# INF552 REPORT

## *Visualization of a dataset describing the proteomic signatures of brain regions affected by tau pathology in early and late stages of Alzheimer's disease*

From September 2023 to December 2023

Auguste Crabeil
Julien Gadonneix
Rémi Pommé

# CONTENTS

# INTRODUCTION

Alzheimer's disease is a progressive neurological disorder that primarily affects memory, thinking skills, and behavior. It's the most common cause of dementia, slowly impairing cognitive functions and eventually interfering with daily life. Alzheimer's is characterized by the buildup of abnormal protein deposits in the brain, which disrupts communication between nerve cells, causing them to degenerate and die. Though there's no cure yet, ongoing research aims to better understand the disease and to develop effective treatments. Henceforth, we chose to visualize a dataset dealing with this terrible disease. We contacted researchers to talk about their paper [1] and we offered to provide them with visualizations about their dataset.

# 1
# DESCRIPTION OF THE DATASET AND PURPOSE OF THE VISUALIZATIONS

The dataset gives information on the proteome of various brain areas closely linked to the memory. The proteome consists in the abundances of the diverse proteins present. In this dataset, which is meant to evolve and increase in size according to the researchers, the four brain areas are the entorhinal, temporal, frontal and parahippocampal cortices. For each of these brain parts, the scientists conducted mass spectroscopies to determine, in the first step, the peptides present and their quantities. Thanks to this information, they were able to determine the abundances of the proteins which is the essential information in this dataset. As for now, they collected the samples on about 30 dead patients suffering from Alzheimer's disease at different stages of the disease. There are 6 of them and they are called *Braak stages*. From these analyses, they could determine the abundances of more than 3000 proteins.

In our visualizations of this dataset, we chose to make it user-friendly with a fancy home page and the list of the categories of data to visualize. There are two of them, one describes the cohort of patients with various graphs and figures explaining their age, sex, and Braak stage in Alzheimer's disease. The other one underpins the evolution of the protein abundances with the evolution of the disease in the different brain regions of interest. In the former section, the goal is to give an overview of the population used for the analyses to ensure that it is relevant for the study. In the latter section, the visualizations aim at giving insights of the changes in the proteomic signatures because of Alzheimer's disease.

# 2
# THE DEVELOPMENT

In order to have a full control of our visualizations and figures and to be able to adapt them to the dataset and to our wishes and ambitions, we developed ourselves the figures using *d3* most of the time, for the graphs.

The first step of the development was to write a short script in python to reshape the data and to slightly process them to be able to extract quickly and efficiently the information for the visualizations. For this script, we used *pandas* to handle and manipulate the data frames correctly.

Then, it was very long to understand how to display an animated 3D brain in our home page, but it was not very technical.

Let's delve deeper into the technical description of the development of the section on evolutions of proteomic signatures due to Alzheimer's disease. As already mentioned, in order to fully control the graphs, we used the package library *d3*. The first and main figure which is a bar chart about the overexpression or under expression of the proteins required to define the positions of the axis, bars and to implement a scale. This scale must translate the changes in the expression of a given protein into a bar size. Then, a point that required some attention was the interactivity because we wanted to provide extra information on a selected protein. This needs to add every element in the page programmatically. The other figures are based on the same principle except from the plots about the repartition of the patients in the different stages for the selected protein. For this one, we had to link the data between the protein abundances and patients' information. Finally, it was also complex to map the positions of the cortices on the brain map divided in the three different point of views.

For the page showing patients' characteristics, we did three graphics. We also put checkboxes and radio buttons to select respectively the brain region and the stages of Alzheimer's disease the user wants to consider. At each change of the buttons, a function is called to ensure a visual transition and to recreate the correct graphs.

The first one is a simple bar chart showing the number of patients according to the stage of Alzheimer they had and their gender. After charging our *.csv* file containing the patients' information, we filtered it in order to only keep the patients corresponding to the brain region and the stages the user selected. We created 2 arrays named *countsmen* and *countswomen* which correspond respectively to the number of man and women suffering from Alzheimer's disease at the same stage. We then created a sub-*svg* element which will contain our graphs and created different rectangle to form our bar chart. The rectangles $x$ position was determined by the stage of Alzheimer's disease considered and the height of each rectangle was proportional to the number of persons at the stage considered. We also added the value at the top of the rectangles and a legend for clarity.

The second graph was also a bar chart, showing the number of patients according to the gender and the age. We also created two arrays to count the patients of each age range. We created our rectangles, the $y$ position corresponding to the age range and the width being proportional to the number of patients of the age range considered. We also wrote the values and added a legend for a better interpretation.

The third graph shows the number of patients depending on the age and the Alzheimer's disease stage, to see if we have a correlation between these two data. We wanted to represent this by a classic graph, the points' size being proportional to the number of patients. We created an array containing both the information about the age range and the stage considered. We created the circles; the $x$ and $y$ positions were determined by respectively the stage and the age range. For the size, we used the length of the list of patients filtered to only keep the correct age range and stage. We also added the values and a legend to make it clear.

# 3
# THE DESIGN AND THE FEATURES

In the following section, we will describe the design of our visualization and its features. From the home page to the different graphs and figures, we will detail our choices.

For the main page of the project, we wanted to contextualize the aim of our project by explaining briefly Alzheimer's disease, the aim of our website and the utility of this work. We also added a 3D brain ( **??**) to illustrate the work done to create the dataset we used, the measures of all the protein expressions in the different regions of the brain. It reminds the visitor of the fact that our work is applied to Biology and that all our graphs are not just numbers but have an importance for the possible development of medicine to cure Alzheimer's disease.

Let's focus on the section about protein changes in expression because of Alzheimer's disease. For the first bar chart, we chose this type of graph because we had to represent a quantitative variable, which is the change in abundance, for different proteins, which is a categorical variable. In order to have a complete overview of this quantitative visualization, we logarithmically transformed the increases or decreases in protein expression because those changes were too diverse among proteins in a selected brain area. In Biology, this operation is called the $log_2$ *FoldChange*. Moreover, we ordered the data sequentially to ease the selection and to increase the selectivity of the bar chart, which means that we had a categorical scale, which divided the data into protein classes, with an ordinal aspect. Indeed, we can immediately detect and select the proteins whose expression underwent the biggest modifications. We also implemented a color hue to make this bar chart more associative and immediately differentiate the proteins whose expression increases from the proteins whose expression decreases. The ordering of the data according to the changes in expression also enabled to increase the discriminability of the bar chart because we can now have a higher, almost infinite, number of proteins and still distinct them by their changes in expression. As some proteins has very small changes in abundance, it creates a discrepancy between the two classes which, together with the colors, contributes to the perception of two distinct groups according to the law of proximity from the *Gestalt Laws of Perceptual Organization*. For the option selection of the brain region and the stages, we chose radio buttons as they are categorical attributes.

All this led to the creation of the following figures with the curves showing the evolution of the selected protein expression with the stages of Alzheimer's disease. We opted for this plot as the stages are similar to a time variable, so we have a type of time series. For the information of the repartition of the patients for the selected proteins, we decided to add more than just error bars on the preceding plot. The dataset is meant to increase, and the repartition of the cohort is crucial for the validity of the study. Thus, the representation of the boxplots and density estimations of the sampling repartition provides the visitor with useful information on the trends about the selected protein.

Concerning the brain maps with the different point of views, this visualization brings a medical approach to the information emphasized. It reminds the visitor of the real application of this study and enables him to perceive the cortico-topographical dimension of the problem. We implemented a color scale because the color saturation is a very efficient way to stimulate the perception according to *Stevens' Psychophysical Power Law.*

To have a great visualization of the patients' characteristics, we wanted to restrain the quantity of characteristics shown in one graph and do several graphs. The feature we wanted to show were the distribution of patients according to the stage of Alzheimer's disease, the distribution of age of the patients and the possible correlation between age and stage of the disease.

First, for the graph showing the number of patients according to the Alzheimer's stage, as we wanted to show a quantity, we thought that a bar chart would be appropriate. Indeed, the size of the rectangles is directly linked to the value we wanted to show and bar charts also facilitate the comparison between the different values. The comparison between the size of the patient batches is interesting to verify that we don't have a bias in our further study of the proteins. We also wanted to separate the number between men and women to have a more detailed visualization, as the protein could differ according to the gender. To distinguish the two groups, we colorized the rectangles in blue for the men and pink for the women, as these are the colors commonly used in the society to characterize the genders. Having a bar chart is also useful to associate the value to the group it belongs to, as the bars are separated according to the different stages.

For the graph showing the number of patients according to the age range, we also created bar charts. However, we separated the bar chart in two, one for the women and one for the men. We used again the pink and blue colors to identify. This time, the bar charts were horizontally inclined because it permits a better readability of the bar chart, in comparison of one bar chart being upside down and one normal. With this disposition, it is still easy to compare the number: we can compare the numbers of men/women for each age range by focusing on bar chart, or we can compare the population of men and women by focusing on a line instead.

Finally, we wanted to do a graph showing the number of patients according to the age range and the Alzheimer's stage. As we had two parameters instead of one, doing another bar chart was not a good solution. We decided to represent this data by drawing the points in the graph. But instead of doing regular circles for each point, the size of the circle represents the number of persons corresponding to the age range and the stage. This graph is selective, by watching only the line or column of the feature that interests the user. The size of the circles makes it ordered. Finally, we can also see a pattern where the patients who suffered from the latest stages of Alzheimer's disease are mainly older than the patients having the earliest stages or the "Control" patients.

# 4
# ILLUSTRATIONS

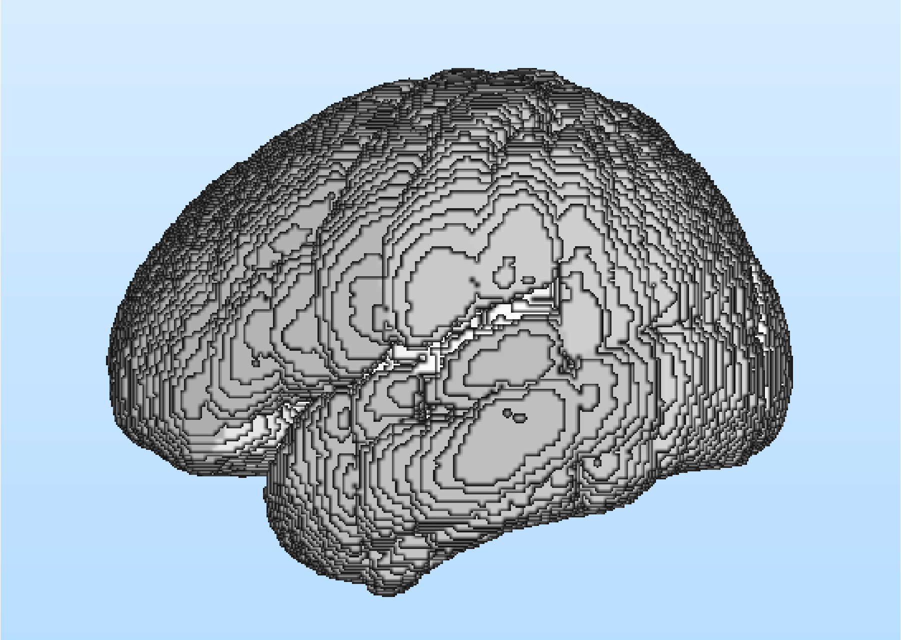
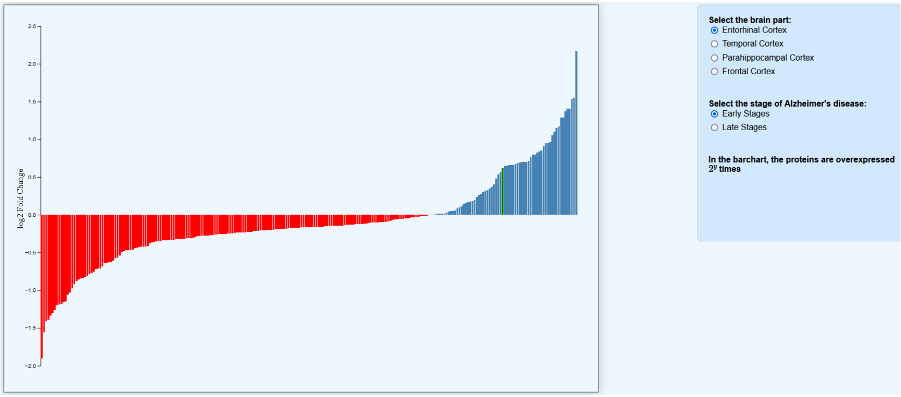


Figure 1: 3D brain from the home page

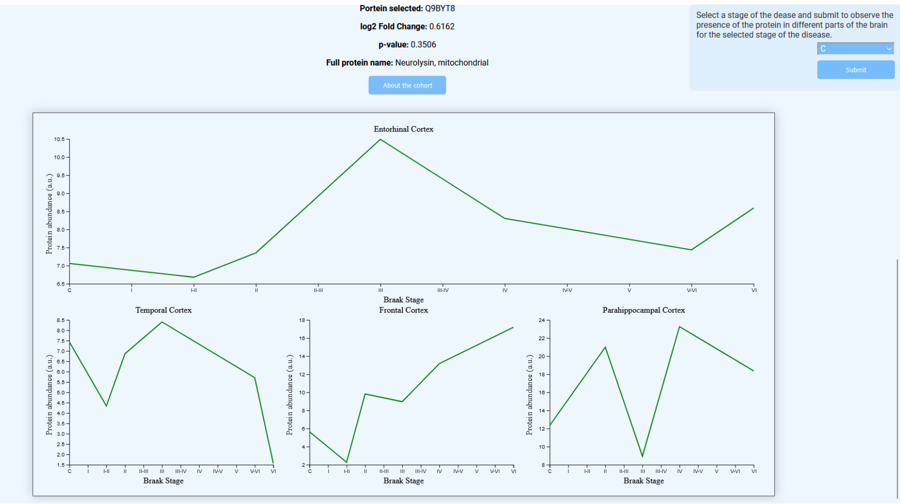


**Select the brain part:**
● Entorhinal Cortex
○ Temporal Cortex
○ Parahippocampal Cortex
○ Frontal Cortex

**Select the stage of Alzheimer's disease:**
● Early Stages
○ Late Stages

**In the barchart, the proteins are overexpressed $2^y$ times**

Figure 2: Expression of different proteins in Alzheimer disease



**Portein selected:** Q9BYT8

**log2 Fold Change:** 0.6162

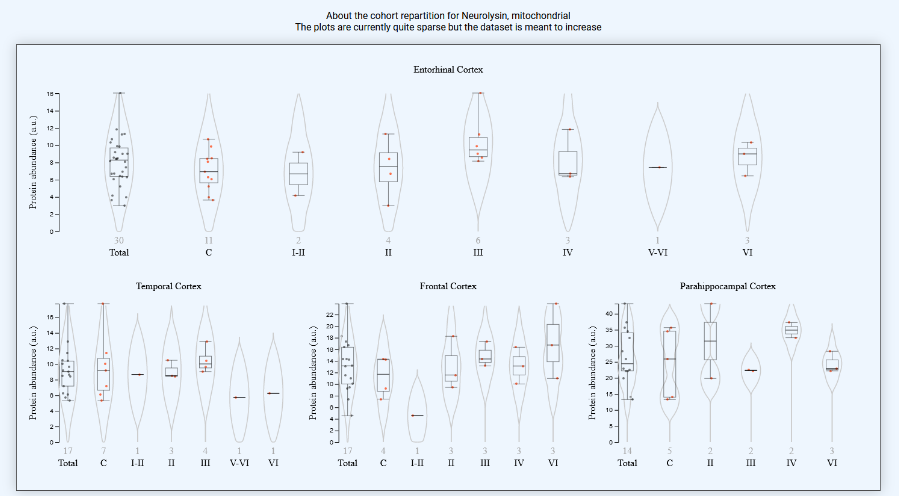
**p-value:** 0.3506

**Full protein name:** Neurolysin, mitochondrial

Select a stage of the dease and submit to observe the presence of the protein in different parts of the brain for the selected stage of the disease.

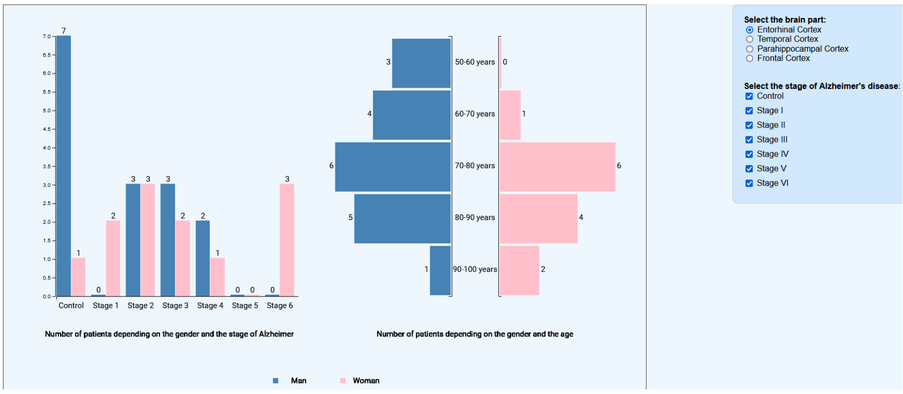
C

Figure 3: Detail of the expression of a single protein



About the cohort repartition for Neurolysin, mitochondrial
The plots are currently quite sparse but the dataset is meant to increase

Figure 4: Expression of a single protein in the different regions of the brain



**Select the brain part:**
● Entorhinal Cortex
○ Temporal Cortex
○ Parahippocampal Cortex
○ Frontal Cortex

**Select the stage of Alzheimer's disease:**
☑ Control
☑ Stage I
☑ Stage II
☑ Stage III
☑ Stage IV
☑ Stage V
☑ Stage VI

Figure 5: Details about the patients from which the dataset was created

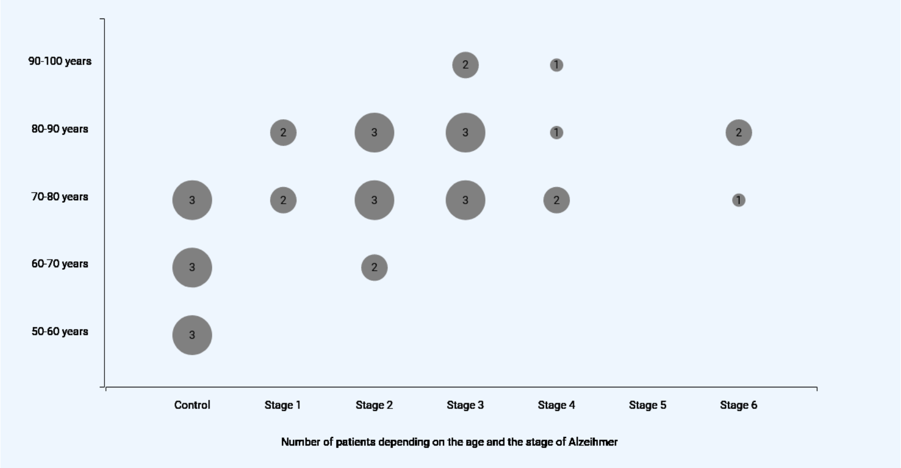


Figure 6: Details about the patients number according to age range and Alzheimer stage

# CONCLUSION

This work was very interesting and permitted us to apply a lot of visualization methods we learned during the course. We realized that the work of showing the content of a dataset was difficult, because the data are vast. As we wanted to do our project of the dataset of the protein expression in Alzheimer's disease, a first step for our project was to understand clearly all the data. As our dataset was very specific, choosing the data we wanted to show and the form of graphs to visualize it was primordial to have a comprehensive website. Finally, the easiest part of the project was to implement our ideas on the website.

# REFERENCES

[1] Clarissa Ferolla Mendonça et al. “Proteomic signatures of brain regions affected by tau pathology in early and late stages of Alzheimer’s disease”. In: *Neurobiology of disease* 130 (2019), p. 104509.