POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

ProteoVAE: a biologically informed Variational AutoEncoder to research new subtypes in type 2 diabetes

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Paolo Serafino Mastroianni**

Student ID: 244784
Advisor: Prof. Francesca Ieva
Co-advisors: Dr. Michela Carlotta Massi
Academic Year: 2024-25

Abstract

Type 2 diabetes (T2D) is an extremely heterogeneous disease, due to different clinical pictures, drug response and disease progression among patients. This study aims to leverage the information about the ongoing biological processes present in the expression of blood serum proteins to explain this heterogeneity by identifying different subtypes of T2D. To perform this task we adapted a previously existing Variational Autoencoder developed for single cell transcriptomics to build ProteoVAE, a biologically interpretable model that can map protein expression into latent representations of the activity of molecular pathways by leveraging the prior knowledge regarding the relationship between these pathways and their proteins.

To validate this framework, we exploited accurate statistical tests and literature review to perform a proof of concept, which showed that the latent features extracted from the model represent, in fact, specific biological processes, some of them strictly related to diabetes.

The extracted latent representations are then fed to an interpretable clustering pipeline involving cutting-edge algorithms for explainable machine learning that led us to find two T2D subtypes presenting significant difference in both biological and clinical picture.

Finally, we present possible technical refinements and a new approach that could further improve the performance of the model and permit a more accurate and biologically meaningful subdivision of diabetic patients, with several possible application in the field of personalized medicine.

Keywords: Proteomics, Type 2 diabetes, Variational Autoencoders, Explainable Machine Learning, Interpretable clustering

Abstract in lingua italiana

Il diabete di tipo 2 (T2D) è una malattia estremamente eterogenea, a causa dei differenti quadri clinici, risposte ai farmaci e progressioni della malattia in diversi pazienti. Questo studio mira a sfruttare le informazioni sui processi biologici attivi presenti nell'espressione delle proteine del siero sanguigno per spiegare questa eterogeneità attraverso l'identificazione di diversi sottotipi di T2D. Per raggiungere questo obiettivo abbiamo adattato un già esistente Autoencoder variazionale sviluppato per la trascrittomica a singola cellula per creare ProteoVAE, un modello biologicamente interpretabile in grado di mappare l'espressione proteica in rappresentazioni latenti dell'attività di pathway molecolari, sfruttando le conoscenze pregresse sulle relazioni tra questi pathway e le proteine che li compongono.
Per validare il modello, abbiamo utilizzato accurati test statistici e una revisione della letteratura per condurre una prova di concetto, la quale ha mostrato come le caratteristiche latenti estratte dal modello rappresentino di fatto specifici processi biologici, alcuni dei quali strettamente correlati al diabete.
Le rappresentazioni latenti estratte sono state quindi utilizzate in una pipeline di clustering interpretabile che impiega algoritmi all'avanguardia di Machine Learning spiegabile, che ci ha condotto all'identificazione di due sottotipi di T2D che presentano differenze significative sia sul piano biologico che sul piano clinico.
Infine, presentiamo possibili miglioramenti tecnici e un nuovo approccio che potrebbe ulteriormente incrementare le prestazioni del modello e permettere una suddivisione più accurata e biologicamente significativa dei pazienti diabetici, con numerose possibili applicazioni nel campo della medicina personalizzata.

Parole chiave: Proteomica, Diabete di tipo 2, Autoencoder variazionali, Machine Learning spiegabile, Clustering interpretabile

Contents

1 | Introduction

In the last years Machine Learning (ML) was increasingly used to analyze biological variables, including genomic data, physiological signals, imaging, metabolomics and proteomics. While providing powerful tools to analyze high-dimensional, noisy and nonlinear data, which is very common in biology, ML faces different challenges in the field of interpretability, the main weak point of Black-Box models.

In particular, talking about **proteomics**, most models are less useful if they are not able to link the extracted features to the underlining biology. In this project, we introduce **ProteoVAE**, a Variational AutoEncoder that maps protein expressions into a biologically informed and interpretable latent space, where each latent variable is forced to represent a specific protein pathway, and one of its possible application regarding **Type 2 Diabetes (T2D)**.

1.1. Type 2 diabetes

Type 2 diabetes is a chronic disease due to persistent high blood sugar levels. It is a very widespread condition; in Europe its prevalence is estimated by the International Diabetes Federation to be 9.8%, corresponding to around 65 million people, and it is expected to show a 10% increase by 2050 [6]. Moreover, if non treated, it can lead to several severe complications as cardiovascular diseases (CVDs), eye conditions like diabetic retinopathy, ulcers that can lead to amputations and many more.

Despite being defined only by glucose in blood, the clinical picture of T2D is very heterogeneous: different patients can present different characteristics, disease progression and drug response. The study by [1] distinguishes between two ways of interpreting T2D heterogeneity. One perspective considers T2D as the result of multiple defects across molecular pathways (the "palette model"), reflecting the true biological complexity of the disease and its mechanistic functions. The other uses easily measurable clinical parame-

ters to subdivide patients into broader groups, a strategy that is more practical for current clinical implementation. As the authors note, these clinical subtypes serve as surrogates for underlying biological mechanisms, because directly quantifying pathway dysfunction is still challenging in routine settings. However, if accurate measurement of pathway dysregulation were achievable, stratifying patients based on specific mechanistic drivers could be particularly valuable: dysregulated pathways naturally suggest targetable processes, offering opportunities for precision treatment strategies such as drug repurposing or pathway-oriented interventions.

1.2. The Importance of Proteomics

Given the above, **pathways** represent an ideal level of biological organization for understanding disease mechanisms. Indeed, they capture coordinated protein interactions that directly regulate cellular function.

Proteins are the main effectors of metabolism, signaling, and immune responses, therefore by summarizing protein behavior into pathway activity profiles, we can obtain interpretable, biologically grounded signals that may explain why different patients develop diabetes differently and respond to treatments in distinct ways, in fact their altered abundance can reveal which biological processes are disturbed.

Proteomics (i.e., the study of the interactions, function, composition, structures, and cellular activities of proteins [2]) provides a powerful window into active mechanisms. In particular, Expression Proteomics usually studies the quantitative and qualitative expression of proteins (meaning how much of a specific protein is produced in a cell, tissue, or organism) to identify disease-related proteins by making comparisons in terms of protein expression between healthy and sick individuals.

1.3. Scope and Outline of the Work of Thesis

The purpose of this thesis is to investigate whether the biological processes underlying type 2 diabetes (T2D) can be captured from blood proteomic data and used to recognize distinct subtypes of the disease. While previous stratification approaches have relied on clinical parameters, we explore whether we can reliably infer pathway activity from circulating proteins, and if this information can provide a mechanistic basis for patient clustering.

Specifically, this work of thesis is structured around two core research questions:

- **RQ1. Pathway signal extraction:** Can a biologically informed variational autoencoder (VAE) reliably extract biologically interpretable latent representations of pathway activation from serum proteomic data?
- **RQ2. Subtype discovery:** Can these learned pathway-level representations be used to identify meaningful subgroups of T2D patients with distinct clinical profiles?

To address these questions, we designed an analytical pipeline combining deep generative modelling and explainable machine learning techniques.

Specifically, we adapted an existing Variational AutoEncoder (VAE) model originally developed for single-cell transcriptomics to the proteomics domain [16]. The model uses a biologically informed latent space, where each latent variable corresponds to a known molecular pathway. This allows the extraction of low-dimensional, interpretable features that summarize the activity of biological processes across individuals. The application of these type of models to serum proteomics and its use for disease subtyping represent a novel approach within the field.

On top of that, we propose a pipeline of interpretable analyses designed to cluster patients and explain the biological basis of the obtained clusters.

As a first step, we test the model's ability to detect well-known biological differences between diabetic and non-diabetic individuals, providing a validation of its capacity to capture relevant signals. We then apply the same framework to the diabetic population only, to investigate whether it can uncover previously hidden heterogeneity within the disease.

The remainder of this thesis is organized as follows:

- **Chapter 2** gives a description of the data on which the analysis has been conducted and the prior biological knowledge used to inform the model, with the preprocessing steps. Moreover the detailed mathematical framework behind the employed strategies is presented, together with their application in the aforementioned context.
- **Chapter 3** shows the computational results, beginning with the model optimization performed to gain better performance, continuing with the proof of concept on the model's ability to capture biological processes inside the latent space and concluding with the identification of the T2D subtypes and a possible clinical explanation for this subdivision.
- **Chapter 4** summarizes the key findings and conclusions of the study, describes the limitations of this approach and proposes potential future directions to expand the

work of this thesis and perform further investigation.

2 | Materials and Methods

This chapter describes the key processes of this study, including the data used and how it was prepared. Moreover, we detail the mathematical framework that serves as the basis for the model and the algorithms used for the analysis and clustering of the extracted variables.

2.1. UK Biobank Pharma Proteomics Project

UK Biobank Pharma Proteomics Project (UKB-PPP) [28] is a massive current project led by UK Biobank, a biorepository that stores human biological samples for use in research. The project consisted in collecting plasma biospecimens from 52736 Uk Biobank applicants, 5474 of whom were diagnosed with type 2 diabetes. The samples were then sent to **Olink**, which through **Proximity Extension Assay (PEA)** [17], a technology where a matched pair of antibodies bind to the respective target protein in a sample, generated raw counts of expression regarding Inflammation, Oncology, Cardiometabolic and Neurologic proteins in blood serum, which were then processed and normalized through software. In detail, a $\log_2$ ratio is computed between the counts in the sample and those in a control sample, and then normalized and adjusted for batch effects to obtain **Normalized Protein eXpression (NPX)**, a relative unit normally distributed under a $\log_2$ scale suitable for statistical analysis.
The resulting dataset therefore consists of 810 continuous features representing protein expression for 52736 patients.

2.2. The Prior Knowledge: KEGG PATHWAY

To answer the first research question regarding the biological latent representations of pathway activations, we need a biologically reliable source from which to extract the prior knowledge that we use to inform the model and thus control the structure of the latent space. In particular, we need to know the pathways of interest and their relationship with the proteins in our dataset, i.e. which proteins belong to which pathways.

These crucial information can be extracted from **KEGG Pathway**, a branch of KEGG [12], which is a database containing a collection of pathway maps of several living organisms divided into *Metabolism*, *Genetic information processing*, *Environmental information processing*, *Cellular processes*, *Organismal systems*, *Human diseases* and *Drug development*, with each map containing genes, proteins, chemical compounds and drug targets among others. These pathway maps are drawn manually and represent the experimental knowledge of various functions of cells and organisms. Since this work is focused on a human disease, we only considered human-reated pathways.

2.3. ProteoVAE Theoretical Background

One of the main issues of the analysis of high dimensional dataset is to reveal underlying patterns not immediately visible and to separate them from measurement-induced noise and redundant information. To tackle this problem, dimensional reduction techniques become crucial, also leading to better performances for all algorithms that fall prey to the curse of dimensionality. In this context, *Variational AutoEncoders* arise as a very useful tool thanks to their flexibility in capturing non-linear relationships and complex manifolds that other algorithms, like PCA [10], usually cannot. In this section, we describe the core ideas behind AutoEncoders, their variational counterparts, and how to make them biologically interpretable like ProteoVAE.

2.3.1. AutoEncoders

AutoEncoders [9] are feed-forward neural networks with the scope of learning low-dimensional representations of input data in an unsupervised fashion. This is done by using the input data as expected output of the network and introducing a low dimensional bottleneck layer that will represent the latent variables extracted from the data.
Introducing some notation, given the input data x, we want to find its latent representation z and two functions (usually non-linear) f_ϕ and g_θ such that $z = f_\phi(x)$ and g_θ minimizes $||x - g_\theta(z)||^2$, where $||\cdot||^2$ is typically the MSE, while ϕ and θ are network parameters learned through backpropagation. However, while this architecture makes autoencoders able to discover low dimensional representation of the data, it usually results in a discontinuous and not easily interpretable latent space. This led to the introduction of a probabilistic framework for the regularized training of this architecture by Kingma et al. [14].

2.3.2. Variational Inference and VAEs

Taking inspiration from Probabilistic Graphical Models [15], Variational AutoEncoders introduce a Bayesian framework to model the relationship between the data x and its latent representation z, which become random variables. The scope is to model the posterior distribution $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ through a *probabilistic encoder* and a *probabilistic decoder*, which are distributions over z and x rather than deterministic functions as in classical autoencoders. The main issue is the intractable computation of $p(x) = \int_z p(x|z)p(z)$ necessary for the computation of the posterior distribution, a problem tackled thanks to *Variational Inference* [4], a Bayesian Learning branch aiming to approximate the posterior distribution $p(z|x)$ with a parametrized family $q_\phi(z|x)$ by seeking, in the optimization process, the optimal parameter ϕ that maximize the log-likelihood of the data. However, it is challenging to compute directly the log-likelihood, so we look for a lower bound to maximize. Formally, by Jensen's inequality [11], we have:

$$\log p_\theta(x) = \log \int_z p_\theta(x,z)dz = \log \int_z p_\theta(x|z)p(z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz \geq$$
$$\geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathbb{E}_{q_\phi(z|x)}[\log \frac{q_\phi(z|x)}{p(z)}] =$$
$$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathbb{D}_{KL}(q_\phi||p)$$

This is the Evidence Lower Bound (ELBO), where θ are the decoder parameters, ϕ are the encoder parameters and $\mathbb{D}_{KL}(q_\phi||p)$ is the Kullback-Leibler divergence, which measures how much the approximating distribution q_ϕ differs from the prior distribution p. By changing the sign of the ELBO we obtain the ELBO loss, that we aim to minimize during the training process.

For what regards the choice of the distribution of ProteoVAE, since we have continuous data, we chose $p(z) = \mathcal{N}(0, I)$ and $p(x|z) = \mathcal{N}(g_\theta(z), \sigma^2 I)$, while $q_\phi(z|x)$ is approximated by a neural network (the encoder) that outputs for each latent variable a mean μ_i and the logarithm of a variance σ_i^2 (for numerical stability) and then samples $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. The problem with the last-mentioned sampling step is the non-differentiability, which makes it impossible to propagate the gradient during the backpropagation in training. To avoid this problem, Kingma et al. introduced the *reparametrization trick* [14], which consists in rewriting the sampling of z as $z = \mu_\phi + \sigma_\phi \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. This is equivalent to the classical sampling but, since ϵ is independent from ϕ it is possible to propagate the gradient through this step during the optimization process.

2.3.3. Decoder architecture for latent space controlling

The main difference between ProteoVAE and a classical VAE is the interpretability of the latent space, obtained through a specific architecture of the decoder inspired by **expiMap** [16].
To easily control the structure of the latent space, its relationship with the reconstruction is represented by a linear combination, i.e. a matrix W.
We want to model the latent space to represent the pathways we extracted from KEGG, therefore we apply a binary mask on the decoder weights. Formally, given d proteins and p pathways, we build a binary matrix M $d \times p$ such that:

$$M_{ij} = \begin{cases} 1 & \text{if protein } i \text{ belongs to pathway } j \\ 0 & \text{otherwise} \end{cases}$$

Then, both in the training and evaluation of the model, the final decoder matrix C is obtained by the element-wise product of the weight matrix and the mask $C = W \odot M$. This ensures that each protein can be reconstructed only from latent variables that represent pathways to which the protein belongs, enforcing each latent variable to activate in correspondence with the activation of that specific pathway in the patient.

2.4. Pre-processing and Feature Selection

To increase the power of dimensionality reduction techniques, it is usually a good practice to select the highly variable proteins in the dataset. This helps to reduce the noise and also to identify the proteins that really discriminate the patients and their ongoing biological processes. To perform this selection we leveraged the procedure described by *Satija et al.* [25] and implemented in the Python package *Scanpy* [31]: for each protein i the mean $\bar{x}_i$ and a dispersion index (variance/mean) of its expression are calculated and all proteins are placed in 20 bins based on their mean. Then, in each bin, each protein expression is z-normalized and a z-cutoff is chosen to select the top 500 highly variable proteins.

For what regards pathways, we exploited the KEGG database by extracting a list of human-related pathways and performing an *OverRepresentation Analysis* (ORA) to find significant overlaps between the proteins in our dataset and those contained in the extracted pathways through **gseapy** [6]. The library performs a Fisher exact test [7] for each pathway in the list to check the hypothesis that the proteins in our dataset are randomly

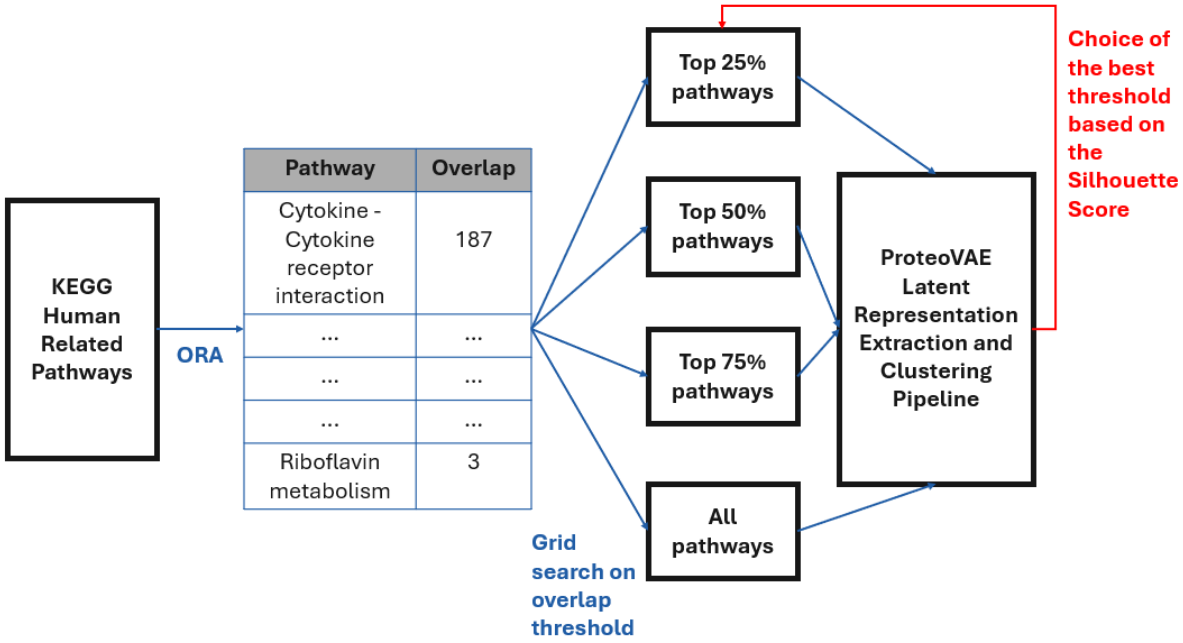

Figure 2.1: Scheme summarizing the pathway selection process

distributed with respect to the proteins in that pathway. The P-values of this tests are then adjusted for multiple testing with the Benjamini-Hochberg correction [3] and only the pathways that presented $P < 0.1$ are retained, for a total of 174.

The selected pathways will be represented by the variables necessary to identify the T2D subtypes, so a more accurate selection is needed to reduce the captured noise. Unfortunately, it is difficult to find an optimal way to perform such selection, so our heuristic approach was to choose the pathways that contained the largest number of proteins in our dataset, by setting a threshold on the number of proteins overlapping with our data that each pathway must have to be included in the model.

Keeping in mind our goal to identify different subtypes of T2D, to find the optimal threshold we performed a grid search with different possible values and chose the optimal one as the threshold that maximized the **Silhouette Score** [23] after performing a clustering of the latent features in diabetic patients. The chosen threshold allowed us to keep only the top 25% of the pathways in terms of the largest protein overlap with our data. The procedure is summarized in Figure 2.1.

At the end of this procedure we removed the proteins that did not appear in the remaining pathways and the final training-validation set for our model consisted of 47262 healthy patients with 332 proteins for each patient and a latent space of 44 pathways. The data, together with the Binary Mask (Figure 2.2), are then fed to the model for the training process.

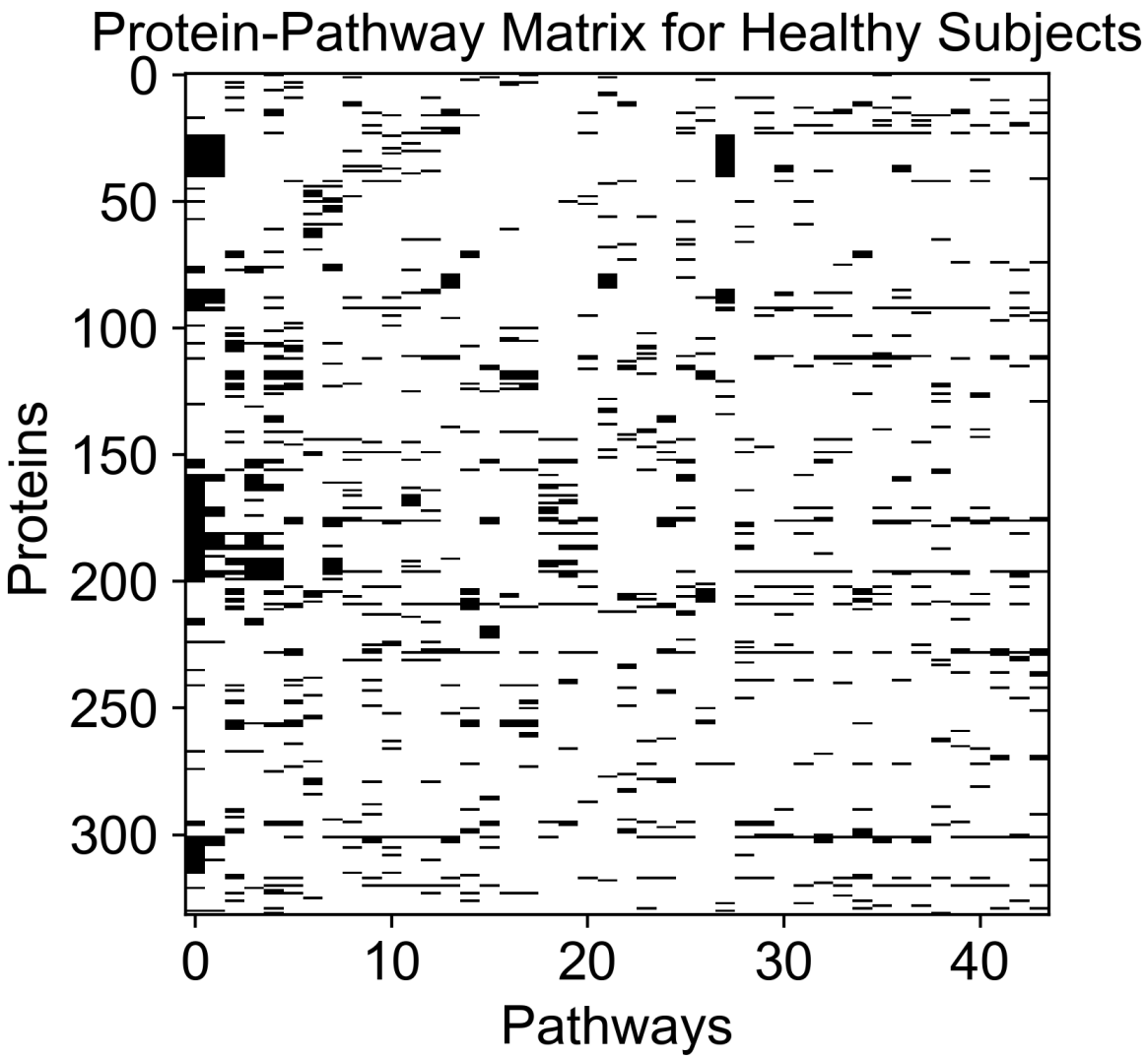

Figure 2.2: Binary Mask: the black cells represent membership of a protein to a pathway, i.e. ones in the implementation

2.5. Model

Inspired by **expiMap** [16] the model has the standard architecture of a Variational AutoEncoder, with some modifications. Here we summarize the architecture of the model and the training process.

The **Encoder** is a Multi-Layer Perceptron with 2 layers of 512 and 256 neurons. Layer Normalization and a ReLu activation function follow each layer. The Encoder extracts the mean μ and the logarithm of the variance σ^2 from which the latent variables are sampled.

The **Latent Space** has the dimension of the selected protein pathways and is sampled

from a Gaussian distribution having as parameters the output of the encoder.
The **Decoder** consists of a linear layer where a binary mask is applied so that only proteins belonging to a specific pathway can be reconstructed from that latent variable. We aim to design the latent space of ProteoVAE in such a way that each variable represents the activation of a determined pathway in a patient. To complete this task, the binary mask is not enough; in fact, we wish to consider a pathway active if the proteins that belong to it are highly expressed in the blood serum of the patient. Therefore, since we are not putting any constraint on the latent variables, we constrain the weights to be non-negative to avoid situations in which a negatively activated pathway, combined with a negative weight, results in a positive reconstruction of a protein.
The **Training** is performed through *Stochastic Proximal Gradient Descent (SPGD)* with *Adam* [13] optimizer to minimize the ELBO loss together with a Group Lasso regularization term:

$$L(\phi, C) = -\mathbb{E}_{q_\phi(z|x)}[\log p_C(x|z)] + \beta \mathbb{D}_{KL}(q_\phi(z|x)||p(z)) + \alpha \sum_j ||C_{:j}||_2$$

The Group Lasso regularization ensures that the pathway-protein connections that are not helpful enough to reconstruct the input are turned off, eliminating noise and reducing overfitting.
α and β are hyperparameters that we chose to fix at $\alpha = 0.3$, since we have a low number of chosen pathways, and $\beta = 0.2$ since we are more interested on a good reconstruction rather than the normality of the latent space. Training is performed for 400 epochs with early stopping only on healthy patients (non-diabetic) to prevent the model from adapting to the disease pattern. The data from ill patients are projected onto the latent space and analyzed later.

2.5.1. Primary model validation through latent space traversal

Before moving on with the clustering of diabetic patients, for the sake of interpretability, it is important to check if the model is distributing the information as we designed. In practice, we want to measure the impact that each latent feature has on the decoder reconstruction, by perturbing one variable at a time and measuring the change in the reconstruction. We decided to perform this analysis through the following pipeline: for each pathway we changed the sign of the corresponding latent variable of every data point and then computed the average reconstruction both for the original latent vector and the perturbed one; then we performed a Ward hierarchical clustering [30] on the absolute

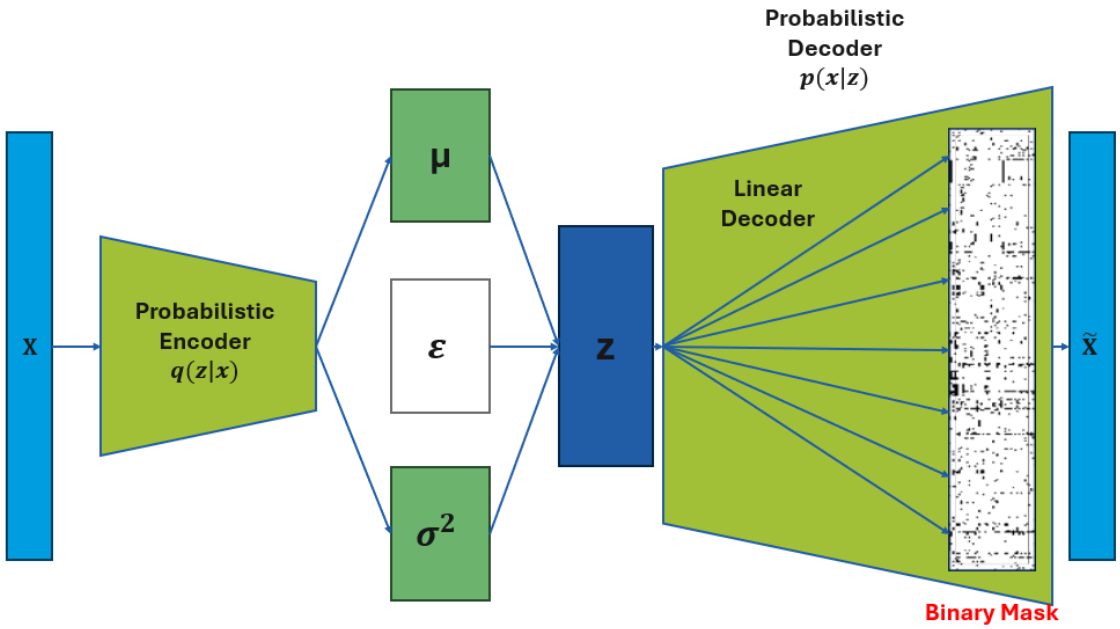

Figure 2.3: Summary diagram of ProteoVAE: the left side is analogous to a standard VAE with the reparametrization trick highlighted, while on the right size we highlight the linear decoder and the binary mask applied before the reconstruction

value of the difference of the two reconstructions and labeled as actively reconstructed the cluster that contained the highest difference. The procedure is summarized in Algorithm 2.1, while the results are presented in Section 3.2:

Algorithm 2.1 Clustering Based Impact on Reconstruction

Require: L batch size, d number of proteins, p number of pathways, z latent space $L \times p$, C decoder matrix

for $i = 1 \dots p$ **do**
  $z' \leftarrow copy(z)$
  $z'^{(i)} = -z^{(i)}$
  $x \leftarrow Cz$
  $x' \leftarrow Cz'$
  $v = \frac{1}{L} \sum_{j=1}^{L} |x_j - x'_j|$
  $max\_id \leftarrow argmax(v)$
  $labels \leftarrow AgglomerativeClustering(v, k = 2)$
  for j = 1 ... L **do**
    if $labels(j) == labels(max\_id)$ **then**
      Protein j actively reconstructed from Pathway i
    end if
  end for
end for

2.6. Interpretable Clustering pipeline for diabetic patients subtyping

Once we assessed the validity of the model, we extracted the biological latent representation for the T2D subtypes identification.
Since we do not have prior information on the existence and the number of these possible subtypes, we opted for a pipeline that involves **Ward agglomerative clustering** [30] and the computation of the **Silhouette score** [23] to individuate the optimal clustering of the data.
To uncover the meaning of the found clusters, i.e. possible biological reasons behind the existence of these groups, we trained a **Random Forest Classifier** [5] with the cluster labels as targets and the latent features as predictors, optimizing its hyperparameters in a Cross-Validation framework. Then we explained the predictions of the Random Forest through **SHapley Additive exPlanations (SHAP)** [18], an algorithm that applies to Machine Learning the concept of **Shapley Values** [26], inherited from Game Theory.

In *Cooperative Game Theory*, the **Shapley value** is a solution that assigns to each player a score measuring how much the player contributes in the coalitions.

Formally, let $N = \{1, 2, \ldots, n\}$ be a set of n players and let $\mathcal{P}(N)$ be the set of possible coalitions the players can create; a game is a vector $v \in \mathbb{R}^{2^N}$ such that for each coalition $A \in \mathcal{P}(N)$ we know its value $v(A)$. For each game v the Shapley value

$$\sigma_i(v) = \sum_{A \in \mathcal{P}(N \setminus \{i\})} \frac{|A|!(n - |A| - 1)!}{n!} [v(A \cup \{i\}) - v(A)]$$

is the only possible solution that satisfies the following properties:

- $\sum_{i \in N} \sigma_i(v) = v(N)$ (Efficiency)
- if $v(A \cup \{i\}) = v(A \cup \{j\})$ $\forall A$ not containing i, j then $\sigma_i(v) = \sigma_j(v)$ (Symmetry)
- if $v(A) = v(A \cup \{i\})$ $\forall A$ then $\sigma_i(v) = 0$ (Null player property)
- $\sigma(v + w) = \sigma(v) + \sigma(w)$ (Additivity)

SHAP, and in particular its version for decision trees **TreeSHAP** [19], takes the concept of Shapley values and applies it to classification models. Let f be a **Decision Tree Classifier** such that, for an input x, $f_x(P)$ denotes the predicted logit of the model for a given label using the set of features P and let $\mathcal{R}$ be the set of all possible orderings in which the N features can be taken into account inside the tree. Then the SHAP value for a tree f, an input x and a feature i takes the form:

$$\Phi_i(f, x) = \frac{1}{N!} \sum_{R \in \mathcal{R}} [f_x(P_i^R \cup i) - f_x(P_i^R)]$$

where P_i^R is the set of features that come before feature i in ordering R.
A such defined SHAP value inherits the aforementioned properties of the Shapley Value; in particular, exploiting the additivity property, we can then compute the SHAP value for a **Random Forest Classifier** as the average of the values computed on each tree.
For each input x the algorithm outputs a score for each feature and each label that indicates how that feature influenced the classifier to classify the input as that label.

Given the framework mentioned above, we used the properties of the SHAP values to explain the clusters found from a biological point of view. On a test set consisting of 20% of the total number of diabetic patients, we used the TreeSHAP algorithm to assign a score to the latent features for each patient. If for a single feature we have a wide range of different SHAP values among different patients, it means that the feature has a high influence on discriminating the patients between the clusters. The results of the described procedure are presented in Section 3.3.

3 | Results and Discussion

In this chapter we present the key results of this work, starting on the optimization that led the model to have its remarkable performances, following with the proof of the model's ability to capture reliable information regarding the biological processes starting from protein expression and concluding with the identification of two T2D subtypes and their possible biological and clinical explanation.

3.1. Model optimization

Although the scope of the project is not strictly related to the reconstruction performance of the model, a good reconstruction of the original proteins from the latent variables indicates a smaller loss of information in the dimensionality reduction. For this reason, to increase the reconstruction performance, we performed fine-tuning on the model hyperparameters. In particular, we tried different configuration regarding encoder architecture, initial learning rate and dropout rate to minimize the validation reconstruction loss.
In this section, we present the results regarding this fine-tuning procedure, also showing the performance in reconstruction obtained by the model that, even if not of primary importance, remains remarkable.

3.1.1. Model architecture

As a Neural Network, ProteoVAE's performances highly depend on its architecture. In particular, since for the sake of interpretability we kept a linear decoder, the entire architecture selection is focused on the non-linear encoder, where we tried different numbers of layers and neurons to balance the model complexity, as is shown in Figure 3.1, leading to the final architecture described in Section 2.5.

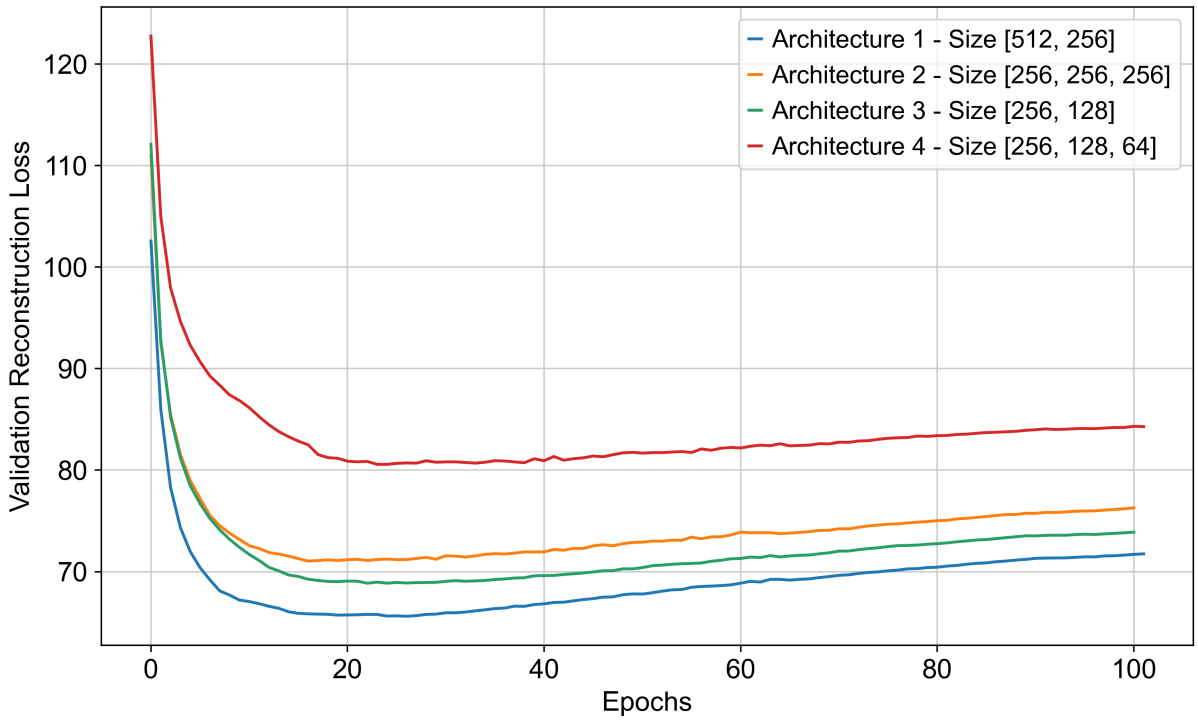

Figure 3.1: Effect of the Architecture Size on the validation reconstruction error during the training process

3.1.2. Dropout Rate

Dropout is a regularization technique introduced by Srivastava et al. [27], that consists in randomly turning off a fraction of the neurons in each layer, controlled by a parameter called dropout rate, preventing units to co-adapt too much and reducing overfitting. We tried different dropout rates, as shown in Figure 3.2, and in our case results show that the loss is lower if no dropout is applied. However, since performances are not significantly lower and our analysis on diabetic patients is entirely performed on unseen data during the training process, we decided to keep a dropout rate of 0.05.

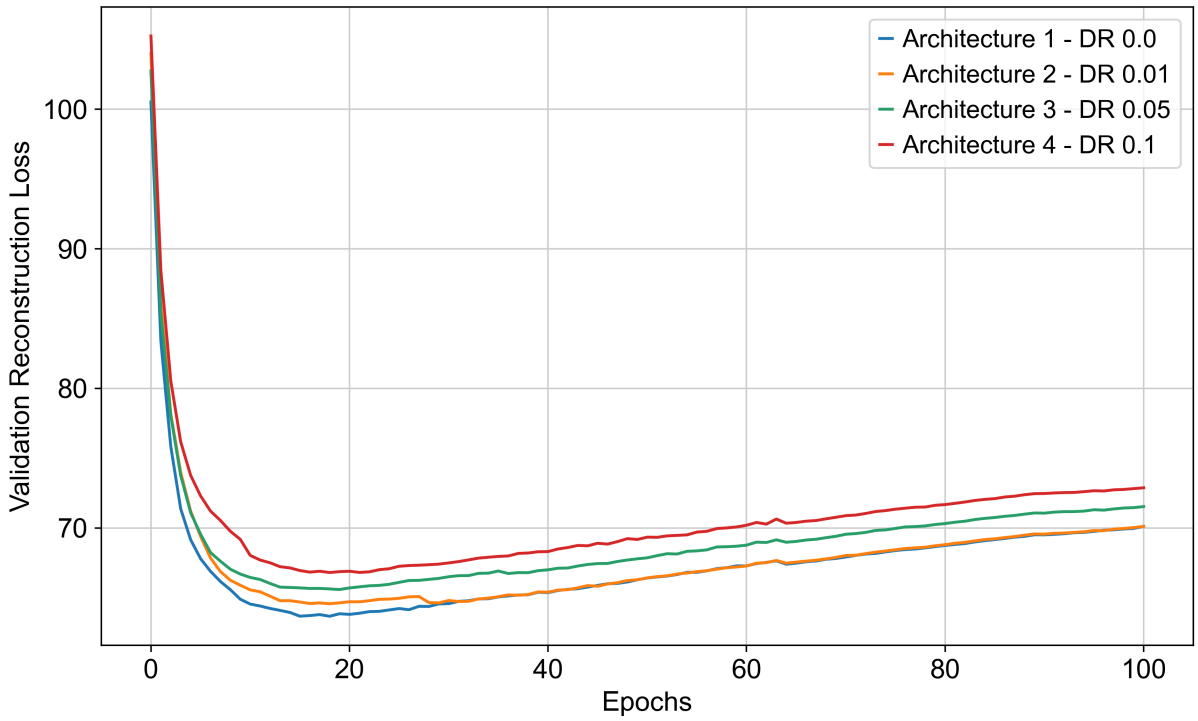

Figure 3.2: Effect of the Dropout Rate on the validation reconstruction error during the training process

3.1.3. Learning Rate

The learning rate is one of the most important hyperparameters of a neural network. It describes how large are the updates of the network parameters at each training step, and a bad choice of learning rate can lead to the divergence of the algorithm, oscillating around the minimum or convergence to local minima. The optimizer Adam adapts dynamically the updates, but the choice of the starting learning rate remains important for a good reconstruction, as shown in Figure 3.3.

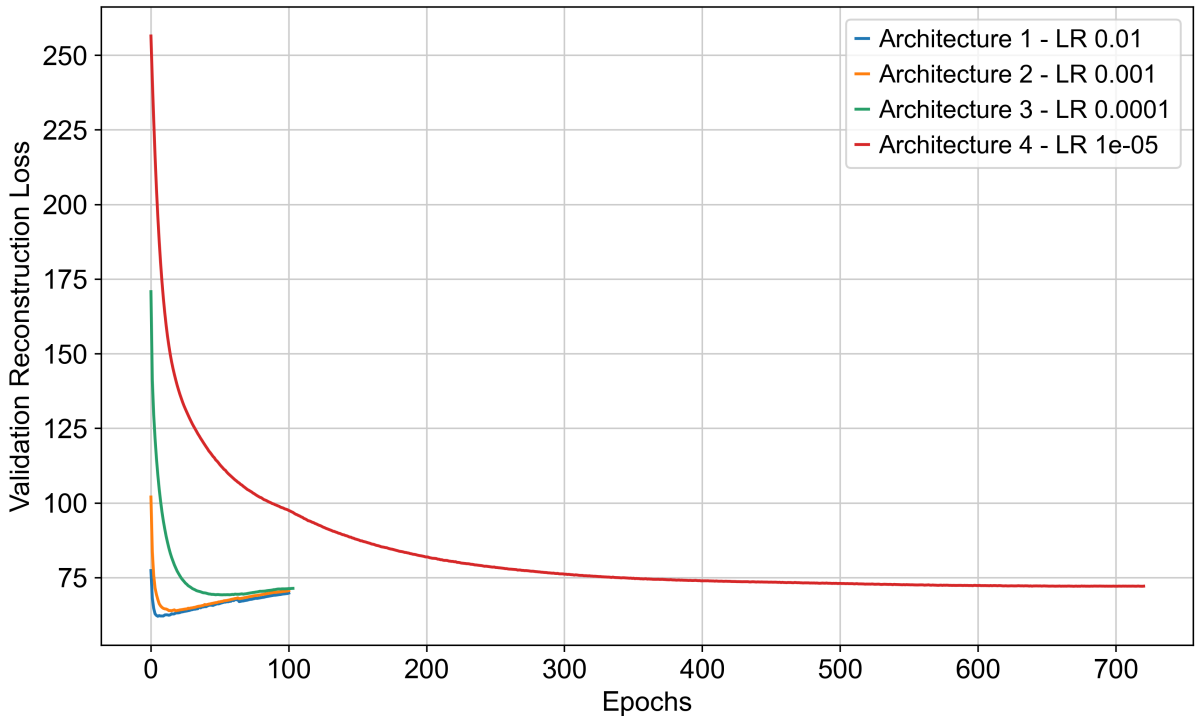

Figure 3.3: Effect of the initial Learning Rate on the validation reconstruction error during the training process; a small initial value leads to a delayed activation of the early stopping regularization, making the training less efficient.

3.2. ProteoVAE Validation

In this section, we report the validation of the model. In particular, we answer the first research question by verifying that the 44 latent variables extracted from the model are able to capture the pathway activation of the patients, including the variations in those activations induced by T2D. To do so, we first present the results of the analysis of latent characteristics of healthy patients used as the training-validation set, followed by the projection of the data regarding diabetic patients into the same latent space and the testing of significantly difference between latent features of healthy and diabetic patients.

3.2.1. Pathway activity profiles in healthy patients

Although not the actual target of this work, healthy patients are extremely useful in assessing the model's ability to capture physiological processes within pathway activity profiles. We start by showing the results of the latent space traversal described in Section 2.5.1. Figure 3.5 shows the impact that each latent variable has on the reconstruction of proteins through hierarchical clustering. As expected, we can see how the pathways with a

greater impact on reconstruction are linked to physiological functions, like the *Cytokine - cytokine receptor interaction pathway*, which is a signaling network that mediates communication between cells of the immune system through proteic molecules called cytokines and their receptors, while pathways linked to specific diseases, like *Measles* or *Yersinia infection*, have a lower impact on the reconstruction, usually with a single protein being actively reconstructed. Figure 3.4 shows also the comparison between the mask we imposed on the decoder weights and the proteins selected as actively reconstructed through the clustering procedure by each pathway. It can be seen how the model is using only the meaningful pathway-protein links thanks to the Group Lasso regularization, avoiding using all the freedom still allowed by the mask to capture noise and overfit data.

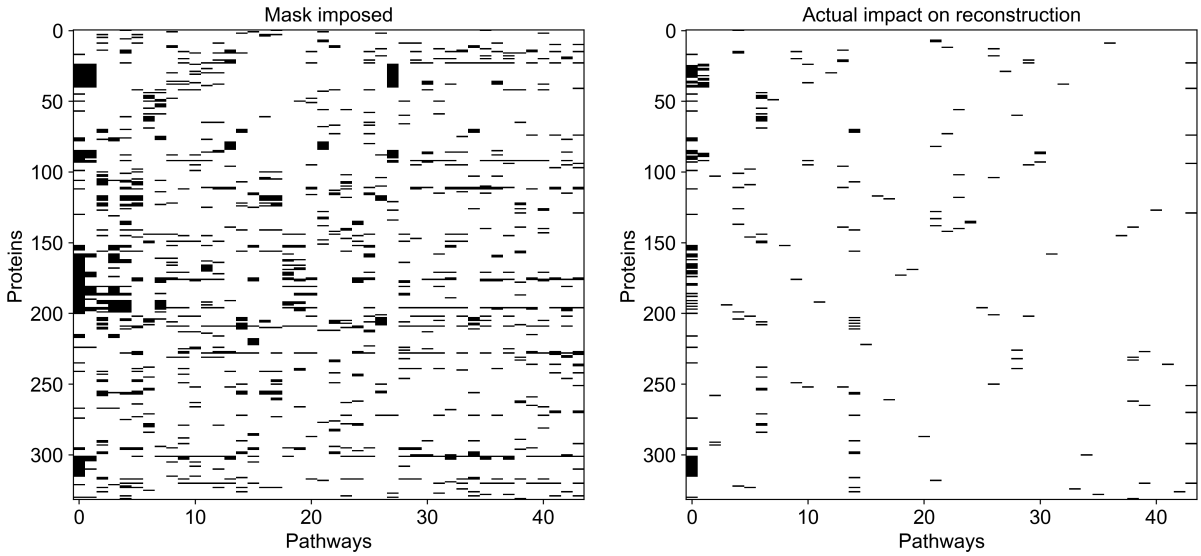

Figure 3.4: Comparison between the imposed mask (left) and the actual protein reconstruction (right); many connections can be considered turned off from the Group Lasso Regularization

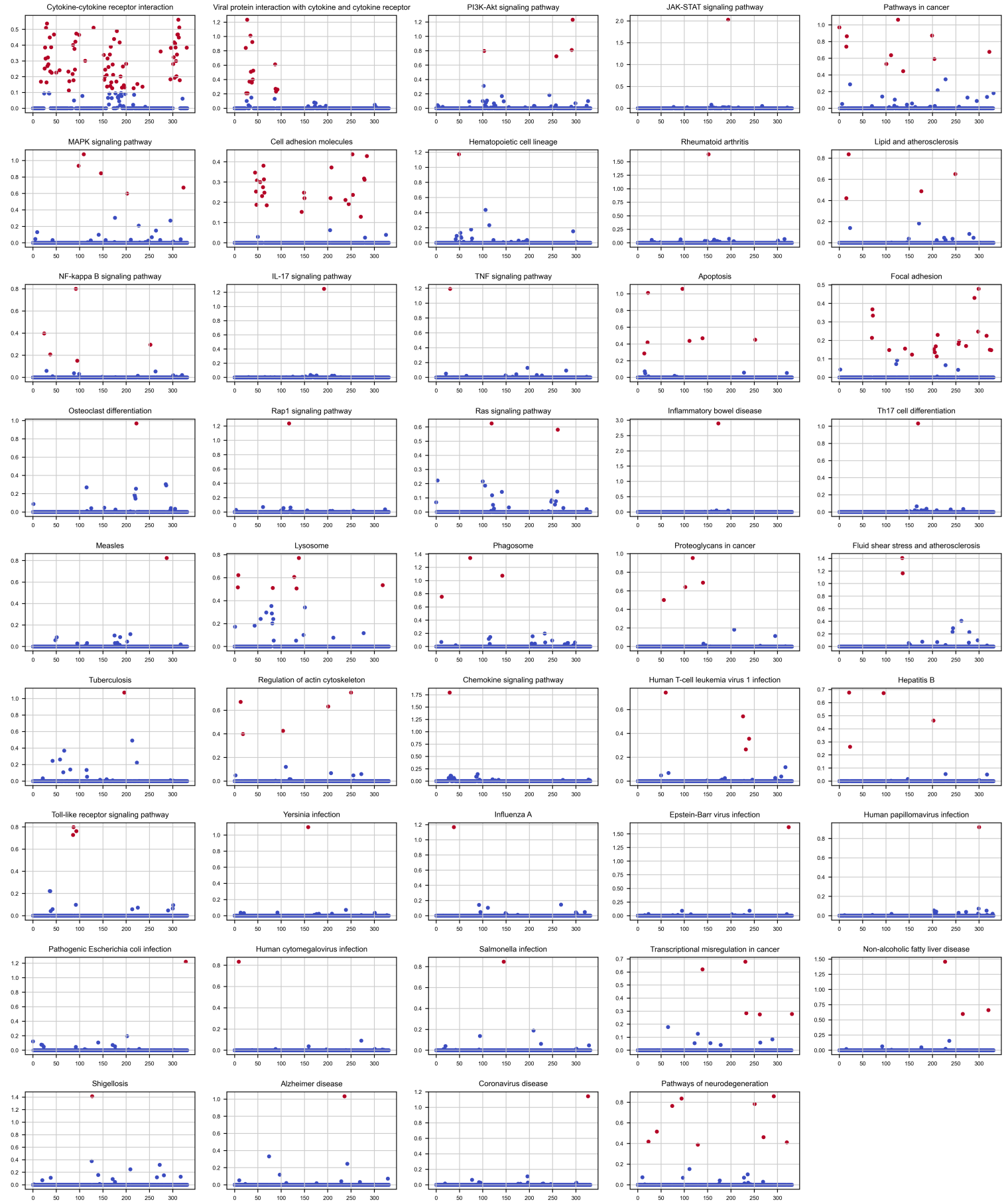

Figure 3.5: Latent Space Traversal

Another important validation step was to check how the model distributed the information between the latent variables. It is important to understand whether the model learned different representations for each variable instead of just capturing the same signal. The correlation matrix (Figure 3.6) shows that most of the latent space appears to be uncorrelated, in particular only three couples of pathways present a correlation higher than 0.8 and involved *Pathways in cancer*, *Apoptosis*, *Proteoglycans in cancer* and *Phagosome*, which are all linked to cell turnover (proliferation and death). This further proves the fact that the representations are capturing different biological processes and that the correlation of the latent features is following the biological correlation of such processes.

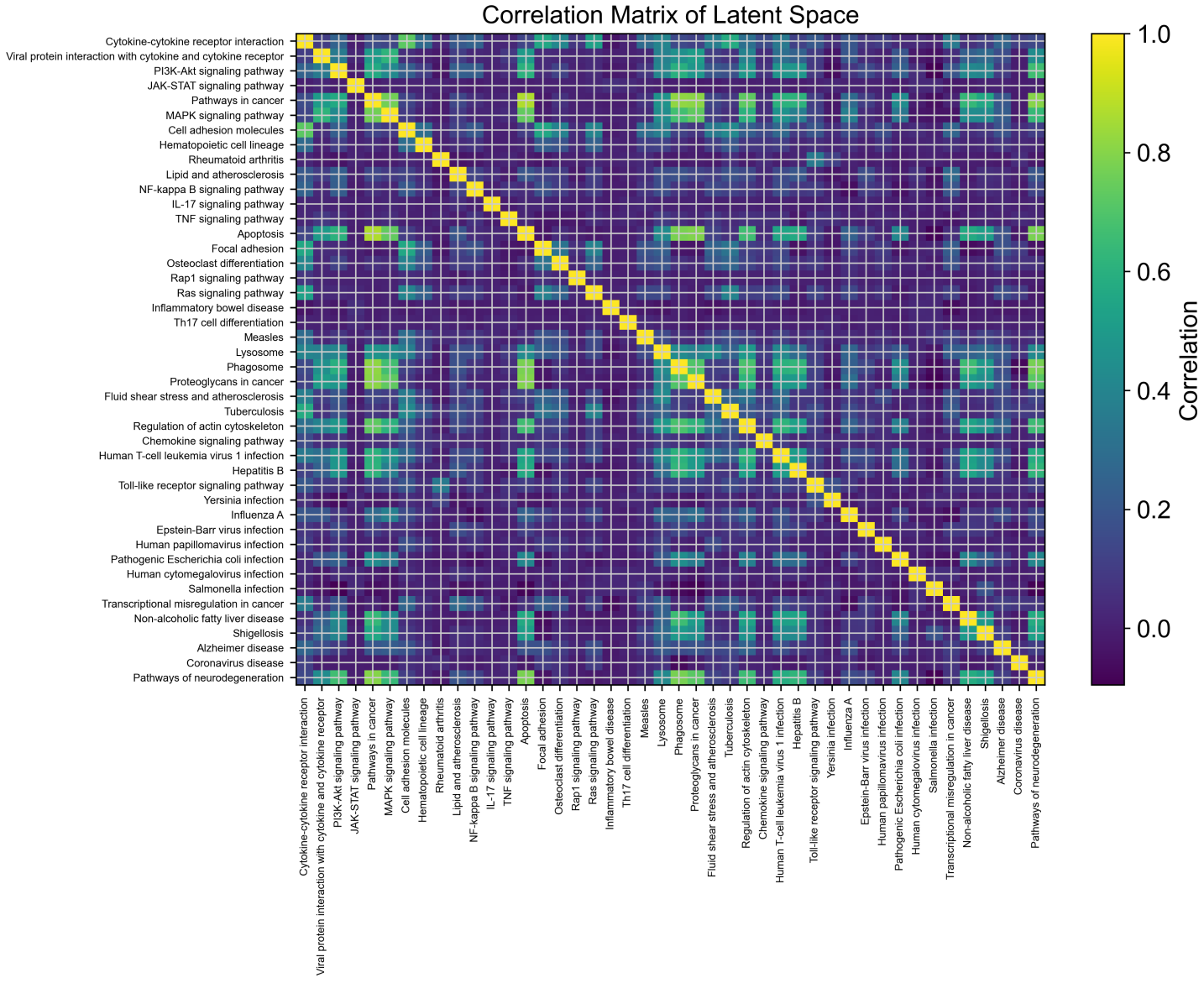

Figure 3.6: Correlation Matrix of the latent variables

3.2.2. Testing diabetes-related biological processes

To further assess the model's ability to uncover latent features from protein data, we decided to test it with completely unseen and slightly different inputs, belonging to diabetic patients, which we projected into the latent space to analyze the variation of the pathways. The goal of this section is to do a proof of concept to understand if the model is able to capture reliable biological processes related to T2D. For an initial visual exploration, we exploited **Principal Component Analysis (PCA)** [10] and plotted the first two scores of healthy and diabetic patients, the last divided into prevalent and incident cases (Figure 3.7).

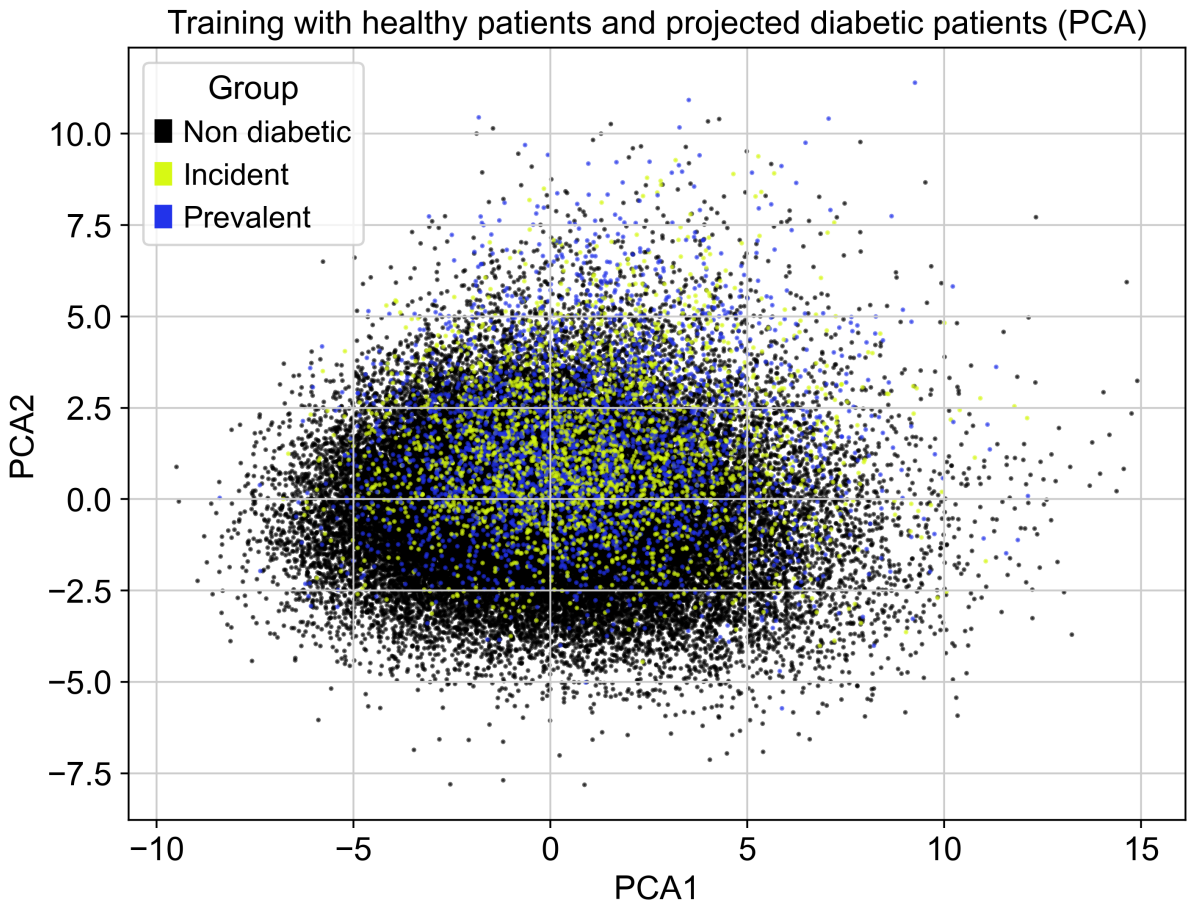

Figure 3.7: First two Principal Components of the latent space coloured according to the status of the disease; prevalent cases are patients that had the disease at the moment of the measurements while incident cases are patients that developed it afterwards.

The PCA shows a difference in the projection of healthy and sick patients, but to understand which variables were mostly responsible for this variation and to check their

biological validity, we leveraged the high sample size to perform simultaneous t-tests and bayesian enrichment tests presented in Figure 3.8 and Figure 3.9.

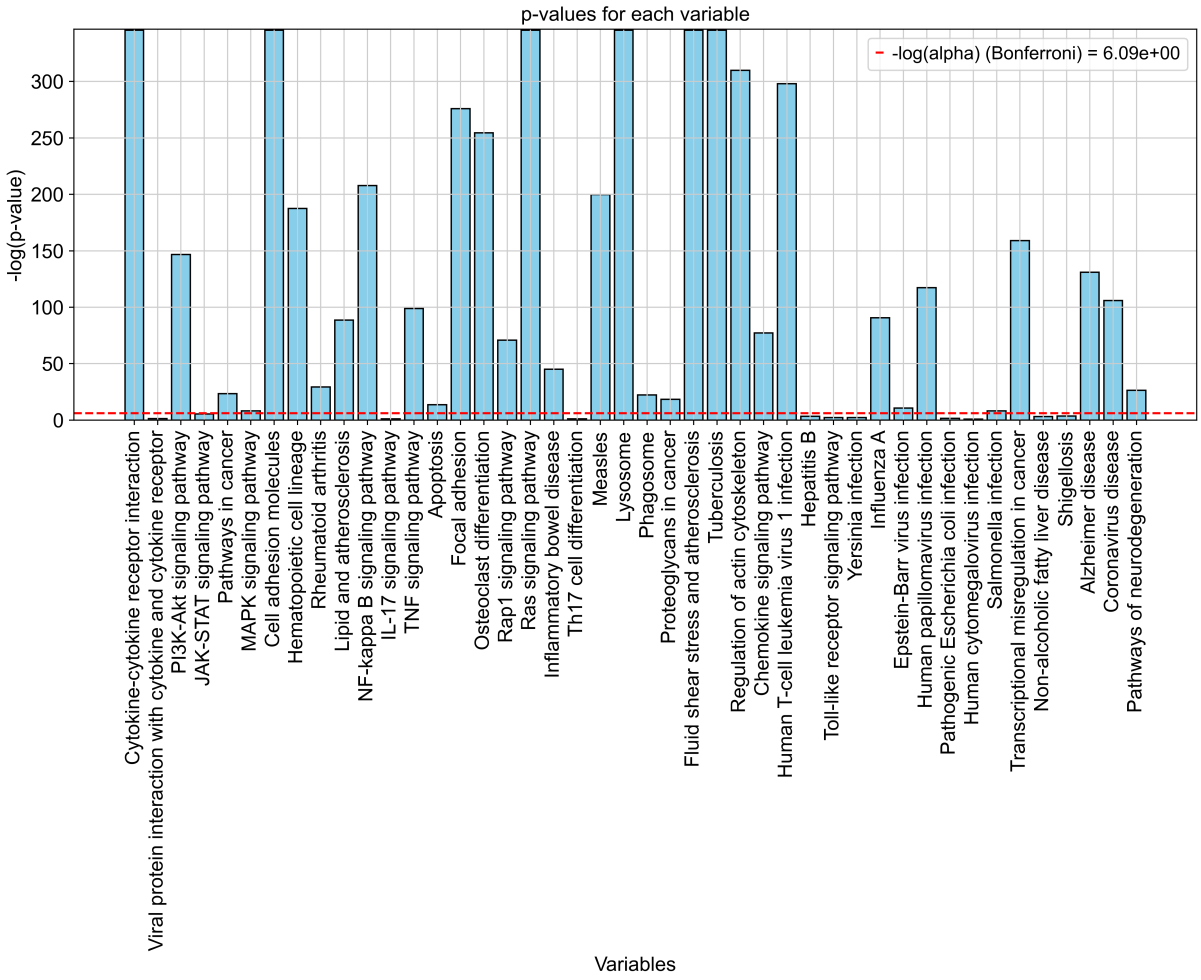

Figure 3.8: t-test p-values: to better visualize the difference at low order of magnitude we plotted the negative logarithm of the p-values and adjusted α with the Bonferroni correction for multiple testing

Both tests suggest some pathways that are significantly different between diabetic and healthy patients. The scope of this analysis is not to make any biological claims, but to understand if the variables extracted by our model significantly different in diabetic patients represent some biological processes involved in the disease. In particular, from Figure 3.9 we see some variables that stand out compared to the others:

- *Cytokine-cytokine receptor interaction* is a pathway responsible for mediating the communication among cytokines and their receptor, which contains proteins called **Suppressors of cytokine signaling (SOCS)** [24] that appear to have an impor-

tant role in the pathological processes leading to type 2 diabetes;

- *Fluid Shear stress* represents the frictional force that the flow of blood exerts at the endothelial surface of the vessel wall and plays a central role in vascular biology and contributes to the progress of *atherosclerosis*. Recent studies [21] have found evidence of correlation between diabetes mellitus and atherosclerosis through chronic inflammation;
- *Cell adhesion molecules (CAM)* are proteins present on the cell surface responsible for the interaction and attachment between neighbor cells. CAMs play a critical role in a wide array of biologic processes that include hemostasis, the immune response, inflammation, embryogenesis, and development of neuronal tissue. A direct connection between CAMs and type 2 diabetes has not been found yet, however a study by *Qiu et al.* [22] found a correlation between elevated circulating CAMs, especially intercellular adhesion molecule-1 (ICAM-1) and E-selectin, and an increased risk of diabetes.

Through these steps, we found strong evidence of the model's capability of extracting biological latent representation of the serum proteins expression, which suggested promising prospects for reaching the goal of this work, regarding the use of these representations to determine different subtypes of T2D.

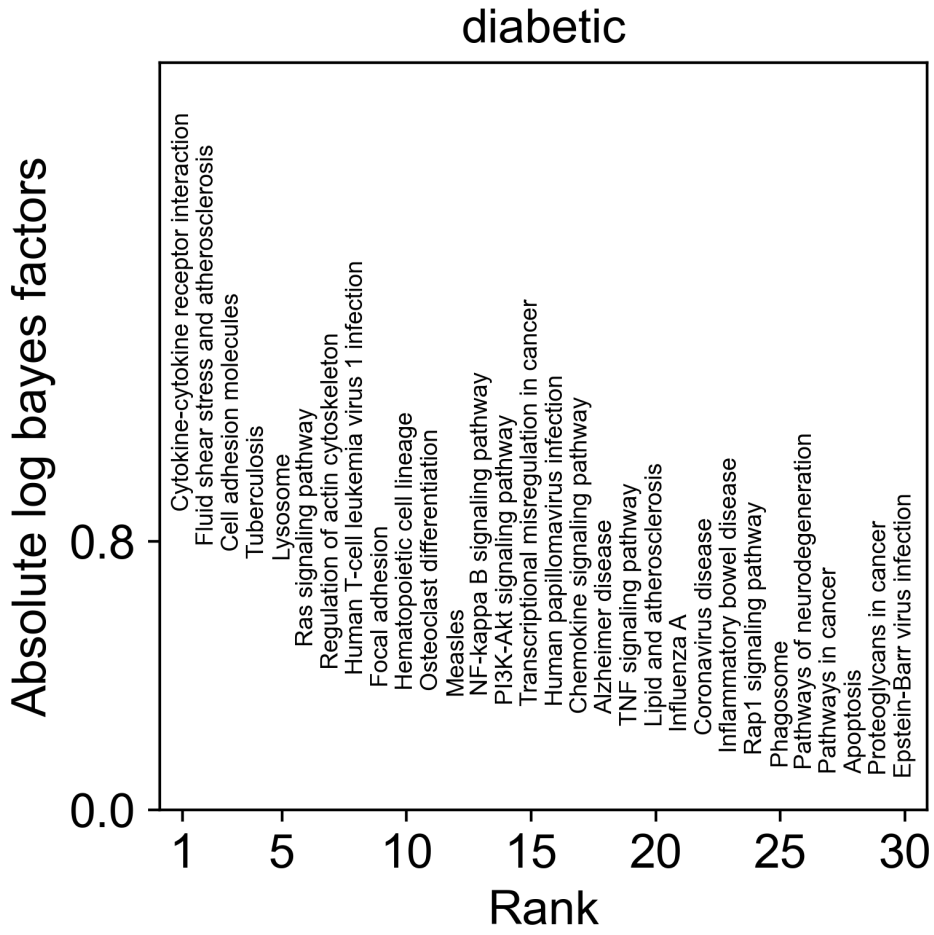

Figure 3.9: Bayes factors: this plot reports for each latent variable i the absolute log bayes factor $|\log \frac{p(H_{0,i})}{p(H_{1,i})}|$ where $H_{0,i} : Z_{i,diab} > Z_{i,health}$ and $H_{1,i} = H_{0,i}^{C}$. Higher values in the plot correspond to more significant difference between the distributions of healthy and sick patients for that variable.

3.3. Identification of T2D subtypes

In this section we present the results regarding the main scope of this thesis, which is the identification of different subtypes of T2D through the extracted latent pathway activation.

As described in Section 2.6, we present in Figure 3.10 the resulting dendrogram of the Ward Hierarchical procedure, along with the values of the silhouette scores computed for the number of clusters k varying in a range of 2 to 9 (Figure 3.11, Table 3.1), which indicates $k = 2$ as the best number of clusters.

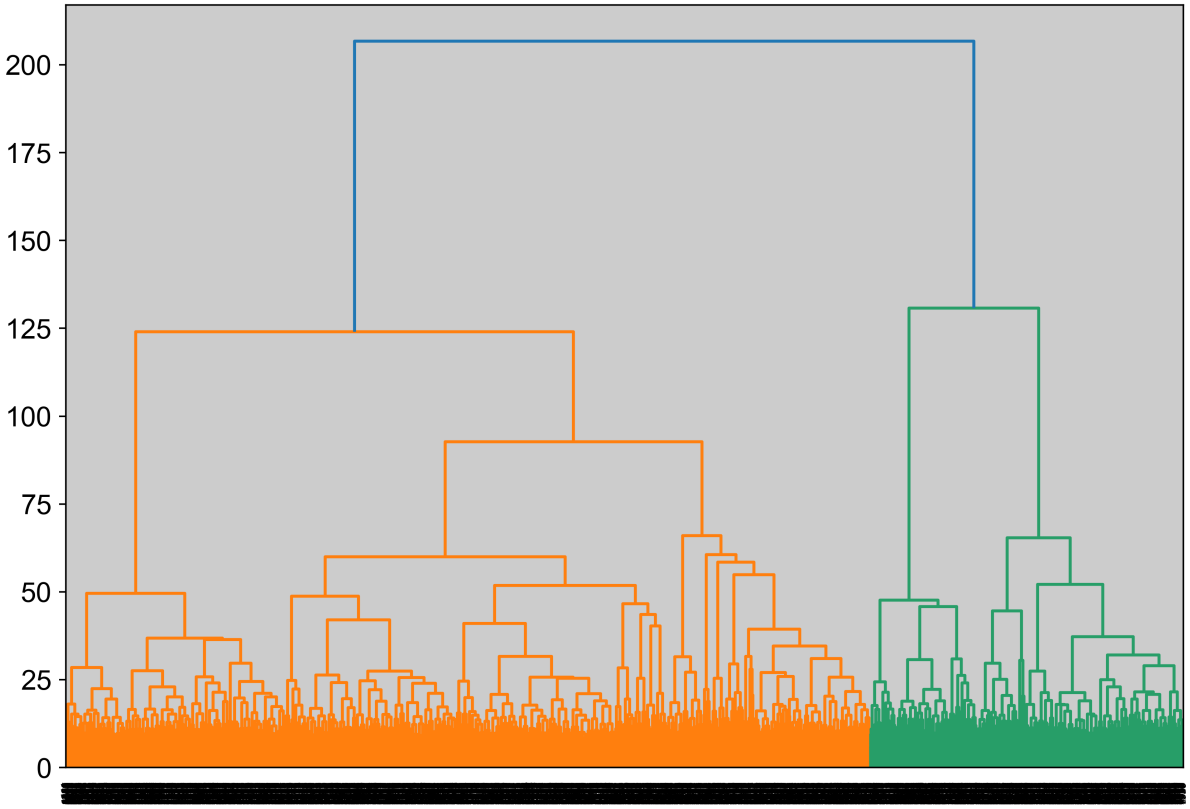

Figure 3.10: Dendrogram of the clustering process. The clusters corresponding to the optimal Silhoutte Score are coloured.

k	Silhouette Score
2	0.1053
3	0.0892
4	0.0323
5	0.0305
6	0.0328
7	0.0251
8	0.0272
9	0.0143

Table 3.1: Silhouette Scores for different values of k

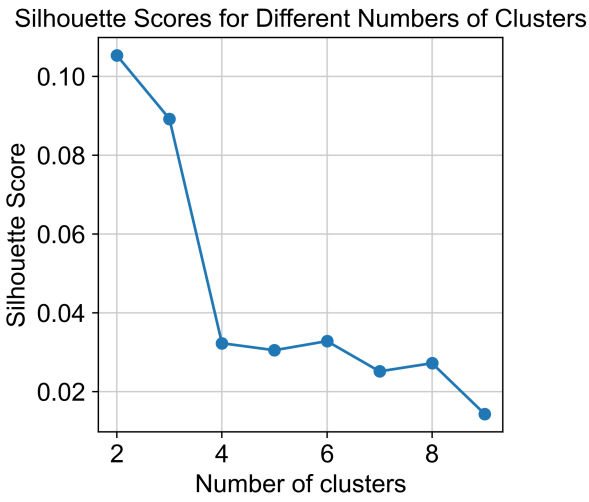

Figure 3.11: Silhouette Scores for different values of k

The subdivision found can be visualized by looking at the first two principal components of the latent features (Figure 3.12).

It is immediate to see a distinction between the two clusters along the first principal component, so a trivial way to spot the latent features that influence the most the membership to a given cluster is given by looking at the loadings of the first PC (Figure 3.13).

To further investigate the influence of the pathways for the clustering, we report the SHAP values of the Random Forest Classification.
Figure 3.14 shows a violin plot of the SHAP values for the classification in *Cluster 1* (in the case of a binary classification the values for the classification in *Cluster 0* are symmetric).

The plot gives insights on how the value of a feature influences the model in classifying every patients as part of the cluster 1, in particular we can see that high values of pathway dysregulation increase the probability of belonging to this cluster. From this analysis, it follows a first possible characterization of Cluster 1 with respect to Cluster 0: it identifies patients with more serious conditions, which could possibly lead to a worse progression of the disease. In particular among the most influential pathways we find once again *Cytokine - cytokine receptor interaction* and *Cell adhesion molecules* along with new ones like *Regulation of actin cytoskeleton*, *Apoptosis* and *Pathways in cancer*. While the last two pathways, taking part in a lot of biological processes, are likely to be noisy, it is interesting to look at the *Regulation of actin cytoskeleton* pathway, responsible for mediating various important cellular processes such as cell structural support, axonal

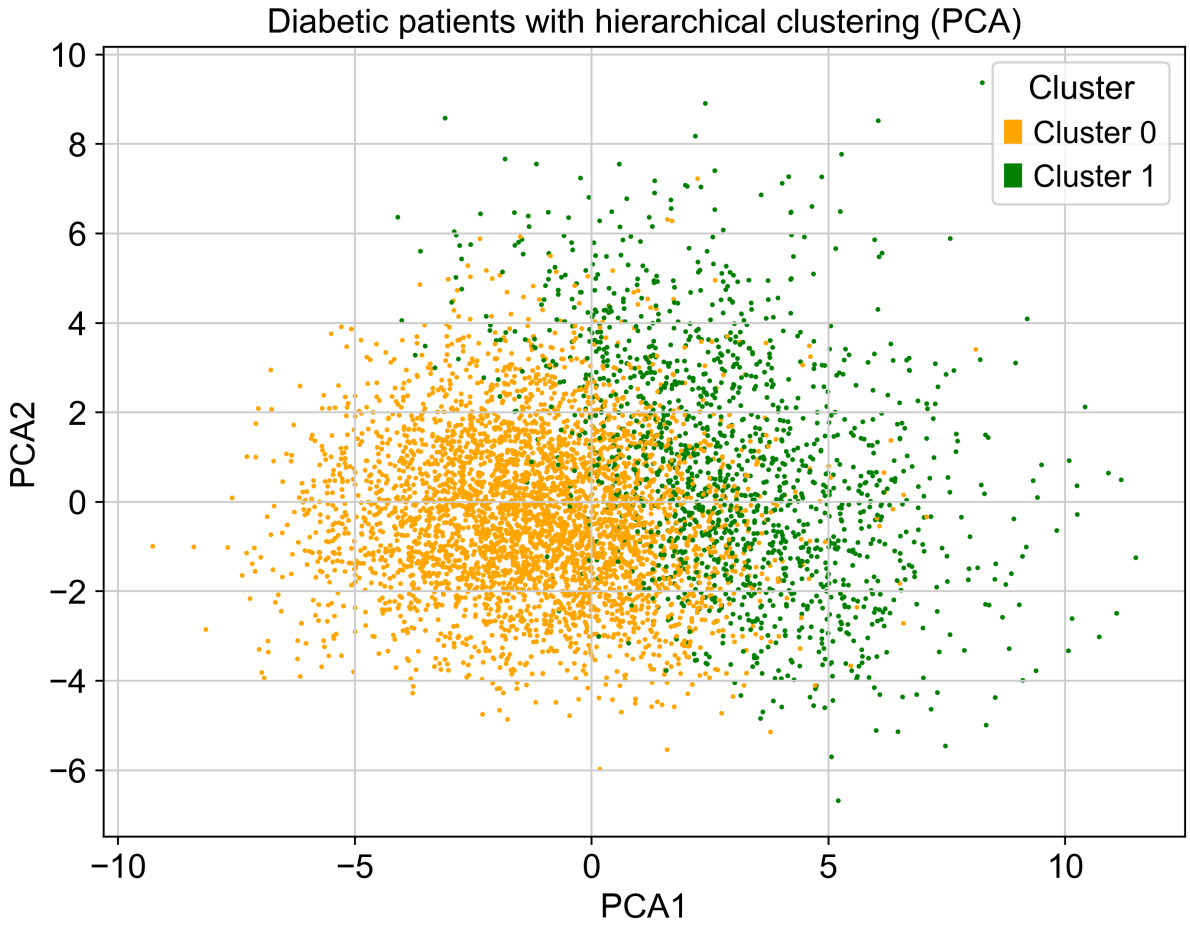

Figure 3.12: First two PC Scores coloured by cluster label

growth, cell migration, organelle transport and phagocytosis. The study [29] found that a defective regulation of the actin cytoskeleton contributes to insulin resistance, a key mechanism in T2D.

To conclude this work, we looked at the clinical picture of the two clusters found. The results of the analysis of the numerical (Figure 3.16) and categorical (Table 3.2) clinical variables confirm what we presumed by looking at the latent pathways: the patients in group 1 seem to have more serious conditions; in particular, they are more likely to have been diagnosed with obesity and hypertension, and also have been subjected to a higher number of treatments. In contrast, variables such as age, sex, and Polygenic Risk Score (PRS) are randomized between the two clusters, confirming that the subtypes likely depend on happening biological processes.

Feature	Cluster 0 Percentage	Cluster 1 Percentage	P-value
Sex_Male	0.602	0.589	0.36308
Obesity_diagnosis	0.074	0.114	0.00001
Obesity_BMI_Based	0.505	0.658	0.00000
Hypertension_diagnosis	0.560	0.658	0.00000

Table 3.2: Categorical clinical features comparison between clusters: the tests are ran for each category exploiting the large sample size and the Central Limit Theorem

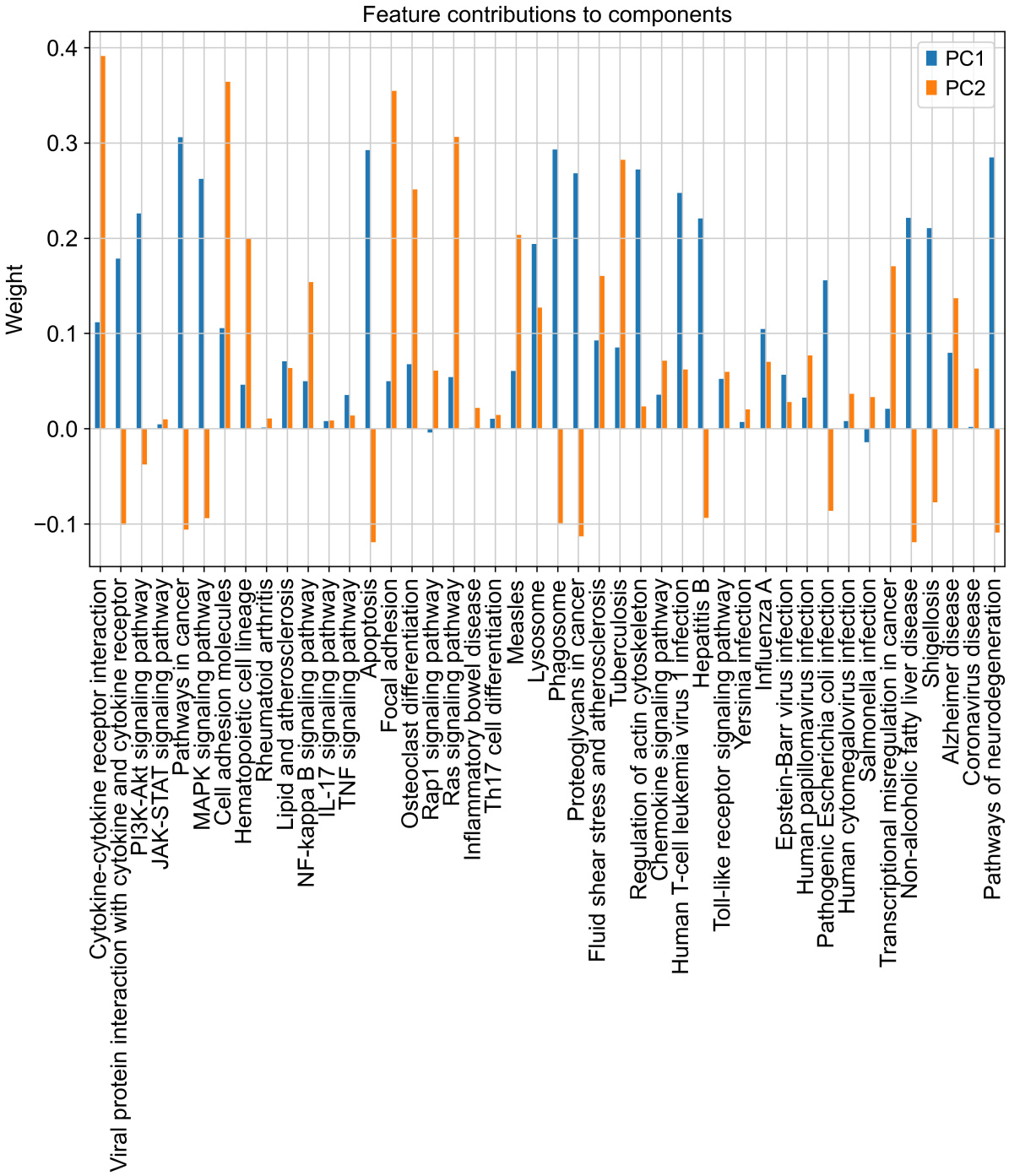

Figure 3.13: Loadings of the first two PC

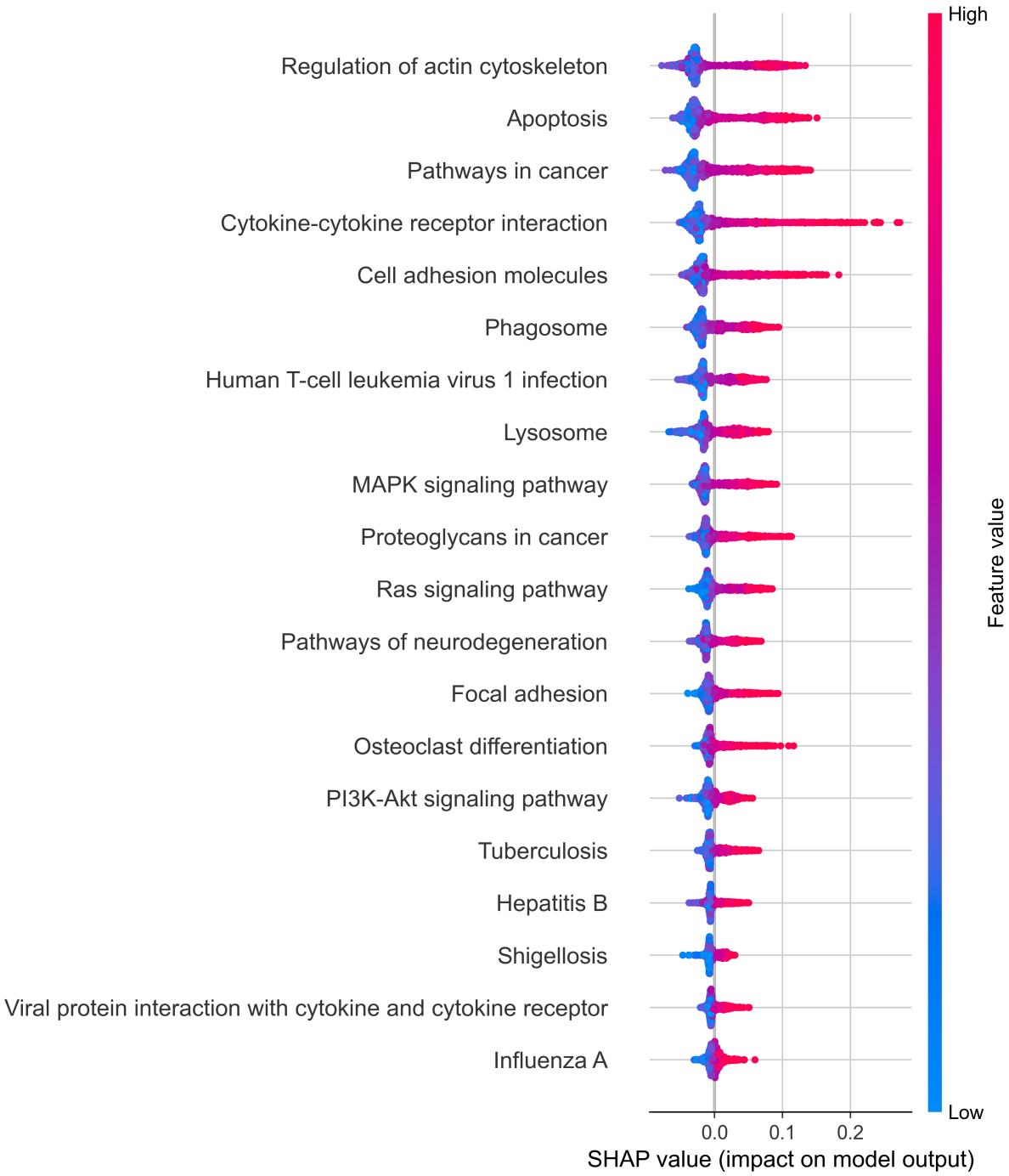

Figure 3.14: Violin plot of the SHAP Values: the color represents the value that the feature took on that specific patient, while the position represents the score of that feature for that patient. An high correlation between color and position in the plot represent an high influence of the feature in the clustering.

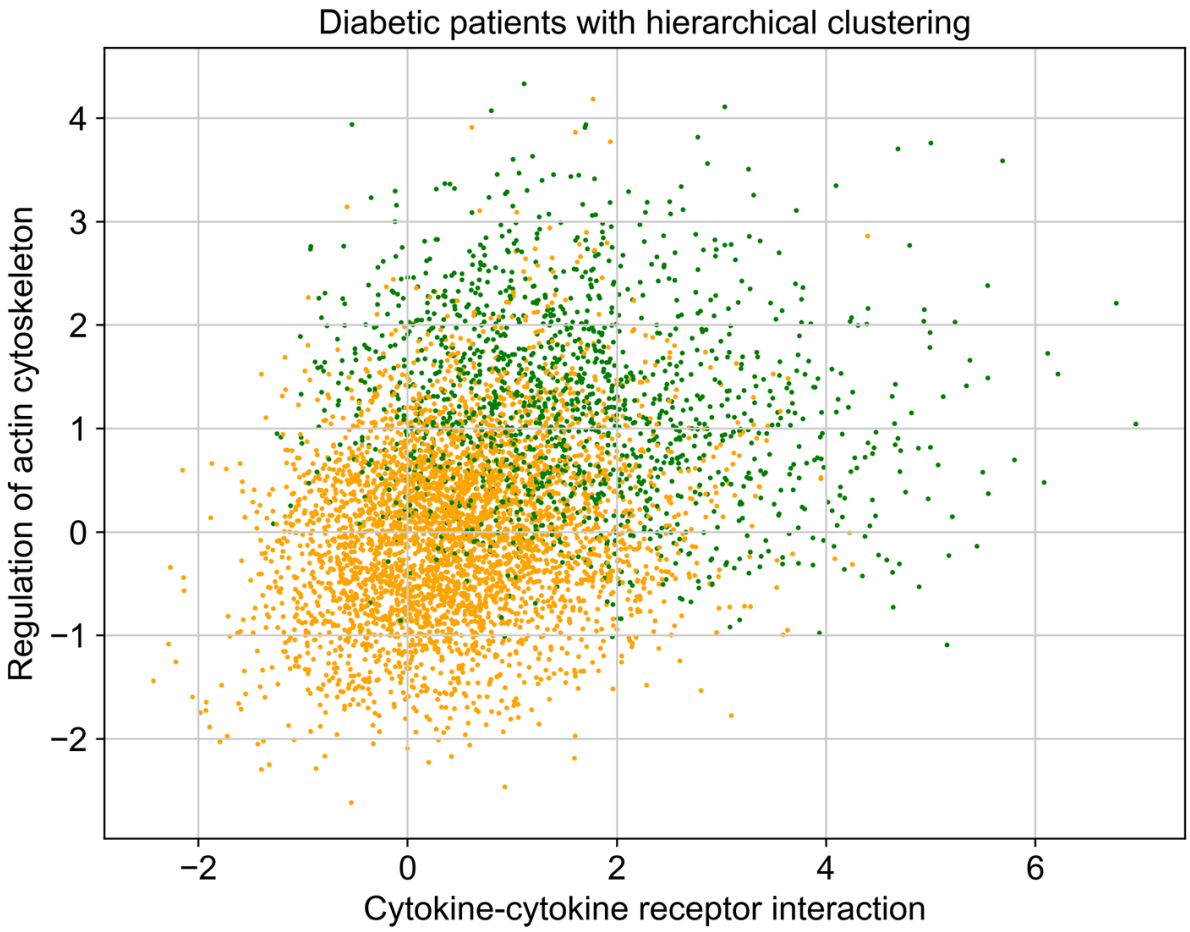

Figure 3.15: Scatter plot of the two most influent pathways for the clustering process

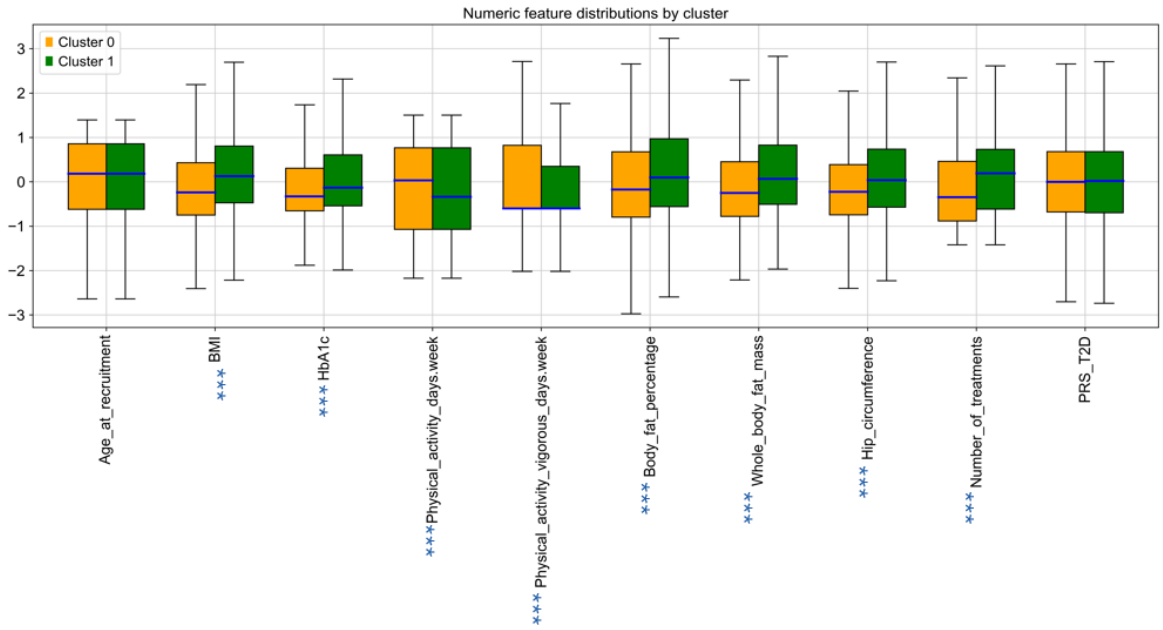

Figure 3.16: Boxplot of numerical clinical variables for each cluster: independent samples t-tests are ran for the mean of each variable between the two clusters and the p-values of such tests are represented by the asterisks
$(*** = p < 0.01, ** = p < 0.05, * = p < 0.1, \quad = p > 0.1)$

4 | Conclusions and future developments

In this thesis, we investigated possible subtypes of type 2 diabetes from a new perspective. In particular, we analyzed ongoing biological processes represented by protein expression within the blood serum. To do so, we introduced ProteoVAE, a biologically interpretable variational autoencoder, which projects protein expressions into a latent space manually controllable through prior biological knowledge inserted in a masked linear decoder. A crucial aspect of the model construction was to retrieve the biological information stored in the manually drawn protein pathways inside KEGG, of which we selected those of interest through statistical tests. We then performed a proof of concept to demonstrate that ProteoVAE was able to leverage this prior knowledge to capture, within the latent variables, the current biological processes occurring within patients, in particular those related to type 2 diabetes.
These latent features are then used in an explainable clustering pipeline, which has allowed us to identify two subtypes of T2D that belong to different risk levels, and understand which factors led to this particular distinction. In detail, patients grouped in cluster 1 are more likely to have additional diseases such as obesity or hypertension, presenting worse disease progression and increased risk, likely caused by inflammatory processes (inflammatory cytokines) and defective regulation of the actin cytoskeleton inside cells [8].

Despite good performance and achievement of the set goal, the model is subject to several limitations. Above all, the entire process is highly dependent on the chosen pathways, and an accurate selection is required beforehand. We still have to find an optimal way to perform this selection, in fact, the method involving the OverRepresentation Analysis is likely to favor pathways containing a large number of proteins and, consequentially, large noise that can mask the important information regarding the processes that we need to perform the clustering.
To tackle this issue in the future, we aim to perform a more accurate pathway selection to select "diabetes-informative" pathways through a literature review from the Reactome

database [20]. In this way our objective is to transform the main weakness of the model into one of its strengths by capturing only biologic processes that are informative to our research, reducing the noise and possibly improve the clustering process by identify additional subtypes.

Moreover, the two identified subtypes have been characterized with a limited number of clinical features; a further characterization with a larger panel could help to better understand the ongoing biological processes that lead to the found subdivision and predict the risks and consequences that belonging to one of these subtypes entails for the patients' disease progression.

Another possible future development will be to test the model in different cohorts. To do this we aim to once again leverage one of expiMap's characteristics [16], making ProteoVAE a conditional variational autoencoder (CVAE) in order to condition the reconstruction on the characteristics shared by the cohorts and also correct eventual batch effects.

In conclusion, this thesis proposed a novel approach to the study of T2D, taking advantage of the combination of cutting-edge deep learning architectures and recent biological knowledge to produce a subclassification of diabetic patients through ongoing biological processes that could provide a framework for further research in personalized medicine.

Bibliography

[1] E. Ahlqvist, R. B. Prasad, and L. Groop. Subtypes of type 2 diabetes determined from clinical parameters. *Diabetes*, 69(10):2086–2093, 2020. doi: 10.2337/dbi20-0001.

[2] S. Al-Amrani, Z. Al-Jabri, A. Al-Zaabi, J. Alshekaili, and M. Al-Khabori. Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, 12(5):57–69, 2021. doi: 10.4331/wjbc.v12.i5.57.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. URL `https://www.jstor.org/stable/2346101`.

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[6] Z. Fang, X. Liu, and G. Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 2022. doi: 10.1093/bioinformatics/btac757.

[7] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. doi: 10.2307/2340521.

[8] B. Hansson, B. Morén, C. Fryklund, L. Vliex, S. Wasserstrom, S. Albinsson, K. Berger, and K. G. Stenkula. Adipose cell size changes are associated with a drastic actin remodeling. *Scientific Reports*, 9(1):1–14, 2019. doi: 10.1038/s41598-019-49418-0. Article number: 12941.

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

[10] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. doi: 10.1037/h0071325.

[11] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906. doi: 10.1007/BF02418571. URL `https://doi.org/10.1007/BF02418571`.

[12] M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura, and M. Ishiguro-Watanabe. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1):D672–D677, 2025. doi: 10.1093/nar/gkae909.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980.

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014. doi: 10.48550/arXiv.1312.6114.

[15] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009. ISBN 9780262013192.

[16] M. Lotfollahi, S. Rybakov, K. Hrovatin, S. Hediyeh-Zadeh, C. Talavera-López, A. V. Misharin, and F. J. Theis. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25(2):337–350, 2023. doi: 10.1038/s41556-022-01072-x.

[17] M. Lundberg, A. Eriksson, B. Tran, E. Assarsson, and S. Fredriksson. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Research*, 39(15):e102, 2011. doi: 10.1093/nar/gkr424. URL `https://academic.oup.com/nar/article/39/15/e102/1024121`.

[18] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777. Curran Associates Inc., 2017. doi: 10.48550/arXiv.1705.07874. URL `https://arxiv.org/abs/1705.07874`.

[19] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[20] M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May, R. Petryszak, E. Ragueneau, K. Rothfels, C. Sevilla,

V. Shamovsky, R. Stephan, K. Tiwari, T. Varusai, J. Weiser, A. Wright, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 2024. doi: 10.1093/nar/gkad1025. URL https://doi.org/10.1093/nar/gkad1025.

[21] A. Poznyak, A. V. Grechko, P. Poggio, V. A. Myasoedova, V. Alfieri, and A. N. Orekhov. The diabetes mellitus–atherosclerosis connection: The role of lipid and glucose metabolism and chronic inflammation. *International Journal of Molecular Sciences*, 21(5):1835, 2020. ISSN 1422-0067. doi: 10.3390/ijms21051835.

[22] S. Qiu, X. Cai, J. Liu, B. Yang, M. Zügel, J. M. Steinacker, Z. Sun, and U. Schumann. Association between circulating cell adhesion molecules and risk of type 2 diabetes: A meta-analysis. *Atherosclerosis*, 287:147–154, 2019. doi: 10.1016/j.atherosclerosis.2019.06.908.

[23] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[24] S. G. Rønn, N. Billestrup, and T. Mandrup-Poulsen. Diabetes and suppressors of cytokine signaling proteins. *Diabetes*, 56(2):541–548, 2007. ISSN 0012-1797. doi: 10.2337/db06-1068. Review.

[25] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015. doi: 10.1038/nbt.3192. URL https://doi.org/10.1038/nbt.3192.

[26] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2 (28):307–317, 1953. doi: 10.1515/9781400881970-018.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[28] B. B. Sun, J. Chiou, M. Traylor, et al. Uk biobank pharma proteomics project (ukb-ppp) and associated publication: Plasma proteomic associations with genetics and health in the uk biobank. https://registry.opendata.aws/ukbppp, 2023. URL https://doi.org/10.1038/s41586-023-06592-6. Accessed: April 2025.

[29] P. Tong, Z. A. Khayat, C. Huang, N. Patel, A. Ueyama, and A. Klip. Insulin-induced

cortical actin remodeling promotes glut4 insertion at muscle cell membrane ruffles. *Journal of Clinical Investigation*, 108(3):371–381, 2001. doi: 10.1172/JCI12348.

[30] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

[31] F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0. URL `https://doi.org/10.1186/s13059-017-1382-0`.

A | Appendix A

Appendix A is dedicated to the pathway selection process, in particular we report the results of the clusterings performed with different thresholds (dendrograms and Silhouette Scores).

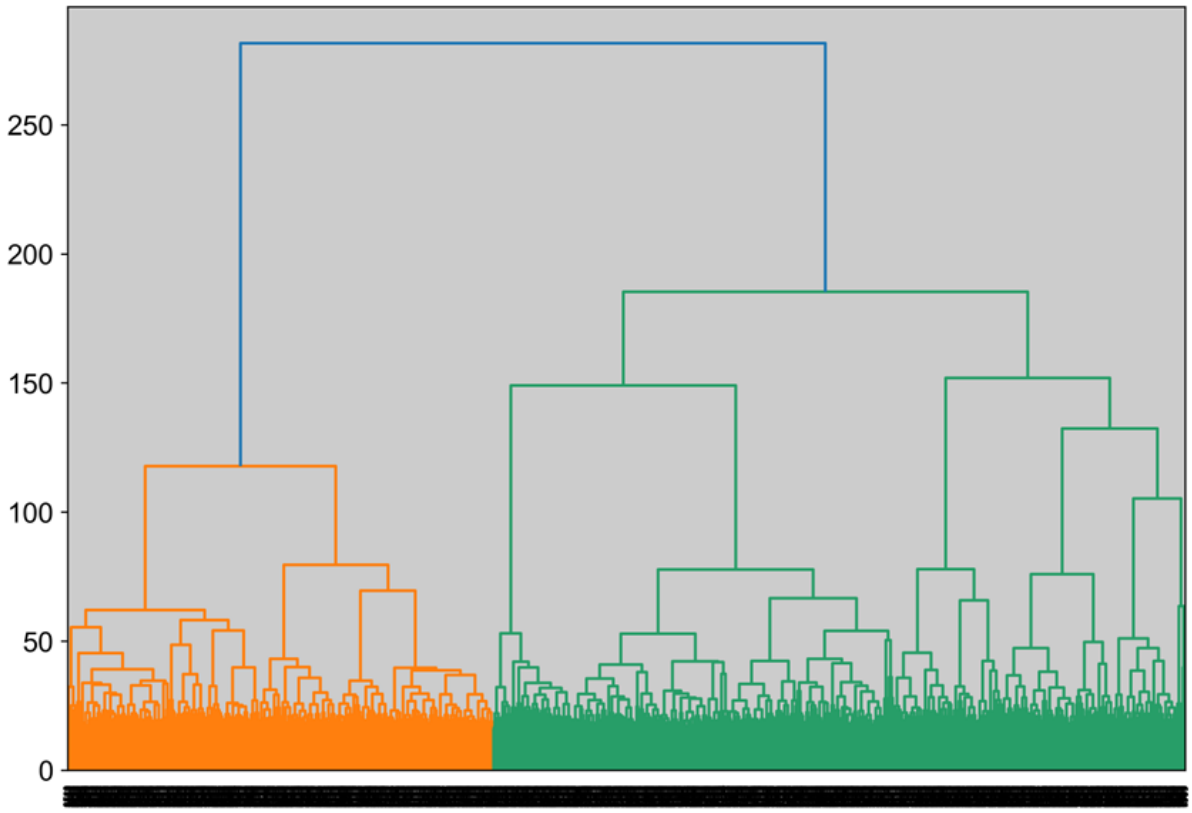

Figure A.1: Dendrogram of the clustering process with all the pathways.

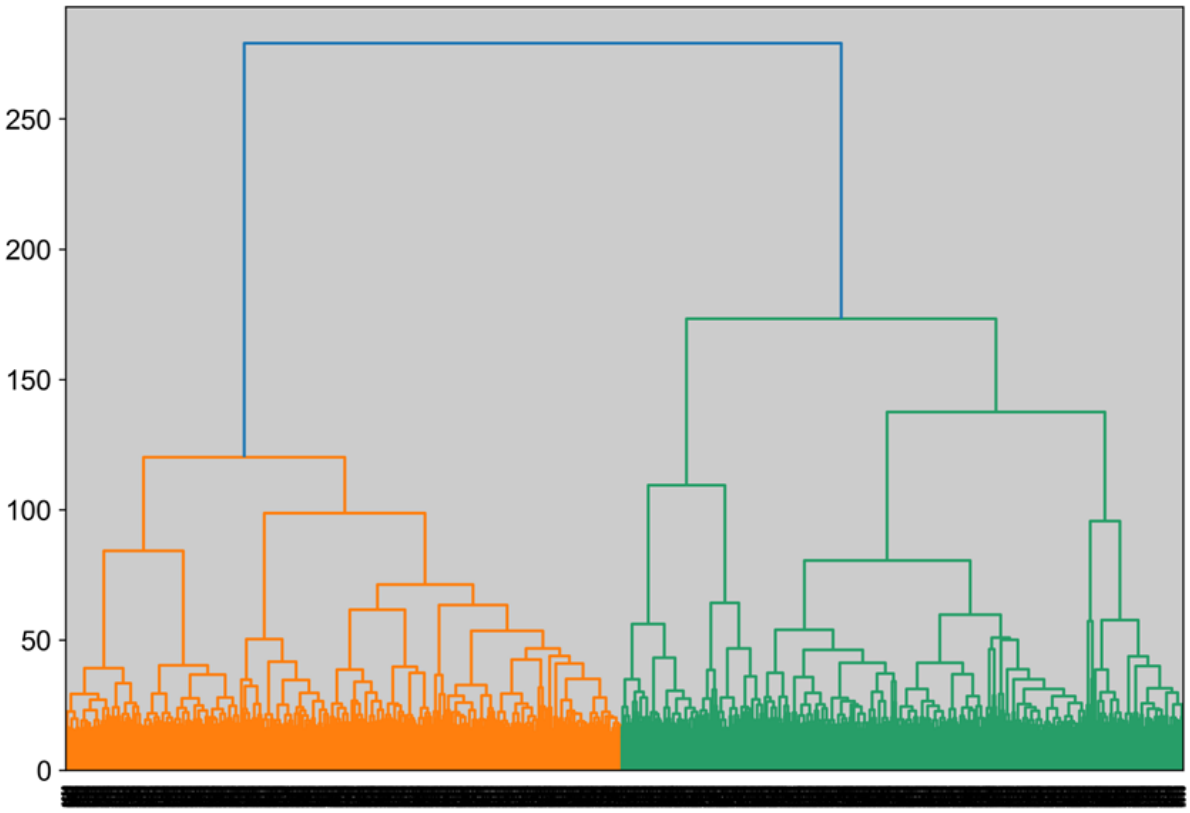

Figure A.2: Dendrogram of the clustering process with the top 75% of the pathways.

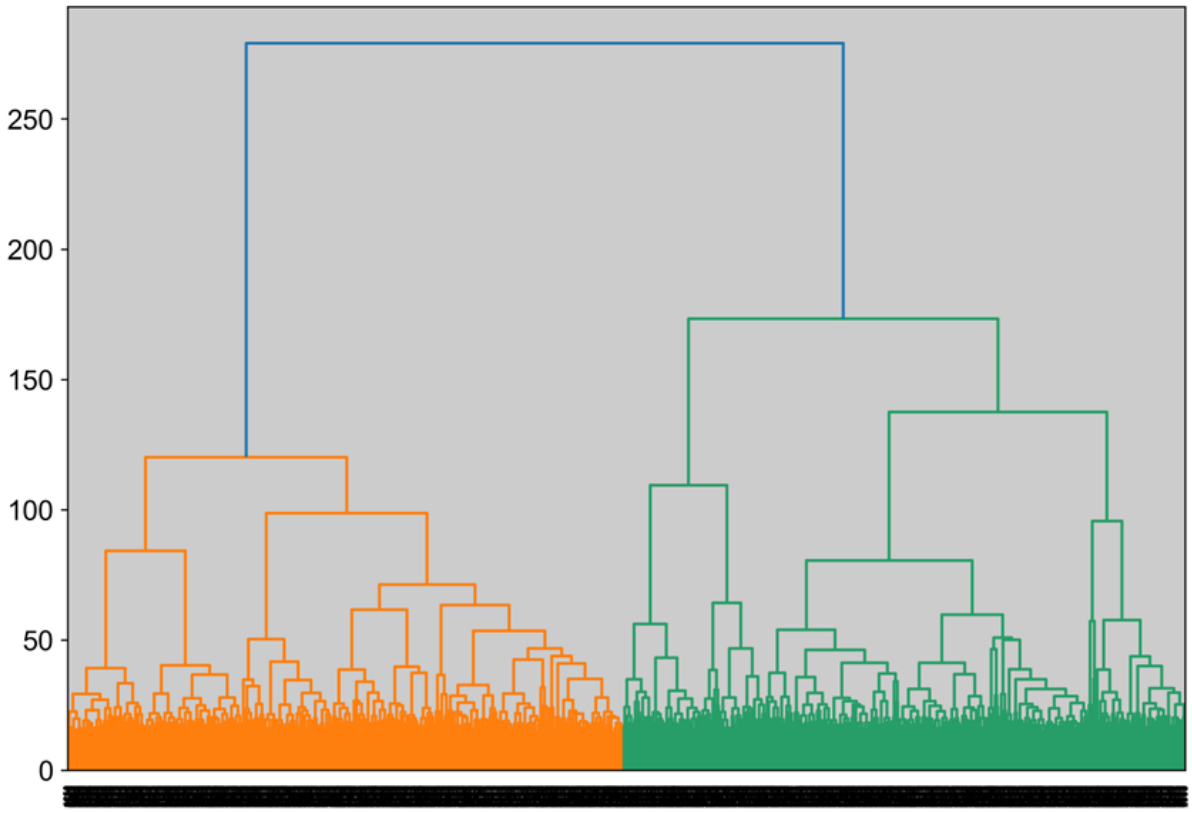

Figure A.3: Dendrogram of the clustering process with the top 50% of the pathways

k	Silhouette Score
2	0.0402
3	0.0303
4	0.0329
5	0.0193
6	0.0222
7	0.0143
8	0.0160
9	0.0100

Table A.1: Silhouette Scores for different values of k using all the pathways

k	Silhouette Score
2	0.0586
3	0.0520
4	0.0311
5	0.0141
6	0.0143
7	0.0140
8	0.0149
9	0.0082

Table A.2: Silhouette Scores for different values of k using the top 75% of the pathways

k	Silhouette Score
2	0.0737
3	0.0648
4	0.0276
5	0.0163
6	0.0179
7	0.0062
8	0.0065
9	0.0081

Table A.3: Silhouette Scores for different values of k using the top 50% of the pathways

B | Appendix B

Appendix B summarizes all the final training parameters of the model.

Parameter	Value
α	0.3
β	0.2
lr	0.01
lr decay	0.1
dr	0.05
epochs	400
early stopping patience	50
input dim	332
encoder dims	$[512, 256]$
latent space dim	44
activations	$ReLU(\cdot)$

Table B.1: ProteoVAE training parameters

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without Fondazione Human Technopole and the Health Data Science Center, which hosted me for the last eight months and provided the data on which all of the analysis was conducted, in addition to a welcoming environment that made me feel immediately part of the group. This incredible opportunity was given to me by prof. Francesca Ieva, who also was my official advisor at Politecnico and to whom I am sincerely grateful for her availability despite having to teach several courses and advise multiple theses. A huge thank you goes to Dr. Michela Carlotta Massi, who welcomed me from the very first day at the HDS center and supervised my entire work, providing extremely useful suggestions whenever I was stuck and putting up with my certainly improvable writing skills.

I also want to thank my parents for giving me the opportunity to study in Milan and for constantly supporting me since the first day of my life.

Finally, I want to thank all my old friends and those I made here in Milan (and during Erasmus in Stockholm) who made the university experience unforgettable. Hoping to not forget anyone I'll mention them by name: Emanuela, Giuseppe, Enrica, Arianna, Francesca, Andrea, Tommaso, Matteo, Giorgio and Daniela.