# Acute Myeloid Leukemia

Applied Statistics Project — Academic Year 2024/2025
*In collaboration with Human Technopole*

Bongiorni Noemi, Nordera Elisa, Redaelli Sara
Rusconi Irene, Zanin Rachele

June 30, 2025

**Abstract**

This study proposes a statistical framework for identifying minimal yet informative subsets of gene and protein expression markers that can accurately signal the presence of Acute Myeloid Leukemia, with the potential to generalize across other genetically influenced diseases.

## 1 Introduction

Acute Myeloid Leukemia (AML) is a severe form of blood cancer that originates in the bone marrow—the soft, spongy tissue at the center of certain bones responsible for producing blood cells. AML develops from the myeloid lineage of hematopoietic stem cells, which normally generate red blood cells, platelets, and most types of white blood cells, excluding lymphocytes.

The disease is marked by the uncontrolled growth and accumulation of immature myeloid cells, known as blasts, which fail to mature into functional blood cells. As these abnormal cells proliferate, they crowd the bone marrow and disrupt the production of healthy blood cells, leading to anemia, increased susceptibility to infections, and bleeding disorders.

AML progresses rapidly and requires immediate medical intervention. Despite progress in treatment—such as chemotherapy, targeted therapies, and stem cell transplantation—the prognosis remains poor for many patients, particularly older individuals and those with high-risk genetic or cytogenetic features. For this reason, scientific research continues to focus on understanding the molecular mechanisms that drive the disease and on developing more effective and personalized treatment strategies.

A significant area of current research involves the study of gene and protein expression in different bone marrow cell populations, as specific molecular alterations can signal the presence and development of AML. In recent years, various leukemia-associated marker genes have been identified, and their expression patterns are now used to support diagnosis and predict disease progression. These molecular markers offer valuable insight into the biological behavior of the disease and its potential response to treatment. However, they are not sufficient on their own to account for the complexity and heterogeneity of

AML, which can present with widely varying genetic mutations, molecular characteristics, and clinical outcomes. This diversity makes accurate diagnosis and risk classification particularly challenging. To address this, researchers are now investigating additional biomarkers and integrating multiple layers of biological information—such as transcriptomic, proteomic, and epigenetic data—to develop more precise and comprehensive tools for diagnosing and treating AML.

Aligning to these considerations and research topics, our project seeks to answer a key question: starting from a large set of genes and proteins expressed in bone marrow cells, is it possible to identify a smaller subset that reliably signals the presence of AML? The goal is to simplify complex molecular data while retaining the most meaningful biological and clinical information, thereby improving the clarity and effectiveness of diagnostic models. In addition to this primary aim, the project also explores whether the same methodological framework can be adapted to study other genetic influenced diseases. Many of such conditions remain poorly characterized and are often diagnosed late. By developing a model capable of recognizing disease-specific molecular signatures across different disorders, we hope to contribute to more accurate diagnoses, better prognostic tools, and potentially the discovery of new therapeutic targets for a broad range of genetic diseases.

# 2 Dataset Presentation

This project is based on two distinct datasets collected from 23 individuals, including 7 adult and 10 pediatric patients diagnosed with AML, as well as 6 healthy donors.

The first dataset contains scRNA-seq (single-cell RNA sequencing) data, consisting of 211,969 rows and 36,601 columns. Each row corresponds to a single bone marrow cell, while each column represents a gene. The value in each cell of the dataset reflects the expression level of that specific gene in the corresponding bone marrow cell.
The second dataset includes CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) data, focusing specifically on Antibody-Derived Tags (ADT). Like the scRNA-seq dataset, it includes 211,969 rows representing individual cells, but only 61 columns, each corresponding to a specific surface protein. The entries in this matrix indicate the abundance of each protein expressed in a given cell.

We start our analysis with the scRNA-seq dataset, applying a comprehensive workflow. Once all steps are completed, we move on to the ADT dataset, following the same approach. Since the scRNA-seq dataset has already been preprocessed to remove outliers and normalized, we begin directly with the exploratory analysis. Various visualizations are generated to provide an initial overview and to better understand the underlying structure of the data.

If we focus for example on the UMAP of patients 9, 12 and 22 (Figure 1),each representing a different group (pediatric AML, adult AML, and healthy donors), with colors indicating cell types, clear differences in the positioning of certain cell types across the groups are immediately noticeable.

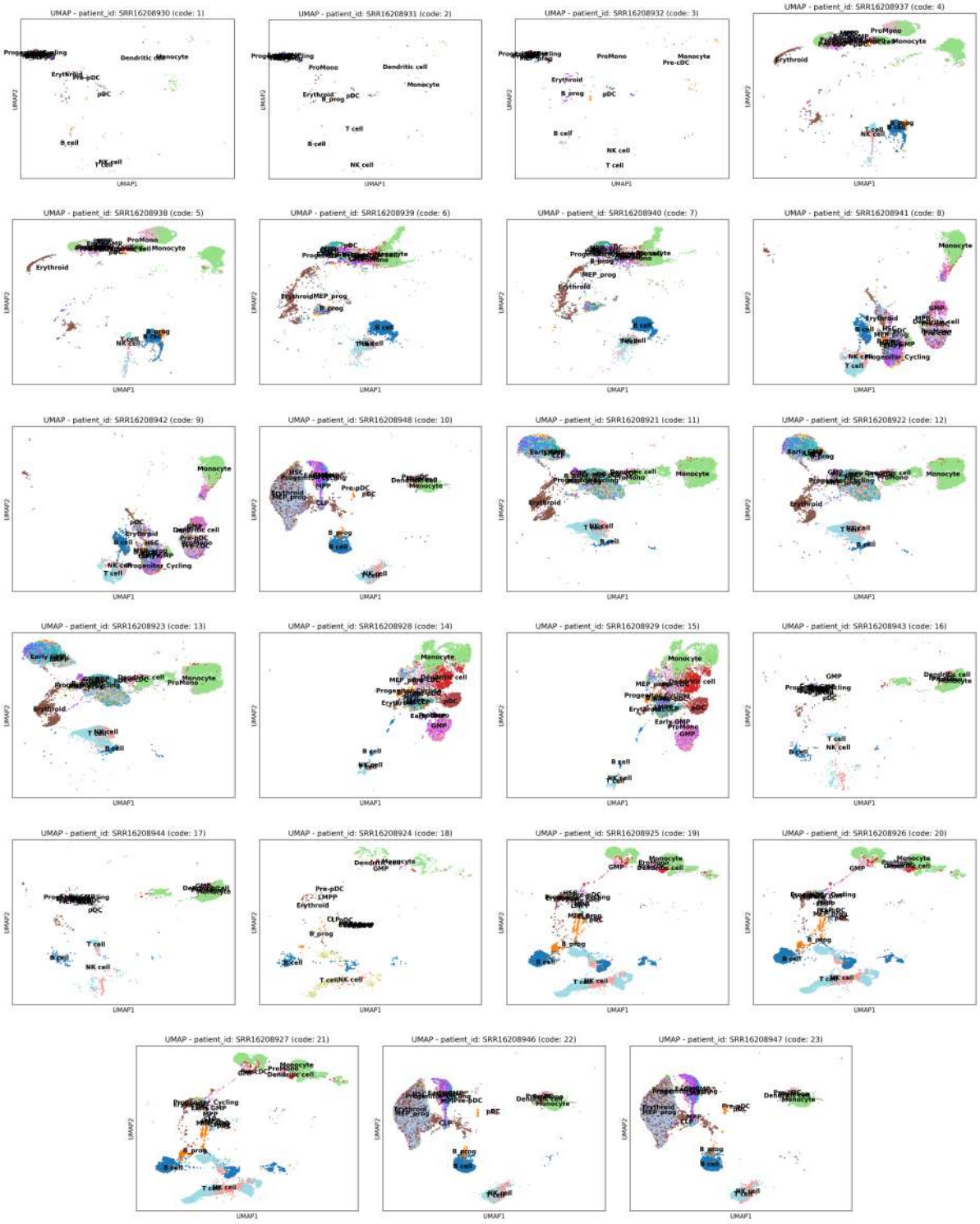


Figure 1: UMAPs of all patients, reporting their original ID, their new local number and their belonging group.

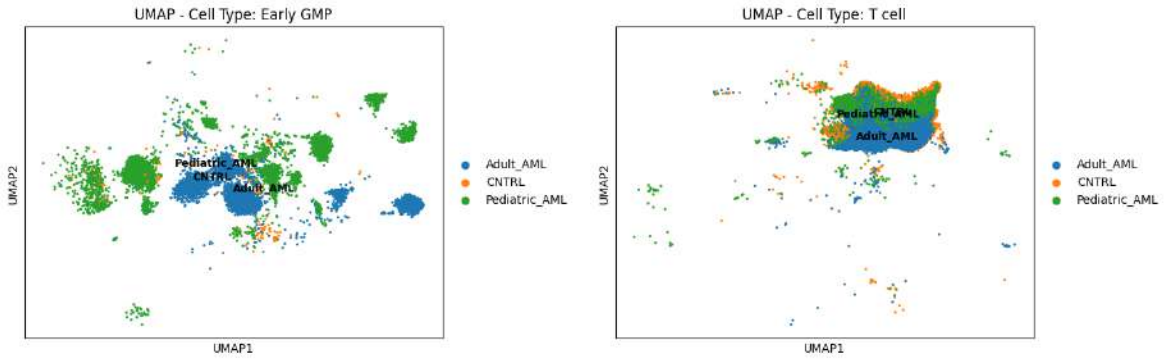


Figure 2: UMAPs for two indicative cell types.

In contrast, UMAP plots in Figure 2 show two different cell types, with cells colored according to group. The positioning of T cells appears relatively consistent across the groups, whereas the Early GMP cells display noticeable differences in their distribution among the three groups.

While reviewing the plots, we also notice that the amount of data available for each patient is not consistent, as some patients have much less data (patients 1,2,3). This lack of data affects our analysis. However, since the patients with the least data are mostly pediatric, the largest group, we still have enough patients with a reasonable amount of data.

Another notable challenge in the analysis is that multiple rows in the dataset may originate from the same patient and cell type. This compromises the assumption of independence between observations, which is essential for many statistical methods. To address this, we aggregate all rows belonging to the same patient and cell type by computing the median expression value. As a result, the dataset is reduced to 434 rows, each representing a unique combination of patient and cell type, which we can reasonably consider as independent units. The number of columns (genes) remain unchanged at 36,601.
To manage the high dimensionality of the data and prepare it for further analysis, we apply Principal Component Analysis (PCA). This method allows us to reduce the number of variables while preserving the majority of the dataset's variability.

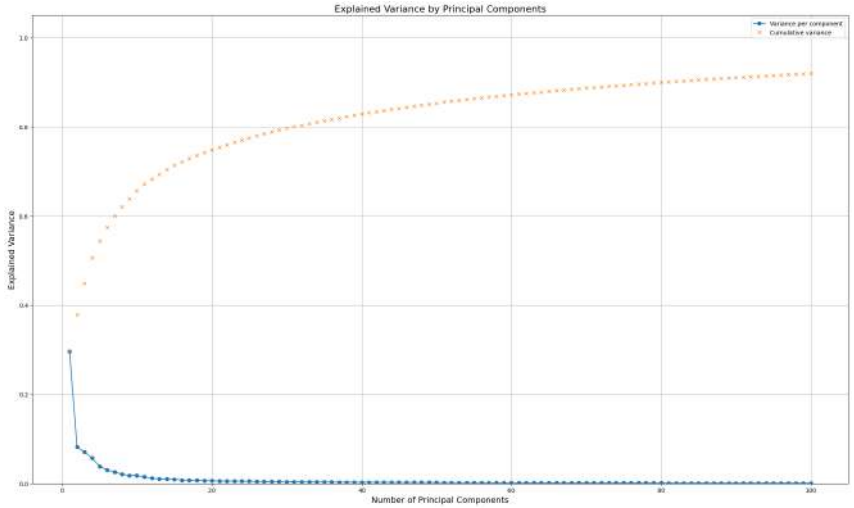


Figure 3: Cumulative variance explained by PCs.

The plot of the explained variance in Figure 3 shows that the first 100 principal components capture most of the variability; consequently, we decide to retain the first 30 components, which together account for approximately 80% of the total variance. This step simplifies the dataset while retaining its essential information. At this stage, the data matrix consists of 434 rows and 30 columns.

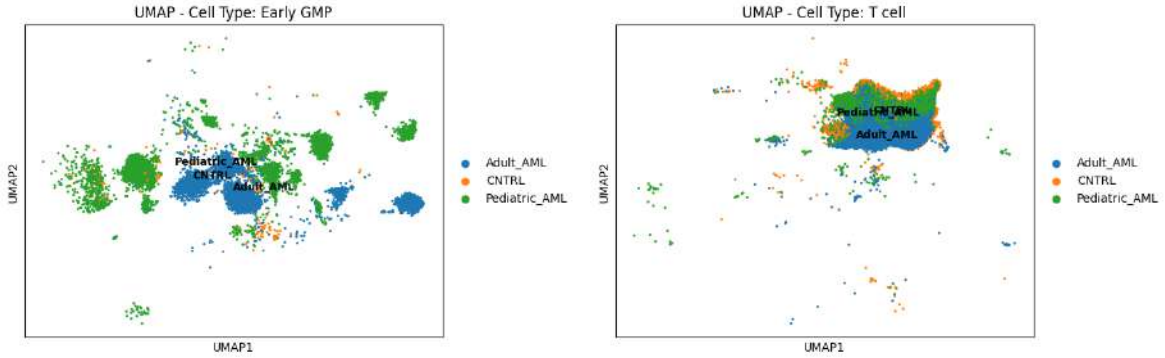


Figure 4: UMAPs for two indicative cell types.

In Figure 4, the UMAP visualization is shown in the new reference system defined by the first 30 principal components focusing on two specific cell types: Early GMP and T cells. These plots highlight how certain cell types, such as Early GMP, exhibit a broad spatial distribution, indicating potential discriminative power in distinguishing between healthy and ill patients. In contrast, other cell types, such as T cells, appear tightly clustered, suggesting limited utility for classification purposes due to the lack of separation between patient groups.

In Figure 5 it is possible to see the same UMAP plotted before for all patients, using however the new reference system obtained by the PCA analysis. It is important to emphasize once again the varying availability of data across patients, which leads to a more precise and reliable analysis for those patients with a larger amount of available data.

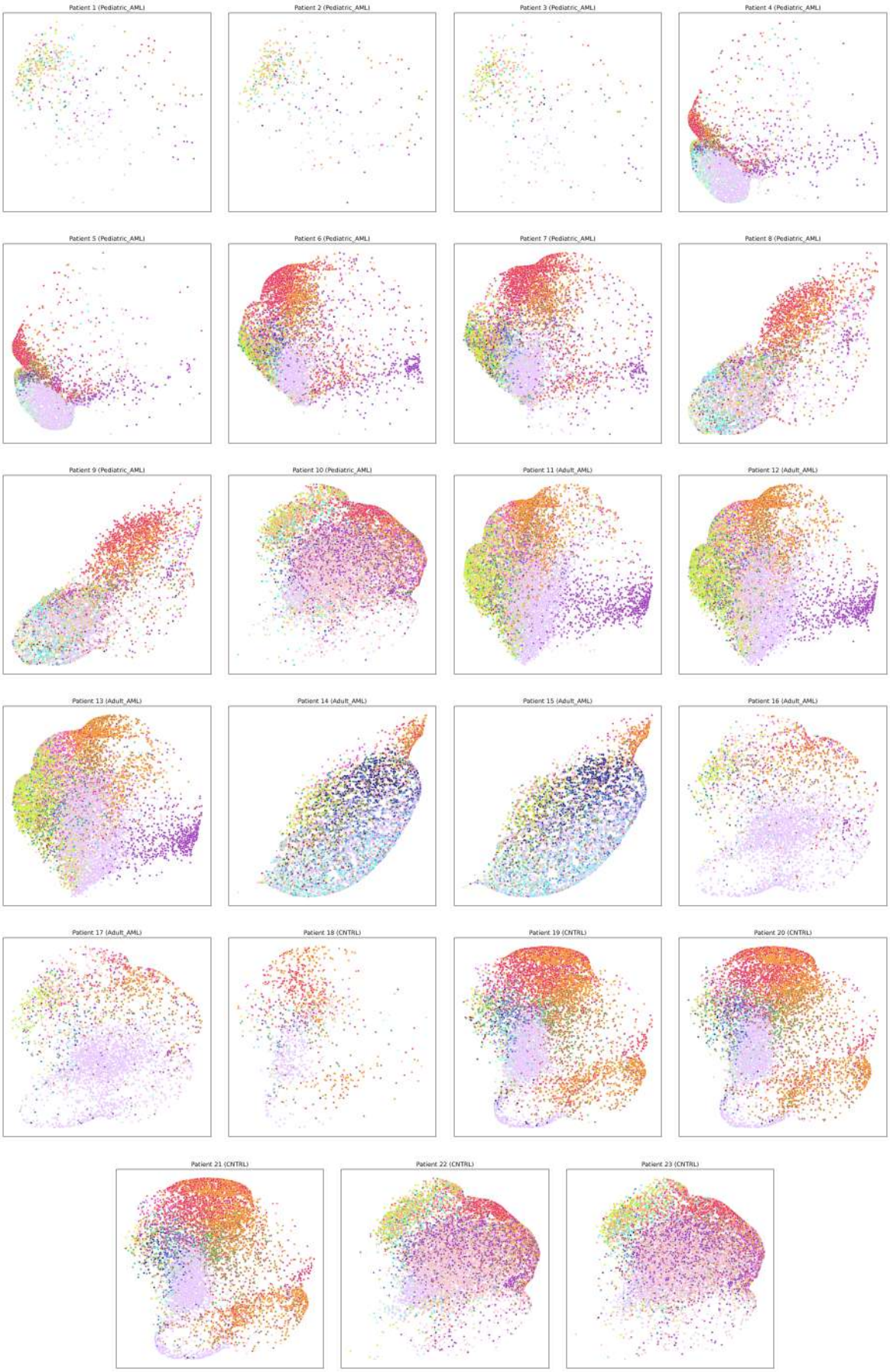


Figure 5: New UMAPs of all patients plotted on the space given by the first 30 PC, reporting their original ID, their new local number and their belonging group.

# 3 Approach to the problem

After PCA has been performed, the preprocessing part comes to an end and we have a new dataset on which we can develop further analysis. In the detail, we decide to adopt two parallel approaches: on the one hand we conduct an unsupervised analysis, as the first dataset we have been provided with does not include the target variable, on the other hand, after receiving the information about the health status of the patient, we focus on a supervised analysis.
In the unsupervised part our aim is to cluster the patients based on the similarities between their cells' gene expression, while in the supervised part the goal is to create a classifier able to assign each patient to the correct group.

Note that, even though the tools used in the unsupervised analysis do not require us to know the true label about the health status of the patients, we decide to take advantage of this knowledge, using it to value the goodness of the methods adopted through the computation of the accuracy.

# 4 Unsupervised Analysis

The starting point of the unsupervised path is the choice of the clustering technique to use. We try different methods - K-Means, hierarchical clustering with different metrics and linkages, DB Scan and Multi Dimensional Scaling - and, as previously mentioned, compute the accuracy of each of them to find the most suitable tool.

## K-Means

First of all, we perform K-Means clustering on the grouped dataset.
Since we know that patients belong to 3 different groups - adult AML patients, pediatric AML patients and control group namely - we set as 3 the number of clusters, so that for each cell type we can detect a cloud for every health status.

We plot the clusters and compare them with the true labels: as also the computation of the accuracy confirms, we notice that K-Means obtains quite poor results as can be noticeable from the cluster's picture and the accuracies computed (less the 0.50 in most of the cases).
This accuracy is computed associating in the most efficient way the three clusters obtained with the three groups.

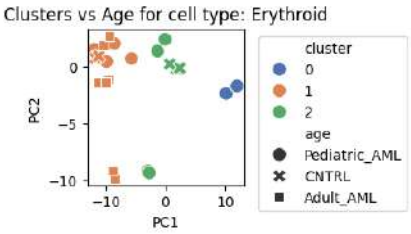


Figure 6: Clusters for the cell type 'Erythroid'.

## Hierarchical clustering

We then move to another clustering tool to investigate whether we could achieve better results.
Since in these types of tasks distances play a crucial role, we focus on finding the best metric and the best linkage to properly cluster the cell. Hence, after grouping the dataset by cell types, we test 10 different distances (Euclidean, square-Euclidean, Manhattan, cosine, correlation, Hamming, Jaccard, Chebyshev, Canberra and Braycurtis) combined with three linkages (complete, average and single). We evaluate all the possible combinations by computing the accuracy and we plot some heatmaps to better understand the results.
What we deduce from these plots is that the combination which performs the best is the one given by the Canberra distance and the complete linkage, which obtains for most of the cell types an accuracy above 0.545.
We also make some separate plots using the Canberra distance and the three types of linkages to visualize the differences between the clustered cells and the true labels.

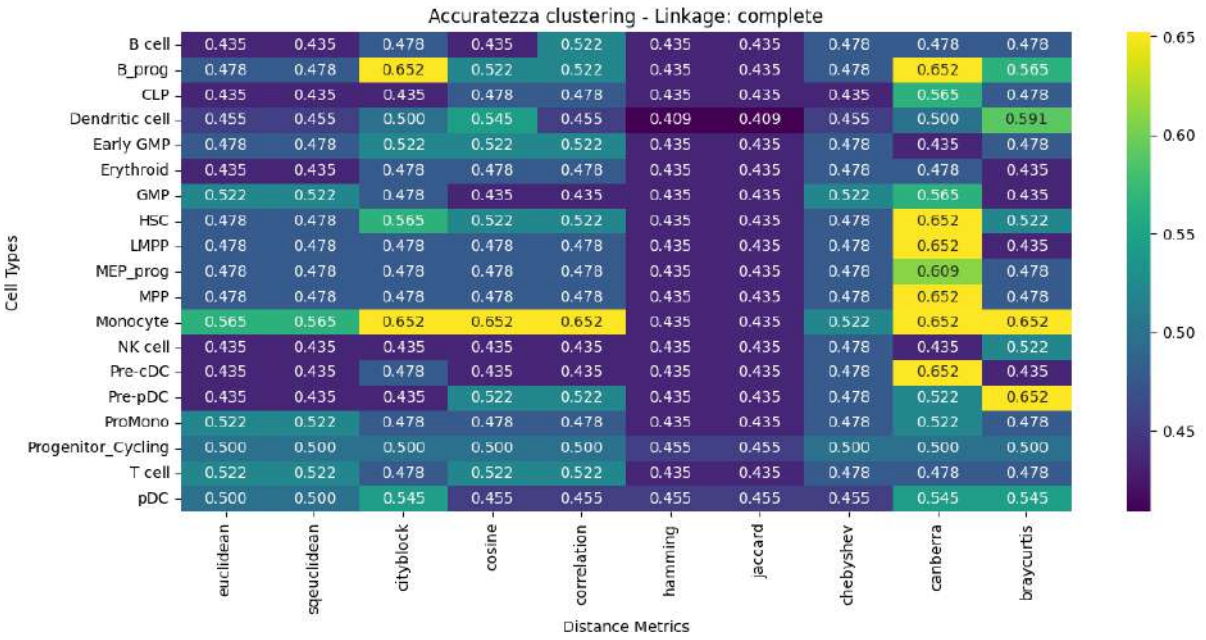


Figure 7: Accuracy obtained for each cell type using several metrics and complete linkage.

Up to this point, it seems to us quite clear that Canberra distance and complete linkage combined together could lead to satisfying results. Hence we decide to evaluate how this hierarchical distance is faithful with respect to the true distance between the cells belonging to the same type.

To reach this purpose we compute the cophenetic distance and the cophenetic correlation coefficient (CPCC): this index, which ranges between 0 and 1, measures the goodness of the dendrograms used in clustering.
In our case almost all the clusterings score higher than 0.838, with peaks of over 0.909 and the lowest CPCC was equal to 0.726. These relatively high values suggest that the hierarchical structure induced by complete linkage on Canberra distance retains much of the original dissimilarity information.

Additionally, to gain a more nuanced understanding of the clustering behavior, we plot a matrix comparing the original Canberra distances to the corresponding cophenetic distances. Points lying on the bisector indicate a faithful representation, while points above the bisector suggest that a given Canberra distance corresponds to a higher cophenetic distance—i.e., the clustering tends to overestimate the distance between merged groups at certain levels of the hierarchy.
These results, particularly when examined through the comparison matrices, confirm that the Canberra distance and the complete linkage are able to coherently capture the true distances between the samples.

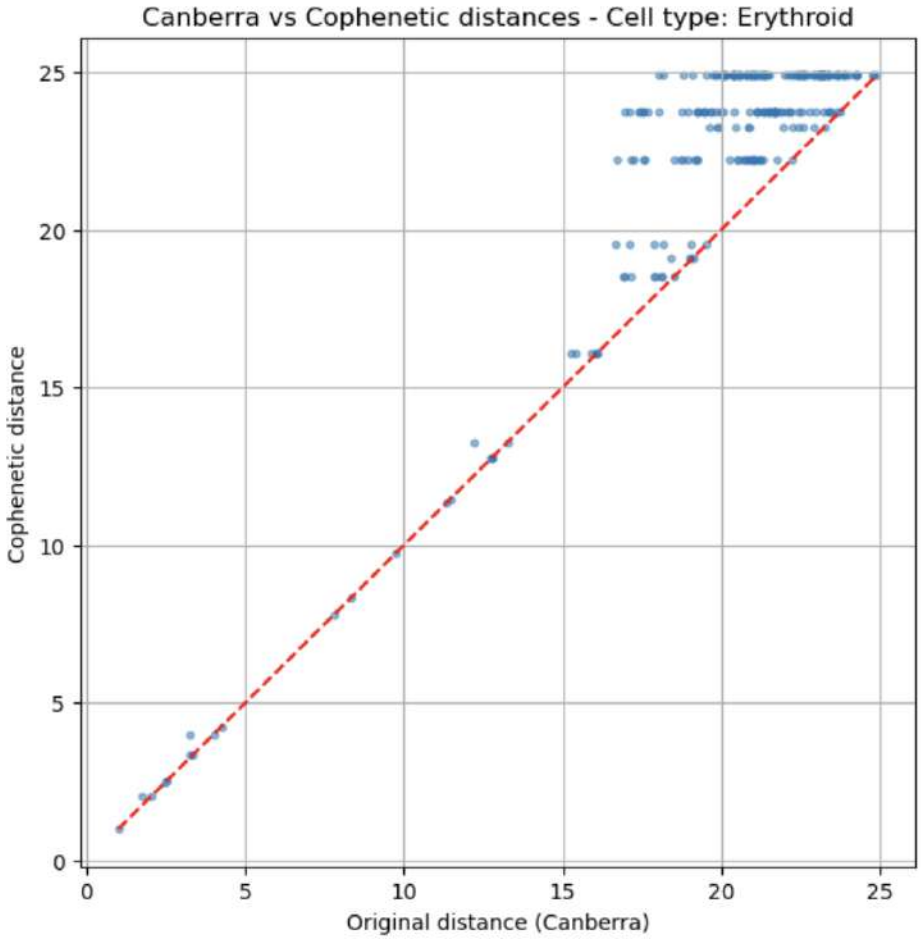


Figure 8: Comparison between Canberra and Cophenetic distance for cell type 'Erythroid'. In this case CPCC is equal to 0.901.

## DB Scan

We take a further step in the exploration of the clustering techniques by developing DB Scan clustering, which has been developed to find regions with a high density of data.
We try to implement this technique with different values for the parameters epsilon and minimum number of samples, but we always use the Canberra metric. However, probably due to the differences in the gene expression of the cells, this technique leads to rather unsatisfying conclusions, as the accuracy for the cell type in which DB Scan best worked is only 0.545.
As we believe we could obtain higher values in the accuracy, we decide to add a Multi

Dimensional Scaling step before the DB Scan in order to reduce the dimensionality of the data to n = 2, but there is no significant increase with respect to what we have already done.

After making all these trials, we come to the conclusion that hierarchical clustering performed through the Canberra distance and the complete linkage is the best choice for our analysis.

## Assembling the results

At this point of the work, we have 19 separate results - one for each cell type - which cluster the data into three groups which do not necessarily coincide, i.e. it is not granted that group 1 in the B cell clustering corresponded to group 1 in the Monocyte clustering. However, since clustering independently of each specific cell type is not enough to determine the group to which a patient belongs to, we need to find a way to combine these results.

A possible idea is to work on the similarities between patients in order to assign them to a specific group.
First of all, we assemble the results of the 19 clusterings into one matrix which, for every patient, keeps track of the cluster each of his/her cells had been assigned to. Then, we compare all the possible couples of patients: we fill a new matrix with 0 and 1 depending on whether, for a specific cell type, those two patients belong to the same cluster. We do this for all the different cell types, obtaining 19 separate matrices which can quantify in some way the closeness between the patients.
After that, we make a weighted sum and thus we obtain a new square matrix in which the element in position (i,j) tells how much the patient i and patient j are similar, based on all the cell types involved in the analysis. More in detail, we define the weights according to literature, because, since we want to detect groups of ill people, we want to give more importance to those cells which discriminate between the presence and absence of AML. So *Early GMP*, *GMP*, *LMPP*, *MEP_prog*, *MPP* and *ProMono* are given the highest weight, *Monocyte*, *Pre-cDC* and *Pre-pDC* are given a medium weight, while all the other cell types are given the lowest weight.

Now we have a matrix which provides us with a new kind of distance between patients, so we make a final cluster to divide the patients into three groups. We try to apply two different methods.
In the first case, starting from our matrix, we normalize the data and build a dissimilarity matrix by subtracting these values to the identity matrix. We then set the diagonal to 0 and then perform the clustering. We do this because for some patients not all cell types are available, which results in non-zero values along the diagonal that are not meaningful. Therefore, we manually set these values to zero for consistency, ensuring that this information is not misleading. When we compute the accuracy, however, the value we obtain is not brilliant (0.478), which leads us to develop a second method.
In the second case, after building the dissimilarity matrix we apply Multi Dimensional Scaling to retrieve Euclidean coordinates and then we perform K-Means clustering setting as 3 the number of clusters we want to have. By comparing the predicted labels with the actual one, we obtain an accuracy of 0.739. We construct the confusion matrix and what we discover is that this algorithm is able to correctly classify all the adult AML patients, while it misclassifies in total 6 subjects belonging to the control and pediatric

AML groups. In particular, pediatric patients 6, 7, 8, 9 are classified as control, while control patients 22, 23 are classified as pediatric. If we look at the gene profile of the misclassified patients, we can see that their cell distributions are quite similar one to another, so it is reasonable to think that it could be hard to discriminate between the two classes.

### Conclusion

We are aware of the fact that an unsupervised approach is less powerful and leads to less robust results than a supervised one, but we believed that it could be useful for the sake of generalisation. Indeed, in case of similar scenarios or whenever you deal with a genomic dataset without target and coming from samples in different groups, it is possible to follow an analogous workflow, building a sort of "similarity distance" that allows to detect correspondences between the gene expressions of different subjects.

## 5 Supervised Analysis

As previously mentioned, the main goal of the supervised part is to build a classifier able to distinguish between the different groups present in our dataset. We consider and compare two classification techniques: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

Both of these methods assume that, for each class, the data follow a multivariate gaussian distribution, which is satisfied by our dataset. The principal theoretical difference in the two techniques lies in the covariance structure of the data. Specifically, LDA assumes that all the classes share the same covariance matrix, whereas QDA allows each group to have its own. We perform Box's M test on our three groups and, based on the results, QDA appears to be more suitable for our dataset. Consequently, we apply QDA to classify the 434 cells, achieving a quite satisfactory accuracy of 0.9080.
Including prior probabilities and misclassification costs in the QDA model would be very meaningful and interesting, in particular to assign a higher cost to classifying an AML patient as a healthy one, an error that is commonly recognized as more severe due to its clinical consequences. Unfortunately, we lack sufficient information to estimate in a reliable way prior probabilities and misclassification costs, so we proceed with a standard QDA, obtaining still a strong performance.

We also consider another method for the supervised analysis: Support Vector Machines (SVM). SVMs work by finding the optimal hyperplane that best separates the data into distinct classes, and they are one of the most widely used supervised learning methods. Using SVMs, we obtain an accuracy of 0.8626, which is slightly lower than the accuracy achieved with QDA (0.9080).

At this point, we have two main options: either select one of the two models based on their performance, or combine them using ensemble methods to potentially improve the overall accuracy. For the sake of simplicity and interpretability of the results, we choose to rely on a single model and decide to continue our analysis using QDA, which has shown the best performance in our previous evaluations.
To provide a more general classification reflecting each patient, instead of all the cell types of every patient, we aggregate the predictions by weighting each cell type according to its

influence on the presence of the illness, as previously done in the unsupervised analysis.

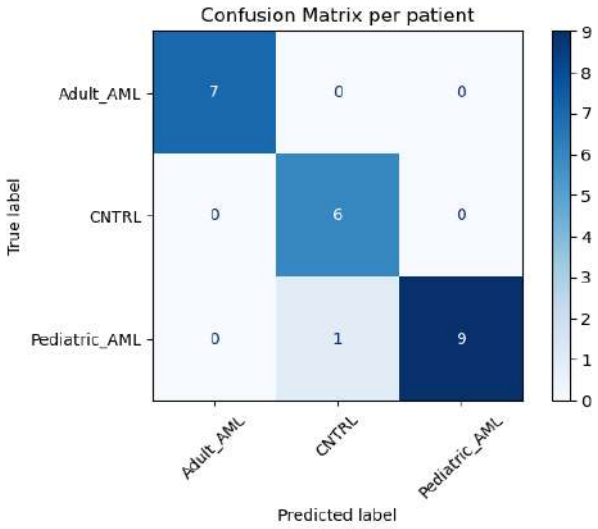


Figure 9: Confusion matrix resulting from the aggregation of the cell type classifications.

As shown in the confusion matrix in Figure 9, the classification result is very satisfactory, with only one misclassified patient.
This patient belongs to the pediatric_AML group but has been predicted as part of the CNTRL group. We can explain this misclassification by comparing the UMAP visualizations of selected patients, which provide a 2D visualization of the cells grouped by cell type. Indeed, we can observe that patient 10, the one misclassified by the QDA, shows a cellular profile that is nearly identical to that of patients 22 and 23, who belong to the CNTRL group. This strong similarity likely explains why the algorithm assigns patient 10 to the CNTRL group. In contrast, the cell profile of patient 10 differs significantly from that of a correctly classified pediatric AML patient, such as patient 9, shown above.

Remembering that one of our main research questions is to identify a smaller group of genes most indicative of the presence of the disease, we aim to trace back from the principal components, which are linear combinations of our original 36,601 genes, to the individual genes themselves. We select the 9 genes with the highest loadings from each principal component, removing the duplicates. This number has been chosen based on empirical trials aimed at balancing dimensionality reduction with the classification performance.
Based on this selection, we build a new dataset restricted to a total of 179 genes. To validate our hypothesis on the significance of this reduced gene subset, we apply QDA to the new dataset (434, 179). The model achieves an accuracy of 0.7126, which, despite being lower than the accuracy achieved using the complete set of genes, as expected due to the reduced dimensionality, still remains a meaningful result.

As in the previous analysis, we then aggregate the classification results at the patient level. This approach leads to a perfect classification across all patients as we can see in the confusion matrix in Figure 10. This improvement is likely due to the reduced noise introduced by irrelevant features and a lower chance of overfitting, thanks to the reduced number of genes included in the model.

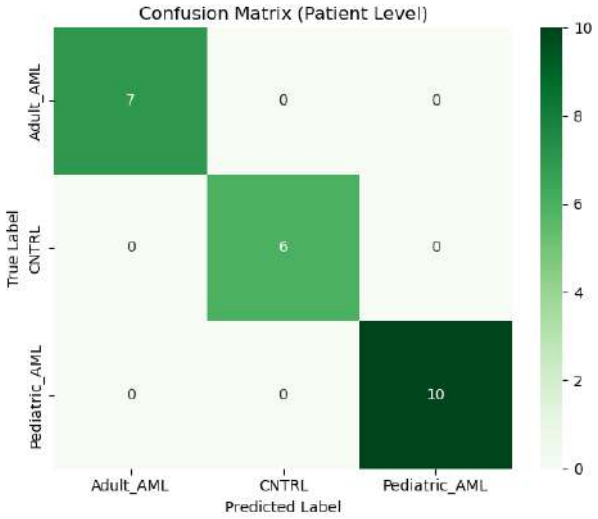


Figure 10: Confusion matrix resulting from the aggregation of the cell type classifications, using only some of the original genes.

## Biological interpretation

We consider that an interesting interpretation of our results could be achieved by comparing them with existing literature on AML, particularly focusing on known marker genes. Marker genes are specific genes whose expression levels serve as reliable indicators of particular biological states or diseases. In the context of AML, these markers act as molecular signatures that facilitate diagnosis and prognosis of disease mechanisms.
Out of the 179 genes identified through our analysis, only one, RUNX1, is a recognized AML marker gene. The other ones are involved in biological processes related to AML, such as inflammation, immune response and cell proliferation, suggesting that they could play an important role in the disease under study. This presents an interesting opportunity for further investigation.

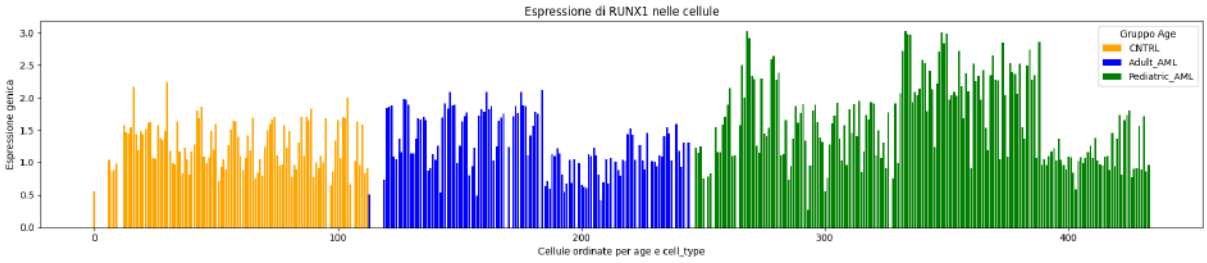


Figure 11: RUNX1 expressions in all cells, divided by group.

## Conclusion

Our findings provide a positive answer to the research question, proving that it is not necessary to consider all genes to accurately detect the disease. This more focused approach, based on selecting a subset of informative genes, appears to be effective while reducing the complexity.
Moreover, this strategy could be extended to other genomic-related pathologies, particu-

larly in cases where samples consist of multiple cell types. In such situations, differences between healthy and diseased individuals may not be evident in every cell type.

# 6 Protein Analysis

As mentioned earlier, the final part of our project focuses on the ADT dataset.
Regarding preprocessing, outliers and NaN values have already been removed.
We now proceed with exploring the dataset, which we expect to differ significantly from the previous one.

The histograms showing the expression frequency of selected genes in Figure 12 indicate that the variables do not follow a normal distribution, as there is no statistical evidence supporting normality.

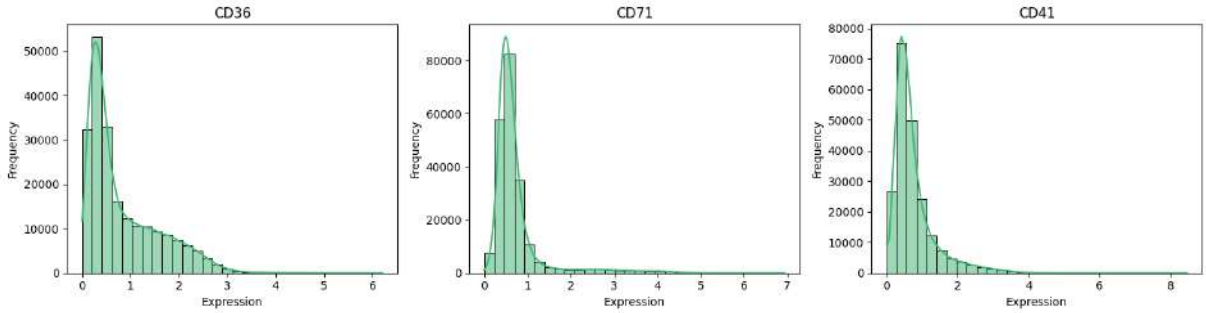


Figure 12: Expression histograms for proteins CD36, CD71 and CD41.

The main goal of our analysis is to find if some proteins are discriminant against AML and to identify them.
Based on the heatmap of sample means across the three groups for each protein (Figure 13) we expect a positive outcome to our research question.

We begin by performing PCA on the ADT dataset and decide to retain the first 40 principal components, which together explain over 90% of the total variability in the data as we can see in Figure 14.

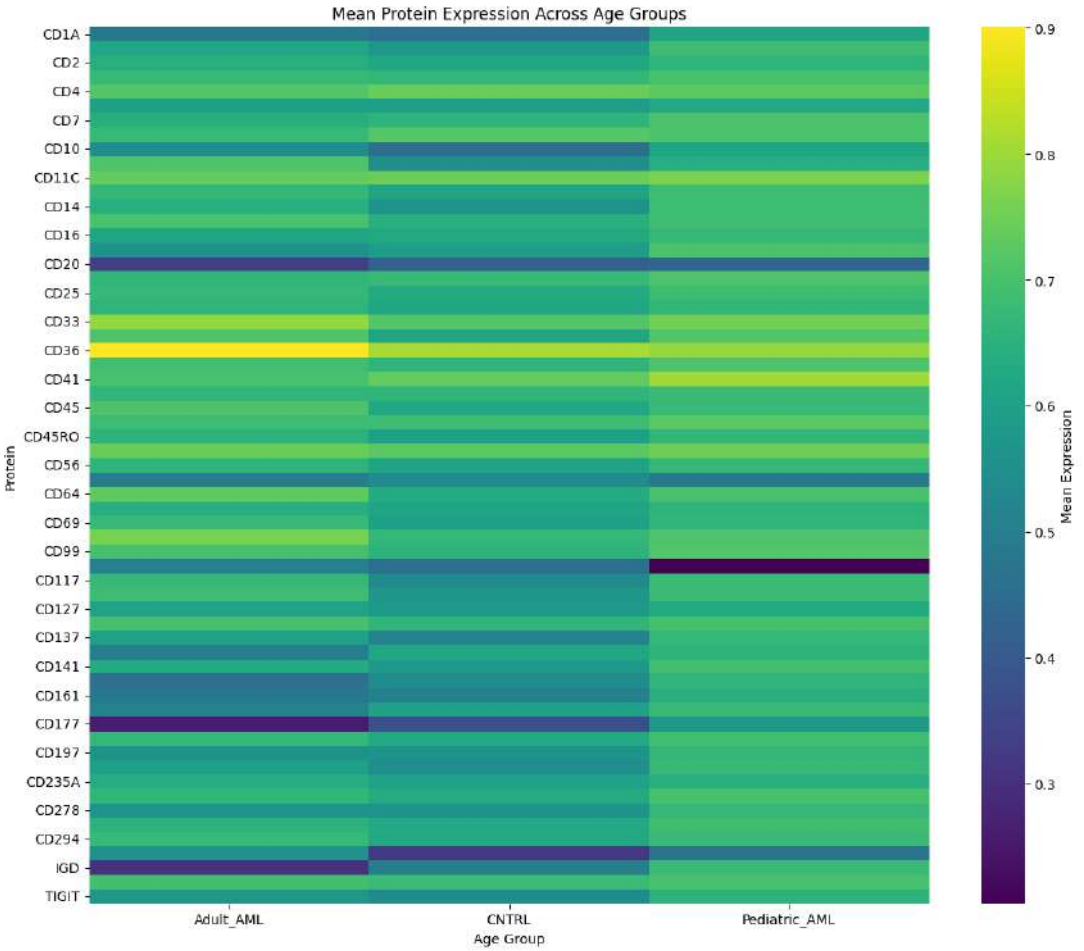


Figure 13: Sample means heatmap for all proteins in the original dataset.

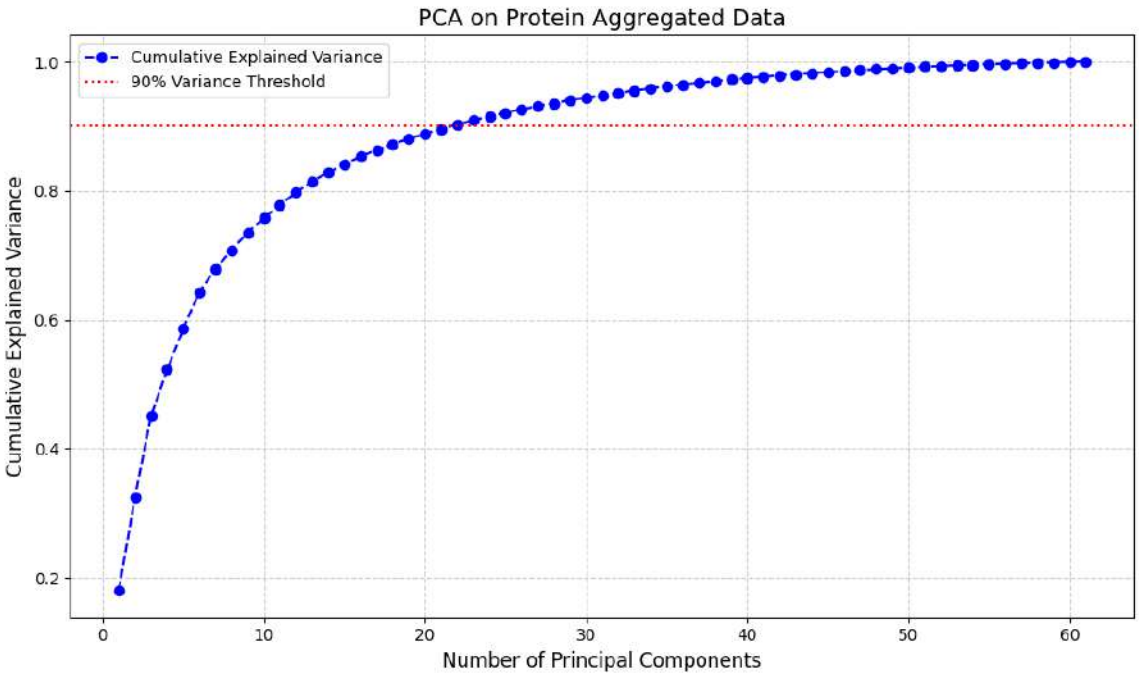


Figure 14: Cumulative explained variance for PCA on the protein dataset.

At this stage, we construct a new dataset, retaining all 211,969 rows and selecting the 40 columns corresponding to the principal components.
Given the lack of normality in the data, we attempt a Box-Cox transformation; however, this does not yield satisfactory results, as normality is not achieved. Consequently, we proceed with a supervised learning approach using methods that do not rely on the assumption of normality.

Although Random Forests are not ideal in terms of interpretability, we choose this method due to its superior accuracy compared to other models, as well as its ability to identify the most discriminative proteins for classification.
Using a test set comprising 30% of the data and a model with 200 trees, we achieve an accuracy of 0.9389.
To further explore the relevance of each principal component, we compute the shapely values and visualize their importance (Figure 15), aiming to assess whether additional dimensionality reduction is feasible.

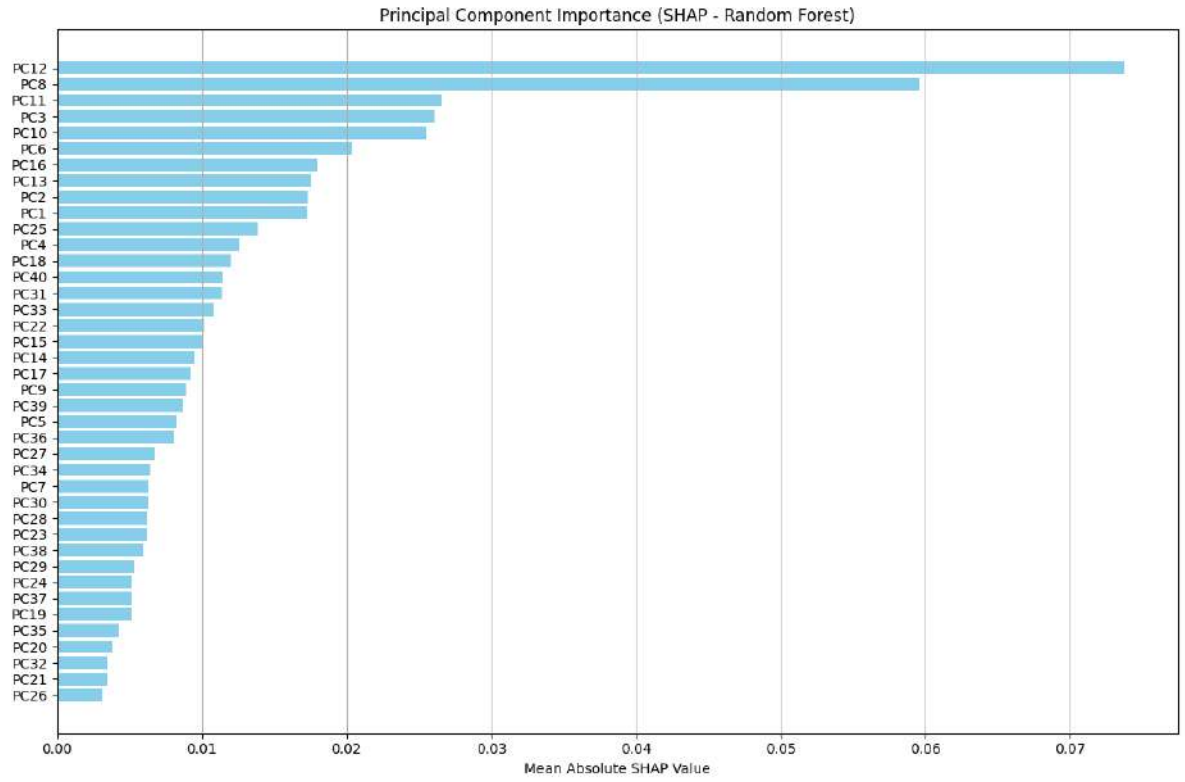


Figure 15: Importance of the principal components in decreasing order, according to Shaply value.

We choose to retain the first principal components and extract the five proteins with the highest loadings in each, removing any duplicates. This results in a reduced dataset with 21 columns, each corresponding to one of the original proteins.
Running the random forest model again on this reduced dataset yields an accuracy of 0.94, suggesting that these 21 proteins are indeed effective in distinguishing AML.
When comparing these results with our previous analysis, we find that several of these proteins are linked to some of the most discriminative genes. Moreover, all 21 proteins are associated with biological functions that are typically disrupted in patients with AML.

# 7 Final conclusions

The reduction in the number of genes and proteins identified as key discriminative features leads to two major benefits.
First, it significantly lowers costs, as future analyses can focus on extracting the expression levels of only a subset of genes and proteins, rather than the full set.
Second, it contributes to the expansion of scientific literature, as researchers and medical professionals can now direct their attention to genes and proteins that were not previously recognized as potential markers for the disease.

Moreover, the methodology applied in this project can be generalized to other diseases. The same analytical framework can be used on different genomic datasets—both in cases where labels are available (supervised learning) and where they are not (unsupervised learning)—to identify new genes or proteins that may serve as discriminative biomarkers for various conditions.