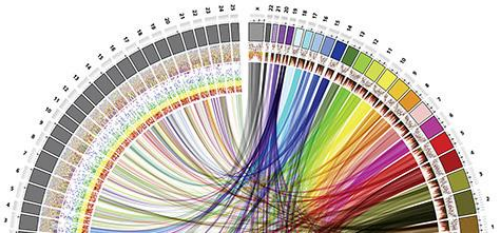

MSc in Bioinformatics – Master thesis

Development of an integrated computational pipeline for the analysis of FTIR data and application of Singular Value Decomposition as dimensionality reduction and noise suppression tool

Juan Torralbo Torrado

3rd July 2024

Departament de Bioquímica i Biología Molecular

Unitat de Biofísica

Supervisors:

Alex Perálvarez-Marín PhD

Mario López Martín PhD

Universitat Autònoma de Barcelona

Signature Page

This thesis has been supervised by Alex Perálvarez-Marín and Mario López Martin (Departament de Bioquímica i Biología Molecular – Unitat de Biofísica – Univertitat Autònoma de Barcelona)

Aproval of the Supervisors:

Alex Perálvarez-Marín

ALEJANDRO PERALVAREZ MARIN - DNI 46688574T

Signat digitalment per ALEJANDRO PERALVAREZ MARIN - DNI 46688574T
Data: 2024.07.02 12:16:29 +02'00'

Mario López Martin

MARIO LOPEZ MARTIN - DNI 48014962R

Firmado digitalmente por MARIO LOPEZ MARTIN - DNI 48014962R
Fecha: 2024.07.02 12:19:58 +02'00'

Author:

Juan Torralbo Torrado

JUAN|TORRALBO|TORRADO

Firmado digitalmente por JUAN|TORRALBO|TORRADO
Fecha: 2024.07.01 20:40:04 +02'00'

Universitat Autònoma de Barcelona

Abstract

Vibrational Spectroscopy (VS) has established as one of the main techniques for the analysis of biological samples in biomedical research. Within VS, infrared (IR) spectroscopy uses IR light to interrogate the biochemical composition of a given sample. IR spectroscopy has evolved over the years, giving birth to Fourier Transform IR (FTIR) spectroscopy, FTIR microspectroscopy (SR-µFTIR) and synchrotron based FTIR microspectroscopy (SR-µFTIR). SR-µFTIR allows to analyze the biochemical and structural properties of a biological sample in a fast, simple and non-destructive way. However, the heterogenous and complex nature of these samples often lessen the analysis power of this technique. To address this issue, data processing is a key step of any SR-µFTIR study. Therefore, this thesis aims to create an integrated pipeline to process, analyze and represent SR-µFTIR data. Also, Singular Value Decomposition (SVD) is explored as a dimensionality reduction and denoising method.

The developed pipeline consists of four different modules, each with a specific purpose. First, the data curation module performs a quality test to remove all non-desired data that could compromise the analysis. Next, the data pre-processing module carries some operations to maximize the extraction of information in latter steps. Then, the exploratory analysis module is used to detect patterns, recognize outliers and to help guide the posterior analyses. Lastly, the statistical analysis module allows to extract valuable information and draw conclusions from the data. This module is evaluated using already analyzed SR-µFTIR data from an experiment that studies postnatal development in the CNS of mice pups.

The results from this study have demonstrated that the developed pipeline is viable and that it will help to conduct future studies within this context. Furthermore, the application of SVD has resulted beneficial for the statistical analysis of the data, proving its value within this context. This opens a new field of investigation for SR-µFTIR studies, as it is expected that SVD will help to decrease data acquisition and processing time, as well as to extract information from noise-affected IR regions.

Index

Figures Index

Tables Index

1. Introduction

1.1. Vibrational spectroscopy

In recent years, vibrational spectroscopy (VS) has emerged and established as a relevant tool in biomedical research for the analysis and evaluation of biological samples (Movasaghi et al., 2008). VS allows to detect electronic changes in the internal vibrational energy levels of biomolecules. Those molecules that contain chemical bonds that are able to vibrate and generate a change in the dipole moment due to their interaction with infrared (IR) radiation are classified as IR active (Martin et al., 2010). These vibrational modes can be measured by IR spectroscopy, and thus being to interrogate and identify a wide variety of biomolecules (Baker et al., 2014).

1.2. IR spectroscopy

IR spectroscopy is a VS technique that measures how the intensity of IR light varies at a given frequency when passing through a sample (Buijs et al., 2004). The wavelength of the IR light absorbed by a vibrating bond depends on the nature of the bond and the inter- and intramolecular interactions of the atoms. This allows the identification of the molecular species present in the sample. Also, since the IR spectrum is additive, all IR active molecules present in a sample will contribute to the final IR spectrum, obtaining a detailed description on the sample's chemistry (Marinkovic and Chance, 2006). However, biological samples are usually heterogenous, and the absorption range of its components tend to overlap, creating highly complex spectra. This supposes a limitation for this technique, as the spectrum must be decomposed in order to be analysed and to identify its individual components. Also, it limits the capability of determining the presence of small concentrations of molecules in a matrix with abundant chemical species, as their signal may be masked by other components (Buijs et al., 2004).

Another important feature of IR spectroscopy is that the absorbance of light depends on the number of molecules contributing to this absorption. This relationship is well described by the Beer-Lambert Law, which stablishes a linear relationship between the absorbance and the product of the concentration, the pathlength and the absorption coefficient of the absorbing molecules allowing quantitative IR analyses (Hardesty and

Attili, 2010). Nevertheless, this law only works for a given range of absorbing molecules. This means that if the concentration or the pathlength are too high, the linearity of this relationship is lost. Therefore, concentration must be controlled, and pathlengths must be very low to properly quantify the samples (Fuwa & Valle, 1963).

1.3. IR spectra

As said before, the frequency of IR light that is absorbed by each molecule depends on its nature. With this, we can classify the analytes according to the frequency in which they absorb IR radiation. The IR spectrum can be divided into 3 subregions, depending on the incident light frequency: far-IR (< 400 cm^{-1}), mid-IR (4000-400 cm^{-1}) and near-IR (13000-4000 cm^{-1}). For the analysis of biological samples, the main used technique is the mid-IR, as it covers most of the fundamental vibrational modes of important biomolecules (Morais et al., 2020). A summary of these peaks can be found at Table 1. Within mid-IR, the fingerprint region, which covers the 1800-900 cm^{-1} range, contains the fundamental vibrations of all compounds needed to characterize any given sample (Magalhães et al., 2021). This region includes absorptions of lipids (C=O symmetric stretching at ~1750 cm^{-1} and CH_2 bending at ~1470 cm^{-1}), proteins (amide I, II and III at ~1650 cm^{-1}, ~1550 cm^{-1} and ~1260 cm^{-1} respectively), carbohydrates (CO-O-C symmetric stretching at ~1150 cm^{-1}) and nucleic acids (PO_2^- asymmetric stretching at ~1240 cm^{-1} and PO_2^- symmetric stretching at ~1085 cm^{-1}) (Movasaghi et al., 2008; Baker et al., 2014). Another important region inside mid-IR is the high region, that expands from 3700 cm^{-1} to 2800cm^{-1}. Here, there is information about proteins (asymmetric N-H stretching at ~3330 cm^{-1} and symmetric N-H stretching at ~3132 cm^{-1}) and fatty acids and lipids (=C-H stretching at ~3005 cm^{-1}, CH_3 asymmetric stretching at ~2970 cm^{-1}, CH_2 asymmetric stretching at ~2942 cm^{-1} and CH_2 symmetric stretching at ~2855 cm^{-1}) (Movasaghi et al., 2008; Paraskevaidi et al., 2017; Matthäus et al., 2008).

Table 1: Vibrational Frequencies (in cm-1) and assignment of the main peaks found in biological samples within the mid-IR region.

High region	Asymmetric N-H stretching	3330
	Symmetric N-H stretching	3132
	=C-H stretching	3005
	CH3 asymmetric stretching	2960
	CH2 asymmetric stretching	2942
	CH2 symmetric stretching	2855
Fingerprint region	C=O symmetric stretching	1750
	CH2 bending	1470
	Amide I	1690-1620
	Amide II	1570-1530
	Amide III	1340-1240
	PO2- asymmetric stretching	1240
	CO-O-C symmetric stretching	1150
	PO2- symmetric stretching	1090

1.4. IR spectrometers

In order to record the IR spectrum, an IR spectrometer is used. IR spectrometers are analytical instruments that use IR radiation to analyse a wide variety of samples. A schematic representation of the functioning of an IR spectrometer is depicted in Figure 1. A global thermal source emits IR radiation over a set range of the IR spectrum. This

radiation passes through the sample and is dispersed by the monochromator. The slits of the monochromator select a narrow frequency range, the energy of which is then measured by the detector. This process is performed sequentially, as the monochromator goes through all the IR spectrum, recording each wavelength at a time. Finally, the spectra are plotted as percentage of transmittance or absorbance as function of the wavelength (White and Roth, 1986).

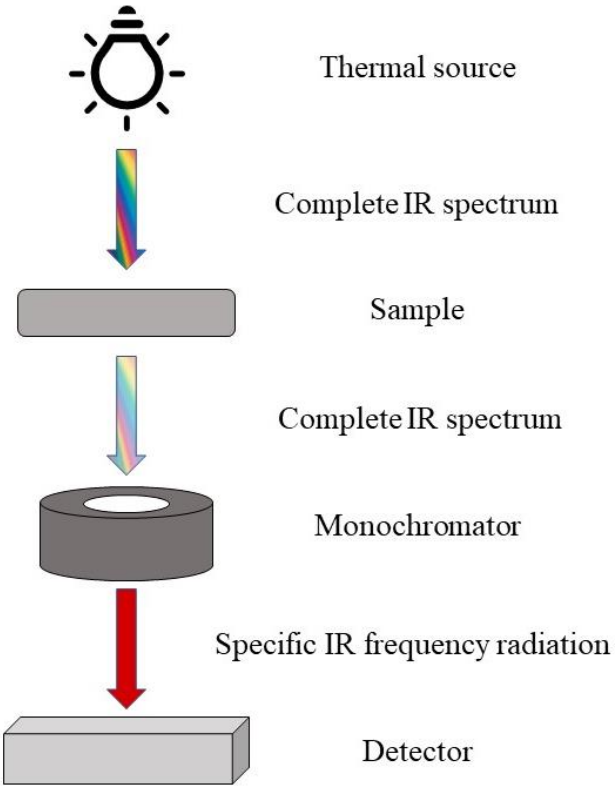

Figure 1: Conceptual scheme of the functioning of an IR spectrometer.

1.5. FTIR spectrometers

The development of new technology and the progress in computational power made able the implementation of the interferometric spectrometers. The interferometric spectrometers, usually called Fourier Transform Infrared (FTIR) spectrometers, incorporate a computer and replace the monochromator with an interferometer. The

interferometer allows to simultaneously collect data from all the infrared spectrum, rather than individual wavelengths, generating an interferogram. Then, the computer applies the Fourier Transform to convert the interferogram into the complete IR spectrum. The development of FTIR spectrometers, along with other improvements along the years, allowed to produce faster measurements, with greater specificity and sensitivity (White and Roth, 1986; Marinkovic and Chance, 2006).

1.6. FTIR microspectroscopy

The next challenge that the spectroscopy field faced was the study of biological heterogeneous structures along with their chemical properties. The goal was to obtain spatially resolved images, analogous to the images obtained in optical microscopy, in a new technique called FTIR microspectroscopy (μFTIR). To accomplish it, an important issue had to be addressed. As said before, the IR absorption spectrum is additive, which implies that probing samples with a beam size larger than the discrete individual elements in the sample produces an average spectrum of all tissues probed by the beam (Marinkovic and Chance, 2006). To overcome this issue, the detected radiation needs to define a specific region of the sample, which may be accomplished with two different strategies: restricting the radiation at the sample plane illuminating only a specific region, the simplest approach, or segmenting the transmitted radiation at the detection plane (Levin and Bhargava, 2005).

The first strategy, referred to as “mapping”, uses an opaque mask with a controlled aperture that restricts the radiation to only the desired area. Then, the transmitted radiation is measured with a single detector. The whole sample is measured by sequentially displacing the beam light along the sample, performing individual measurements from all the points (Levin and Bhargava, 2005). However, this technique presents some major drawbacks. FTIR mapping needs precise positioning, is very time consuming and has a restrictive size limit. The probed area at each single point cannot be less than 15 x 15 μm in order to allow sufficient throughput of radiation. Furthermore, diffraction and stray light effects may compromise the spectral quality of the data (Koenig et al., 2001).

The second strategy is called “imaging” and follows the opposite approach. The whole sample is illuminated and the IR radiation is then separated at the detection plane by IR multichannel detectors called focal plane array (FPA) detectors. FPAs consist of many small single detectors in the same plane following a grid pattern. Each detector is capable of recording the IR absorption from a specific sample region, creating an image similar to optical microscopes that contains the IR spectrum of each point in the sample. This method is much less time-consuming than FTIR mapping, achieving images with near diffraction-limited resolution in just a few minutes (Koenig et al., 2001).

µFTIR offers an in situ, non-invasive technique to analyse the biochemical and structural properties of fixed and nonfixed biological samples. It is relatively simple and reproducible, and only requires a small amount of material, without the need for staining, homogenization or any other manipulation that can affect the nature of the analysed tissue (Pushie et al., 2018). µFTIR provides complete information on the biochemical composition at the molecular level allowing the study of functional groups, bonding types and molecular conformations. The spectral bands from µFTIR are very sensitive, relatively narrow and easy to resolve, allowing the identification of different compounds and the comparison between different samples (Bunaciu et al., 2017). All these reasons make µFTIR a key technique in biomedical research, diagnosis and histopathological studies.

However, due to the nature of the IR radiation used to study biological samples, the resolution that can be achieved fluctuates between 3-30 µm, depending on the numerical amplitude of the objective and the wavelength of the light used (Lasch and Naumman, 2006). This limits the study to the analysis of single cells for most mammalians, leaving unresolved subcellular information along the way (Chan et al., 2020).

1.7. Synchrotron

To overcome this limitation, one possible solution is to use much more potent sources of radiation then the usual thermal sources like, for example, a synchrotron. Synchrotrons are facilities that produce high-flux electromagnetic radiation on a small point by high-speed electrons orbiting close to the speed of light. They produce a continuous spectrum from infrared through visible light and UV to X-ray radiation. This

can be attached to a FTIR microspectrometer, stablishing the synchrotron-based µFTIR (SR-µFTIR) (Marinkovic and Chance, 2006). The reduced beam size and the much greater brightness that synchrotron produces, along with new optical technology allows to improve greatly spatial resolution to sub-diffraction levels, decrease data acquisition time and increase signal-to-noise ratio (Nasse et al., 2011; Findlay et al., 2015; Tobin et al., 2018).

1.8. Previous studies of SR - µFTIR in the biomedical context

In recent years SR- µFTIR has been successfully used in biomedical research to study a wide variety of tissues, such as cardiac tissue (Olkowski et al., 2020), skin (Elmi et al., 2022), epiretinal membrane (Andjelic et al., 2023), lung tissue (Mazarakis et al., 2020), hair (Sandt and Borondics, 2021), dental tissue (Seredin et al., 2021), muscle fibres (Kasprzyk, 2023) and many others. Also, the great resolution of µFTIR enables to analyse cells individually, such as single live cells (Tobin et al.m 2010), stem cells (Qian et al., 2022), bystander T cells (Lipiec et al., 2015), osteocytes (Portier et al., 2020) and cancer cells (Pijanka et al., 2009; Gazi et al., 2005; Nuez-Martínez et al., 2021; Chatchawal et al., 2020), gaining insights about subcellular structure and chemical properties. Furthermore, µFTIR has been utilized to analyse and study several diseases and pathologies (Grzelak et al., 2018; Kreuzer et al., 2020; Seredin et al., 2020). However, within all these applications, one of the most interesting and that has benefited more from SR- µFTIR is the study of the central nervous system (CNS). The detailed biochemical information that this technique provides, coupled with its great spatial resolution makes this technique key for modern neuroscientists (Pushie et al., 2018). In this context, several studies about the CNS have been conducted in the recent years using this technique. The study of neurodegenerative diseases such as Alzheimer's disease, Parkinson’s diseases, multiple sclerosis saw a great advance with the introduction of SR- µFTIR to the routine pathological detection techniques, as is explained and reviewed in detail by Caine et al. (2012). Also, other pathological states like Traumatic Brain Injury (TBI), epilepsy and different types of cancer has been studied in detailed (Guo et al., 2020; Dudala et al., 2012; Nuez-Martinez et al., 2021). Furthermore, normal healthy central nervous tissue can be analyzed with this method. Different regions of the CNS can be easily distinguished using spectral data based on the presence or absence of myelin, a lipid-rich sheath that surrounds axons (Caine et al., 2012). This difference can be clearly seen

between white matter (WM) and grey matter (GM), as the former contains much higher amount of myelin when compared to the latter. Based on this, several studies have been conducted comparing these two areas of the CNS in different conditions and developmental stages. Sanchez-Molina et al. (2020) compared the biochemical properties of WM and GM on non-pathological human and mouse brains. Sanchez-Molina et al. (2021) evaluated the influence of age over microglial cells in WM and GM areas in two transgenic lines of mice. Moreover, Peris et al. (2023) studied the biochemical and structural changes during postnatal development of WM and GM areas of mouse brain and cerebellum. In summary, the use of SR-μFTIR has proven especially useful for nervous system tissue because of the differential WM *versus* GM lipidic and protein composition.

1.9. Data processing

As already said, biological samples are heterogeneous, complex and highly variable in composition and concentration, which often lessen the identification and quantification power of biological components. Furthermore, the development of better instrumentation creates higher resolution data. It also allows to measure biological samples in a more native state, which implies the presence of more contaminants and undesired compounds. Moreover, the differences between analysed samples are usually very discrete and not obvious when looking at raw data. All of this results into highly complex spectra difficult to analyse (Gautam et al., 2015). In this regard, data processing is crucial for the success of the analysis. The aim of this procedure is to reduce the contributions to the spectra of signals that are not related to the target of the study, and thus improving the accuracy and precision of qualitative and quantitative analyses (Morais et al., 2020). It also has the purpose of transform the spectral data to the best fit condition to ensure optimal performance in subsequent steps in the analysis.

Data processing involves many different steps that will transform raw data. These steps vary between studies depending on the nature of the data and the aim of the study, but they can be grouped as follows: quality control, data pre-processing, exploratory analysis and further procedures. The quality control step aims to eliminate all spectra that do not met certain quality criteria set for the study. It is a way of removing non-desired data due to low quality measurements, experimental errors or poor sample handling

(Lasch, 2012). The pre-processing step is perhaps the most important part of data processing and data analysis in spectroscopic studies. It involves many different procedures that aim to correct baseline slopes, remove sample contaminants, correct for uneven sample thickness, remove noise, reduce the contribution of signals of unwanted compounds, enhance discriminating features and other important operations (Trevisan et al., 2012). The exploratory analysis provides an initial assessment of the data that allows the scientist to detect patterns, recognize outliers and draw some conclusions that will help to conduct posterior analysis (Morais et al., 2019). Then, further procedures refer to all further steps applied in order to extract conclusions from the data, that will vary greatly among different studies, depending on their objective. These procedures may be statistical tests, classification models, diagnostic tools and many others.

Compared to other FTIR techniques, data processing involved in FTIR imaging takes most of the experimental time (Koenig et al., 2001). Furthermore, the improvement in technology over the years, which has allowed to improve spatial resolution and reduce data acquisition time, has resulted into more complex and abundant data. Within this context, the establishment of an integrated and automatic pipeline that can conduct data processing and analysis can result very useful for present and future spectroscopic studies by reducing experimental time.

1.10. Singular Value Decomposition

Moreover, here the use of Singular Value Decomposition (SVD) is explored as a method for noise reduction in FTIR spectral data. SVD is a mathematical procedure to decompose a matrix into two sets of orthogonal vectors, represented as matrices, that describe the variance of the original matrix. These two sets of vectors are connected by a diagonal matrix composed by singular values, which indicates the importance of the vectors and their influences over the original data matrix (Golub and Kahan, 1965). Using SVD and a low-rank approximation, a considerable noise reduction can be achieved (Scharf, 1991). Even though SVD is widely used in several domains such as image compression, geophysics, electrocardiograms and data mining, it is still not a common procedure in spectroscopic data processing (Laurent et al., 2019).

Therefore, in this thesis, a new computational pipeline to process and analyse FTIR spectral data is established. This pipeline is distributed in modules and pretends to

be customizable, so the scientist can better apply it to diverse studies with different needs. Also, the automatization and reproducibility of FTIR data processing is aimed to improve with this workflow. Finally, the application SVD to FTIR data is explored as a noise removal and dimensionality reduction method.

2. Objectives

- To automate and harmonise the processing, analysis and representation of μFTIR data by creating an integrated computational pipeline.

- To integrate Singular Value Decomposition (SVD) into the pipeline as a dimension reduction and denoising method to improve the analysis outcome.

- To validate the pipeline by data reanalysis.

- To decrease spectral data time acquisition by improving and automating data processing.

3. Material and methods

3.1. Previous procedures

3.1.1. µFTIR Data Acquisition

Data for this experiment was obtained from C57BL/6 mouse pups P0 to P28 of both sexes as described in (Peris et al., 2023). In the study individuals from three different groups were included: wildtype (WT), transgenic mice with overexpressed interleukin 6 ($IL6^+$) and transgenic mice with overexpressed interleukin 10 ($IL10^+$). Then, µFTIR based on synchrotron radiation was performed at the MIRAS beamline of the ALBA Synchrotron light source (Catalonia, Spain) (Yousef et al., 2017). The measurements, performed as in (Sanchez-Molina et al., 2020) were as follows. A Hyperion 3000 microscope with a 36X objective coupled to a Vertex 70 spectrometer was used. Spectra were collected in transmission mode at 4 cm^{-1} resolution and 10 x 10 µm aperture. 128 scans were recorded for the brain samples, and 64 scans for the cerebellum samples using Opus 7.5 software. Measurements were performed in two cerebral regions, the brain and the cerebellum, and in two different tissues, white matter (WM) and grey matter (GM). In the brain, the measurements were recorded at the corpus callosum for the white matter (BR-WM), and at the cortex for the grey matter (BR-GM). Meanwhile, in the cerebellum the measurements were recorded at arbor vitae for white matter (CB-WM) and at the granular and molecular layers for the GM (CB-GMG and CB-GMG). For each region of interest, approximately 50 spectra with a step size of 30 µm x 30 µm were acquired. Each one of these will be referred to as *measurements.* All the samples coming from the same individual and from the same cerebral region are said to pertain to the same *sample.*

Finally, to extract all the data into the proper format to be analysed by the pipeline, Unscrambler X software (CAMO Software, Oslo, Norway) was used to convert the files into text files containing the information in ASCII format.

3.2. The proposed pipeline

The proposed pipeline in this article is depicted in Fig. 2, and consists in four main modules: (i) data curation module, (ii) data pre-processing module, (iii) exploratory analysis and (iv) statistical analysis module. Each module consists in multiple operations

that are applied sequentially. These operations can be modified or modulated to optimize the analysis for each study, depending on the nature and the origin of the data.

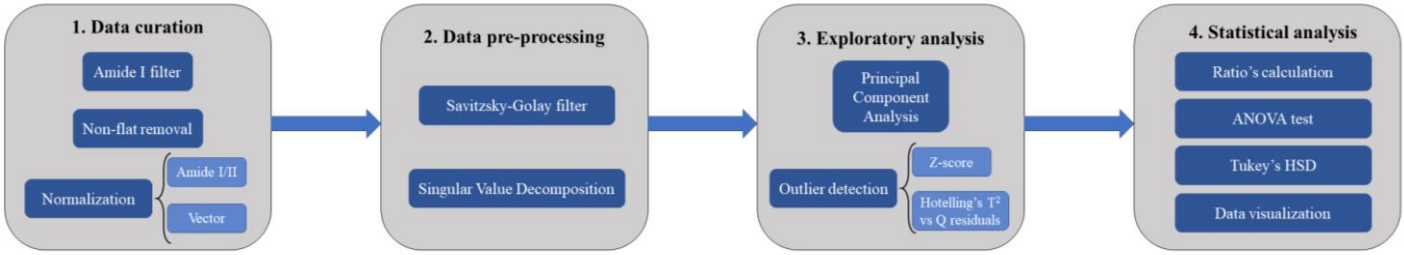

Figure 2: Schematic illustration about the proposed pipeline.

3.2.1. Data curation

This module performs a quality test in order to remove all non-desired data due to low quality measurements, experimental errors or extreme values. Usually, this part of the process is done manually by the scientist by looking at the data and evaluating which of the recordings must be eliminated. However, the idea of this pipeline is to automate the process for the analysis to be more efficient and reproducible, and less time consuming. This quality control is achieved in different steps that may vary within different studies.

3.2.1.1. Amide I filter

The first operation to perform over the spectral data is filter measurements according its absorbance at 1656 cm^{-1}. This peak is assigned to the amide I, which correspond to amide bonds between peptides, so it can be used as a measurement for the protein content (Movasaghi et al., 2008). This peak is fundamental for the analysis of biological samples, as it is ubiquitous for any kind of sample. All measurements below 0.1 and above 1 at 1656 cm^{-1} peak must be removed. Measurements below 0.1 indicate some kind of experimental error, such as a measuring on the wrong spot or bad sample

preparation. Measurements above 1 can be a sign of instrumental errors or wrong sample preparation, for example, excessive sample thickness. In this case, linearity is lost in the Beer-Lambert law, causing the disruption of the Amide I and Amide II bands, and leading to spurious conclusions (Grdadolnik, 2003).

3.2.1.2. Non-flat samples removal

By the nature of the biological samples, there are regions in the spectra that are flat. For samples in the CNS, those regions are comprised between 2000 and 1800 cm^{-1} and between 2800 and 2380 cm^{-1}. Any measurement that presents a clear "belly-shaped" signal within these two regions is because some experimental error has occurred, and therefore must be removed from the experiment. For this purpose, a filter is applied to all the samples in the two aforementioned regions. This filter interpolates the spectra in a given region to a quadratic equation that fits its shape. Then, the area under the curve of this equation is calculated using the trapezoidal rule. For flat measurements, this value will be close to 0. Therefore, all the samples that surpass a given threshold will be considered as non-flat, and consequently removed from the study. The recommended threshold value is 0.1 but it can be modulated to better fit the data. The optimal way to do it is by visualizing the data after the non-flat measurements are removed, while visualizing the samples that are being removed as well. By doing this, the scientist can optimize the threshold and eliminate all non-flat measurements while keeping the rest of them. Furthermore, the regions of the spectra where this function acts can be adjusted to better fit the shape of the spectra.

3.2.1.3. Normalization

Normalization is a common preprocessing step in spectral analyses. It allows to correct for confounding factors such as different sample thickness and concentration that may distort the analysis (Baker et al., 2014). However, normalization must be applied carefully and only when needed, since it may hide important features among samples or introduce non-linearities to the data (Morais et al., 2019).

The most commonly used normalization techniques are the Amide I/II normalization and the vector normalization. Amide I/II normalization is used when the Amide I and II peaks are consistently present in all the dataset and they are not an important feature in the data, whereas vector normalization is applied when Amide I and II peaks are relevant for the analysis and cannot be used for normalization (Lilo, 2022). Usually Amide I/II normalization is used after baseline correction as a preprocessing step, while vector normalization is used after differentiation of spectra as a data curation module, as described below (Baker et al., 2014).

Amide I/II normalization

To perform Amide I or Amide II peak normalization, the absorbance of the chosen peak is set to have a maximum value of 1 for all the spectra in the dataset, scaling all the remaining peaks accordingly, maintaining the initial proportions. To do this, the whole spectra is divided by the value of the chosen peak.

Vector normalization

To perform vector normalization, each sample or measurement in the dataset will be treated as an independent vector, called vector X. There are different types of vector normalization, but the most used for this purpose is the 2-norm or Euclidean normalization. This method converts a vector into a unit vector X^^ with length 1, that preserves the direction of the original vector X. This is achieved by calculating the 2-norm |X| of the vector X, and dividing each component of X by |X|, as seen in equation 1. Usually, vector normalization is applied after differentiation, when there is no apparently consistent peak across the spectra (Trevisan et al., 2012).

$$X\text{^^} = \frac{X}{|X|} = \left(\frac{x_1}{|X|}, \frac{x_2}{|X|}, \ldots \frac{x_i}{|X|}\right) \quad (1)$$

3.2.2. Data pre-processing

Once the dataset has been curated, it can be pre-processed in order to maximize the extraction of information in latter steps. This module applies two key operations that smooth the spectra and remove noise from the data.

3.2.2.1. Savitzky-Golay filter

The Savitzky-Golay (SG) filter is the most used technique for smoothing and removing noise from spectral data (Morais et al., 2020). This filter calculates the second derivative of the data and fits successive sub-sets of datapoints by a polynomial equation in using linear least squares (Savitzky and Golay, 1964). SG filter allows to remove large instrumental noise while preserving important features of the spectra. It also enhances narrow bands and eliminates baseline contributions. However, its major disadvantage is that it can introduce some distortions and "smooth out" important features of the data if not used properly (Trevisan et al., 2012). For this reason, it is very important to carefully choose the right parameters when applying this filter. The main parameters for the SG filter are: polynomial order and window size. For the polynomial order, the recommendation is to select a polynomial order similar to the shape of the spectra (Morais et al., 2020). Usually, a second or third polynomial order is selected. While a third polynomial order better preserve features of the spectra and gives a smoother result, it may introduce more artifacts than the second order (CITA, es de chat gpt). For the window size, the only requirements are that it must be an odd number (as a central point is needed for the smoothing process, and it must be at least equal to the polynomial order (Morais et al., 2020). Then, the window size can be modulated freely, usually between 3 and 21, until the optimal size is found for the given set of spectra. Selecting a window size too small will retain a major part of the noise, while selecting a window size too large may distort the spectral shape (Morais et al., 2020). There are no optimal and universal parameters for the SG filter. They have to be carefully chosen given the needs of the study and the characteristics of the data. Therefore, the selection of the right parameters relays on the scientist conducting the analysis.

3.2.2.2. Singular Value Decomposition

Given a matrix *A* representing spectral data with *m* rows and *n* columns, Singular Value Decomposition (SVD) can be applied, decomposing *A* into the product of three other matrices *U*, Σ and V^T, as represented in Equation 2, where matrix *U* is an orthogonal *m x m* matrix, matrix Σ is a diagonal m x n matrix, and *V* is an orthogonal *n x n* matrix. The columns of *U* are called left singular vectors, and similarly, the columns of *V* are called right singular vectors. These left and right singular vectors represent the variation of *A*. Moreover, the elements in the diagonal of Σ are non-negative values called singular values (Golub and Kahan, 1965). Each singular value σ_{ii} corresponds to the degree in which u_i and v_i contribute to the data matrix *A*. As the singular values are sorted in a decreasing order, *A* can be approximated as A_k (Eq. 3), where k is the number of singular values used in the reconstruction. In the case of noise-containing data, the real signal, which contributes the most to the variation in A, is captured in low-k values, while noise signal is captured in high-k values (Laurent et al., 2019). Using this approximation, a significant denoising may be achieved by selecting the appropriate k value.

$$A = U\Sigma V^T \tag{2}$$

$$A \approx A_k = U_k \Sigma_k V_k^T \tag{3}$$

In order to choose an adequate k value, a common procedure is to plot the singular values σ_{ii} against *i*, as depicted in Figure X. Then, look for the *i* in which the slope changes from steep to not steep. This point is called “elbow”, and usually is a good approximation of which k is able to separate de real signal from the noise (Schanze, 2017).

Another feature of SVD that is explored is weighting. In a common SVD model, all the components of the original matrix (i.e., the wavenumbers in the IR spectrum, or individual samples in the data matrix) have the same importance for the reconstruction.

For some applications of SVD this is good approach, as all the components of the original data may be equally important. Nevertheless, IR data from biological samples do not fit this criterion. Over the IR spectrum there are flat regions with no signal, where only noise is being captured, that are not important for the study. Moreover, following the same approach, noisy data could alter the SVD reconstruction, as the variation introduced by the noise is being captured with the same importance than the variation from the real signal. Therefore, giving the same importance to all the components in the SVD could lessen its denoising power. To overcome this, weighted SVD (WSVD) can be used.

WSVD introduces a new matrix W that represents the importance of each component in the matrix A. Depending on the type of weighting being applied, the matrix W will have different features. For a row-wise weighting, that in this case would correspond to different samples having different weights, the matrix W_r is a $m \; x \; m$ matrix with weights w_{ri} on the diagonal, indicating the relative importance of each sample in the original matrix A. For a column-wise weighting, that would correspond to different wavenumbers having different weights, the matrix W_c is a $n \; x \; n$ matrix with the weights w_{ci} on the diagonal, indicating the importance given to each of the wavenumbers of the IR spectrum. These types of weighting can be applied separately or combined, as shown in Equations 4-6.

Then, to recover the real proportions of the data and ensure that the reconstructed matrix represents the original matrix correctly, the weighting must be undone. For this, the reconstructed A_k matrix is multiplied by the matrix W^{-1}, that corresponds to a matrix with the same proportions of W and the inverse values of the weights w_i along the diagonal (Eq. 7). By applying WSVD the important components of the original data matrix are emphasized and a more robust denoising may be accomplished (Srebro and Jaakkola, 2003).

$$A_w = W_r A = U_w \sum_w V_w^T \tag{4}$$

$$A_w = AW_c = U_w \sum_w V_w^T \quad (5)$$

$$A_w = W_r A W_c = U_w \sum_w V_w^T \quad (6)$$

$$A \approx A_k = W_r^{-1} A_{wk} W_c^{-1} = U_{wk} \sum_{wk} V_{wk}^T \quad (7)$$

To determine the importance of each component, there are two strategies. The first, which is appropriate for the column-wise weighting, is to manually determine the weight of each component, in this case, each wavenumber. Usually, a value of 1 for the more important wavenumbers and 0.001 for the less important wavenumbers are used. The regions with higher and lower importance are decided by the scientist, depending majorly on the shape of the spectra and the objectives of the study. The second strategy, suitable for the row-wise weighting is to set a weight value depending on the signal to noise ratio (SNR) of each sample. For this, the SNR of all samples is calculated, and then normalized between 0 and 1. By doing this, the sample with the highest SNR will have a weight of 1, while the sample with the lowest SNR will have a value of 0. In order to avoid "dead" measurements, those that have a weight below 0.1, will be set to 0.1.

The calculation of the SNR for a given sample is quite straightforward (Eq. 8). SNR of a sample equals the standard deviation for a signal region divided by the standard deviation of a noise region. Therefore, first the mean value for a noise region must be calculated, as well as for a signal region. The 1900-1800 cm^{-1} region is taken as noise, and the Amide I peak (1656 cm^{-1}) is taken as the signal region. Then, for each sample the standard deviation is calculated by subtracting the mean value to the of that sample, and dividing it by the number of individuals in the population minus 1.

$$SNR = \frac{\sigma_{signal}}{\sigma_{noise}} \quad (8)$$

3.2.3. Exploratory analysis

After pre-processing, an exploratory analysis is recommended in order to detect patterns and observe tendencies in the dataset. From the exploratory analysis some conclusions can be drawn that will help to guide posterior analysis (Morais et al., 2019). Then, outlier detection techniques can be used to recognize possible outliers in the dataset and study their origin.

3.2.3.1. Principal Component Analysis

The most common exploratory analysis tool is Principal Component Analysis (PCA), as it is unsupervised and unbiased. In PCA, the dataset is decomposed into a few factors called Principal Components (PCs) which represent most of the variance of the original dataset. The PCs are orthogonal to each other, and generated in decreasing order of explained variance, so the first PC explains most of the variance, followed by the second PC and so on (Bro & Smilde, 2014). These PCs can be seen as a new set of axes in which the data can be projected to help the visualization and interpretation of the data. The mathematical decomposition is represented in Equation 9, where X is the pre-processed dataset, T represents the PCA scores, P represents the PCA loadings and E the PCA residuals.

$$X = TP^T + E \quad (9)$$

The scores are the coordinates of the original data in the new space composed by the PCs. Each score corresponds to a transformed data point in this new space. The scores can be plotted in order to observe patterns or cluster in the dataset in an easy and simple way. The loadings are the coefficients of the linear combinations of the original variables that creates each PC. They represent the contribution of each variable to the variation observed in the dataset. They can be used to detect which of the original variables, in this case, wavenumbers, have the highest importance for the patterns observed in the score plot. The residuals represent the differences between the original data and the reconstructed data points in the new space. In other words, they represent the variance that is not being captured by the PCs. They can be used to detect to identify experimental errors or to detect outliers. PCA residuals should be low and randomly distributed. Otherwise, they can indicate experimental bias or erroneous processing of the data (Morais et al., 2019; Lilo, 2022).

3.2.3.2. Outlier detection

The next step in the exploratory analysis is the outlier detection. The idea is to remove all measurements that are anomalous, also called outliers. Outlier measurements behave differently from the majority of the data, and can be produced by errors, mislabelling or because they are extreme cases that are not representative of the real world (Rousseeuw and Hubert, 2017). In all cases, these outliers must be detected and removed from the dataset, as they can spoil the downstream analysis. This filter consists in two different operations that have the same final purpose. These filters may be used independently or combined, depending on the needs of the study and the characteristics of the dataset.

Z-score

The first filter used for outlier detection is the Z-score. The Z-score determines how many standard deviations an individual measurement differs from the population's mean (Curtis et al., 2016). It determines the degree of difference between an individual measurement and the average spectra. To calculate the z-score, first the average spectra is calculated. Then, the z-score is calculated for each measurement, by subtracting the

average to each measurement, and dividing it by the standard deviation of the whole population.

This operation is performed over each of the wavelengths of the IR spectrum, getting a Z-score for each wavelength for all the individual measurements. Then, the Z-score for the whole measurement is calculated by taking the mean of all the Z-scores across the spectra of a given measurement. This final value gives an estimation on how different an individual recording is from the rest. It is important to point out that the Z-score can have positive or negative value, so the absolute value should be taken into account when performing the average Z-score. Otherwise, it could be severely underestimated, as the negative values would cancel out the positive values.

This Z-score can be used to determine whether a measurement should be removed or not from the dataset. For this, a threshold of 2 is recommended. A Z-score of 2 means that a given measurement differs 2 standard deviations from the mean populations. Assuming a normal distribution for the variable, 97.7% of the individuals will have a Z-score below 2 (Curtis et al., 2016). So, any measurement with a Z-score greater than 2 will be considered as an outlier measurement, and therefore removed from the dataset. The threshold can be modulated to better adjust the analysis to the data. For a less restrictive analysis, a Z-score of 3 may be used.

Furthermore, the region of the spectra used to calculate the average Z-score for each measurement can be changed to adjust the shape of the spectra. As there are flat regions in the spectra, the Z-score for those regions should be low, as there is no signal. Therefore, the final Z-score will be underestimated if those regions are included.

Hotelling's T^2 vs Q residuals

The second filter for the outlier detection is the Hotelling's T^2 versus Q residuals test. This filter is one of the most used and visually intuitive techniques for outlier detection in FTIR data (Morais et al., 2020). In this test, a chart is created representing the Hotelling's T^2 for each measurement in the x-axis, and the Q residuals in the y-axis. Both of these metrics rely on fitting the data into a PCA. The first metric represents the sum of the normalized squared scores of the PCA, which is the distance from the multivariate mean to the projection of the measurement in to the PCs (Morais et al., 2019). A Hotelling's T^2 value near 0 indicates that the measurement has an average score for all

the PCs (which would be the average of the spectra). A high Hotelling's T^2 value indicates that it has extreme scores for the PCs, indicating that it differs greatly from the rest of the data. The final value for the Hotelling's T^2 is arbitrary, and will depend on many factors. However, the more dimensions that the PCA takes into account, the higher the T^2 will be. As more information is represented in the PCA, the projections tend to be further from the multivariate mean.

The second metric represents the sum of squares of each measurement in the error matrix, giving an estimation of the difference between a measurement and its projection in to the PCs (Morais et al., 2019). A low Q residual value, close to zero, indicates that the measurements are being properly represented by the PCs, while a high Q residuals value indicates that the representation with the PCs misses out in accurately describing the data. In other words, a high Q residuals value means a poor fit of the PCA model. This happens because the measurement does not follow the tendency of the rest of the data, which is common for outliers. As contrary to the Hotelling's T^2, Q residuals tend to be lower when the number of dimensions in the PCA increases. As more information is represented, the representation in the PCs is more accurate, therefore the missed information is lower.

When representing the chart, all measurements far from the origin are considered candidate for outliers. They stand out from the majority of the data, which should be near the origin of the chart. Outliers must be investigated and removed from the analysis. Measurements should be removed one by one, as PCA models are highly sensitive to outliers, and the result may be distorted (Morais et al., 2020). It is recommended to adjust the region of the spectra where this analysis is performed, as flat regions could distort the results and underestimate both Hotelling's T^2 and Q residuals values.

Once outliers have been detected, it is recommended to visually inspect them, try to identify their origin and remove them from the dataset. If the scientist considers it, outliers may be included in posterior analysis, or they might undergo an independent analysis to perform an in-depth study of them.

3.2.4. Further analyses

This last module is intended to extract information and draw conclusions from the spectral data. The way to do this is to use statistical tests to proof the hypothesis that arise along the process, especially in the previous module. For this purpose, there are several statistical tests that may be used, depending on the characteristics of the data. The data used in this study pretends to detect differences in the IR spectrum of individuals with different conditions. These conditions allow to classify the individuals in groups. So, to compare those groups, the most suitable statistical method is the ANOVA test.

3.2.4.1. Ratios' calculation

To statistically analyze IR spectra, the ratios between different peaks are used, in order to remove possible variations in the sample thickness or in the IR absorption efficiency. This way the changes in biochemical composition are less biased and more faithful. The main ratios utilised in this study are shown in Table 2, alongside what they represent. Also, this can be modified in order to calculate any desired ratio. To calculate each peak, the wavenumber associated with a given bond plus the adjacent wavenumbers are taken and averaged. Then the ratios are calculated for each of the measurements.

3.2.4.2. ANOVA test

To compare the differences in the mean of two independent groups the Student's t-test is usually applied. However, if more than two groups are compared, the procedure is not as straightforward. One may apply several t-tests, comparing the multiple groups in pairs, until every possible comparation is made. The problem comes from the increase in Type 1 error (false positive) probability that this would carry (Kim, 2014). To avoid this, the alternative is to use ANOVA test. This test is made to compare the means for more than two independent groups of individuals (Ståhle & Wold, 1989). When performing this test, a p-value is given, which indicates the probability that the analysed data arises while the null hypothesis (the means of the analysed groups are equal) is true. So, to decide if a given hypothesis is true or not, a threshold for the p-value is stablished. If the p-value for the comparison is lower that the set threshold, it is assumed that the

means of the analysed groups are not equal and, in fact, there is a difference between them.

Table 2: Main ratios used for the statistical analysis alongside their interpretations and the frequencies used to calculate them.

Ratios	Biological interpretation	Frequencies (cm^{-1})
CH_2/Amide I	Changes in lipid / protein ratio	2852 / 1656 2921 / 1656
C=CH/CH_2	Changes in lipid composition	3012 / 2852
C=O/CH_2		1743/ 2852
CH_2/CH_3		2852 / 2960
Amide I (β) / Amide I (α)	Changes in protein secondary structure	1637 / 1656

Furthermore, this comparison method allows also to test for more than one independent variable. When the different groups of individuals are made only from one independent variable, the so-called one-way ANOVA is used. When the groups of individuals are based on two independent variables, the two-way ANOVA is used. This allows to examine the effect of both factors over the population, as well as the interaction between them (Wilcox, 2021).

However, an ANOVA test only allows to determine if the means of all the compared groups are equal or not. If the null hypothesis is rejected, the model does not have the ability to determine which of the means differs from the other and in which degree. To assess this limitation, post-hoc tests are used. Post-hoc tests perform additional comparisons using all possible pairs, in order to determine which group differs from which other group (Kim, 2014).

The simplest way to control the Type 1 error while performing all pairwise comparisons is using Tukey's honestly significant difference (HSD) test. This test uses the Q statistic critical values calculated from the number of groups and number of individuals within groups to determine if the difference between a given pair of groups is significant (Kim, 2015). This test assumes that all groups have an evenly distributed size. When this is not the case, the modified Tukey-Kramer HSD test is applied to overcome this limitation (Kramer, 1956). This test I may be applied to 1-way or 2-way ANOVA equally. For all statistical analyses done in this study, a p-value of 0.0001 has been used to determine the significancy of the interactions.

3.2.4.3. Data visualization tools

Usually, after performing the proper statistical tests, data visualization is a common procedure, as it is a clear way to understand, obtain and represent conclusions. In this case, boxplots or bar plots are used to represent the mean of the ratios for each of the conditions analysed, plus the standard error means. This is way to easily visualize the tendency of the data for a given region of the CNS, or for a given developmental stage.

Then, a correlogram is used to visually represent the results from the Tukey's HSD test. This plot displays all the conditions included in the analysis along the x and y axes. The intersection of two given conditions in the plot will represent using a two-colour system whether this interaction is statistically significant or not.

Table 3: Summary for all the measurements included in the dataset, classified by experimental group (WT, IL6$^+$, IL10$^+$), developmental stage (P0, P7, P14, P21, P28) and region of the CNS (CB-WM, CB-GMM, CB-GMG, BR-WM, BR-GM)

	Stage	CB-WM	CB-GMM	CB-GMG	BR-WM	BR-GM	Total
WT	P0	144	144	144	144	144	720
	P7	144	144	96	192	192	768
	P14	192	192	192	192	192	960
	P21	144	192	192	96	96	720
	P28	192	192	176	144	144	848
	Total	816	864	800	768	768	4016
IL6$^+$	P0	192	182	144	144	144	806
	P7	144	144	96	96	96	576
	P14	144	144	144	144	144	720
	P21	192	193	192	144	144	865
	P28	192	192	192	144	144	864
	Total	816	807	768	720	720	3831
IL10$^+$	P0	96	96	96	144	144	576
	P7	144	145	144	176	192	801
	P14	192	192	192	144	144	864
	P21	48	48	48	96	96	336
	P28	192	192	192	144	144	864
	Total	672	673	672	704	720	3441
		2304	2344	2240	2192	2208	11288

4. Results and Discussion

This pipeline was applied to the SR-μFTIR spectral data described and used in Peris et al. (2023) as a method of validation for the pipeline. Also, similar data from two transgenic groups was included in the analysis. In total, the dataset initially includes 11288 measurements before carrying on any procedure. The samples were interrogated with IR radiation in the range of 3200-950 cm^{-1}. A summary table of these measurements can be found in Table 3. Also, Figure 3 contains a graphical representation of the initial dataset.

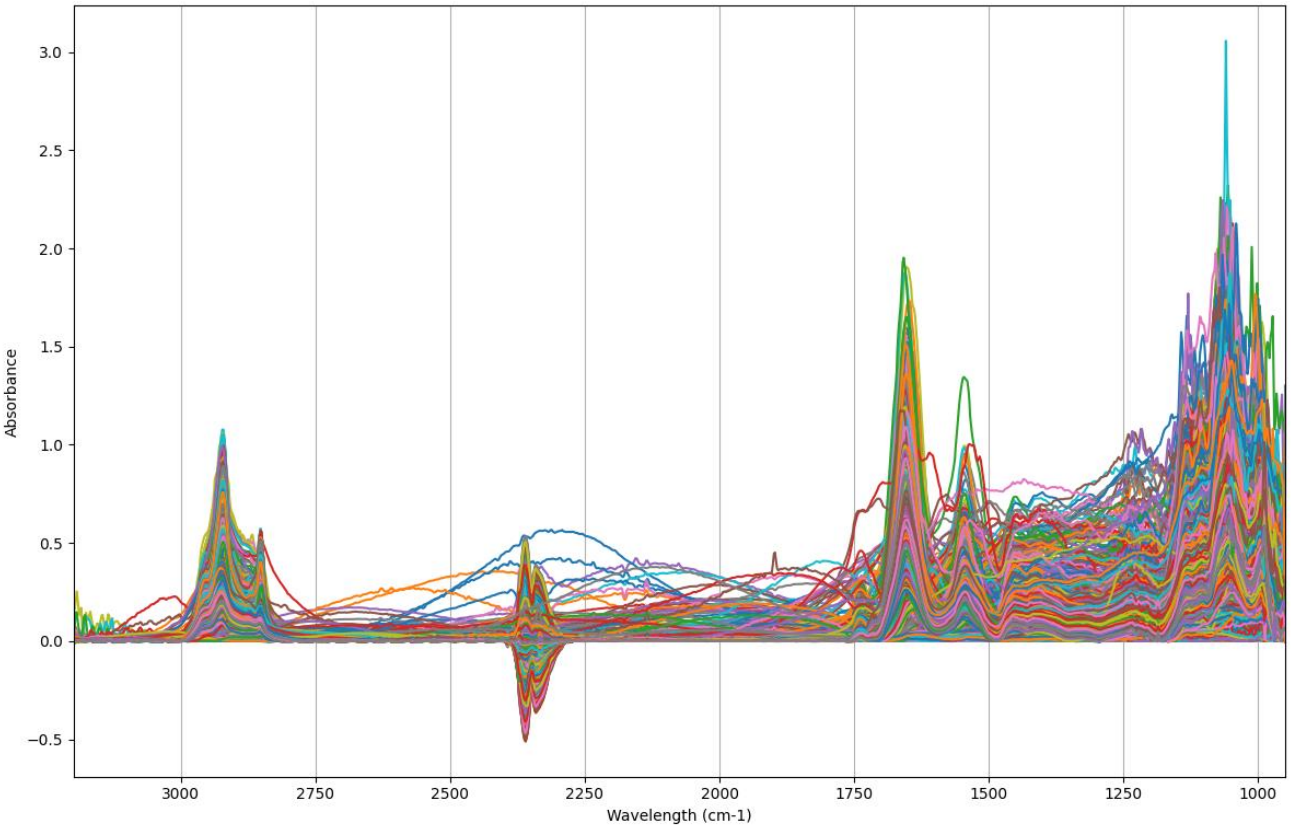

Figure 3: Graphical representation of the IR absorbance of all the measurements present in the dataset before undergoing any processing step.

The first step of the analysis is the data curation module. Within this module, the Amide I filter and the non-flat measurements filter are applied, to remove low-quality spectra and experimental errors. From the first of these filters, 446 measurements are removed from the dataset, and labelled as low-quality data (Fig. 4). Then, the second filter is applied over the remaining 10842 measurements. This filter examines the frequencies inside the 2800-2380 cm^{-1} range, and the 2000-1800 cm^{-1} range. It interpolates a quadratic function that approximates to the spectrum shape and calculate the area under the curve of that interpolated function. The threshold for this area is 0.18 and 0.08 respectively for the two ranges. 48 measurements are labelled as non-flat for the first range, 133 for the second and 34 for both. In total, 10627 measurements remain after the data curation module, that are represented in Figure 5. No normalization step is applied, as the original study did not used any normalization method. However, spectral data may be normalized by Amide I peak normalization for a better visual representation (Peris et al., 2023).

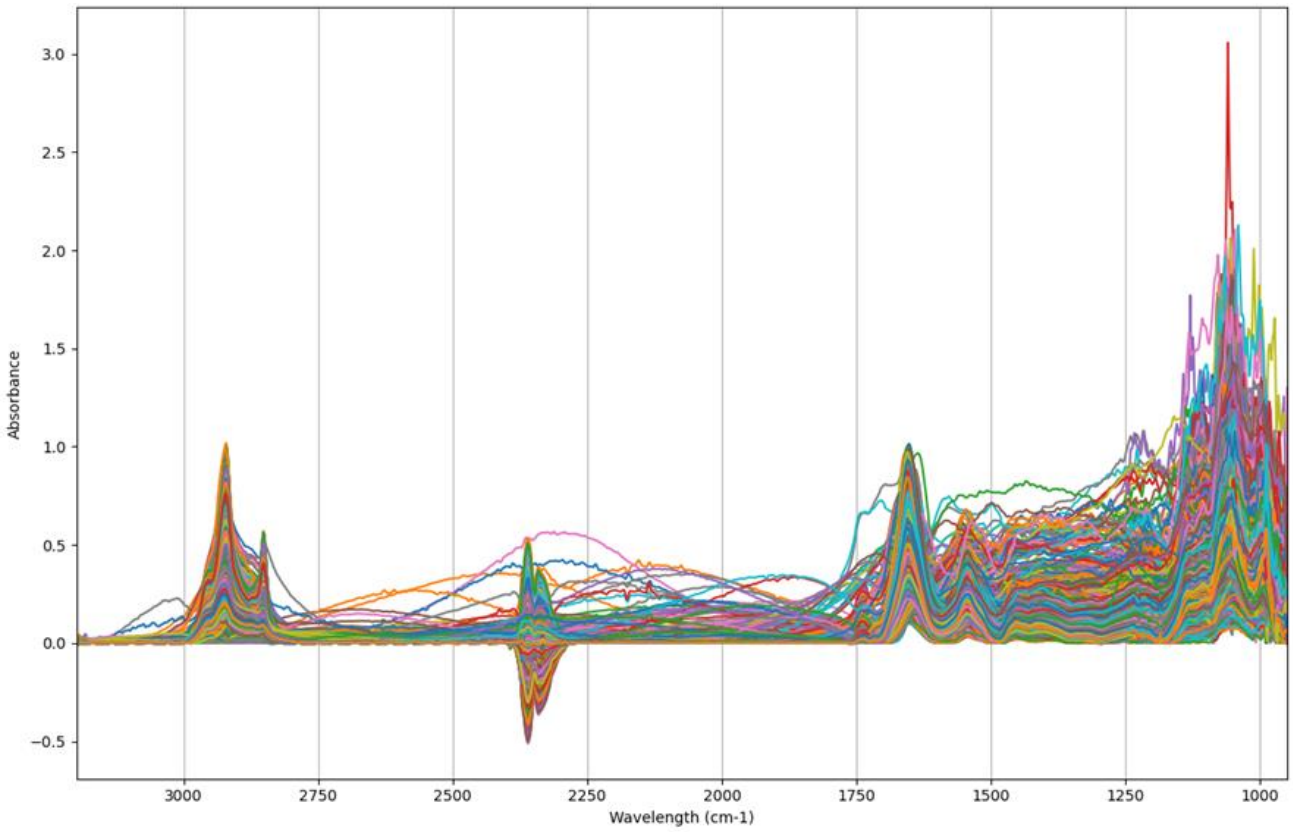

Figure 4: Graphical representation of the IR absorbance of all the measurements remaining in the dataset after the Amide I filter is applied.

After data curation, the dataset undergoes the data pre-processing module. Firstly, the Savitzky-Golay filter is applied to correct baseline contributions, enhance narrow bands and remove instrumental noise. The parameters used for this filter must be carefully chosen. In the original study, Peris et al. (2023) performed Savitzky-Golay filter with an 11-points window size, and 3rd polynomial order. However, as this pipeline integrates the application of SVD as a noise reduction tool, 13-points, 9-points and 7-points window size will also be tested, to see which one has the best performance. All spectra, with the different parameters used are shown in Figure 6.

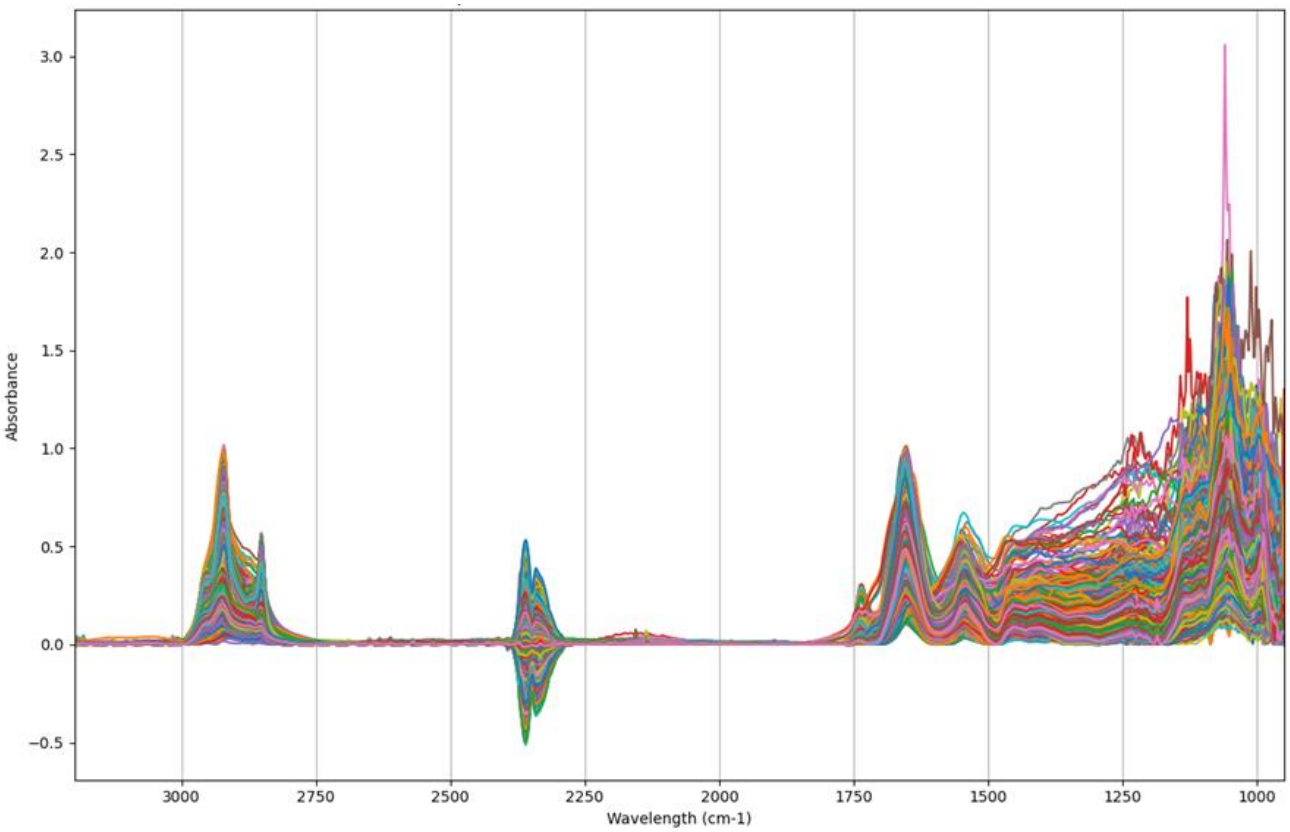

Figure 5: Graphical representation of the IR absorbance of all the measurements remaining in the dataset after the non-flat measurements are removed.

After the differentiation of the spectra with the SG filter, SVD is applied to the spectral data, which is represented as a *A*, which is a 10627 x 1171 matrix given the 10627 measurements and the 1171 wavenumbers. *A* is converted into the product of 3 matrices,

U, S and V^T, as shown in Eq. 2. Using the approximation in Eq. 3, A_k is calculated by taking the k first singular values of S. To choose the value of k, all singular values σ_{ii} are plotted against i, as shown in Figure 7. The appropriate value of k can be inferred by looking for the "elbow" of this plot. This pipeline includes an option that calculates this k value, based on the angle formed by the x-axis and a straight line connecting two consecutive points in the plot. This procedure is done for all 4 possibilities of the SG filter. For the 13-points window, as well as for the 9-points window, the appropriate k is set to 7. For the 11-points k is set to 7, and for the 7-points window k is set to 10.

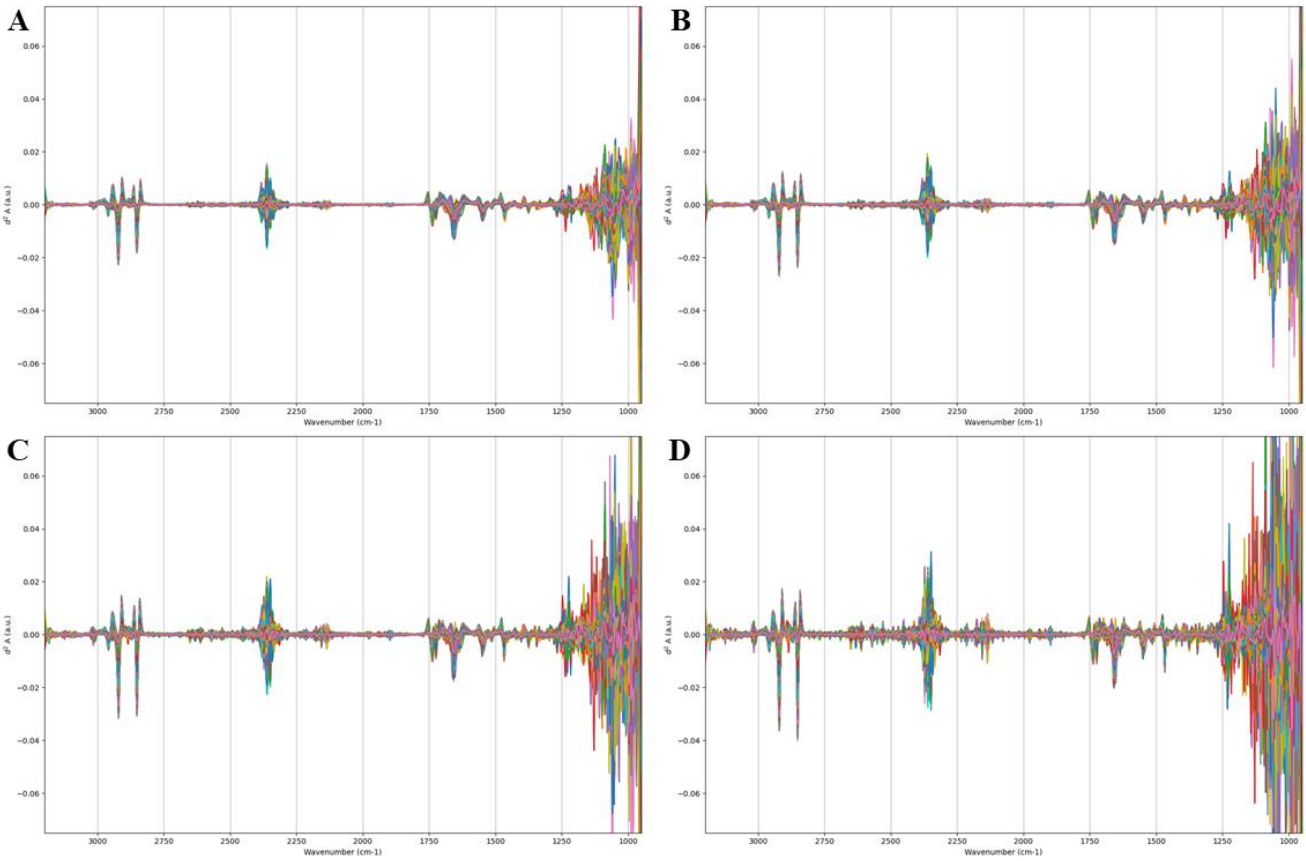

Figure 6: Representation of the spectral data after applying the Savitzky-Golay filter with the following parameters: second derivative, third polynomial order and 13 (A), 11 (B), 9 (C) and 7 (D) points for the window size.

As explained in Section 2.2.2.2., the application of WSVD includes two types of weighting, a column-wise weighting corresponding to the wavenumbers, and a row-wise

weighting corresponding to the measurements. For the column-wise, the regions with higher importance must be defined. In this case, the spectrum presents two clear regions where the relevant signals appear. Those regions are the 3000-2800 cm^{-1} region and the 1750-1100 cm^{-1} region, which will be the weighted regions. Then, for the row-wise weighting, the SNR is calculated for each measurement and normalized between 0 and 1. These values will be used to construct the W_r matrix (Eq. 6). The resulting spectra after the low-rank approximation by WSVD are depicted in Figure 8. Based on the spectral shape of all four possible parameters for the SG filter, from now on the analysis will be conducted with the 9-points window size. The reason for this is that this has a similar amount of noise than the 13 and 11 points, but in theory it collects more features than these two. The 7-points window size is disregarded as the noise is more prominent than in the 9-points.

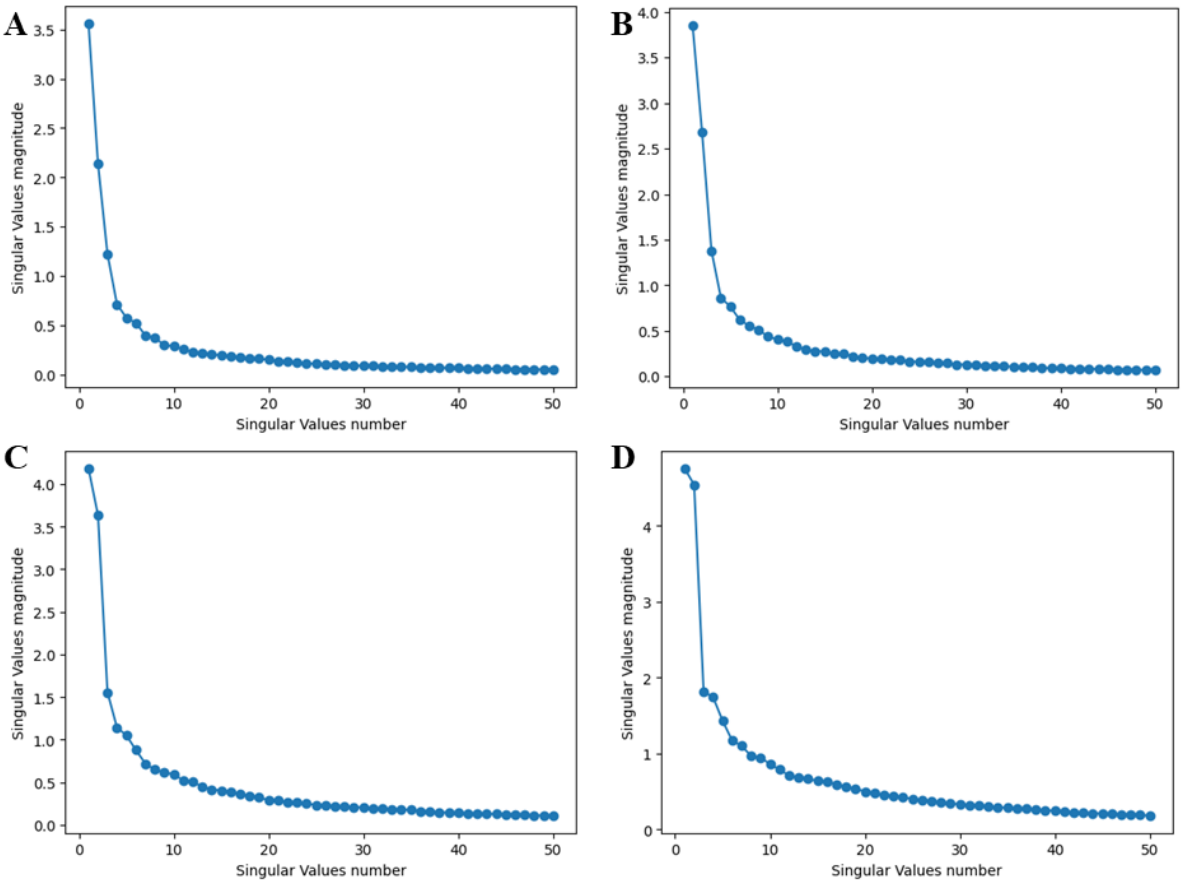

Figure 7: Graphical representation of the singular values magnitude vs. the singular values order of the spectral data after Savitzky-Golay filter is applied (A: 13-points window size, B: 11-points window size, C: 9-points window size, D: 7-points window size)

Now, an exploratory analysis is conducted. Firstly, a PCA featuring the fingerprint region (1800-1300 cm^{-1}) of the IR spectrum is done, in order to detect any possible deviation or strange behaviour. Two independent PCAs are conducted for the brain and for the cerebellum measurements. This procedure is done first with all the measurements belonging to the WT group, which is the data used in Peris et al. (2023). Then, the same steps will be conducted with the other two groups, $IL6^{+}$ and $IL10^{+}$, and the results will be compared (Fig 9).

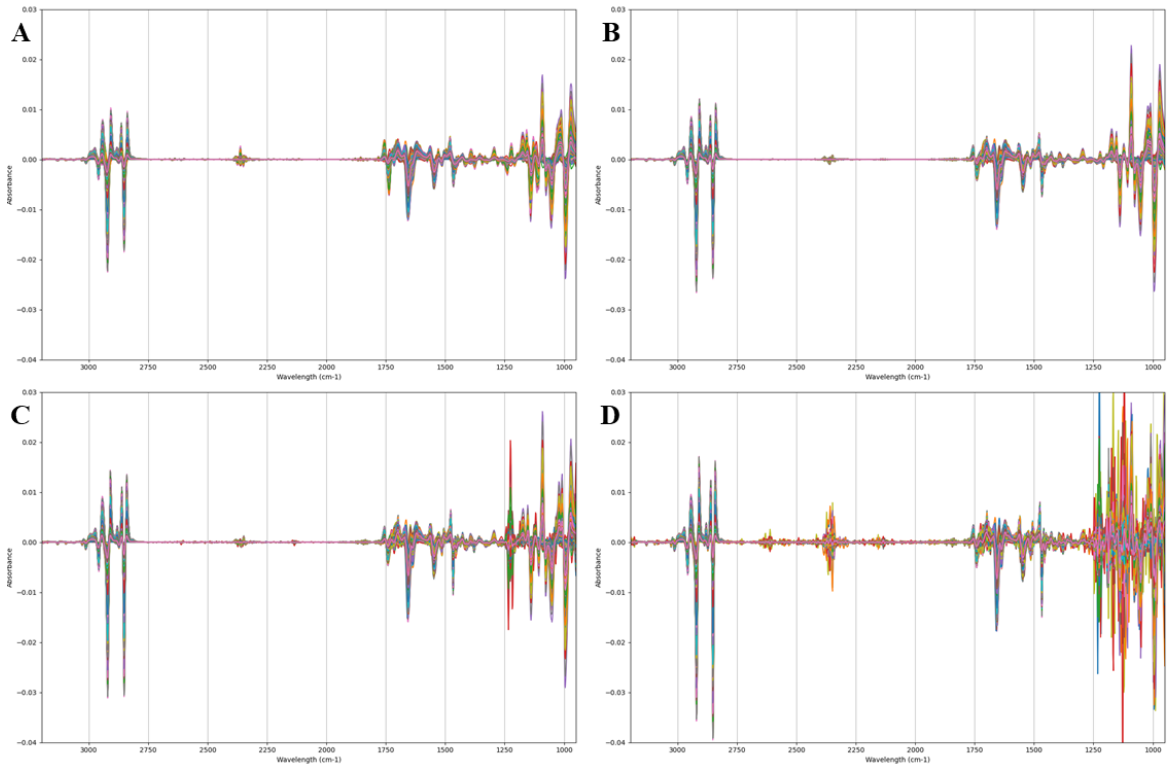

Figure 8: Representation of the spectral data after low-rank WSVD approximation for the 13 (A), 11 (B), 9 (C) and 7 (D) points window size data. The number of singular values taken in the approximation are 4 for A, B and C, and 6 for D. The weighted regions are 3000-2800 cm-1 and 1750-1100 cm-1.

The PCAs for the WT group shows a similar distribution as presented in Peris et al. (2023). For the brain's PCA (Fig. 9A), the measurements from white matter (WM) tissue, at P21 and P28, form a main cluster separated from the rest of the data at the top

of the plot. Then, the rest of the data is located at a central cluster, with some dispersion for the measurements from grey matter (GM) at early developmental stages (P0, P7, P14). The cerebellum's PCA (Fig. 9B) shows a similar top cluster with the late developmental stages (P21, P28) of the WM measurements. Then, the rest of the data points are located in a central cluster. However, the P14 measurements of WM appear in a third small cluster between the former two. This behaviour is also similar at the one seen in Peris et al. (2023).

Regarding the other two groups, the distribution is very similar to the observed in the PCAs for the WT group. However, one difference stands out. In the PCA for the cerebellum of the $IL6^+$ group (Fig. 9D), the measurements for P14 in WM are located at the top cluster. Furthermore, there are several samples from WM at P0 that appear in the top cluster composed by the late stages of WM.

To further analyze the distribution of the data and any possible outlier, the Hotelling's T^2 vs Q residuals method was applied within the fingerprint region. Just as with the PCAs, this technique is applied separately for the brain and cerebellum, and for each group independently. In general terms, the results were as expected. No major outlier measurement was detected in any of the cases. However, within the $IL6^+$-CB dataset, a systematic deviation was observed. When looking at the origin of these measurements, it was seen that many of them belonged to the same individual, and to the same region of the cerebellum. Specifically, these measurements were taken from a P0 pup, identified as N150, in the (WM) region. In Figure 10A the Hotelling's T^2 vs Q residuals plot is shown with the measurements in question highlighted in red. In this case, the more probable explanation for this behaviour is that in this sample there was some kind of experimental error at the data acquisition step. Therefore, all 42 measurements from this sample were removed from the dataset. 10585 measurements remain in the dataset and will undergo the further analyses.

Once outliers have been evaluated and removed, the next step is to perform statistical tests. These tests study the ratios of specific key peaks and how they change along groups and conditions. Therefore, the ratios have to be previously calculated. The ratios used in this study are shown in Table 2.

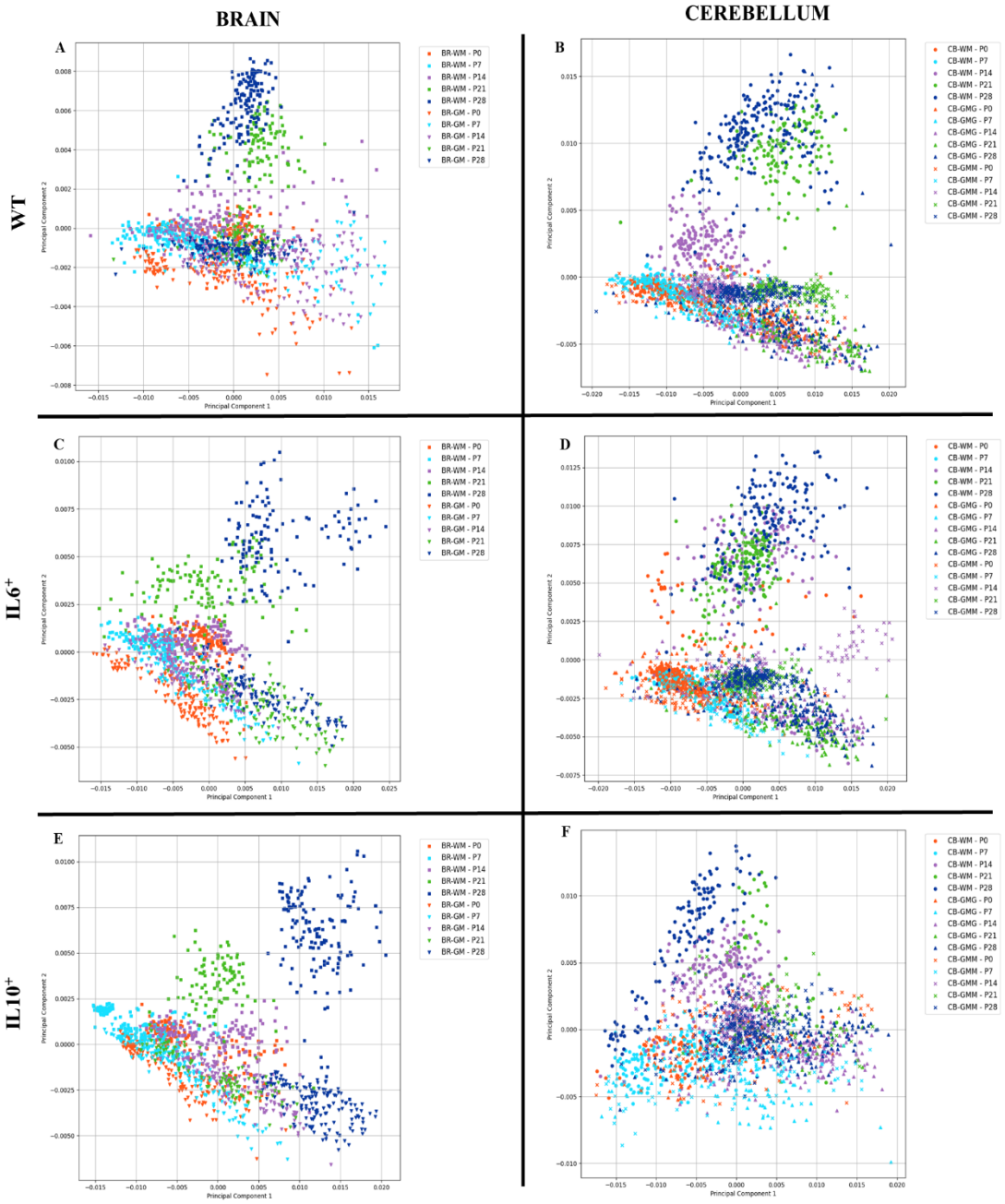

Figure 9: PCA in the second derivative of µFTIR data on the fingerprint region after WSVD low-rank approximation. Score plots for WT (A, B), IL6$^+$ (C, D) and IL10$^+$ (E, F) of brain (A, C, E) and cerebellum (B, D, F).

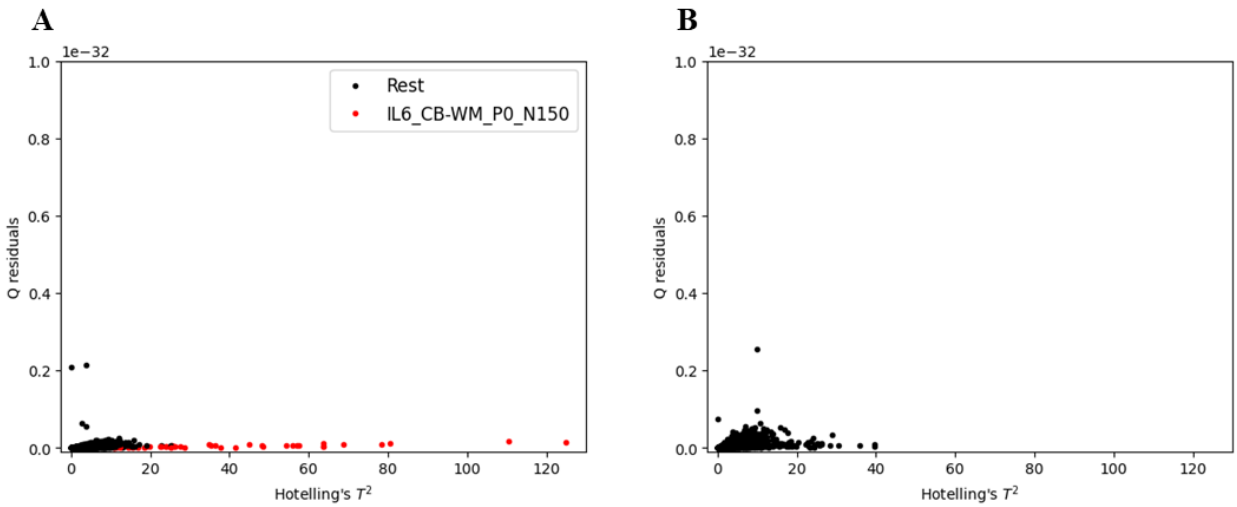

Figure 10: Hotelling's T^2 vs Q residuals plot for the measurements of the cerebellum from IL6$^+$ individuals. A shows the plot with all measurements; B shows the plot after removing the measurements from the IL6_CB-WM_P0_N150 outlier sample.

After the calculation of the ratios, a 2-way ANOVA test is performed to see the influence of the postnatal development and the different cerebral regions over the ratios. This test is done independently for the 3 groups. Then, to account for the influence of the groups, more tests must be performed. In order to make truthful comparisons between groups, the compared measurements must belong to the same cerebral region, and have to be in the same developmental stage. Therefore, two new sets of ANOVA tests have to be done. The first set will evaluate the influence of the groups plus the cerebral region. For this, 5 2-way ANOVAs are performed accounting for the group and the region, while data is taken only from one developmental stage in each of the tests. Similarly, the second set of tests will evaluate the effect of the group and the developmental stage. For this, another 5 2-way ANOVAs are performed accounting for the group and the stage, taking the data from only one region in each of the tests. The statistical tests must be done for each ratio separately. In total, 13 2-way ANOVAs for each ratio are performed, which allows to make every possible comparison in the dataset.

To facilitate the understanding of the statistical analyses, 2 different plots are used. On the one hand, a bar plot represents the mean value for each of the interaction groups

for the two analyzed conditions. On the other hand, a heatmap is used to represent significance of the interaction between all the pairs of the analyzed conditions.

At this point, this pipeline is done, and the scientist is responsible of extracting any possible conclusions from the data. In this case, the main objective was to implement SVD within the data processing of µFTIR data. In order to evaluate if this has any effect on the analysis, the original results that do not include SVD will be compared with the ones obtained with this pipeline. To do this, the WT group will be used, as it is the one analyzed in Peris et al. (2023). Therefore, 5 different 2-way ANOVA tests will be performed, one for each ratio, evaluating the effect of the developmental stage and the CNS region.

4.1. CH_2/Amide I ratio

Comparing the original results with the new results, there are no visible changes in the CH_2/Amide I ratio ($d^2\ A_{2852}$ / $d^2\ A_{1656+1637}$) when graphically representing the data (Fig. 11A, 11B; 12A, 12B). Measurements from the same region at different developmental stages show the same tendency in both datasets. The same behaviour can be observed between measurements at the same stages from different regions. Another feature that can be observed is that the standard error means of some conditions are lower in the new dataset. This means that measurements show less intragroup variation when applying WSVD approximation, suggesting that this procedure is beneficial and may improve statistical results. If we look at the Tukey's HSD test results, there are some minor changes. In the ANOVA test made for the cerebellum's measurements, there is no change in the significancy of the interaction pairs (Fig. 11C, 11D) (Table 4). Contrarily, in the ANOVA for the brain, there are two pairs that are affected in the new dataset (Fig. 12C, 12D) (Table 5). On the one hand, the group P14:BR-GM with the group P21:BR-GM was found not-significant with the original dataset, but significant in the new one. On the other hand, the group P0:BR-WM and the group P28:BR-GM changed from significant to not-significant with the new procedure.

4.2. CH_2/CH_3 ratio

In the case of the CH_2/CH_3 ratio ($d^2 A_{2852} / d^2 A_{2960}$) the situation is very similar to the previous one. The ratios do not change drastically (Fig. 13A, 13B; 14A, 14B), but there are some changes in the Tukey's HSD results, as well as in the standard error means. In the analysis for the cerebellum, there are 6 interactions that change from not-significant to significant (Fig. 13C, 13D) (Table 4), all of them featuring the group P7:CB-GMM. Also, we can see that this specific group has a major decrease in the standard error mean after SVD application. This suggests that the group P7:CB-GMM was heavily affected by noise, which compromised the statistical results. Applying SVD removed this noise and improved the conclusions drawn from the analysis. For the brain analysis, 6 interactions changed from not-significant to significant, while one interaction did the opposite. Also, the group P28:BR-GM reduced its standard error with the SVD application (Fig. 14C, 14D) (Table 5).

4.3. Amide I (β) / Amide I (α) ratio

Looking at the Amide I (β) / Amide I (α) ratio ($d^2 A_{1637} / d^2 A_{1656}$) the tendencies of the data are pretty similar between the original and the new datasets (Fig. 15A, 15B; 16A, 16B). But, just as before, the standard error means decrease, and the Tukey's HSD test did show differences. For the cerebellum analysis, there are 9 pairs that become significant with the SVD application, while 4 become not-significant (Fig. 15C, 15D) (Table 4). For the brain analysis, those numbers are 5 and 3 respectively (Fig. 16C, 16D) (Table 5).

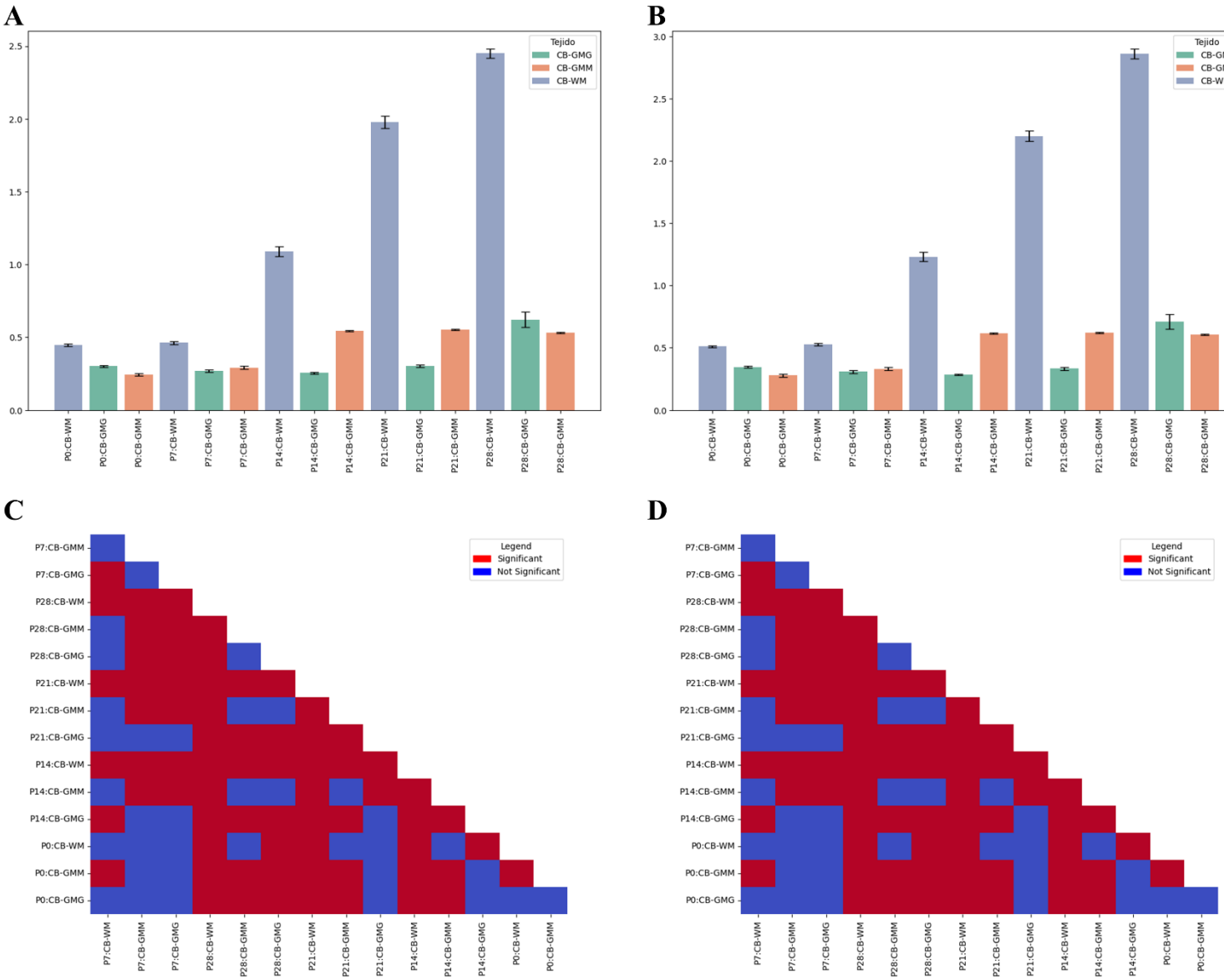

Figure 11: Bar plot representing the mean value and standard mean error for the CH_2/Amide I ratio for WT:CB measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of CB in the original (C) and new (D) datasets.

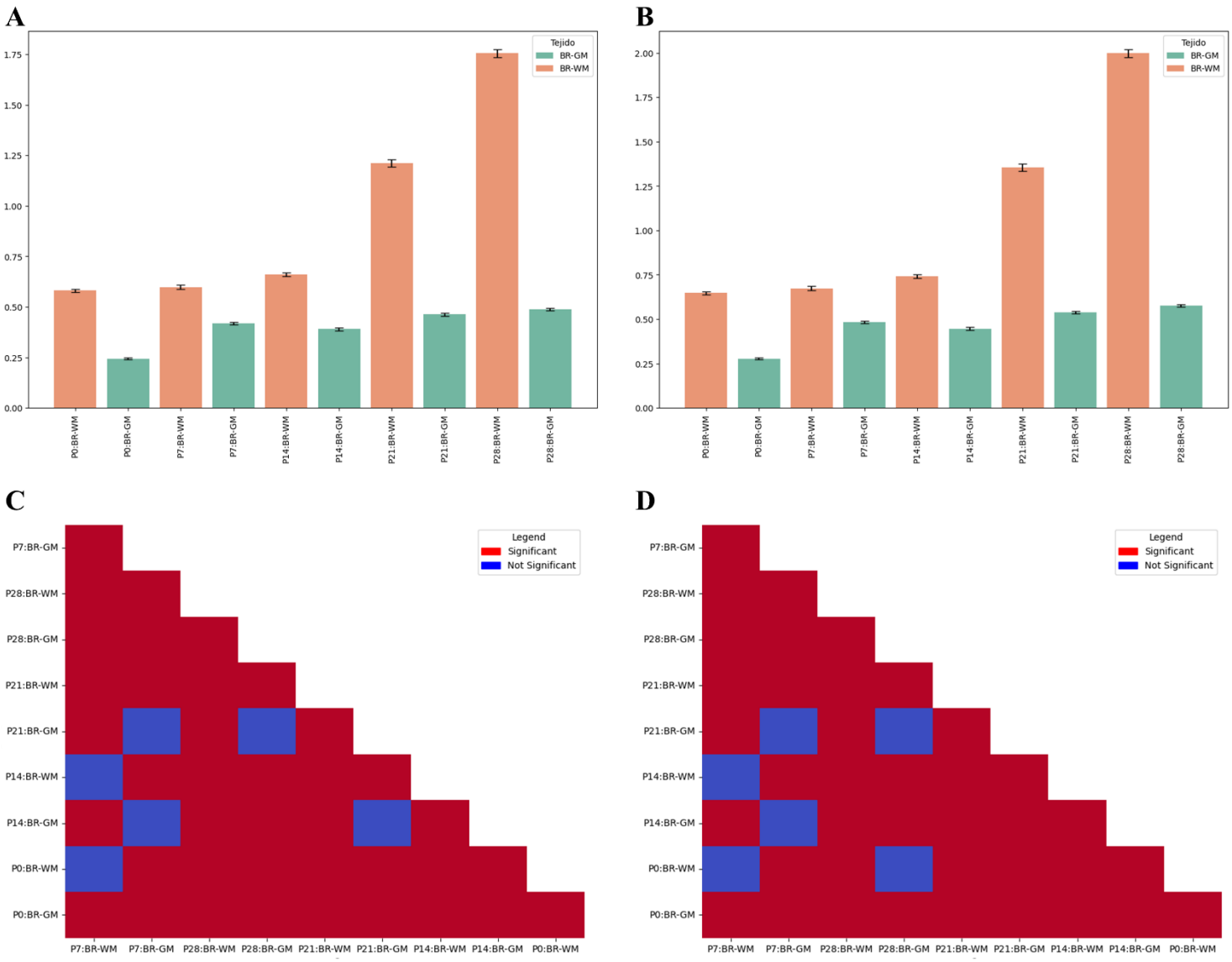

Figure 12: Bar plot representing the mean value and standard mean error for the CH_2/Amide I ratio for WT:BR measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of BR in the original (C) and new (D) datasets.

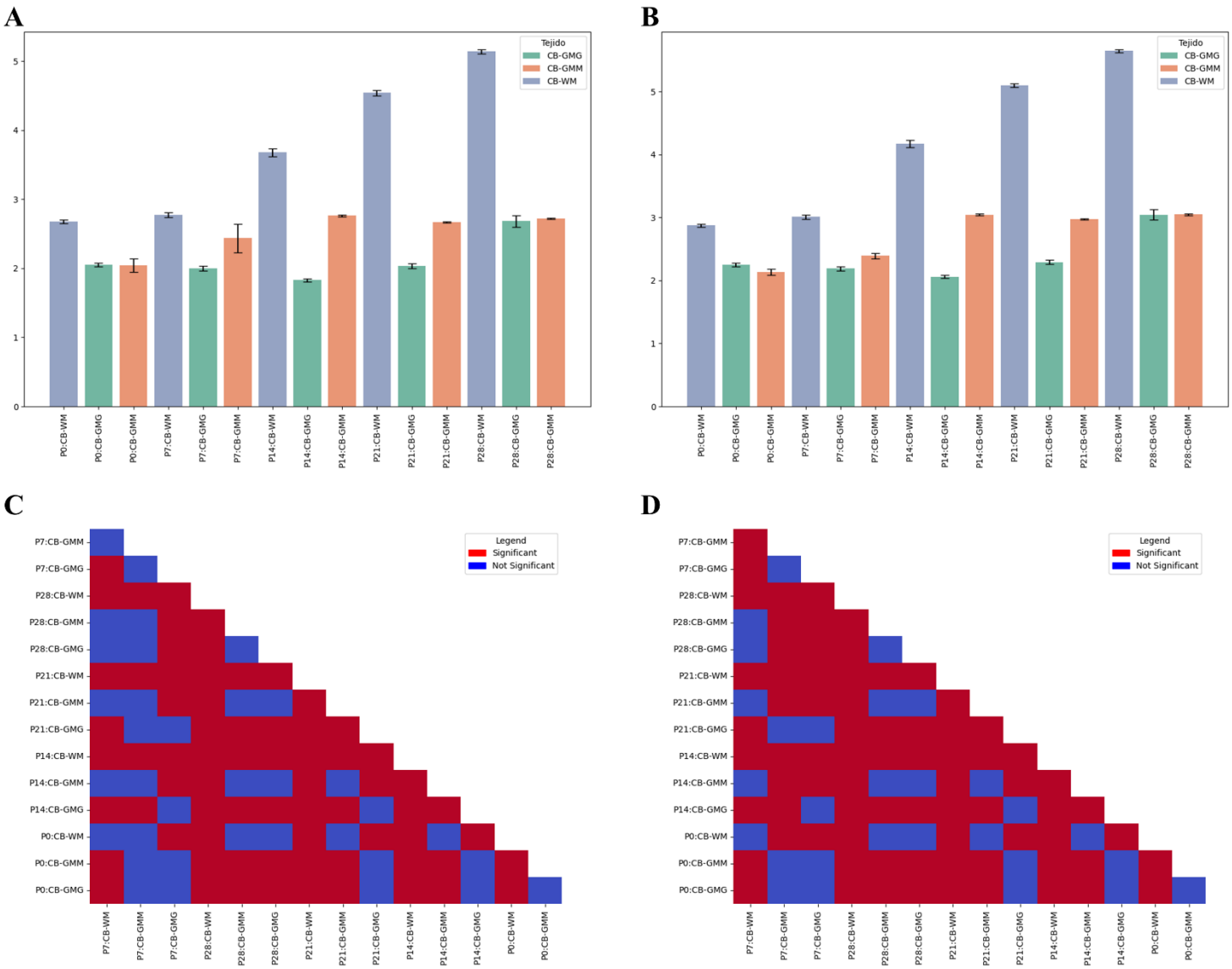

Figure 13: Bar plot representing the mean value and standard mean error for the CH_2/CH_3 ratio for WT:CB measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of CB in the original (C) and new (D) datasets.

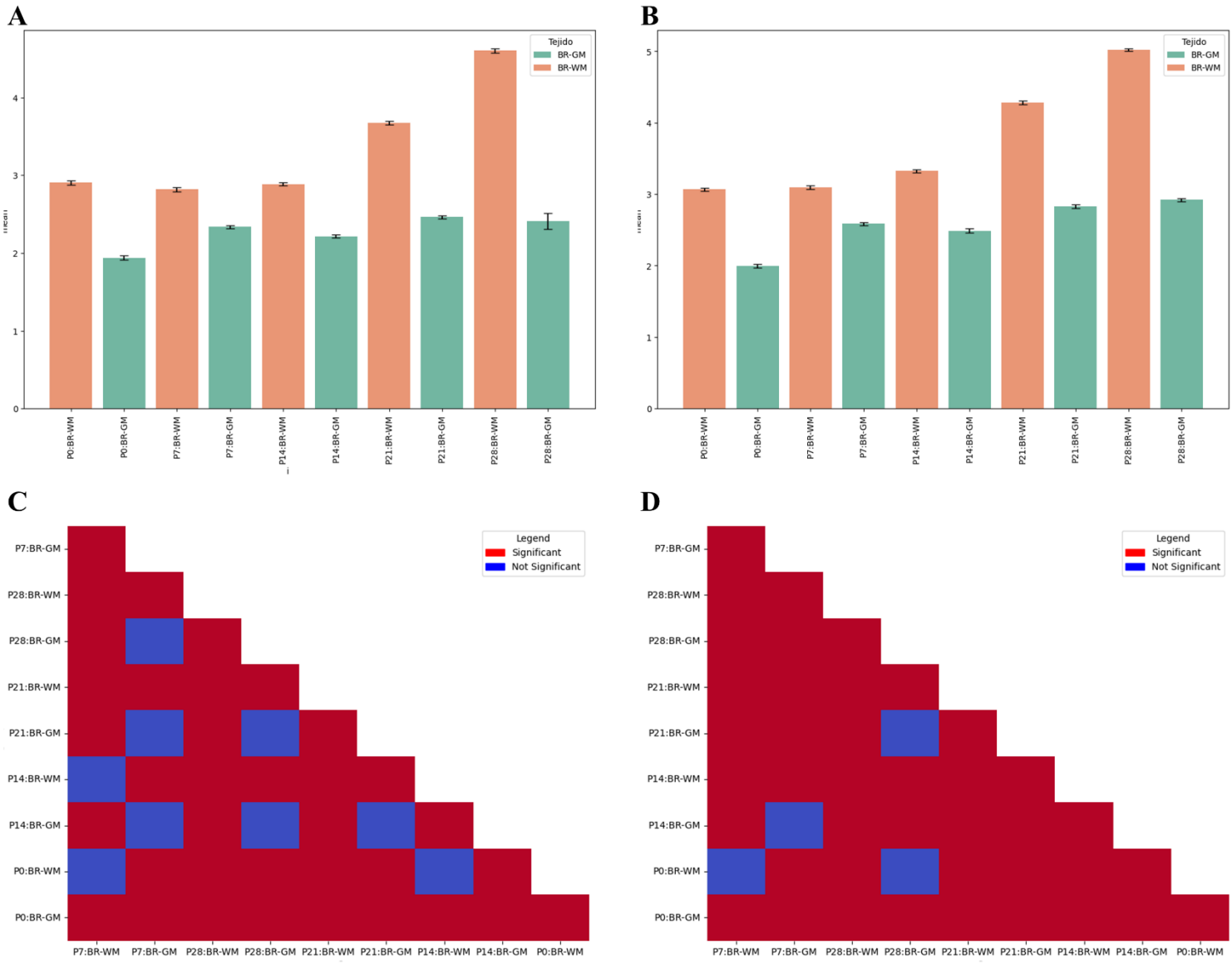

Figure 14: Bar plot representing the mean value and standard mean error for the CH_2/CH_3 ratio for WT:BR measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of BR in the original (C) and new (D) datasets.

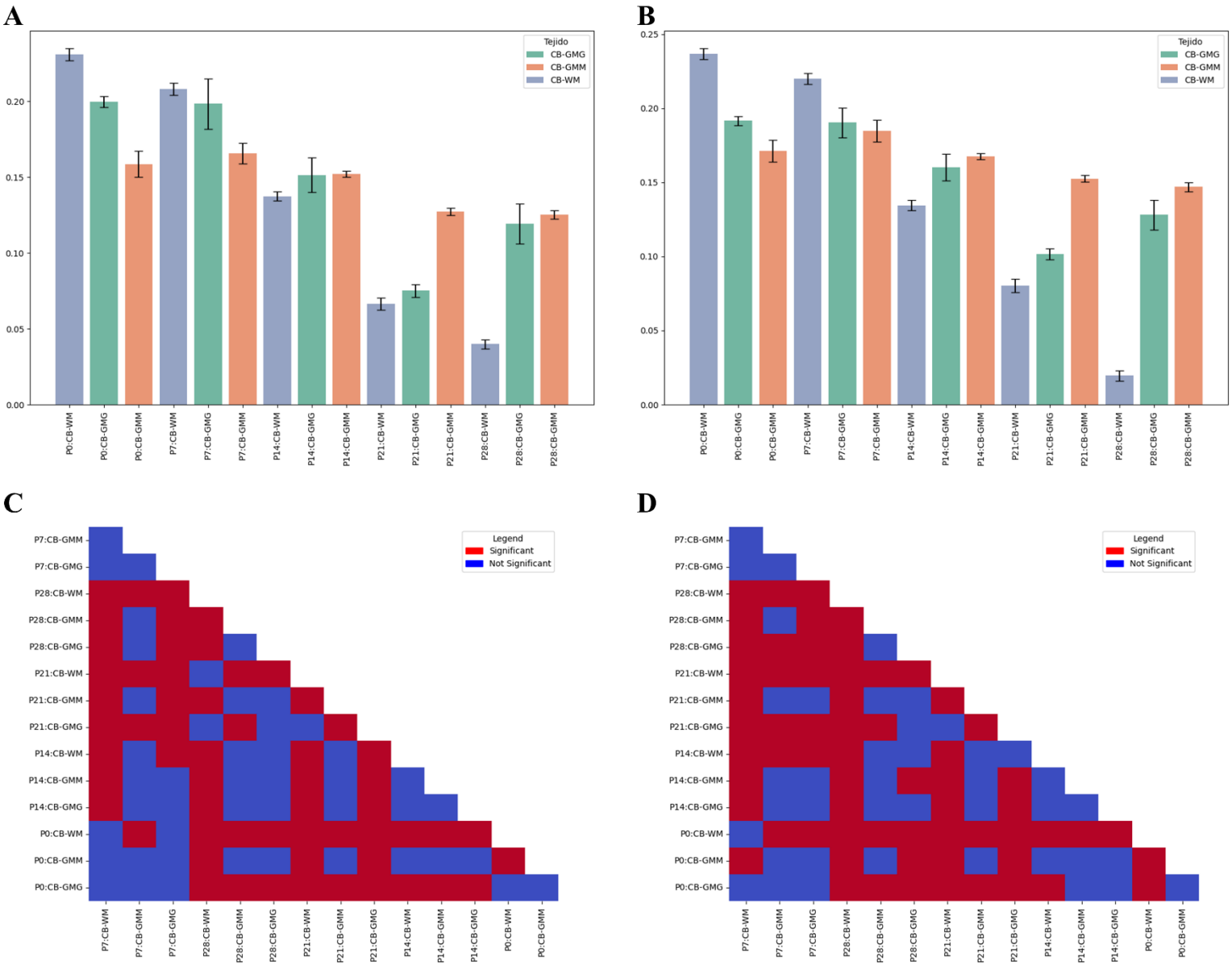

Figure 15: Bar plot representing the mean value and standard mean error for the Amide I (β) / Amide I (α) ratio for WT:CB measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of CB in the original (C) and new (D) datasets.

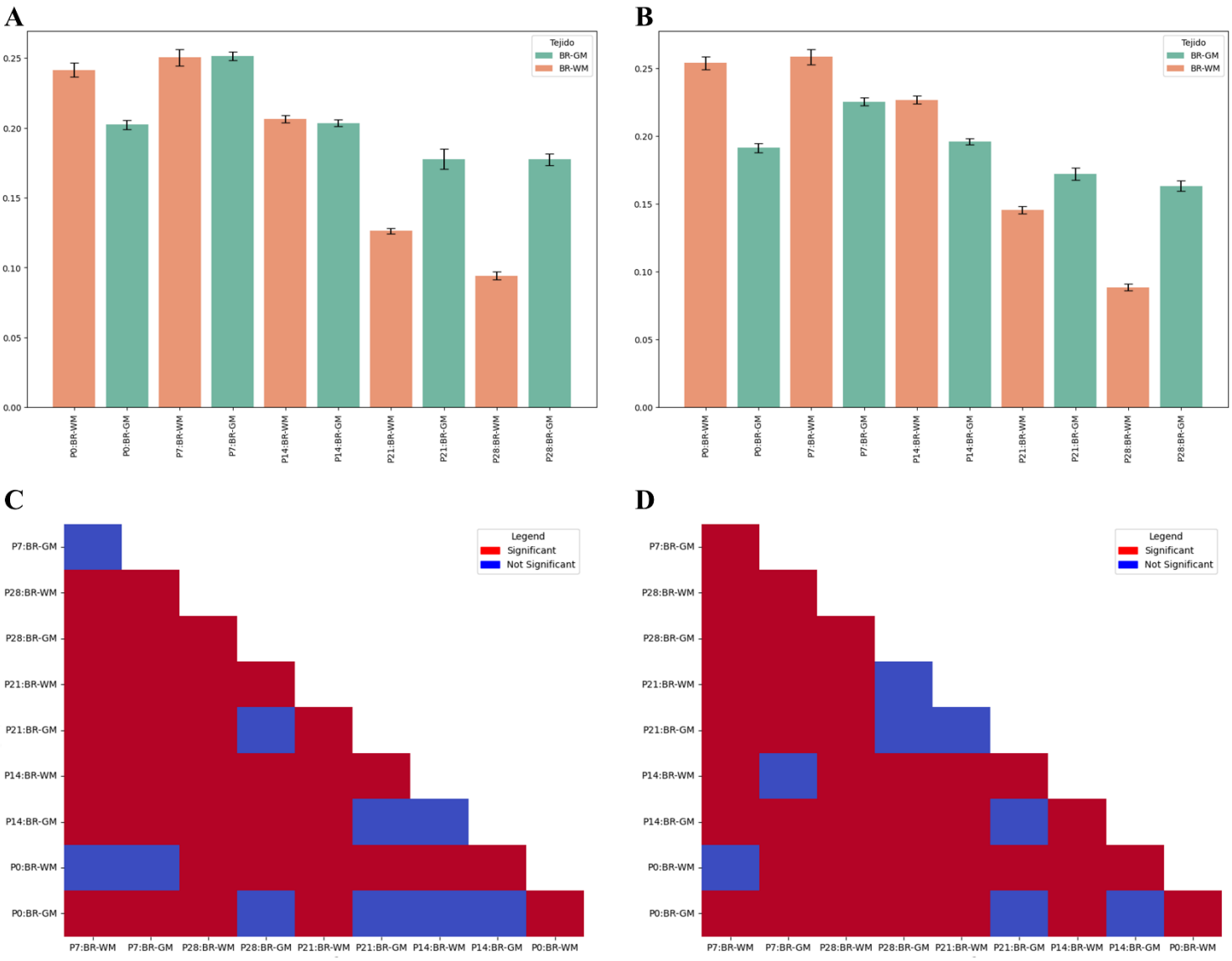

Figure 16: Bar plot representing the mean value and standard mean error for the Amide I (β) / Amide I (α) ratio for WT:BR measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of BR in the original (C) and new (D) datasets.

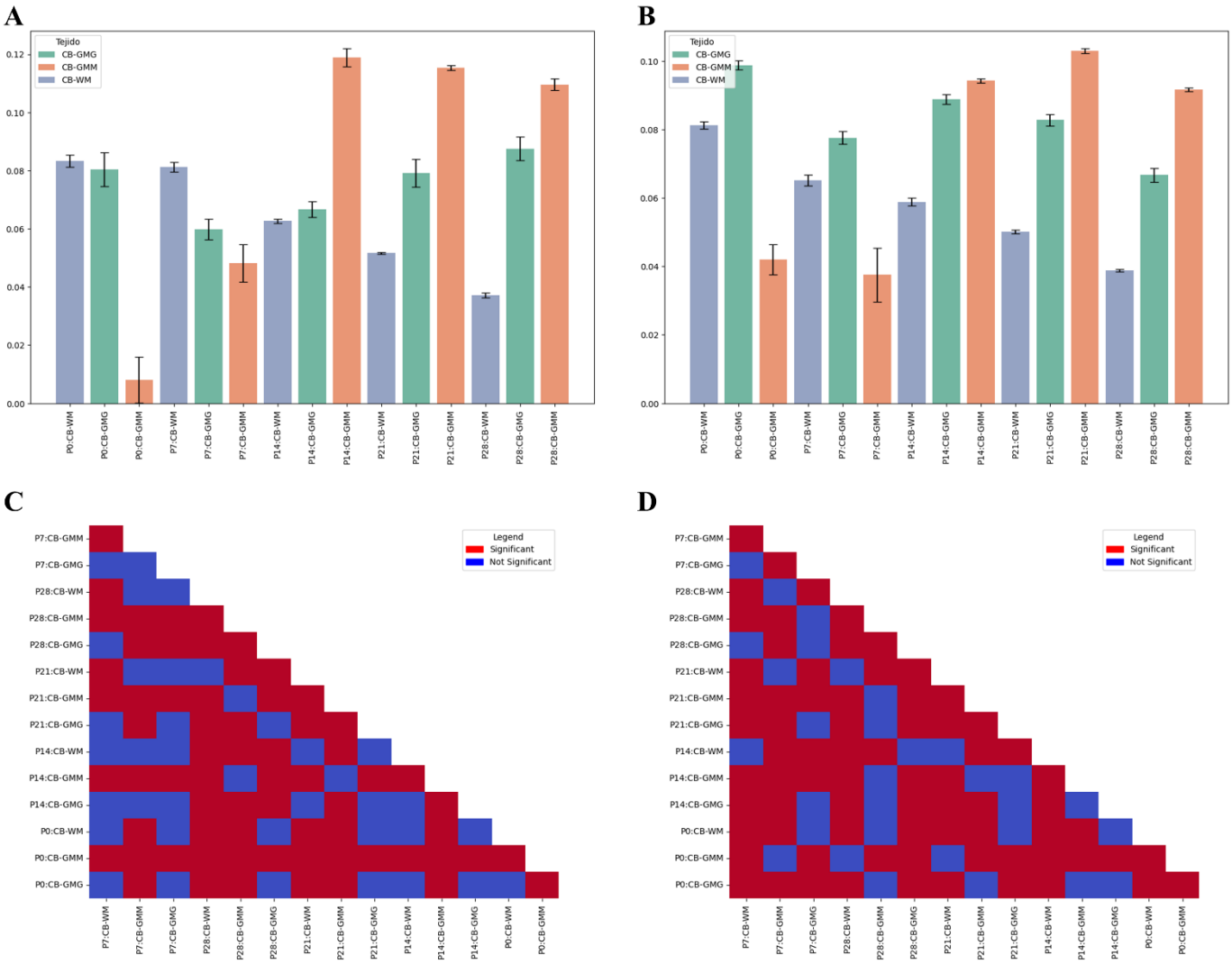

Figure 17: Bar plot representing the mean value and standard mean error for the C=CH/CH₂ ratio for WT:CB measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of CB in the original (C) and new (D) datasets.

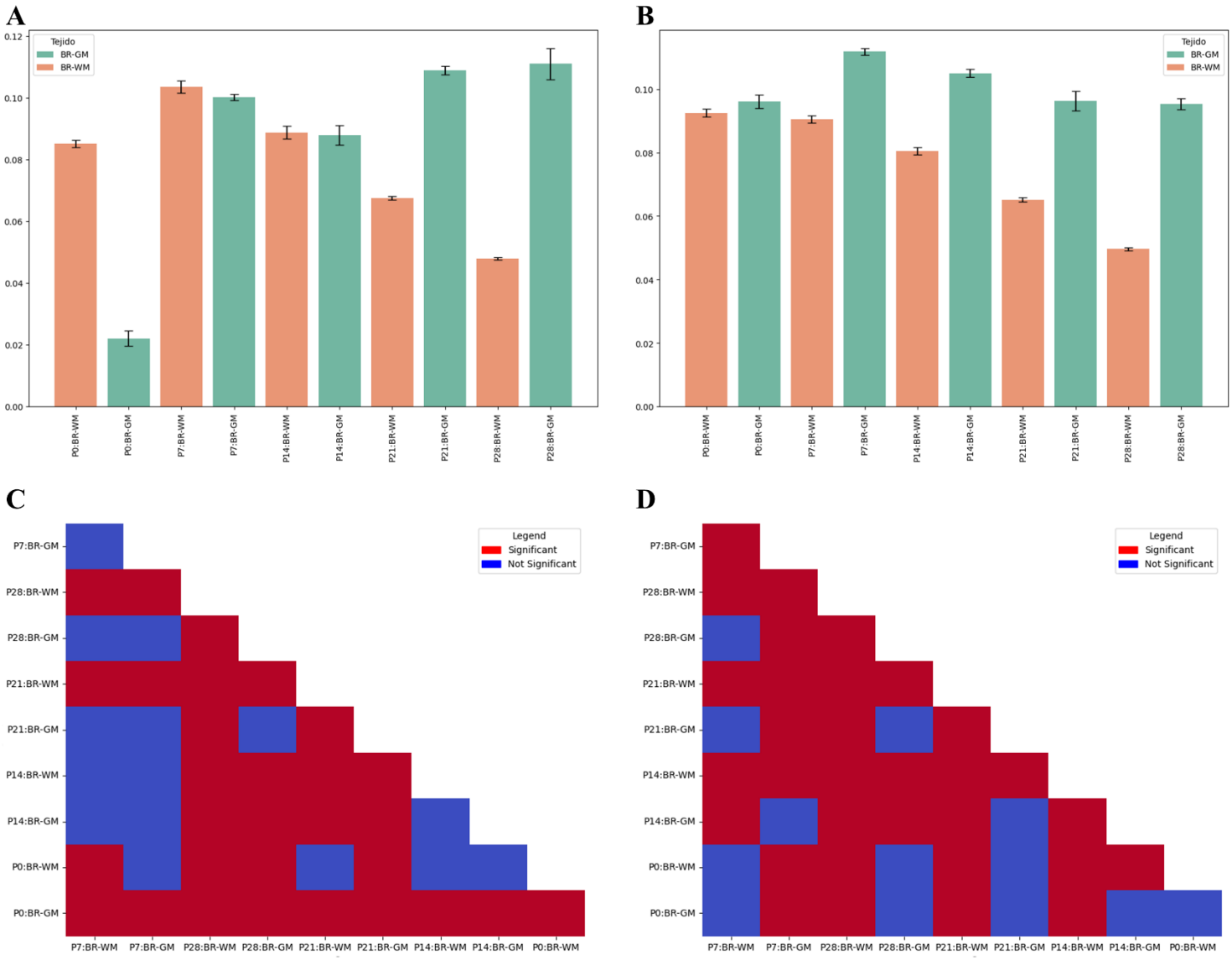

Figure 18: Bar plot representing the mean value and standard mean error for the C=CH/CH_2 ratio for WT:BR measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of BR in the original (C) and new (D) datasets.

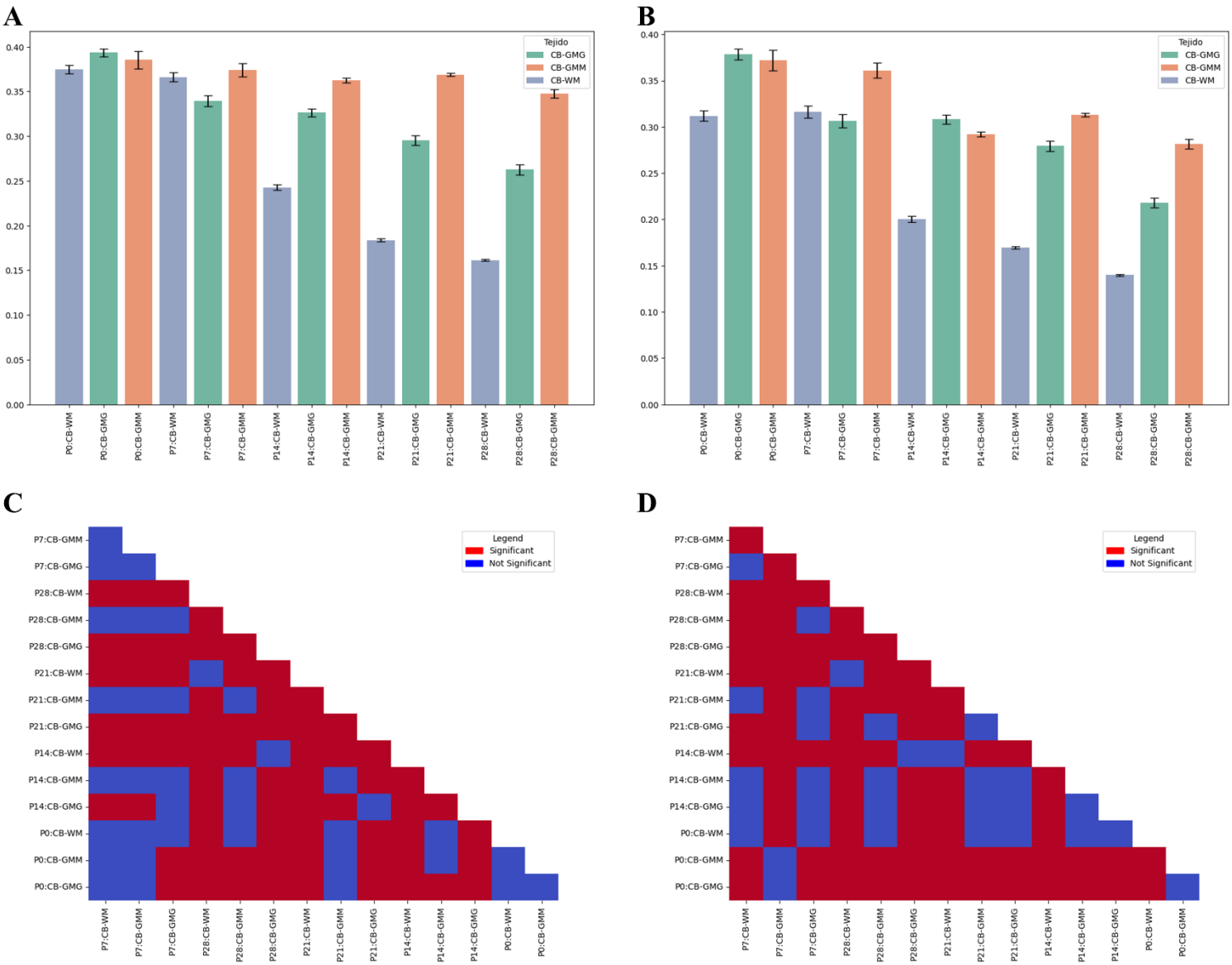

Figure 19: Bar plot representing the mean value and standard mean error for the C=O/CH_2 ratio for WT:CB measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of CB in the original (C) and new (D) datasets.

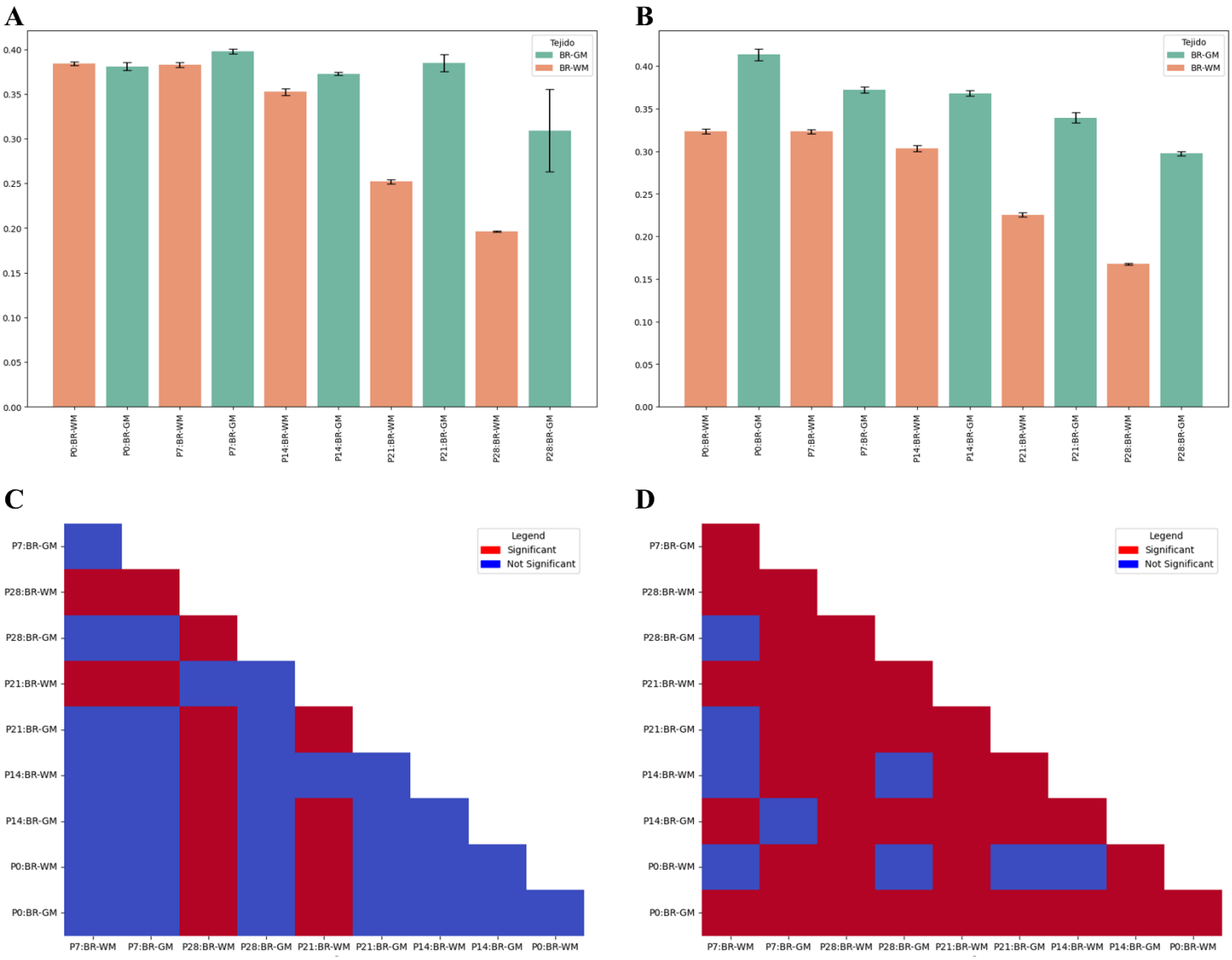

Figure 20: Bar plot representing the mean value and standard mean error for the C=O/CH_2 ratio for WT:BR measurements in the original (A) and the new (B) datasets. Correlogram representing the statistically significant interactions between different conditions of BR in the original (C) and new (D) datasets.

4.4. C=CH/CH_2 ratio

Now, by looking at the C=CH/CH_2 ratio (d^2 A_{3012} / d^2 A_{2852}), numerous differences appear between pre and post SVD analysis. For the cerebellum, the group P0:CB-GMM increases drastically with the SVD application. Also, P7:CB-WM slightly decreases with respect other CB-WM measurements. Finally, the CB-GMG group increases considerably with respect other regions. Regarding standard errors, most of the groups has a lower standard error mean with the new procedure, specially groups like P0:CB-GMM and all CB-GMG groups, which are groups that has a notable change post-SVD. However, the P7:CB-GMM had a little increase in standard error (Fig. 17A, 17B). All these changes led to a big difference regarding posthoc comparisons. 21 pairs changed to significant with SVD, while 14 did the other way around (Fig 17C, 17D) (Table 4). For the brain analysis, the biggest difference can be seen at the P0:BR-GM group. This group went from an almost neglectable value prior to SVD, which did not match the rest of the data, to having a similar value with the groups of the same region (Fig. 18A, 18B). This drastic change, into a more plausible result reinforces the case for the need of applying SVD in this type of analysis. Another change to notice is the slight increase of P0:BR-WM to match similar levels than P7:BR-WM. With this, BR-WM shows a stable C=CH/CH_2 ratio during the first week of postnatal development followed by a decrease in time, rather than increasing from P0 to P7 and then decreasing onwards. Moreover, there are various groups that has a lower standard error mean after SVD application. Regarding Tukey's HSD, 11 interactions acquired statistical significancy, most of them featuring P7:BR-GM. In contrast, 9 interactions became not-significant, mostly due to the P0:BR-GM change (Fig. 18C, 18D) (Table 5).

4.5. C=O/CH_2 ratio

Within the C=O/CH_2 ratio (d^2 A_{1743} / d^2 A_{2852}) analysis, there are also some notable changes. For the cerebellum's ANOVA test, the major change was noticed at the CB-GMM region. While in the original analysis this region seems unaffected by the postnatal development, after applying SVD a decrease in the C=O/CH_2 ratio was observed after P7. Also, a more pronounced decrease after P0 was observed in the CB-GMG region. Then, CB-WM remained the same, even though P0 suffered a significant decrease compared to the other 2 regions (Fig. 19A, 19B). In the Tukey's HSD test 15 interactions changed to

be significant, primarily interactions involving CB-GMM groups. Also, 10 interactions changed to not-significant in the new analysis (Fig. 19C, 19D) (Table 4). For the analysis of the brain regions, 2 things stand out at first sight. Even though both WM and GM maintain the tendencies within each region, after the SVD the WM measurements suffer a notable decrease with respect the GM measurements. Also, the group P28:BR-GM presents a great standard error in the original dataset, while in the new dataset the standard error is quite low (Fig. 20A, 20B). In the Tukey's HSD results there are also numerous changes. In total, 22 conditions become significant after SVD application, mostly the early stages of GM. Conversely, no condition changes to not-significant (Fig. 20C, 20D) (Table 5).

All these analyses were also performed for the $IL6^+$ and $IL10^+$ individuals, following the same procedures. The aim of this project is not to acquire biological knowledge about these mutations and how they play a role in postnatal development of the CNS in mice. Instead, this project aims to evaluate the integration of WSVD within a data processing pipeline for µFTIR. Therefore, the most interesting findings will be commented, as they are also useful to see the power of WSVD.

Within the $IL6^+$ group, P0:CB-GMM suffers a big change in the $C{=}CH/CH_2$ ratio, and a drastic decrease in the standard error mean in the CH_2/CH_3 ratio. Also, P0:CB-WM heavily decreases, adopting a value similar to groups of the same region. Then, within the $IL10^+$ group, the standard error mean for the P0:CB-WM measurements decreased a lot for the CH_2/CH_3 ratio. For the rest of the conditions, the behaviour shown was similar to the one seen at the WT group (Data not shown).

Table 4: Change in significancy from before to after SVD application of interaction between conditions for all ratios in the CB measurements.

	From not-significant to significant	From significant to not-significant
CH_2/Amide I	---	---
CH_2/CH_3	P7:CB-GMM/P7:CB-WM, P28:CB-GMM/P7:CB-GMM, P28:CB-GMG/P7:CB-GMM, P21:CB-GMM/P7:CB-GMM, P14:CB-GMM/P7:CB-GMM, P0:CB-WM/P7:CB-GMM	---
Amide I (β) / Amide I (α)	P28:CB-GMG/P7:CB-GMM, P21:CB-WM/P28:CB-WM, P21:CB-GMG/P28:CB-WM, P14:CB-WM/P7:CB-GMM, P14:CB-GMM/P28:CB-GMG, P0:CB-WM/P7:CB-GMG, P0:CB-GMM/P7:CB-WM, P0:CB-GMM/P28:CB-GMG, P0:CB-GMG/P0:CB-WM	P21:CB-GMM/P7:CB-GMG, P14:CB-WM/P21:CB-GMG, P0:CB-GMG/P14:CB-GMM, P0:CB-GMG/P14:CB-GMG
$C=CH/CH_2$	P7:CB-GMG/P7:CB-GMM, P28:CB-WM/P7:CB-GMG, P21:CB-WM/P7:CB-GMG, P21:CB-GMG/P7:CB-WM, P21:CB-GMG/P28:CB-GMG, P14:CB-WM/P7:CB-GMM, P14:CB-WM/P7:CB-GMG, P14:CB-WM/P21:CB-GMG, P14:CB-GMG/P7:CB-WM, P14:CB-GMG/P7:CB-GMM, P14:CB-GMG/P21:CB-WM, P14:CB-GMG/P14:CB-WM, P0:CB-WM/P7:CB-WM, P0:CB-WM/P28:CB-GMG, P0:CB-WM/P14:CB-WM', 'P0:CB-GMG/P7:CB-WM, P0:CB-GMG/P7:CB-GMG, P0:CB-GMG/P28:CB-GMG, P0:CB-GMG/P21:CB-GMG, P0:CB-GMG/P14:CB-WM, P0:CB-GMG/P0:CB-WM	P28:CB-GMM/P7:CB-GMG, P28:CB-GMG/P7:CB-GMG, P21:CB-GMG/P28:CB-GMM, P14:CB-WM/P28:CB-GMG, P14:CB-GMM/P21:CB-GMG, P14:CB-GMG/P28:CB-GMM, P14:CB-GMG/P14:CB-GMM, P0:CB-WM/P28:CB-GMM, P0:CB-GMM/P7:CB-GMM, P0:CB-GMM/P28:CB-WM', P0:CB-GMM/P21:CB-WM, P0:CB-GMG/P28:CB-GMM, P0:CB-GMG/P21:CB-GMM, P0:CB-GMG/P14:CB-GMM
$C=O/CH_2$	P7:CB-GMM/P7:CB-WM, P7:CB-GMG/P7:CB-GMM, P28:CB-GMM/P7:CB-WM, P28:CB-GMM/P7:CB-GMM, P21:CB-GMM/P7:CB-GMM, P21:CB-GMM/P28:CB-GMM, P14:CB-GMM/P7:CB-GMM, P0:CB-WM/P7:CB-GMM, P0:CB-GMM/P7:CB-WM, P0:CB-GMM/P21:CB-GMM, P0:CB-GMM/P14:CB-GMM, P0:CB-GMM/P0:CB-WM, P0:CB-GMG/P7:CB-WM, P0:CB-GMG/P21:CB-GMM, P0:CB-GMG/P0:CB-WM	P21:CB-GMG/P7:CB-GMG, P21:CB-GMG/P28:CB-GMM, P21:CB-GMG/P21:CB-GMM, P14:CB-WM/P21:CB-WM, P14:CB-GMM/P21:CB-GMG, P14:CB-GMG/P7:CB-WM, P14:CB-GMG/P21:CB-GMM, P14:CB-GMG/P14:CB-GMM, P0:CB-WM/P21:CB-GMG, P0:CB-WM/P14:CB-GMG

Table 5: Change in significancy from before to after SVD application of interaction between conditions for all ratios in the BR measurements.

	From not-significant to significant	From significant to not-significant
CH_2/Amide I	P14:BR-GM/P21:BR-GM	P0:BR-WM/P28:BR-GM
CH_2/CH_3	P28:BR-GM/P7:BR-GM, P21:BR-GM/P7:BR-GM, P14:BR-WM/P7:BR-WM, P14:BR-GM/P28:BR-GM, P14:BR-GM/P21:BR-GM, P0:BR-WM/P14:BR-WM	P0:BR-WM/P28:BR-GM
Amide I (β) / Amide I (α)	P7:BR-GM/P7:BR-WM, P14:BR-GM/P14:BR-WM, P0:BR-WM/P7:BR-GM, P0:BR-GM/P28:BR-GM, P0:BR-GM/P14:BR-WM	P21:BR-WM/P28:BR-GM, P21:BR-GM/P21:BR-WM, P14:BR-WM/P7:BR-GM
C=CH/CH_2	P7:BR-GM/P7:BR-WM, P28:BR-GM/P7:BR-GM, P21:BR-GM/P7:BR-GM, P14:BR-WM/P7:BR-WM, P14:BR-WM/P7:BR-GM, P14:BR-GM/P7:BR-WM, P14:BR-GM/P14:BR-WM, P0:BR-WM/P7:BR-GM, P0:BR-WM/P21:BR-WM, P0:BR-WM/P14:BR-WM, P0:BR-WM/P14:BR-GM	P14:BR-GM/P21:BR-GM, P0:BR-WM/P7:BR-WM, P0:BR-WM/P28:BR-GM, P0:BR-WM/P21:BR-GM, P0:BR-GM/P7:BR-WM, P0:BR-GM/P28:BR-GM, P0:BR-GM/P21:BR-GM, P0:BR-GM/P14:BR-GM, P0:BR-GM/P0:BR-WM
C=O/CH_2	P7:BR-GM/P7:BR-WM, P28:BR-GM/P7:BR-GM, P21:BR-WM/P28:BR-WM, P21:BR-WM/P28:BR-GM, P21:BR-GM/P7:BR-GM, P21:BR-GM/P28:BR-GM, P14:BR-WM/P7:BR-GM, P14:BR-WM/P21:BR-WM, P14:BR-WM/P21:BR-GM, P14:BR-GM/P7:BR-WM, P14:BR-GM/P28:BR-GM, P14:BR-GM/P21:BR-GM, P14:BR-GM/P14:BR-WM, P0:BR-WM/P7:BR-GM, P0:BR-WM/P14:BR-GM, P0:BR-GM/P7:BR-WM, P0:BR-GM/P7:BR-GM, P0:BR-GM/P28:BR-GM, P0:BR-GM/P21:BR-GM, P0:BR-GM/P14:BR-WM, P0:BR-GM/P14:BR-GM, P0:BR-GM/P0:BR-WM	---

5. Conclusions

In conclusion, a new computational pipeline for the processing, analysis and representation of µFTIR data has been successfully developed. This establishes an automatic and customizable workflow that scientist can adapt to their needs in terms of the origin and processing needs of the data and the objectives of the study.

Furthermore, Singular Value Decomposition has been integrated as a key step of this pipeline to act as a dimension reduction and denoising method, allowing to improve statistical analyses and increase the extraction of information from spectral µFTIR data.

The viability of this pipeline and the veracity of the results have been validated by conducting the analysis of already studied and published data. This data has also allowed to compare and validate the usefulness of integrating Singular Value Decomposition.

All these results have translated in the reduction of experimental time, by improving and automatizing the data analysis.

As future perspectives, this study proposes to reduce data acquisition time, as the noise reduction may allow to use less scans to accomplish a viable measurement. Moreover, this work encourages scientist to extract new information from noise-affected IR regions that, until now, were problematic and difficult to analyze.

6. Bibliography

Andjelic, S., Kreuzer, M., Hawlina, M., & Lumi, X. (2023). Characterization of Different Types of Epiretinal Proliferations by Synchrotron Radiation-Based Fourier Transform Infrared Micro-Spectroscopy. International Journal Of Molecular Sciences, 24(5), 4834. https://doi.org/10.3390/ijms24054834

Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., Hughes, C., Lasch, P., Martin-Hirsch, P. L., Obinaju, B., Sockalingum, G. D., Sulé-Suso, J., Strong, R. J., Walsh, M. J., Wood, B. R., Gardner, P., … Martin, F. L. (2014). Using Fourier transform IR spectroscopy to analyze biological materials. Nature protocols, 9(8), 1771–1791. https://doi.org/10.1038/nprot.2014.110

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical Methods, 6(9), 2812-2831. https://doi.org/10.1039/c3ay41907j

Buijs, H. L., Rochette, L., & Chateauneuf, F. (2004). Evolution of FTIR technology as applied to chemical detection and quantification. Proceedings Of SPIE, The International Society For Optical Engineering/Proceedings Of SPIE. https://doi.org/10.1117/12.516519

Bunaciu, A. A., Hoang, V. D., & Aboul-Enein, H. Y. (2017). Vibrational Micro-Spectroscopy of Human Tissues Analysis: Review. *Critical reviews in analytical chemistry*, *47*(3), 194–203. https://doi.org/10.1080/10408347.2016.1253454

Caine, S., Heraud, P., Tobin, M. J., McNaughton, D., & Bernard, C. C. (2012). The application of Fourier transform infrared microspectroscopy for the study of diseased central nervous system tissue. NeuroImage, 59(4), 3624-3640. https://doi.org/10.1016/j.neuroimage.2011.11.033

Chan, K. L. A., Altharawi, A. I., Fale, P., Song, C. L., Kazarian, S. G., Cinque, G., Untereiner, V., & Sockalingum, G. D. (2020). Transmission Fourier Transform

Infrared Spectroscopic Imaging, Mapping, and Synchrotron Scanning Microscopy with Zinc Sulfide Hemispheres on Living Mammalian Cells at Sub-Cellular Resolution. Applied Spectroscopy, 74(5), 544-552. https://doi.org/10.1177/0003702819898275

Chatchawal, P., Wongwattanakul, M., Tippayawat, P., Jearanaikoon, N., Jumniansong, A., Boonmars, T., Jearanaikoon, P., & Wood, B. R. (2020). Monitoring the Progression of Liver Fluke-Induced Cholangiocarcinoma in a Hamster Model Using Synchrotron FTIR Microspectroscopy and Focal Plane Array Infrared Imaging. Analytical Chemistry, 92(23), 15361-15369. https://doi.org/10.1021/acs.analchem.0c02656

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. *Aorta*, *4*(04), 124-130.

Dudała, J., Janeczko, K., Setkowicz, Z., Eichert, D., & Chwiej, J. (2012). The use of SR-FTIR microspectroscopy for a preliminary biochemical study of the rat hippocampal formation tissue in case of pilocarpine induced epilepsy and neutroprotection with FK-506. Nukleonika, 615-619. http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.baztech-article-BUJ8-0023-0058

Ellis, D. I., & Goodacre, R. (2006). Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. The Analyst, 131(8), 875–885. https://doi.org/10.1039/b602376m

Elmi, M. M., Elmi, F., & Feizi, F. (2022). Synchrotron FTIR microspectroscopy study of the diabetic rat skin wound healing with collagen+glycolipoprotein-90 treatment. Vibrational Spectroscopy, 118, 103335. https://doi.org/10.1016/j.vibspec.2022.103335

Findlay, C. R., Wiens, R., Rak, M., Sedlmair, J., Hirschmugl, C. J., Morrison, J., Mundy, C. J., Kansiz, M., & Gough, K. M. (2015). Rapid biodiagnostic ex vivo imaging at 1 μm pixel resolution with thermal source FTIR FPA. Analyst

(London. 1877. Online)/Analyst, 140(7), 2493-2503. https://doi.org/10.1039/c4an01982b

Fuwa, K., & Valle, B. L. (1963). The Physical Basis of Analytical Atomic Absorption Spectrometry. The Pertinence of the Beer-Lambert Law. Analytical Chemistry, 35(8), 942-946. https://doi.org/10.1021/ac60201a006

Gautam, Rekha & Vanga, Sandeep & Ariese, Freek & Umapathy, Siva. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. EPJ Techniques and Instrumentation. 2. 10.1140/epjti/s40485-015-0018-6.

Gazi, E., Dwyer, J., Lockyer, N., Miyan, J., Gardner, P., Hart, C., Brown, M., & Clarke, N. (2005). A study of cytokinetic and motile prostate cancer cells using synchrotron-based FTIR microspectroscopic imaging. Vibrational Spectroscopy, 38(1-2), 193-201. https://doi.org/10.1016/j.vibspec.2005.02.026

Grdadolnik, J. (2003). Saturation effects in FTIR spectroscopy: Intensity of amide I and amide II bands in protein spectra. Acta Chimica Slovenica, 50, 777-788.

Grzelak, M., Wróbel, P., Lankosz, M., Stęgowski, Z., Chmura, Ł., Adamek, D., Hesse, B., & Castillo-Michel, H. (2018). Diagnosis of ovarian tumour tissues by SR-FTIR spectroscopy: A pilot study. Spectrochimica Acta. Part A, Molecular And Biomolecular Spectroscopy, 203, 48-55. https://doi.org/10.1016/j.saa.2018.05.070

Golub, G., & Kahan, W. (1965). Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal Of The Society For Industrial And Applied Mathematics. Series B, Numerical Analysis, 2(2), 205-224. https://doi.org/10.1137/0702016

Guo, Y., Chen, T., Wang, S., Zhou, X., Zhang, H., Li, D., Mu, N., Tang, M., Hu, M., Tang, D., Yang, Z., Zhong, J., Tang, Y., Feng, H., Zhang, X., & Wang, H. (2020). Synchrotron Radiation-Based FTIR Microspectroscopic Imaging of

Traumatically Injured Mouse Brain Tissue Slices. ACS Omega, 5(46), 29698-29705. https://doi.org/10.1021/acsomega.0c03285

Hardesty, J. H., & Attili, B. (2010). Spectrophotometry and the Beer-Lambert Law: An important analytical technique in chemistry. Collin College, Department of Chemistry, 1-6.

Kasprzyk, P. (2023). Biomolecular and elemental micro-analysis of the skeletal muscle in the quest of new tissue markers of neuromuscular diseases (Doctoral dissertation, Jagiellonian University Medical College).

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. Korean Journal Of Anesthesiology, 70(1), 22. https://doi.org/10.4097/kjae.2017.70.1.22

Kim, H. (2015). Statistical notes for clinical researchers:post-hocmultiple comparisons. Restorative Dentistry & Endodontics, 40(2), 172. https://doi.org/10.5395/rde.2015.40.2.172

Koenig, J. L., Wang, S., & Bhargava, R. (2001). Peer Reviewed: FTIR images. Analytical Chemistry, 73(13), 360 A-369 A. https://doi.org/10.1021/ac012471p

Kramer, C. Y. (1956). Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. Biometrics, 12(3), 307. https://doi.org/10.2307/3001469

Kreuzer, M., Dučić, T., Hawlina, M., & Andjelic, S. (2020). Synchrotron-based FTIR microspectroscopy of protein aggregation and lipids peroxidation changes in human cataractous lens epithelial cells. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-72413-9

Lasch, P., & Naumann, D. (2006). Spatial resolution in infrared microspectroscopic imaging of tissues. Biochimica Et Biophysica Acta. Biomembranes, 1758(7), 814-829. https://doi.org/10.1016/j.bbamem.2006.06.008

Lasch, P. (2012). Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemometrics And Intelligent Laboratory Systems, 117, 100-114. https://doi.org/10.1016/j.chemolab.2012.03.011

Laurent, G., Woelffel, W., Barret-Vivin, V., Gouillart, E., & Bonhomme, C. (2019). Denoising applied to spectroscopies – part I: concept and limits. Applied Spectroscopy Reviews, 54(7), 602-630. https://doi.org/10.1080/05704928.2018.1523183

Levin, I. W., & Bhargava, R. (2005). FOURIER TRANSFORM INFRARED VIBRATIONAL SPECTROSCOPIC IMAGING: Integrating Microscopy and Molecular Recognition. Annual Review Of Physical Chemistry, 56(1), 429-474. https://doi.org/10.1146/annurev.physchem.56.092503.141205

Lilo, T. (2022) Predicting Meningioma Recurrence Using Spectrochemical Analysis of Tissues and Subsequent Predictive Computational Algorithms. Doctoral thesis, University of Central Lancashire. http://doi.org/10.17030/uclan.thesis.00047977

Lipiec, E., Bambery, K. R., Lekki, J., Tobin, M. J., Vogel, C., Whelan, D. R., Wood, B. R., & Kwiatek, W. M. (2015). SR-FTIR Coupled with Principal Component Analysis Shows Evidence for the Cellular Bystander Effect. Radiation Research, 184(1), 73-82. https://doi.org/10.1667/rr13798.1

Magalhães, S., Goodfellow, B. J., & Nunes, A. (2021). FTIR spectroscopy in biomedical research: how to get the most out of its potential. Applied Spectroscopy Reviews, 56(8–10), 869–907. https://doi.org/10.1080/05704928.2021.1946822

Marinkovic, N. S., & Chance, M. R. (2006). Synchrotron infrared microspectroscopy. Encyclopedia Of Molecular Cell Biology And Molecular Medicine. https://doi.org/10.1002/3527600906.mcb.200500021

Martin, F. L., Kelly, J. G., Llabjani, V., Martin-Hirsch, P. L., Patel, I. I., Trevisan, J., Fullwood, N. J., & Walsh, M. J. (2010). Distinguishing cell types or populations based on the computational analysis of their infrared spectra. Nature protocols, 5(11), 1748–1760. https://doi.org/10.1038/nprot.2010.133

Matthäus, C., Bird, B., Miljković, M., Chernenko, T., Romeo, M., & Diem, M. (2008). Chapter 10 Infrared and Raman Microscopy in Cell Biology. En Methods in cell biology (pp. 275-308). https://doi.org/10.1016/s0091-679x(08)00610-9

Mazarakis, N., Vongsvivut, J., Bambery, K. R., Ververis, K., Tobin, M. J., Royce, S. G., Samuel, C. S., Snibson, K. J., Licciardi, P. V., & Karagiannis, T. C. (2020). Investigation of molecular mechanisms of experimental compounds in murine models of chronic allergic airways disease using synchrotron Fourier-transform infrared microspectroscopy. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-68671-2

Morais, C. L. M., Lima, K. M. G., Singh, M., & Martin, F. L. (2020). Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nature protocols*, *15*(7), 2143–2162. https://doi.org/10.1038/s41596-020-0322-8

Morais, C. L. M., Paraskevaidi, M., Cui, L., Fullwood, N. J., Isabelle, M., Lima, K. M. G., Martin-Hirsch, P. L., Sreedhar, H., Trevisan, J., Walsh, M. J., Zhang, D., Zhu, Y. G., & Martin, F. L. (2019). Standardization of complex biologically derived spectrochemical datasets. *Nature protocols*, *14*(5), 1546–1577. https://doi.org/10.1038/s41596-019-0150-x

Movasaghi, Z., Rehman, S., & ur Rehman, D. I. (2008). Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, *43*(2), 134-179. https://doi.org/10.1080/05704920701829043

Nasse, M. J., Walsh, M. J., Mattson, E. C., Reininger, R., Kajdacsy-Balla, A., Macias, V., Bhargava, R., & Hirschmugl, C. J. (2011). High-resolution Fourier-

transform infrared chemical imaging with multiple synchrotron beams. Nature Methods, 8(5), 413-416. https://doi.org/10.1038/nmeth.1585

Nuez-Martínez, M., Pedrosa, L., Martinez-Rovira, I., Yousef, I., Diao, D., Teixidor, F., Stanzani, E., Martínez-Soler, F., Tortosa, A., Sierra, À., Gonzalez, J. J., & Viñas, C. (2021). Synchrotron-Based Fourier-Transform Infrared Micro-Spectroscopy (SR-FTIRM) Fingerprint of the Small Anionic Molecule Cobaltabis(dicarbollide) Uptake in Glioma Stem Cells. International Journal Of Molecular Sciences, 22(18), 9937. https://doi.org/10.3390/ijms22189937

Olkowski, A. A., Wojnarowicz, C., & Laarveld, B. (2020). Pathophysiology and pathological remodelling associated with dilated cardiomyopathy in broiler chickens predisposed to heart pump failure. Avian Pathology, 49(5), 428-439. https://doi.org/10.1080/03079457.2020.1757620

Paraskevaidi, M., Morais, C. L. M., Lima, K. M. G., Snowden, J. S., Saxon, J. A., Richardson, A. M. T., Jones, M., Mann, D. M. A., Allsop, D., Martin-Hirsch, P. L., & Martin, F. L. (2017). Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. Proceedings of the National Academy of Sciences of the United States of America, 114(38), E7929–E7938. https://doi.org/10.1073/pnas.1701517114

Peris, M., Benseny-Cases, N., Manich, G., Zerpa, O., Almolda, B., Perálvarez-Marín, À., González, B., & Castellano, B. (2023). Roadmap for Postnatal Brain Maturation: Changes in Gray and White Matter Composition during Development Measured by Fourier Transformed Infrared Microspectroscopy. *ACS chemical neuroscience*, *14*(17), 3088–3102. https://doi.org/10.1021/acschemneuro.3c00237

Pijanka, J. K., Kohler, A., Yang, Y., Dumas, P., Chio-Srichan, S., Manfait, M., Sockalingum, G. D., & Sulé-Suso, J. (2009). Spectroscopic signatures of single, isolated cancer cell nuclei using synchrotron infrared microscopy.

Analyst (London. 1877. Online)/Analyst, 134(6), 1176. https://doi.org/10.1039/b821112d

Portier, H., Jaffré, C., Kewish, C., Chappard, C., & Pallu, S. (2020). New insights in osteocyte imaging by synchrotron radiation. Journal Of Spectral Imaging. https://doi.org/10.1255/jsi.2020.a3

Pushie, M. J., , Kelly, M. E., , & Hackett, M. J., (2018). Direct label-free imaging of brain tissue using synchrotron light: a review of new spectroscopic tools for the modern neuroscientist. The Analyst, 143(16), 3761–3774. https://doi.org/10.1039/c7an01904a

Qian, J., Gao, X., Wang, Y., Li, X., Hu, J., & Lü, J. (2022). Synchrotron Infrared Microspectroscopy for Stem Cell Research. International Journal Of Molecular Sciences, 23(17), 9878. https://doi.org/10.3390/ijms23179878

Rousseeuw, P.J., & Hubert, M. (2017). Anomaly detection by robust statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8.

Sanchez-Molina, P., Kreuzer, M., Benseny-Cases, N., Valente, T., Almolda, B., González, B., Castellano, B., & Perálvarez-Marín, A. (2020). From Mouse to Human: Comparative Analysis between Grey and White Matter by Synchrotron-Fourier Transformed Infrared Microspectroscopy. *Biomolecules*, *10*(8), 1099. https://doi.org/10.3390/biom10081099

Sandt, C., & Borondics, F. (2021). A new typology of human hair medullas based on lipid composition analysis by synchrotron FTIR microspectroscopy. Analyst (London. 1877. Online)/Analyst, 146(12), 3942-3954. https://doi.org/10.1039/d1an00695a

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36(8), 1627–1639. doi:10.1021/ac60214a047

Schanze, T. (2017). Removing noise in biomedical signal recordings by singular value decomposition. Current Directions In Biomedical Engineering, 3(2), 253-256. https://doi.org/10.1515/cdbme-2017-0052

Scharf, L. L. (1991). The SVD and reduced rank signal processing. Signal Processing, 25(2), 113-133. https://doi.org/10.1016/0165-1684(91)90058-q

Seredin, P., Goloshchapov, D., Ippolitov, Y., & Vongsvivut, J. (2021). Engineering of a Biomimetic Interface between a Native Dental Tissue and Restorative Composite and Its Study Using Synchrotron FTIR Microscopic Mapping. International Journal Of Molecular Sciences, 22(12), 6510. https://doi.org/10.3390/ijms22126510

Seredin, P., Goloshchapov, D., Ippolitov, Y., & Vongsvivut, J. (2020). Development of a new approach to diagnosis of the early fluorosis forms by means of FTIR and Raman microspectroscopy. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-78078-8

Srebro, N., & Jaakkola, T. (2003). Weighted low-rank approximations. International Conference On Machine Learning, 720-727. http://people.csail.mit.edu/tommi/papers/SreJaa-icml03.pdf

Ståhle, L., & Wold, S. (1989). Analysis of variance (ANOVA). Chemometrics And Intelligent Laboratory Systems, 6(4), 259-272. https://doi.org/10.1016/0169-7439(89)80095-4

Tobin, M., Vongsvivut, J., Martin, D., Sizeland, K., Hackett, M., Takechi, R., Fimorgnari, N., Lam, V., Mamo, J., Carter, E., Swarbrick, B., Lay, P., Christensen, D., Perez-Guaita, D., Lowery, E., Heraud, P., Wood, B., Puskar, L., & Bambery, K. (2018). Focal plane array IR imaging at the Australian Synchrotron. Infrared Physics & Technology, 94, 85-90. https://doi.org/10.1016/j.infrared.2018.06.022

Tobin, M. J., Puskar, L., Barber, R. L., Harvey, E. C., Heraud, P., Wood, B. R., Bambery, K. R., Dillon, C. T., & Munro, K. L. (2010). FTIR spectroscopy of single live cells in aqueous media by synchrotron IR microscopy using microfabricated sample holders. Vibrational Spectroscopy, 53(1), 34-38. https://doi.org/10.1016/j.vibspec.2010.02.005

Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D., & Martin, F. L. (2012). Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *The Analyst*, *137*(14), 3202–3215. https://doi.org/10.1039/c2an16300d

White, J. L., & Roth, C. B. (1986). Infrared spectrometry. En Soil Science Society of America book series (pp. 291-330). https://doi.org/10.2136/sssabookser5.1.2ed.c11

Wilcox, R. R. (2021). Two-way ANOVA: Inferences about interactions based on robust measures of effect size. British Journal Of Mathematical & Statistical Psychology/British Journal Of Mathematical And Statistical Psychology, 75(1), 46-58. https://doi.org/10.1111/bmsp.12244

Yousef, I., Ribó, L., Crisol, A., Šics, I., Ellis, G., Ducic, T., Kreuzer, M., Benseny-Cases, N., Quispe, M., Dumas, P., Lefrançois, S., Moreno, T., García, G., Ferrer, S., Nicolas, J., & Aranda, M. (2017). MIRAS: The Infrared Synchrotron Radiation Beamline at ALBA. Synchrotron Radiation News, 30(4), 4-6. https://doi.org/10.1080/08940886.2017.1338410