genes.

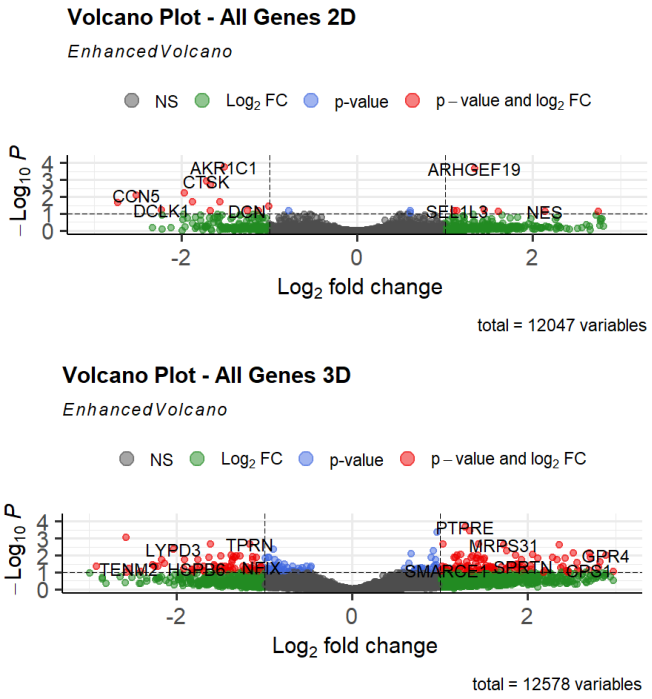

Figure 4.9. Volcano Plots for 2D and 3D data.

Considerations

The difference between 2D and 3D results suggests some considerations, as well as the presence of :

- Enhanced biological relevance: Spheroid culture may better represent the real cellular environment, allowing biological mechanism to manifest more clearly than in traditional culture.

- Amplified signal detection: The 3D environment appears to amplify transcriptional differences that remain below detection thresholds in 2D culture. This reflects in a higher number of significant genes.

- Disease insights: The predominance of up-regulated genes in 3D conditions (67 % of significant genes) suggests that Ehlers-Danlos syndrome may involve up-regulation of cellular pathways.
- The consistency of the outlier sample *RR* across multiple analytical approaches, from heatmaps to PCA plots, reinforces the robustness of these observations. This suggests the presence of biological or technical factors that have influenced the sample in consideration.

4.2 Proteomics

Missing value imputation

Before starting the statistical analysis, missing values in the proteomic datasets need to be imputed. A evaluation of different imputation methods was conducted using a masking approach, where known values were artificially hidden and then predicted to assess method performance. The parameter used to compared them is the RMSE. In Table 4.2, the result is shown for each method applied: Bayesian PCA, K-Nearest Neighbors, Probabilistic PCA and Singular Value Decomposition.

Method	bPCA	KNN	pPCA	SVD
RMSE	0.45	0.76	0.47	0.58

Table 4.2. Comparison of missing value imputation methods.

Bayesian PCA performs as best method, achieving the lowest RMSE of 0.45, followed closely by Probabilistic PCA (0.47). These error values highlight the effectiveness of matrix factorization approaches over distance-based methods (as KNN) for proteomic data imputation. For this reason, bPCA was selected for imputing missing values in the final dataset.

Boxplots and density plots

The boxplots reveal consistent protein abundance distributions across samples after log transformation (Figures 4.10 and 4.11). Both 2D and 3D datasets exhibit similar median values with comparable interquartile ranges, indicating successful normalization. The presence of outliers in both conditions reflects the heterogeneity of protein expression levels, with some proteins showing extremely high or low abundances. In addition, 3D data shows slightly more variability in the lower quartiles, suggesting that spheroid culture may introduce additional biological complexity in protein expressions.

The density plots (Figures 4.12 and 4.13) demonstrate approximately normal distributions for most samples after log transformation, validating the preprocessing approach. Both 2D and 3D datasets show similar distribution shapes with peaks centered around the same values. Nevertheless, the distributions tend to exhibit long right tails, indicating

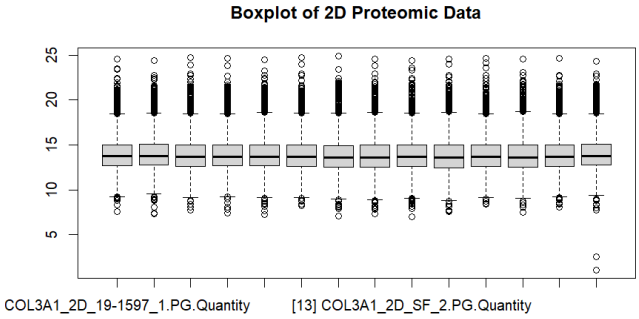

Figure 4.10. Boxplots of protein abundances for 2D data.

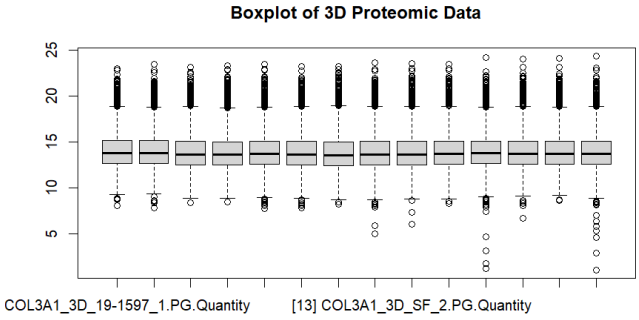

Figure 4.11. Boxplots of protein abundances for 3D data.

the presence of a small number of genes with relatively high values even after log transformation. In addition, the *RR* sample (last in the first row) appears slightly different from the others, displaying a pronounced left tail.

Statistical assumptions validation

Shapiro-Wilk tests were performed to evaluate the normality assumption required for parametric statistical analyses. In Table 4.2 the results are shown.

Data	Total Proteins	Non-normal Proteins	Percentage
2D	6024	610	11 %
3D	5666	638	12 %

Table 4.3. Shapiro-Wilk test results for normality check.

The values indicate that around 89 % of proteins in both datasets follows normal distributions, supporting the use of parametric statistical methods for the majority of the

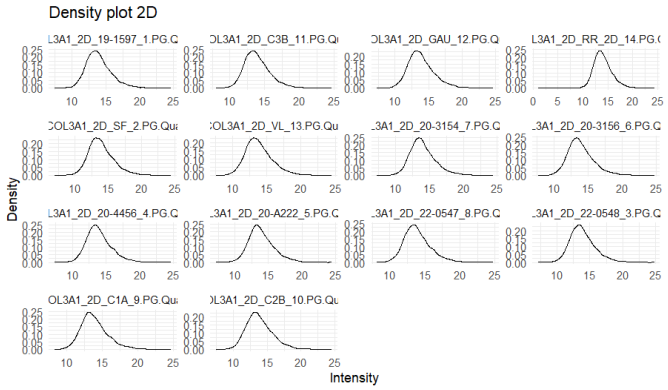

Figure 4.12. Density plots of protein abundances for 2D data.

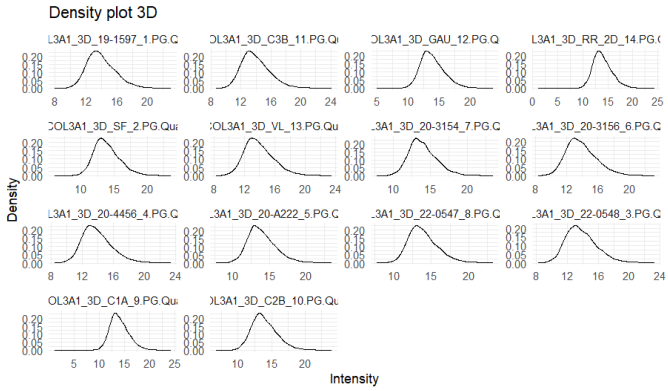

Figure 4.13. Density plots of protein abundances for 3D data.

dataset. The similar proportions of non-normal proteins between 2D (11 %) and 3D (12 %) conditions suggest that culture dimensionality does not affect distributional properties.

Levene's tests evaluate variance homogeneity across experimental groups. The results obtained through the tests are shown in Table 4.2.

Data	Total Proteins	Heteroscedastic Proteins	Percentage
2D	6024	285	5 %
3D	5666	385	7 %

Table 4.4. Levene's test results for homoscedasticity check.

The homoscedasticity assumption is satisfied for 95 % of proteins in 2D culture and 93% in 3D culture, indicating that ANOVA and limma approaches are appropriate for the

vast majority of proteins. The slightly higher proportion of heteroscedastic proteins in 3D conditions (7 % vs 5 %) may reflect increased biological variability introduced by the more complex spheroid environment.

PCA plots

The PCA analysis of proteomic data reveals distinct clustering patterns between culture conditions (Figures 4.14 and 4.15).

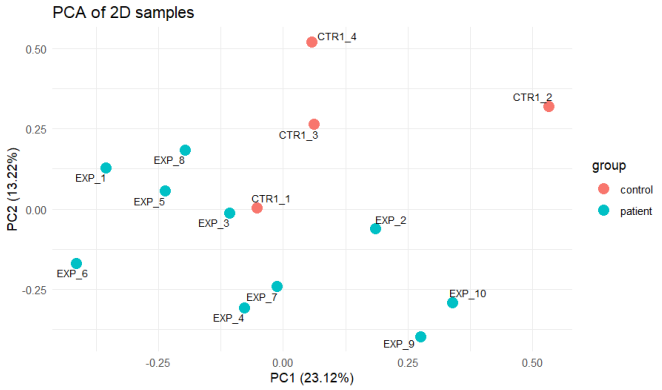

Figure 4.14. PCA plot of 2D proteomic samples.

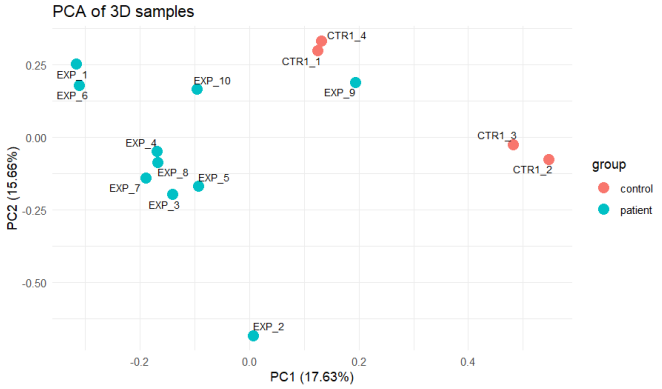

Figure 4.15. PCA plot of 3D proteomic samples.

In the 2D proteomic data, the first two components PC1 and PC2 capture 36.34 % of total variance, with moderate separation between patients and controls. The control samples show more clustered behavior, while patient samples display greater dispersion across both principal components.
Regarding the 3D data, PCA presents strong separation along the first component PC1, dividing quite good the patients on the left side and controls on the right part. The total

variance retained by the first two components PC1 and PC2 is similar to the 2D case, with 33.29 % of variance captured. The separation between patients and controls is more pronounced than in 2D culture, with controls forming a more distinct cluster in the right portion of the plot and patients predominantly occupying the left region.
Additionally, the *RR* sample, which behaves as an outlier in the transcriptomics analysis, does not display the same trend here.
The enhanced separation in 3D proteomic data is consistent with the transcriptomic results, reinforcing the hypothesis that spheroid culture amplifies disease biological mechanism at both transcriptional and proteomic levels.

Statistical results

The results of the statistical analyses performed with ANOVA, LIMMA and Wilcoxon tests are summarized respectively in Tables 4.2, 4.2 and 4.2. Overall, the number of significant genes identified is quite limited, particularly in the 2D dataset, where with most thresholds on the adjusted p-value obtains very few or no significant features. The adjusted p-value levels appear largely clustered, providing limited discriminatory power for the current analysis.

Dataset	p. adj < 0.1	$p_{adj} < 0.25$
2D	0	15
3D	193	901

Table 4.5. Number of significant genes identified with ANOVA.

Dataset	p. adj < 0.1	$p_{adj} < 0.25$
2D	0	90
3D	219	935

Table 4.6. Number of significant genes identified with LIMMA.

Dataset	p. adj < 0.1	$p_{adj} < 0.25$
2D	0	0
3D	0	287

Table 4.7. Number of significant genes identified with Wilcoxon test.

This lack of strong significance is likely due to the small and unbalanced sample size (10 patients vs. 4 controls), which strongly reduces the robustness of classical statistical tests such as the ones applied. Consequently, these results alone are insufficient to draw reliable biological conclusions.
What is interesting to notice instead is that the 3D dataset consistently shows a larger number of significant genes across all three statistical tests. This suggests that the 3D

culture may capture more detailed gene expression patterns and relevant biological mechanisms than the bulk data.
Given the limited significance observed in the single-omics proteome analyses, further investigation using multi-omics integration approaches is necessary to uncover more robust and biologically meaningful results.

4.3 Multi-omics integration

Following the preprocessing and individual analysis of transcriptomics and proteomics datasets, multi-omics integration was performed to identify coordinated molecular patterns across both data layers. The integrated dataset comprised 28 samples (14 samples $\times$ 2 culture conditions) with matched transcriptomics and proteomics measurements for each sample. Principal Component Analysis was applied on the merged datasets, as first step of the initial exploration. Through PCA, the combined dataset revealed distinct patterns across both omics layers and culture conditions.

The transcriptomic PCA (Figure 4.16) does not show clear clustering patterns by condition nor by group. The 3D samples (triangles) are characterized by more dispersed distribution compared to 2D samples (circles), suggesting that spheroid culture introduces additional transcriptional variability.

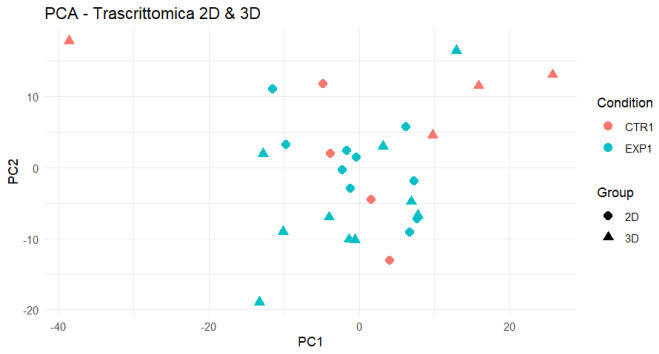

Figure 4.16. PCA of combined transcriptomic data (2D and 3D).

On the contrary, the proteomic PCA (Figure 4.17) reveals more pronounced separation patterns. Control samples cluster predominantly in the lower-left quadrant for both culture conditions, while patient samples distribute more broadly across the upper part of the plot space.

4.3.1 MOFA results

Multi-Omics Factor Analysis (MOFA) was employed as first integration approach to identify latent factors that capture coordinated patterns of variation across transcriptomic and proteomic datasets. The integration was performed for 2D and 3D culture conditions, each

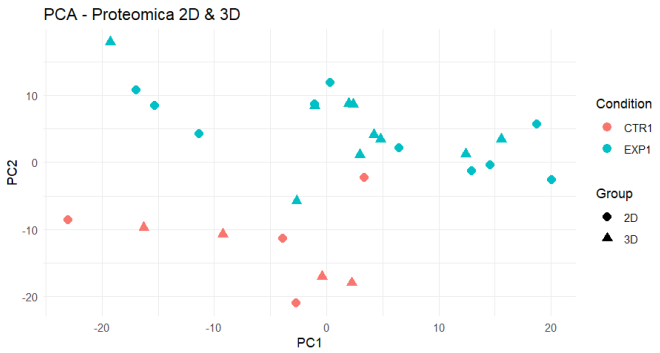

Figure 4.17. PCA of combined proteomic data (2D and 3D).

comprising 14 samples with transcriptomic data (11 418 features) and proteomic data (5 050 features) (Figure 4.18).

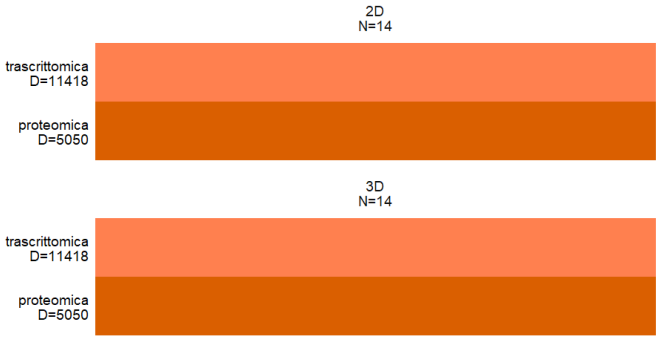

Figure 4.18. MOFA integration layout.

MOFA identified four latent factors for both culture conditions, with distinct patterns of variance explained across the two omics layers (Figure 4.19). In the 2D culture system, Factor 3 emerged as the most informative, explaining approximately 20% of the variance in the proteomic data, while the transcriptomic layer showed more distributed variance across factors. Notably, the 3D culture condition demonstrated superior information retention, particularly in the transcriptomic layer, where Factor 1 captured approximately 20% of the total variance. This observation aligns with the increased biological complexity of three-dimensional culture systems, which may preserve more natural cellular states and gene expressions compared to traditional cultures.

The percentage of variance retained in each factor by the omics layers can be visualized through Table 4.3.1, giving the precise values of what we noticed in the previous plot (Figure 4.19).

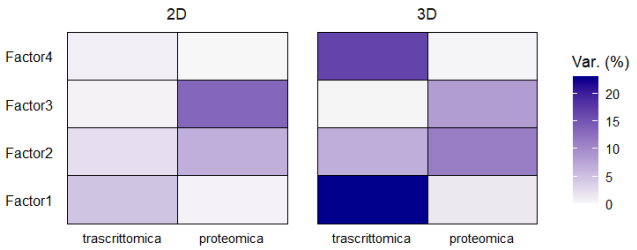

Figure 4.19. Total variance retained per view and per factor.

Factor	Transcriptomic 2D	Proteomic 2D	Transcriptomic 3D	Proteomic 3D
1	4.68 %	0.42 %	23.10 %	1.45 %
2	2.26 %	6.53 %	6.64 %	11.27 %
3	0.41 %	13.33 %	0.25 %	8.22 %
4	0.70 %	0.03 %	16.69 %	0.39 %

Table 4.8. Variance explained by each factor in transcriptomic and proteomic data.

This finding supports the hypothesis that 3D culture systems better explain in vivo conditions and preserve more biologically relevant molecular patterns.

The distribution of individual factors across experimental conditions revealed distinct clustering patterns between control and patient groups (Figure 4.20). Factor 1 showed the outlier sample (RR) to retain the majority of variance along its axis, with the other control samples and patients generally exhibiting similar factor values. Factor 2 demonstrated moderate discriminatory power, with all the controls showing higher values than patients. Instead, Factors 3 and 4 showed more overlapping distributions between conditions.

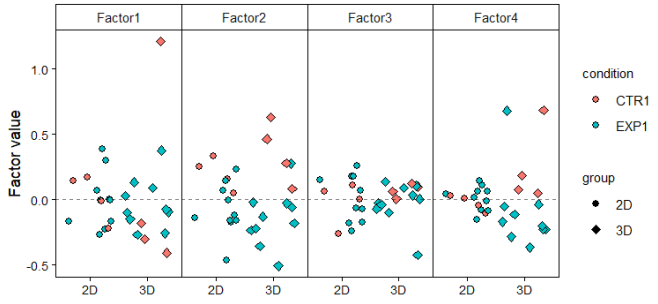

Figure 4.20. Single factors plot.

Going on with the analysis, we represent the pairwise scatter plot matrix of all four factors,

which revealed the multidimensional structure of the data and the relationships between different latent components (Figure 4.21). In most cases, the visualization demonstrated overlapping clustering patterns, even though in some plots the clusters are more clear, with patient and control samples forming distinct groups in the factor space. This is noticeable, for example, when combining Factor 2 and 1 or Factor 2 and 3, showing a stronger separation along the first axis and the diagonal. In general, Factor 2 provides the best differentiation among the groups, as shown also in the corresponding density plot, in which the patients and controls distribution are quite well separated.

Figure 4.21. Combined factors plot.

To understand the molecular basis of the identified latent factors, we examined the top 20 contributing genes for each omics layer across the most informative factors.
First of all, we noticed that transcriptomic data primarily contributed to Factor 1 and Factor 4, whereas proteomic features dominated Factors 2 and 3. Factor 1 was strongly influenced by transcriptomic features, particularly genes involved in diverse biological processes, suggesting that it captures metabolic and stress-response signatures. In contrast, Factor 4 appeared to be mainly associated with cellular maintenance and regulatory processes.
At the protein level, the contribution of proteomic features was generally weaker compared to transcriptomic data. Nevertheless, the concordance between transcriptomic and proteomic loadings for certain factors validates the biological relevance of the identified latent components and demonstrates successful integration of the two omics layers.

4.3.2 iClusterPlus results

The iClusterPlus method was applied with varying numbers of latent variables (k = 1, 2, and 3) to identify optimal clustering structures in the integrated multi-omics space.

The simplest clustering solution with k = 1 partitions the 28 samples (14 samples repeated with bulk and spheroid data) into two clusters (Figure 4.22).
The two-cluster solution shows highly unbalanced clustering, with Cluster 2 containing the majority of samples (22/28). This asymmetric distribution suggests that a smaller subset (Cluster 1) exhibits distinct molecular characteristics. However, this cluster does not contain only controls patients, but it is composed of different patients and only one

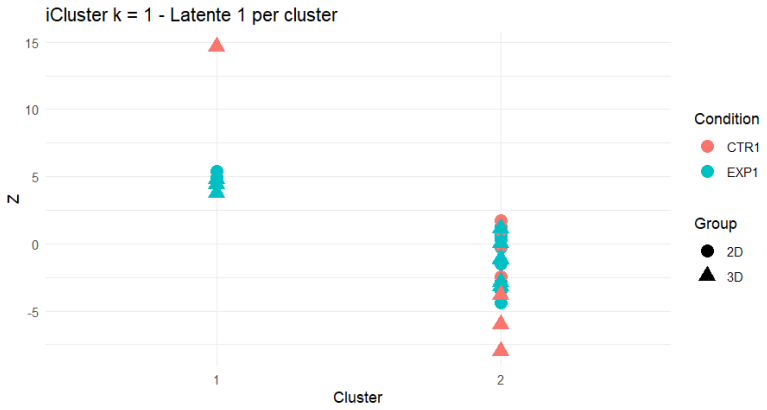

Figure 4.22. Clustering with k = 1.

Cluster	Sample Count	% of controls
1	6	16 %
2	22	32 %

Table 4.9. Clustering with k = 1 (two clusters).

control (specifically, *RR* sample). This clustering does not show relevant distinction, neither by group nor by condition. In the Table 4.3.2 the number of samples grouped in each cluster is reported, with the additional information about their condition.

Going on with the analysis, we increased the number of latent variable to k = 2, that yields to three clusters.
The three-cluster solution provides a somewhat more balanced partitioning compared to fewer clusters. As reported in Table 4.3.2, Cluster 2 contains 6 out of the 8 control samples, effectively grouping most controls together, but it also includes 5 patient samples, indicating that some mixing between conditions remains. In general, the division is not effective for well separating controls from patient samples.

Cluster	Sample Count	% of controls
1	13	7 %
2	11	55 %
3	4	25 %

Table 4.10. Clustering with k = 2 (three clusters).

The most complex solution with k = 3 generates four clusters, that can be visualized in three-dimensional latent space, as in Figure 4.24, or in multiple two-dimensional projections, as in Figure 4.25, showing the position of each sample in respect to two components

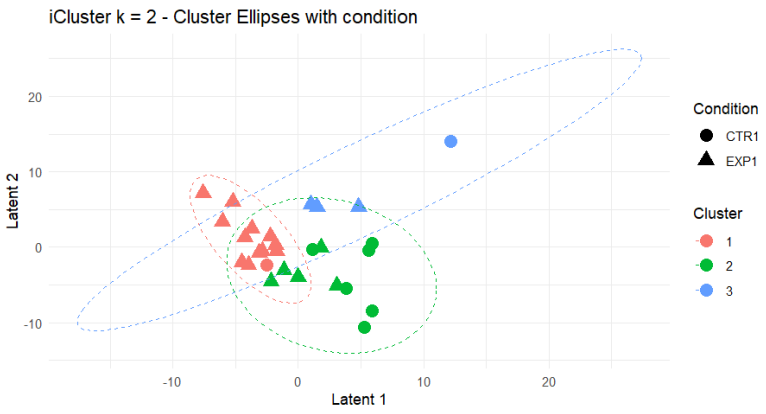

Figure 4.23. Clustering with k = 2.

per plot: Latent 1 vs 2, Latent 1 vs 3 and Latent 2 vs 3.

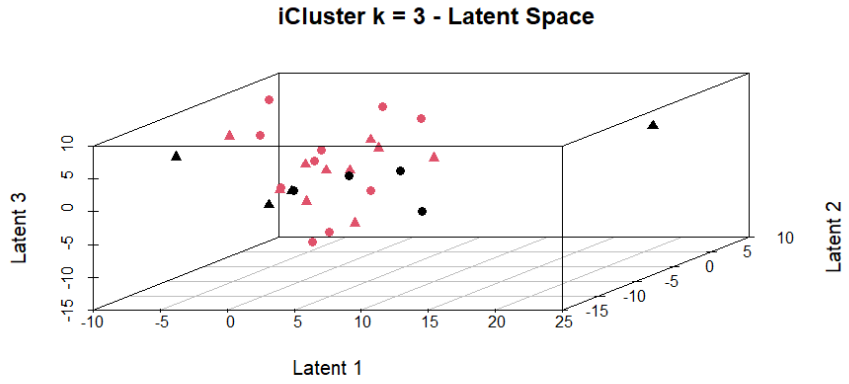

Figure 4.24. Clustering with k = 3 in 3-dimensional latent space.

The four-clusters solution highlights complex multi-dimensional relationships among the samples. First thing we can notice is the *RR* outlier: in the 3-dimensional plot it is completely separated, positioned in the upper-right region, while in the 2-dimensional projections it consistently remains distant from the rest along the first latent variable. Consequently, Cluster 4 (purple) corresponds to a singleton containing only this outlier. The remaining clusters show variable sizes and spatial distributions across the three latent dimensions. Cluster 1 (red) appears relatively dense and compact across all components, whereas Cluster 2 (green) is clearly separated from the others along the third latent variable.

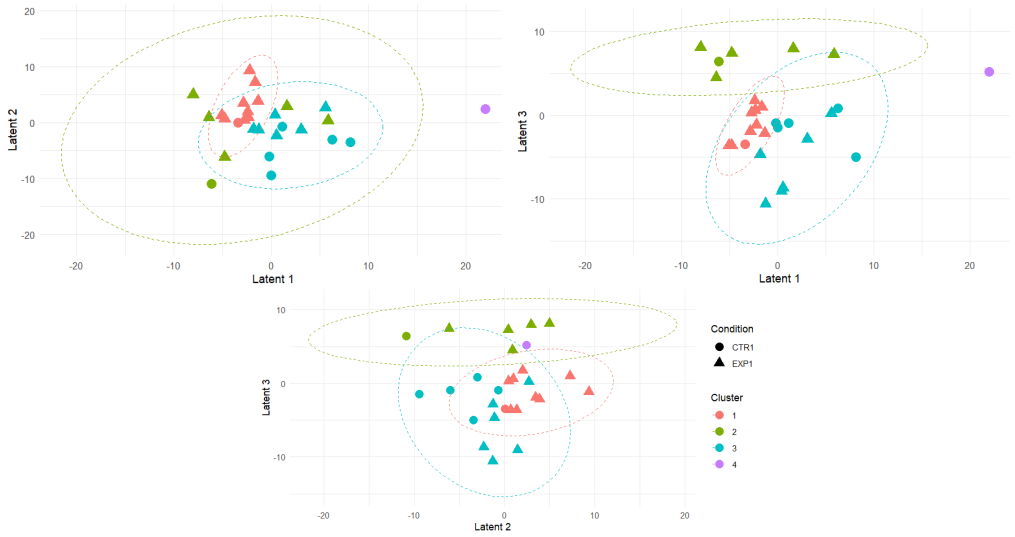

Figure 4.25. iClusterPlus k = 3: Latent variables 1 vs 2 projection.

In the Table 4.11, the clustering results are reported, highlighting the strong isolation of Cluster 4 as a singleton (100 % control, corresponding to the *RR* outlier) and the uneven distribution of controls across the remaining clusters, with Cluster 3 grouping the largest fraction of them (45 %, corresponding 5 out of the 8 controls).

Table 4.11. Clustering with k = 3 (four clusters).

Cluster	Sample Count	% of controls
1	10	10 %
2	6	16 %
3	11	45 %
4	1	100 %

4.3.3 SNF results

SNF was implemented to create fused similarity networks to integrate transcriptomic and proteomic data layers. The analysis generated both global fusion results and modality-specific similarity matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$.

Starting with the similarity matrices heatmaps, it is possible to visualize complementary views of sample relationships, in order to notice possible affinity within controls group or patients group (Figure 4.26).
The transcriptomic affinity matrix shows high relation for the first samples (right upper part, with dark red) and moderate similarity patterns with gradual transitions between sample groups (orange to yellow). In contrast, the proteomic affinity matrix demonstrates

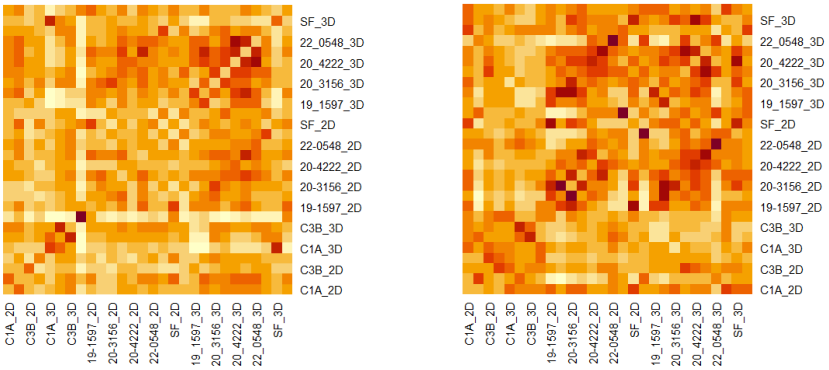

Figure 4.26. Heatmaps of affinity matrices: transcriptomics $W^{(1)}$ and proteomics $W^{(2)}$.

sharper boundaries and more pronounced block structures, indicating that protein-level similarities are more robust to differentiate groups.
Similarly, we visualize the heatmap of the fused matrix $\mathbf{P}$ in Figure 4.27.

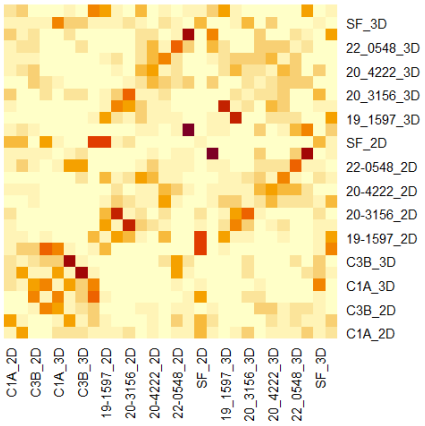

Figure 4.27. Heatmap of fused similarity matrix P.

The fused similarity heatmap reveals block-diagonal structures that suggest natural sample groupings. In addition, a clear affinity among the control samples can be observed in the lower-left region, highlighted by the orange and red colours.

After constructing the fused matrix, the SNF network reveals coherent sample clustering patterns when projected into a two-dimensional space (Figure 4.28).
The fused space visualization reveals improved separation between patients and controls compared both to single-omics analyses and to alternative integration approaches such as MOFA and iClusterPlus. Patient samples (blue) cluster predominantly in the upper and

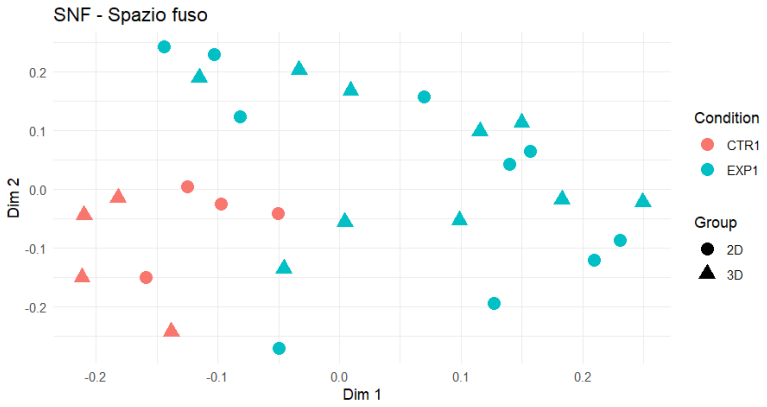

Figure 4.28. Samples in the fused similarity space.

right regions, while control samples (red) occupy the lower-left quadrant. Although culture condition (2D vs. 3D) introduces additional variability, the two culture types remain intermixed, confirming that the dominant source of separation comes from the disease condition. This positive result strongly supports the robustness of the SNF method, especially for the rare disesase case where samples are limited.

According to these results, further analyses will focus on identifying the genes that most strongly drive the two main axes of separation, as well as the biological mechanisms they regulate, since these are important for guiding the improved clustering observed here.

Before moving to the biological insights of these genes, we will show the SNF clustering performed with $k = 2$ and $k = 3$, with k being the number of clusters, in the Table 4.3.3 and 4.3.3.

Cluster	Sample Count	% of controls
1	16	50 %
2	14	0 %

Table 4.12. Clustering with $k = 2$ (second clusters).

The solution with two clusters obtains balanced division, with all the controls grouped in Cluster 1. This suggests a good grouping of control samples, however, this cluster also includes 8 patient samples, indicating only partial discrimination.
When increasing to $k = 3$, the separation improves: all controls remain grouped in Cluster 1, but the number of patients misclassified within this cluster decreases from 8 to 3. This reflects a stronger affinity among control samples and a clearer boundary between disease and control conditions.

Cluster	Sample Count	% of controls
1	11	73 %
2	7	0 %
3	10	0 %

Table 4.13. Clustering with $k = 3$ (three clusters).

In conclusion, the three approaches provide similar insights into the integrated data. However, SNF stands out for its ability to discriminate between the two conditions and to capture more coherent sample groupings. By leveraging network-based similarity relationships, it highlights the underlying structure of the data more effectively, making it a particularly powerful method for biological interpretation in the Ehlers-Danlos syndrome context.

Chapter 5

Conclusions

This study demonstrates the strength of multi-omics integration approaches to investigate rare genetic diseases, specifically Ehlers-Danlos syndrome, through comparative transcriptomics and proteomics analysis under different culture conditions. Single-omics and multi-omics analysis reveals a series of important findings regarding the disease's molecular mechanism.

The transcriptomic analysis revealed differences between 2D and 3D culture conditions, with spheroid cultures generating more differentially expressed genes (218 vs. 23 significant genes). This observation suggests that three-dimensional culture systems better represent the in vivo cellular environment, allowing disease-specific signatures to manifest more clearly. Most of significant genes were up-regulated genes under 3D conditions (67 % of significant genes), indicating that EDS could be caused predominantly by activation, and not inhibition, of transcriptional processes.
The proteomic comparison was more challenging due to the small sample sizes for research in rare disease work. The limited number of proteins that were statistically significant, particularly in 2D cultures, are a reflection of the statistical power limitation when working with 10 patients and 4 controls. However, the overall tendency toward increased significance in 3D conditions with all three statistic methods (ANOVA, limma, and Wilcoxon) supports the hypothesis that spheroid cultures preserve more biologically relevant patterns at the molecular level.
The presence of the RR outlier sample in many analyses (transcriptomics heatmaps, PCA plots and then integration techniques) demonstrates the robustness of our analysis pipeline for capturing potential batch effects or biological variation. For this analysis, we preferred to keep all samples in the data because of the already small number of available controls. However, further study may consider removing this outlier to gain a less biased overview of the datasets.

The comparative evaluation of three integration strategies highlights distinctive strengths and limitations of each method. MOFA exhibited better variance decomposition capacity, particularly in the detection of factors explaining coordinated variation between omics

layers. The ability to handle different data types by employing correct likelihood functions was advantageous for merging discrete count data (transcriptomics) and continuous abundance measures (proteomics).
iClusterPlus provided insights into latent variable structures and sample clustering, though the results were more sensitive to the number of latent variables chosen. Solutions to clustering were less distinguishing by patient versus control populations with significant mixing between conditions.
SNF proved to be the top-performing method in sample discrimination, with clear discrimination of controls from patients, with only 3 misclassified patients in the optimal clustering solution. The network-based method effectively used complementary knowledge across omics levels without losing the uniqueness of each type of data. The ability of the method to construct patient similarity networks and update these iteratively according to cross-omics information was particularly effective in rare disease cases where sample relationships are complex and sample size are limited.

Several limitations need to be mentioned in deriving these conclusions. The sample size is limited (n=14), therefore statistical power and generalizability of findings decrease. Even though this limitation is inherent to studies of rare conditions, it particularly affects the robustness of statistical tests and the stability of clustering algorithms. Larger datasets in future studies would be appreciated, even if this is complicated with the rare diseases, such as EDS. Absence of external validation limits confidence in molecular signatures identified. Cross-validation strategies within this dataset provide qualitative assessment of method stability, yet independent validation cohorts would add biological relevance of findings.
Technical limitations include the different feature sets captured by transcriptomic and proteomic platforms, which necessitated working with the intersection of detected genes. This restriction may have excluded biologically relevant molecules that are detectable by only one technology.

There are several paths for expanding this research that are worthy to be investigated. The application of other multi-omics integration methods, such as Multi-Omics Network Analysis (MONA), Joint and Individual Variation Explained (JIVE), or more recent deep learning platforms like autoencoders, would provide complementary insights into the structure of the data.
Validating the identified molecular signatures on independent groups of EDS patients would confirm the general applicability of the findings and support their translation into clinical practice. Functional validation through pathway analysis and gene ontology enrichment of key genes would add biological results behind the patterns found. Moreover, incorporation of additional omics levels, such as epigenomics or metabolomics, could provide a more integrated systems-level view of EDS, by integrating more layers of omics data.
Applying these analytical frameworks to other rare genetic diseases would establish their general applicability and identify trends for this type of data. These comparative analyses would allow for the development of standardized approaches to rare disease multi-omics analysis.

Even though this study is an initial analytical investigation, the identified molecular signatures might have some clinical implications. The proteins and genes contributing most to sample discrimination may be potential biomarkers for EDS quantification or diagnosis. Their utility would require rigorous validation and regulatory approval before application in the clinic.
The observation that 3D culture conditions reveal more significant disease signatures suggests that spheroid-based assays have the potential to improve diagnostic accuracy for EDS and for rare diseases studies.

Bibliography

[1] Debabrata Acharya and Anirban Mukhopadhyay. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Briefings in Functional Genomics*, 23(5):549–560, 04 2024. ISSN 2041-2657. doi: 10.1093/bfgp/elae013. URL https://doi.org/10.1093/bfgp/elae013.

[2] Russ Altman, Mike Gosink, Michael Gribskov, Jared Roach, and Kevin Warnick. *Computational Genomics with R*. Chapman and Hall/CRC, 2020. URL https://compgenomr.github.io/book/.

[3] Ricard Argelaguet. Mofa2: training a model in r. Tutorial online, hosted on GitHack, September 2020. URL https://raw.githack.com/bioFAM/MOFA2_tutorials/master/R_tutorials/getting_started_R.html#2_What_is_MOFA. Accesso diretto alla sezione "What is MOFA?" del tutorial.

[4] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis: a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018. doi: 10.15252/msb.20178124.

[5] Efi Athieniti and George M. Spyrou. A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal*, 21:134–149, 2023. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2022.11.050. URL https://www.sciencedirect.com/science/article/pii/S200103702200544X.

[6] Marcus Bantscheff, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965, 2012. doi: 10.1007/s00216-012-6203-4.

[7] CD Genomics. Microarray vs rna sequencing: Advantages of rna-seq over microarray technology, 2023. URL https://www.cd-genomics.com/resource-microarray-vs-rna-sequencing.html. Accessed: 2025-07-20.

[8] Prabhakar Chalise, Deukwoo Kwon, Brooke L. Fridley, and Qianxing Mo. Statistical methods for integrative clustering of multi-omics data. In *Methods in Molecular*

Biology, volume 2629, pages 73–93. Springer, 2023. doi: 10.1007/978-1-0716-2986-4_5. Author manuscript available in PMC 2024-03-19.

[9] M. Collotta. Le scienze "omiche", ovvero: dal dogma centrale della biologia alle scienze omiche. la biologia molecolare al servizio del clinico. *Rivista Società Italiana di Medicina Generale*, 2018.

[10] M. Collotta. Trascrittomica: la differenza che conta. *Rivista Società Italiana di Medicina Generale*, 2018.

[11] Xiaofeng Dai and Li Shen. Advances and trends in omics technology development. *Frontiers in Medicine*, Volume 9 - 2022, 2022. ISSN 2296-858X. doi: 10.3389/fmed.2022.911861. URL https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.911861.

[12] Fondazione Telethon. Sindrome di ehlers-danlos, 2022. URL https://www.fondazionetelethon.it/cosa-facciamo/ricerca/malattie-studiate/sindrome-di-ehlers-danlos/. Accessed: 2025-07-17.

[13] Laurent Gatto. Chapter 8 quantitative proteomics data analysis. Omics Data Analysis, WSBIM2122, 2025. URL https://uclouvain-cbio.github.io/WSBIM2122/sec-prot.html. Online chapter, Université catholique de Louvain.

[14] Jessica Gliozzo, Mauricio A. Soto Gomez, Arturo Bonometti, Alex Patak, Elena Casiraghi, and Giorgio Valentini. miss-snf: a multimodal patient similarity network integration approach to handle completely missing data sources. *Bioinformatics*, 41 (4):btaf150, 2025. doi: 10.1093/bioinformatics/btaf150. URL https://academic.oup.com/bioinformatics/article/41/4/btaf150/. Published 4 April 2025.

[15] Harvard Chan Bioinformatics Core (HBC). Introduction to differential gene expression analysis. Online workshop materials, Harvard Chan Bioinformatics Core, 2025. URL https://hbctraining.github.io/Intro-to-DGE/schedule/links-to-lessons.html. Includes links to lessons on workflow setup, QC, DESeq2, visualization, and functional analysis.

[16] Harshad Hegde, Neel Shimpi, Aloksagar Panny, Ingrid Glurich, Pamela Christie, and Amit Acharya. Mice vs ppca: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, 17:100275, 2019. doi: 10.1016/j.imu.2019.100275. URL https://www.sciencedirect.com/science/article/pii/S2352914819302783.

[17] IBM. What is principal component analysis (pca)? https://www.ibm.com/think/topics/principal-component-analysis, 2023. Accessed: 2025-07-24.

[18] IBM. What is principal component analysis (pca)? — a pca plot is a graphical representation, typically showing two principal components (pc1, pc2) on the x- and y-axis. https://www.ibm.com/think/topics/principal-component-analysis, 2023. Accessed: 2025-08-07.

[19] Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, and Ihsan Ali. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, 171:114591, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.114591. URL https://www.sciencedirect.com/science/article/pii/S0957417421000324.

[20] Abhishek Jain. Everything about density plot. https://medium.com/@abhishekjainindore24/everything-about-density-plot-3e74bb664d98, 2024. Medium article; Accessed: 2025-08-05.

[21] Liang Jin, Yingtao Bi, Chenqi Hu, Jun Qu, Shichen Shen, Xue Wang, and Yu Tian. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11:1760, 2021. doi: 10.1038/s41598-021-81279-4.

[22] Piyush Kashyap. Handling missing values in data: A beginner's guide to knn imputation. https://medium.com/@piyushkashyap045/handling-missing-values-in-data-a-beginner-guide-to-knn-imputation-30d37cc7a5b7, 2024. Medium article; Accessed: 2025-08-05.

[23] Weijia Kong, Harvard Wai Hann Hui, Hui Peng, and Wilson Wen Bin Goh. Dealing with missing values in proteomics data. *Proteomics*, 22(23–24):e2200092, 2022. doi: 10.1002/pmic.202200092.

[24] Michael I. Love, Wolfgang Huber, and Simon Anders. Deseq2: rlog function documentation, 2024. URL https://rdrr.io/bioc/DESeq2/man/rlog.html. Accessed: 2025-08-05.

[25] Michael I. Love, Simon Anders, and Wolfgang Huber. Analyzing rna-seq data with deseq2. Bioconductor vignette, July 25 2025. URL https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html. DESeq2 package version 1.49.3.

[26] Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. Transcriptomics technologies. *PLOS Computational Biology*, 13:e1005457, 05 2017. doi: 10.1371/journal.pcbi.1005457.

[27] Gianluca Malato and Brennan Whitfield. An introduction to the shapiro-wilk test for normality. https://builtin.com/data-science/shapiro-wilk-test, 2025. Accessed: 2025-08-07.

[28] mixOmics team. Sparse principal component analysis — mixomics. https://mixomics.org/methods/spca/, 2024. Accessed: 2025-07-17.

[29] Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S. Chan, and Susan G. Hilsenbeck. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxx017.

[30] NBIS Sweden. Similarity network fusion – multi-omics integration workshop. https://nbisweden.github.io/workshop_omics_integration/session_nmf/SNF_main.html, 2024. Accessed: 2025-07-24.

[31] Nist. Levene test for equality of variances. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm, 2025. Accessed: 2025-08-16.

[32] Number Analytics. The ultimate guide to heatmaps in bioinformatics. https://www.numberanalytics.com/blog/ultimate-guide-heatmaps-bioinformatics, 2023. Accessed: 2025-08-07.

[33] Martin Olivier, Reto Asmis, Gail A. Hawkins, Tim D. Howard, and Lisa A. Cox. The need for multi-omics biomarker signatures in precision medicine. *International Journal of Molecular Sciences*, 20(19):4781, 2019. doi: 10.3390/ijms20194781. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6801754/.

[34] pcaMethods package maintainers. bpca: Bayesian pca missing value estimation — pcamethods documentation. https://www.rdocumentation.org/packages/pcaMethods/versions/1.64.0/topics/bpca, 2020. Accessed: 2025-08-07.

[35] pcaMethods package maintainers. bpca: Bayesian pca missing value estimation — pcamethods documentation. https://rdrr.io/bioc/pcaMethods/man/bpca.html, 2020. Accessed: 2025-08-07.

[36] Juan Carlos Pineda. Removing outliers based on cook's distance. Medium, June 2020. URL https://medium.com/@jcpineda/removing-outliers-based-on-cooks-distance-8d5d8913c8eb.

[37] R Documentation contributors. pheatmap: Pretty heatmaps — r documentation. https://www.rdocumentation.org/packages/pheatmap/versions/1.0.13/topics/pheatmap, 2022. Accessed: 2025-08-07.

[38] R Documentation contributors. missforest: Nonparametric missing value imputation using random forest — r documentation. https://www.rdocumentation.org/packages/missForest/versions/1.5/topics/missForest, 2022. Accessed: 2025-08-07.

[39] Thierry Rabilloud and Catherine Lelong. Two-dimensional gel electrophoresis in proteomics: a tutorial. *Journal of Proteomics*, 74(10):1829–1841, 2011. doi: 10.1016/j.jprot.2011.03.040.

[40] Peifeng Ruan, Ya Wang, Ronglai Shen, and Shuang Wang. Using association signal annotations to boost similarity network fusion. *Bioinformatics*, 35(19):3718–3726, 2019. doi: 10.1093/bioinformatics/btz124.

[41] Sartorius. What is principal component analysis (pca) and how it is used. https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186, 2020. Accessed: 2025-08-07.

[42] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 09 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp543. URL https://doi.org/10.1093/bioinformatics/btp543.

[43] Steven R. Shuken. An introduction to mass spectrometry-based proteomics. *Journal of Proteome Research*, 22(7):2151–2171, 2023. doi: 10.1021/acs.jproteome.3c00002.

[44] Gordon K. Smyth, Matthew E. Ritchie, Natalie Thorne, James Wettenhall, Wei Shi, and Yifang Hu. *limma: Linear Models for Microarray and RNA-Seq Data — User's Guide*. Bioinformatics Computational Biology Division, The Walter and Eliza Hall Institute of Medical Research, first edition: 2 december 2002; last revised: 14 april 2025 edition, 2025. URL https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf.

[45] Statistics Canada. Statistics: Power from data! https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214890-eng.htm, 2021. Accessed: 2025-08-05.

[46] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 9780521784504.

[47] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333–337, 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2810.

[48] Dankyu Yoon, Eun-Kyung Lee, and Taesung Park. Robust imputation method for missing values in microarray data. *BMC Bioinformatics*, 8(Suppl 2):S6, 2007. doi: 10.1186/1471-2105-8-S2-S6. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-S2-S6.

[49] Marinka Zitnik, Finale Doshi-Velez Nguyen, Bohan Wang, Jure Leskovec, Anna Goldenberg, and Michael M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019. doi: 10.1016/j.inffus.2018.09.012. URL https://www.sciencedirect.com/science/article/pii/S1566253518304482.