Integrating Machine Learning with Metabolic Models for Precision Trauma Care: Personalized Endotype Stratification and Metabolic Target Identification: Supplementary material 3

Marin de Mas I.*[1,2], Moura L.[3], Antunes F.L.M.[3], Guerrero J.M.[4,5], and Johansson P.I.[1,2]

[1]CAG Center for Endotheliomics, Copenhagen University Hospital, Rigshospitalet, 2100, Copenhagen, Denmark
[2]Department of Clinical Medicine, Copenhagen University Hospital, Rigshospitalet, 2100, Copenhagen, Denmark
[3]Department of Electrical Engineering, Federal University of Ceará, Fortaleza, 60430-160, Ceará, Brazil
[4]Center for Renewable Energy and Microgrids, Huanjiang Laboratory, AAU Energy, Zhejiang University, 314423, Zhejiang, China
[5]Center for Research on Microgrids (AAU CROM), AAU Energy, Aalborg University, 9220, Aalborg , Denmark

1 Introduction

Previous studies have identified four shock-induced endotheliopathy (SHINE) phenotypes [1]. These phenotypes are associated with significant differences in the metabolic profiles from blood samples, which correlate with variations in mortality rates and are independent of the severity of the injury. Using Genome-scale metabolic model (GEM) and flux balance analysis [2] on data from 95 trauma patients, we derived a constraint space of solutions that describe their reaction fluxes. This section goal is to identify patterns in the patients' metabolic flux profiles associated with the different endotypes, allowing for patient stratification and the identification of potential therapeutic targets by applying. To this aim a pipeline combining different approaches based on linear multivariate analysis is developed. The pipeline comprises different steps: i.

*Corresponding author: Marin de Mas I. Email: igor.bartolome.marin.de.mas@regionh.dk

data preprocessing, ii. dimensionality reduction [3] and iii. clustering to identify potential patterns (Figure 1). Tasks in this section were performed using MATLAB (version R2022a) and Python (version 3.9.12) with the packages numpy (1.21.5) and scikit-learn (1.0.2).

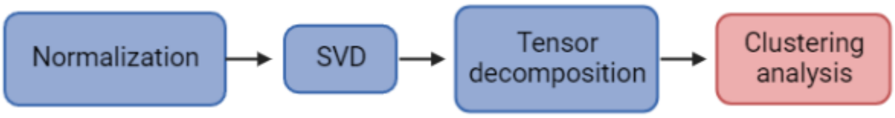


Figure 1: First method summary

2 Source Data

This study employs a mechanistic genome-scale metabolic model (GEM) integrated with flux balance analysis (FBA) to analyze metabolic flux profiles of trauma patients. We use the iEC3006 GEM, which includes 3006 reactions, to generate a baseline model. By sampling the baseline model the initial lower and upper flux boundaries of the exchange reactions were determined (Figure 2). Next, these boundaries are adjusted to reflect patient-specific conditions using metabolic concentration data collected from 95 trauma patients relative to a control group of healthy individuals. Patient-specific flux boundaries are calculated by determining the ratio between metabolite concentrations in patients and the control group. For example, if a patient exhibits twice the metabolite secretion compared to healthy individuals, the corresponding reaction's minimum and maximum flux boundaries in the baseline model are doubled. This procedure is applied to all reactions in the baseline model to generate trauma-specific metabolic boundaries. To account for variability in the control group, we create three sets of boundaries per patient using the mean, minimum, and maximum concentrations. FBA is performed under steady-state conditions, where the mass of substrates equals the mass of products ($A = C + D$; Figure 2). Mathematically, this condition is expressed as $S \cdot v = 0$, where S is the stoichiometric matrix and v is the flux vector constrained by the adjusted boundaries. Since large metabolic models contain more reactions than metabolites, this system has infinite feasible solutions rather than a single solution. To characterize this solution space, we perform random sampling of steady-state flux solutions, generating a matrix where rows represent reaction fluxes (3006 reactions) and columns correspond to sampled solution points. This sampling defines the constrained solution space of metabolic fluxes for each patient. Each patient's data is stored as three MATLAB structure arrays, corresponding to the flux boundaries derived from the minimum, mean, and maximum values of the control group. Each structure contains the lower and upper flux boundaries, reaction names, and the sampling output. The resulting matrix for each patient has 3006 reaction fluxes and 18036 ($6012 \cdot 3$ randomly sampled solution points, optimized to represent the solution space effectively. In total, we generate 95 patient-specific models, each with three variations, resulting in 285 matrices. These matrices provide a comprehensive representation of reaction fluxes and their constrained solution spaces, enabling detailed metabolic analysis of trauma patients.

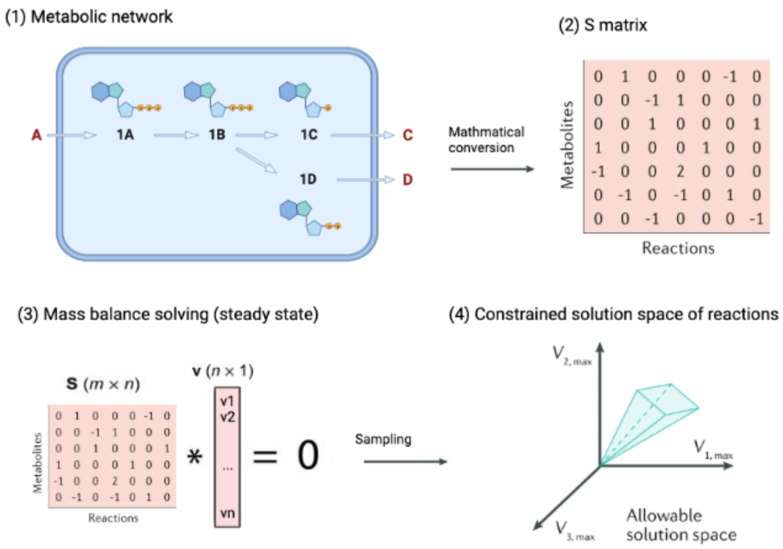


Figure 2: Sampling Solution Points of Reaction Fluxes Using FBA: (1). Modeling of Reactions: The blue line represents the cell membrane, with A (in red) as the substrate, and C and D as the products. (2). Stoichiometric Matrix: This matrix shows the production (1) and elimination (-1) of different metabolites across multiple reactions. (3). Steady State Condition: At steady state, ($S \cdot v = 0$) defines the flux through each reaction. (4). Post-Sampling Matrix: After sampling, we obtain a new matrix (reactions by solution points) that defines the solution space of the reactions for the patient.

3 Pre-processing

3.1 Normalization

After loading the solution points matrices from each patient, we need to normalize the data since the reaction fluxes have different boundaries. Using a standard scaler, we standardize the reaction fluxes by dividing each value by the standard deviation (scaling to unit variance) and subtracting the mean (centering the distribution around 0). We then concatenate the three matrices of each patient and normalize them using the standard scaler. This process results in a normalized matrix for each patient, such as the one shown in Figure 2 (Normalized matrix of a patient), which has 3006 reaction fluxes as rows and 18036 solution points (max, mean, and min) as columns.

3.2 Impact of the variance of the healthy patients

To eliminate the variability within the control group, we compute the ratio of the trauma patients' values against the minimum, mean, and maximum values of the healthy individuals. This approach results in three solution points matrices for each patient. Performing a PCA on the reaction fluxes reveals that the

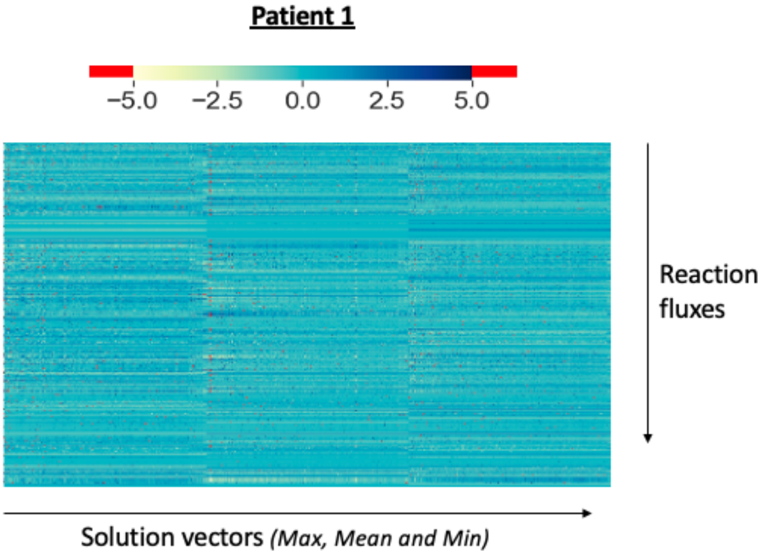


Figure 3: Normalized matrix of a patient

minimum, mean, and maximum models form three distinct groups for every patient in the cohort, as shown in Figure 4. This indicates that all three models are essential to account for the variability within the control group.

3.3 Singular value decomposition (SVD)

The solution points are intended to define the solution space of the reaction fluxes. However, due to sampling, some points may be highly correlated with others. Singular value decomposition (SVD) helps summarize these solution points into components that explain most of the variance. For the average of the 95 patients, 350 SVD components account for more than 95% of the variance. Consequently, we obtain a reduced matrix of 3006 reaction fluxes by 350 SVD components for each patient (Figure 5). To compare different patients and identify patterns, we need to organize our data into a 3-way tensor. We combine each patient's matrix into a tensor (as shown in Figure 6) with three dimensions:

- Patients
- Reaction fluxes
- Solution points

Originally, since the solution points are randomly distributed, the columns in the different patient's matrix doesn't explain the same thing. Using SVD, instead of the solution points, we will have components that are sorted by their explained variance which means that each patient's matrix has columns sorted from the component that explained most of the variance to the one that explained the less. This way the solution points dimension isn't randomly distributed from a patient to another anymore and we can concatenate all the matrices in one 3-way tensor.

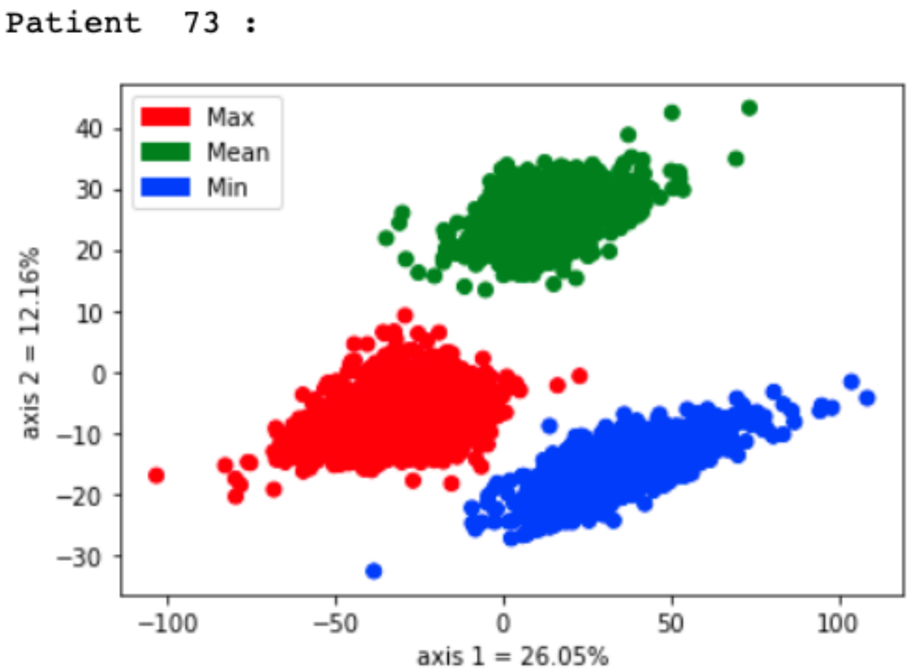


Figure 4: Example of distribution of the solution points (patient 73)

3.4 Tensor decomposition

Once we have our tensor, we reduce it into multiple components to identify patterns (feature extraction) in each dimension. This method has been previously studied on simulated metabolomics data (low dimension) [3] and has shown that we can effectively distinguish different groups of patients using the decomposed vectors. To determine the optimal number of components, we compute the reconstructed tensor and then calculate the percentage difference between the original and reconstructed tensor (Figure 7). We evaluate this reconstruction error on the tensor with 350 SVD components for various numbers of tensor decomposition components, as shown in Figure 8. Our results indicate that approximately 99% of the data is explained by the model when the number of tensor decomposition components matches the number of SVD components (e.g., the model with 350 tensor decomposition components explains 98% of the tensor with 350 SVD components).

4 Results

4.1 Clustering analysis

As shown in Figure 7, the tensor decomposition yields three vectors (representing the dimensions of our tensor) with multiple components. From a tensor decomposition with 350 components, we obtain three vectors, each with 350 components. Using this preprocessed data, we can perform a clustering analysis on the patient vectors to identify any correlations among patients. By grouping

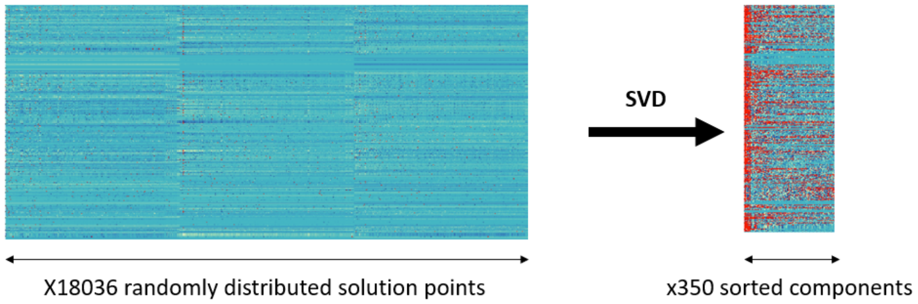


Figure 5: SVD 350 components

these patients, we can then analyze the reaction fluxes to uncover pathways that explain their correlations.

4.2 Compute the number of clusters

Using the K-means clustering method, we computed the SSE (Figure 9) and identified an optimal number of clusters at 3 (elbow point). The silhouette curve (Figure 9) also indicates that the highest similarity within each cluster occurs at 3 clusters. However, this optimal number of clusters does not align with the original number of metabolic groups (4), suggesting that our distribution does not accurately represent the original groups.

4.3 Compare computed clusters with the original metabolic groups

The original clinical data contained 4 clusters (metabolic groups). When we compare them to the distribution of our patient vectors on 2 components using PCA for visualization (Figure 10 – a), we see that they appear randomly distributed. To compare the metabolic groups, we need to compute 4 clusters (instead of 3, which was the optimal number of clusters) and find the permutation with the highest percentage of identity. With the K-means clusters sorted as [1; 2; 0; 3], we achieve 37.89% identity with the metabolic groups (Figure 10 – b shows the distribution of the computed clusters with this permutation). As shown in Figure 10 – (c) Distribution of the clusters, this permutation results in different distributions for the computed clusters and the metabolic groups. However, it is homogeneous when we look at the identity within each cluster (Figure 10 – (d) Patients with identical clusters). From these results, we can conclude that the identity between the two sets of clusters is random, indicating no correlation between our distribution of patients from the tensor decomposition and the metabolic groups. Instead, this distribution is correlated with the explained variance of the SVD for each patient. In Figure 11 - Visualization of the patients (350 SVD components), the clusters represented are computed (using K-means) from the list of variances explained by the SVD with 350 components for each patient. The first component of the PCA (x-axis on the plot) explains more than 80% of the variance, and there is a correlation between the clusters of variances explained for each patient and this axis. In Figure 12 - Visualization of the patients (500 SVD components), we observe the same cor-

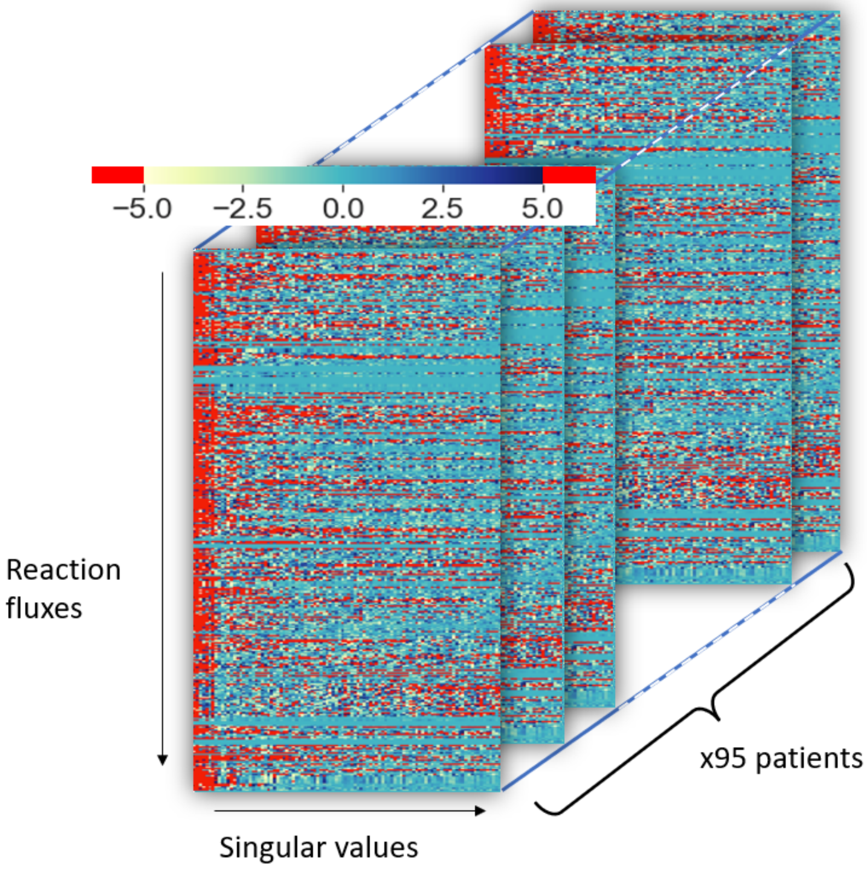


Figure 6: 3 dimensions tensor (patients, reaction fluxes and solution points)

relation with 500 SVD components, which explain on average more than 99% of the variance for each patient compared to 95

5 Conclusion

In this study, we utilized a genome-scale metabolic model (GEM) integrated with flux balance analysis (FBA) to analyze the metabolic flux profiles of 95 trauma patients. By adjusting the baseline model with patient-specific metabolic data, we generated a comprehensive set of patient-specific models. The sampled metabolic flux profile of the pateints were normalized and reduced using singular value decomposition (SVD) to facilitate pattern recognition and clustering analysis. Our analysis revealed that the variability within the control group necessitated the use of three distinct models (minimum, mean, and maximum) for each patient. This approach ensured a robust representation of the metabolic flux profiles. The tensor decomposition and subsequent clustering analysis identified three optimal clusters, although this did not align with the

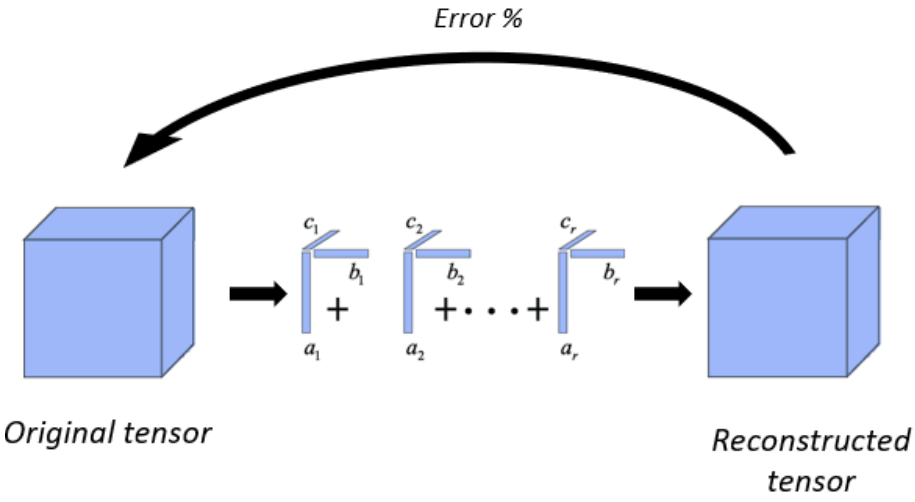


Figure 7: Reconstruction error

original four metabolic groups identified in clinical data. This discrepancy suggests that the metabolic profiles derived from the GEM and FBA approach may capture different aspects of patient variability compared to traditional clinical classifications. Furthermore, the clustering analysis indicated that the distribution of patients was more closely correlated with the explained variance of the SVD components rather than the original metabolic groups. This finding underscores the importance of considering the underlying data structure and variance when interpreting clustering results. Overall, our study demonstrates the potential of combining GEM, FBA, and advanced multivariate analysis techniques to uncover metabolic patterns in trauma patients. These insights could inform patient stratification and the identification of therapeutic targets, ultimately contributing to improved clinical outcomes.

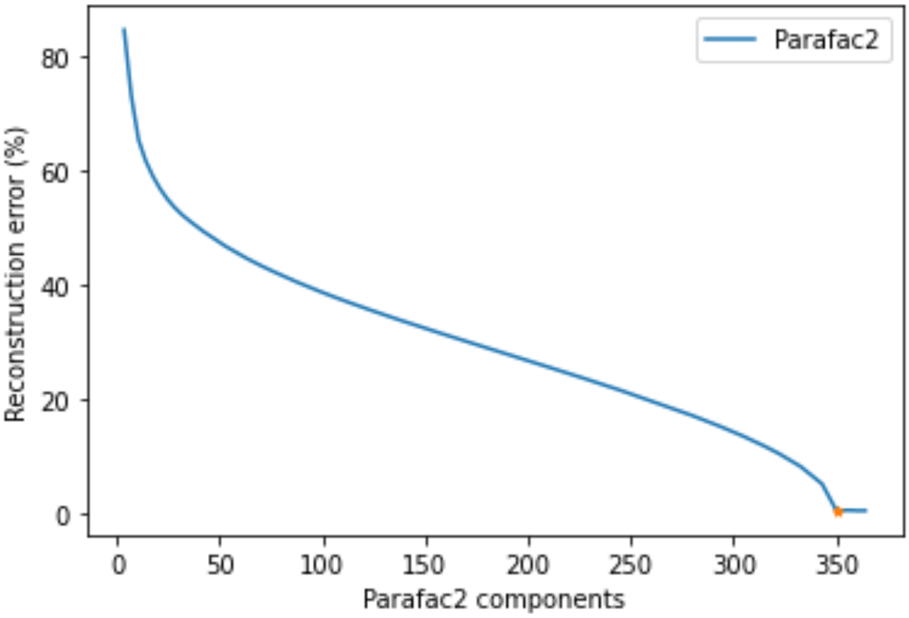


Figure 8: Find the optimal number of tensor decomposition components

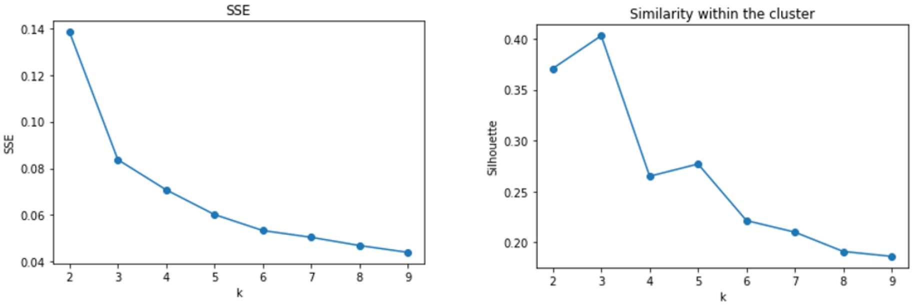


Figure 9: Kmean optimal number of clusters

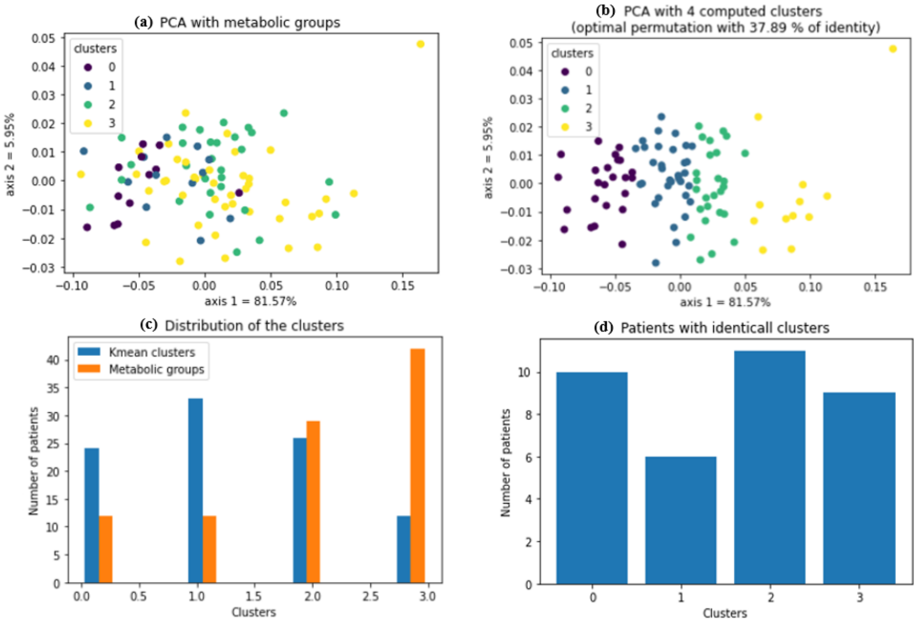
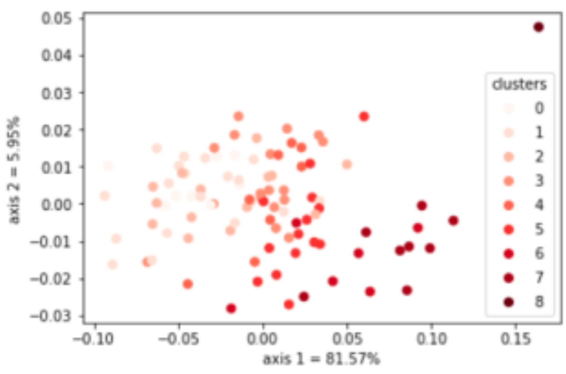
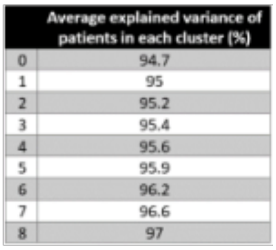


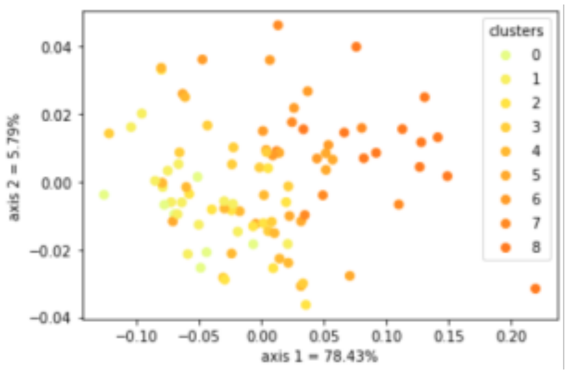
Figure 10: Distribution of the 95 patients

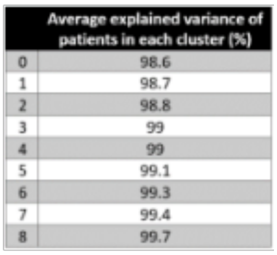




	Average explained variance of patients in each cluster (%)
0	94.7
1	95
2	95.2
3	95.4
4	95.6
5	95.9
6	96.2
7	96.6
8	97

Figure 11: Visualization of the patients (350 SVD components)





	Average explained variance of patients in each cluster (%)
0	98.6
1	98.7
2	98.8
3	99
4	99
5	99.1
6	99.3
7	99.4
8	99.7

Figure 12: Visualization of the patients (500 SVD components)