**Information required from inventors to enable a patent application**

1. **Title of Invention**

Raman Data analysis

1. **Inventor Names**

Paulo Gomes

1. **Background to the invention**
   1. Does the invention address a problem and/or provide a benefit? If so, what is the problem or benefit?

The problem this software attempts to solve is related to how Raman data analysis is performed. Here we centralise multiple tools that permit quick data analysis with simple clicks and with complicated operations made easy. The centralised Raman analysis packages were inspired by how multiple authors analysed their Raman data. A compilation of the most used techniques was implemented into this software. Not only does it plot the data, but there is also a multivariable analysis that takes care of the Principal Component Analysis (PCA) part (used in almost all Raman analysis papers) and a Neural Net option Self-Organising Map (SOM) (used in some Raman analysis papers).

Moreover, an automated multipeak Lorentzian fit was implemented, gathering all the fitted peaks from the plots, PCA loadings and SOM activations. The peak fitting results are stored in a .csv file and compared with other .csv files with the same format, facilitating a peak comparison between different samples by a single button press. Furthermore, each user can have their peak library to use and share it with other scientists. These libraries can also be compared with user-built .csv files from literature peaks. The matching feature also outputs an intensity ratio comparison for matched peaks making the script a quantitative and quantitative analysis tool for Raman spectra.

Overall, the code created for this effect solves most Raman data analysis used in standard Raman papers in the literature:

* Integrated code helps with efficiency, one code all solutions.
* Code operates multiple Raman plot options and multivariable data analysis with PCA and SOM functions.
* Multi Lorentzian peak fitting option that generates users own peak libraries to use and share with other researchers.
  1. What attempts have been made in the past to address this problem or provide the same or similar benefits, by you or by others?

No other software integrated plot options with multivariable analysis (PCA and SOM) with peak finding features.

There are, however, individual tools that do make plots and some data analysis like Originlab (<https://www.originlab.com> ), QtiPlot (<https://www.qtiplot.com/>) and Quasar (<https://quasar.codes/download/>), among others. For multivariable analysis, there are PCA (python (<https://www.python.org/> ) with the ‘sklearn’ package and R-studio(<https://www.rstudio.com/> ) with ‘pcomp’ package) and SOM there is python (<https://github.com/JustGlowing/minisom>) and Java (<https://github.com/carlbanbury/raman-tools>) solutions. For peak finding, we have Originlab (<https://www.originlab.com> ), Wire (<https://www.renishaw.com/en/raman-software--9450>). Wire is by far the most completed program listed here. There are peak finding/fitting tools and libraries that match these peaks with existing samples in Wire. Wire also has PCA.

Compared to the software we propose, ours can potentially have an unlimited peak library, many more plot options, SOM analysis and in detail PCA analysis. Which far surpasses what we can do in other applications. Since what we propose goes more in-depth into the specifics of Raman analysis.

* 1. Are there any publications (including patents, papers, conference proceedings, books, journals etc) describing attempts to address the problem or describing ways to achieve the same or similar benefit?

No previous work has, to our knowledge, provided an easy integrated Raman data analyser that can make paper ready images and data analysis with a throughout peak analysis procedure and matching. Also, it allows users to create their own Raman peak databank, which can be used and shared among the community.

* 1. How does your research work differ from these previous attempts to address the problem or provide the benefit?

This work can theoretically create any customised peak library for any compound without relying on the limited library choices installed with other software packages. Also, integrated multivariable data analysis for your data is always a benefit. The script has been optimised for the least number of inputs to run smoothly.

1. **Summarise the invention**
   1. What do you see as the key points of the invention?

Integrated Raman data analysis script with the following key features: single Raman plot, stacked Raman plot, heatmap Raman plot, 3D stacked Raman plot, 2D PCA, 3D PCA, Raman dilutions plot, Raman spectra peak finder/matching, Raman spectra PCA loadings peak finder/matching, Raman spectra SOM neuron activation peak finder/matching.

* 1. Are there any features that you think are key to the performance of the invention?

The multipeak code is design to perform local Lorentzian fits to individual peaks. These peaks have been previously identified by locating zero values for the gradient of the Raman spectra. These are called major peaks. Furthermore, all major Lorentzian peaks are collected and summed, and the difference between the initial Raman spectra is calculated. This residue is subsequentially Lorentzian fitted again. When the residue is positive and a peak, the fit is performed and considered a minor positive peak. If the residue is negative and there is a crest, the peak is fitted and considered a minor negative peak. All these peaks are written onto a .csv file. However, the analysis is mostly done over the major peaks.

The same operation also helps to explain the peak fitting for the PCA loadings. The Raman analysis literature uses the PCA to show the cluster separation but usually skip to show the loading values. In this script, we performed the multiLorentzian peak analysis for the Loadings ‘spectra’, which helps explain the separation, making it easier to interpret the PCA clustering. All information related to the Loading ‘spectra’ fit is stored in a .csv file and can be peak matched with other .csv files.

The SOM script also outputs all activation features that help to interpret the SOM clustering. Also, a peak finding is used to compare activation spots between different samples, making it an ideal tool for Raman analysis.

One of the plot features, the stack plot, has a unique way to stack plots. Here we calculate the minimum difference between adjacent line plots and use that as the separation plus the standard deviation of the latest, making it always appealing to the eye.

Another important feature is the ability to generate a peak file from any kind of compound. Usually, authors use bioanalytes that are too complex to have a specific library installed on software. Thus, we offer a solution for the users to generate peak files from complex analytes and then use them to match other samples.

* 1. Are there any features that you think can enhance the performance of the invention?

Peak finding itself can be further perfected and optimised to guarantee full overlap between real and fitted peaks. Moreover, the peak finding option for SOM activations and PCA loadings and then peak matching from tabulated values will further be optimised from the prior.

* 1. Which features of the invention have each of the named inventors contributed to?

I am the main author of this work.

I did solve most of my questions using google (<https://www.google.com/>) and StackOverflow (<https://stackoverflow.com/>).

Some other predefined modules were used and further adapted by me. All the models used were miniSOM (<https://github.com/JustGlowing/minisom>) for the SOM mathematics, Circlify (<https://pypi.org/project/circlify/>) for the SOM spheres arrangement.

The peak finder was inspired by the work of

<https://github.com/emilyripka/BlogRepo/blob/master/181119_PeakFitting.ipynb> and then further modified by me.

Finally, the asymmetric least square smoothing used for the baseline was used from

<https://gist.github.com/perimosocordiae/efabc30c4b2c9afd8a83>.

Also, to have an improved label system on plots, the AdjustText package

<https://github.com/Phlya/adjustText>

1. **Describe the invention**
   1. Provide an overview of the invention and how it works.

The general idea this code presents to solve is when the user has a known analyte X and an unknown but related to X, analyte Y+X. Typical examples in the literature of interest extract saliva from healthy (Raman spectra: X) and unhealthy patients (Raman spectra: Y+X). The goal is to find how different X+Y is from X, thus obtaining how the Raman spectra of Y would look. Y would be represented by the absence or presence of multiple different Raman peaks, and Ys Raman fingerprint would be specific for cancerous patients.

Another example in the bioanalysis world could be extracted cancerous (X+Y) and non-cancerous (X) cells of patients’ colons. The same procedure would also apply here.

However, we are not limited to binary cases (cancerous[X+Y] vs healthy[X]), but in general, Y can be multiple stages of the illness or concentration or any state that is changing. For instance, some authors analyse the difference between cancerous (X+Y) against benign tumours (X+Y’) gains healthy people (X).

This multitude of options on dealing with and analysing the Raman data is the overall goal of the script. Let us describe now all features used to achieve these goals from the script.

The script is supposed to be executed with minimum input from the user. The minimum input required is of the raw Raman data in .txt format. Then individual one parameter option can solve most of the Raman data analysis. The algorithm has been perfected to have minimal input from the user. Here we mostly show the visual results obtained from the script with a short description of what to press and what the result represents.

Furthermore, all images can be saved as .svg using pythons interface and subsequentially processed using Inkscape or any other image processing software.

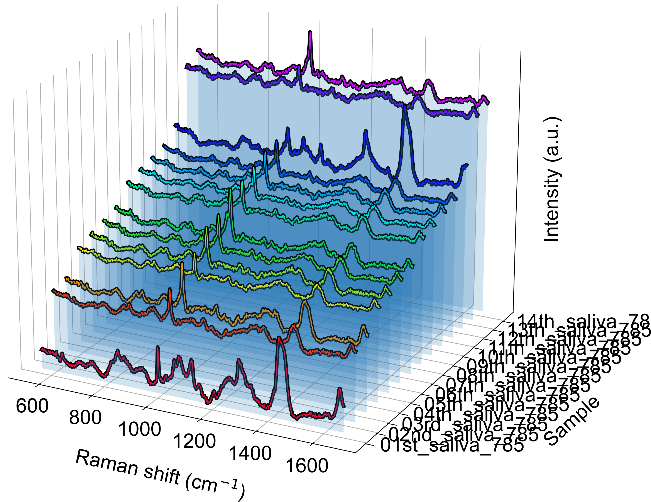
First, we show the graphical user interface (GUI).

Graphical user interface, application

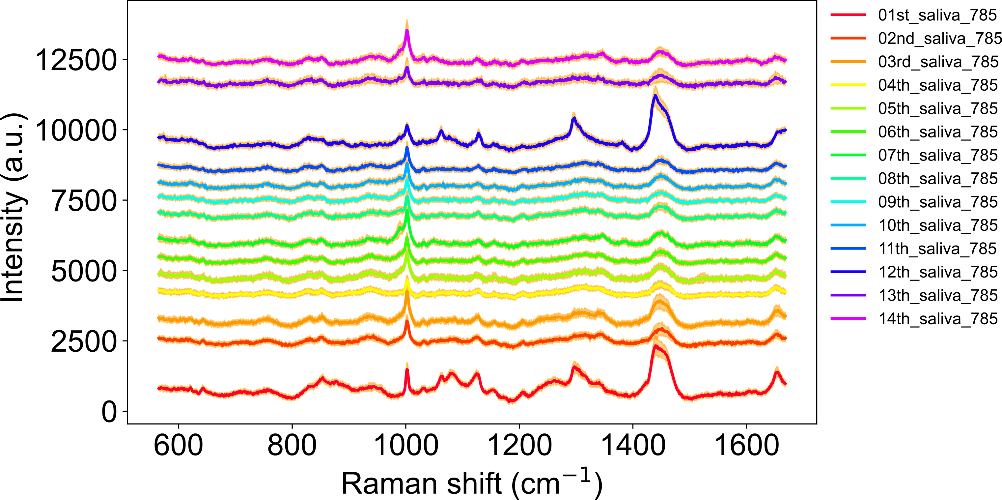
Description automatically generated

It has multiple options that can be accessed. The first step is to press the ‘Browse’ button and select a .txt file with your raw Raman data. You can select any additional processing to your data (Baseline, standard deviation, normalisation) by ticking the boxes and press then one of the four buttons. Plot, Stackplot, Heatplot and Waterfallplot. The commands are very straightforward. Just for the Stackplot, a more efficient way of presenting the plots has been devised. Instead of the constant increment, a more adaptive addition was implemented to maximise the spacing between adjacent plots. This command is used for the Waterfall plot and the Stackplot.

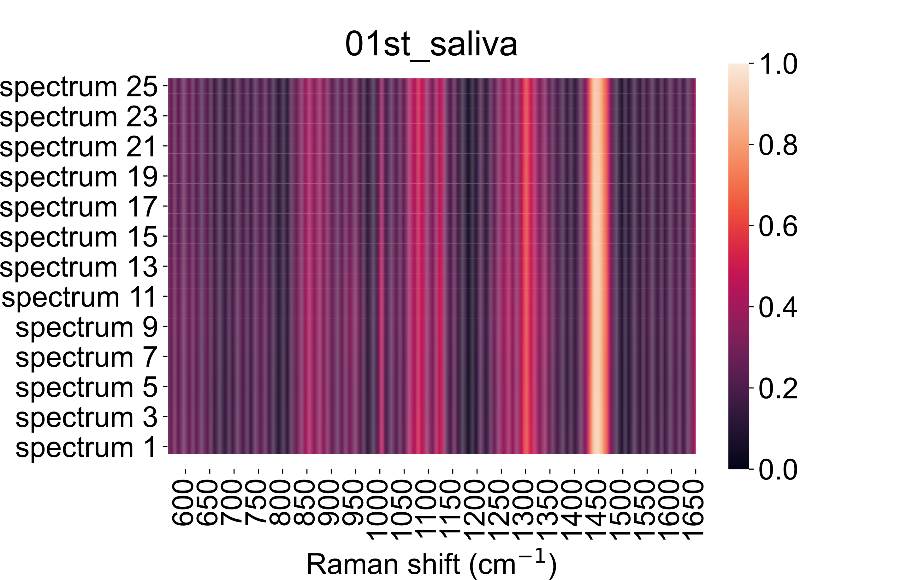
3D waterfall plot



Adaptive Stackplots



Individual Heatmaps



A dedicated Principal Component Analysis (PCA) was designed to plot the multiple outcomes from a PCA analysis. The data is still loaded from the Browser option, and any box can be ticked, and both the PCA and the 3D PCA buttons can be pressed.

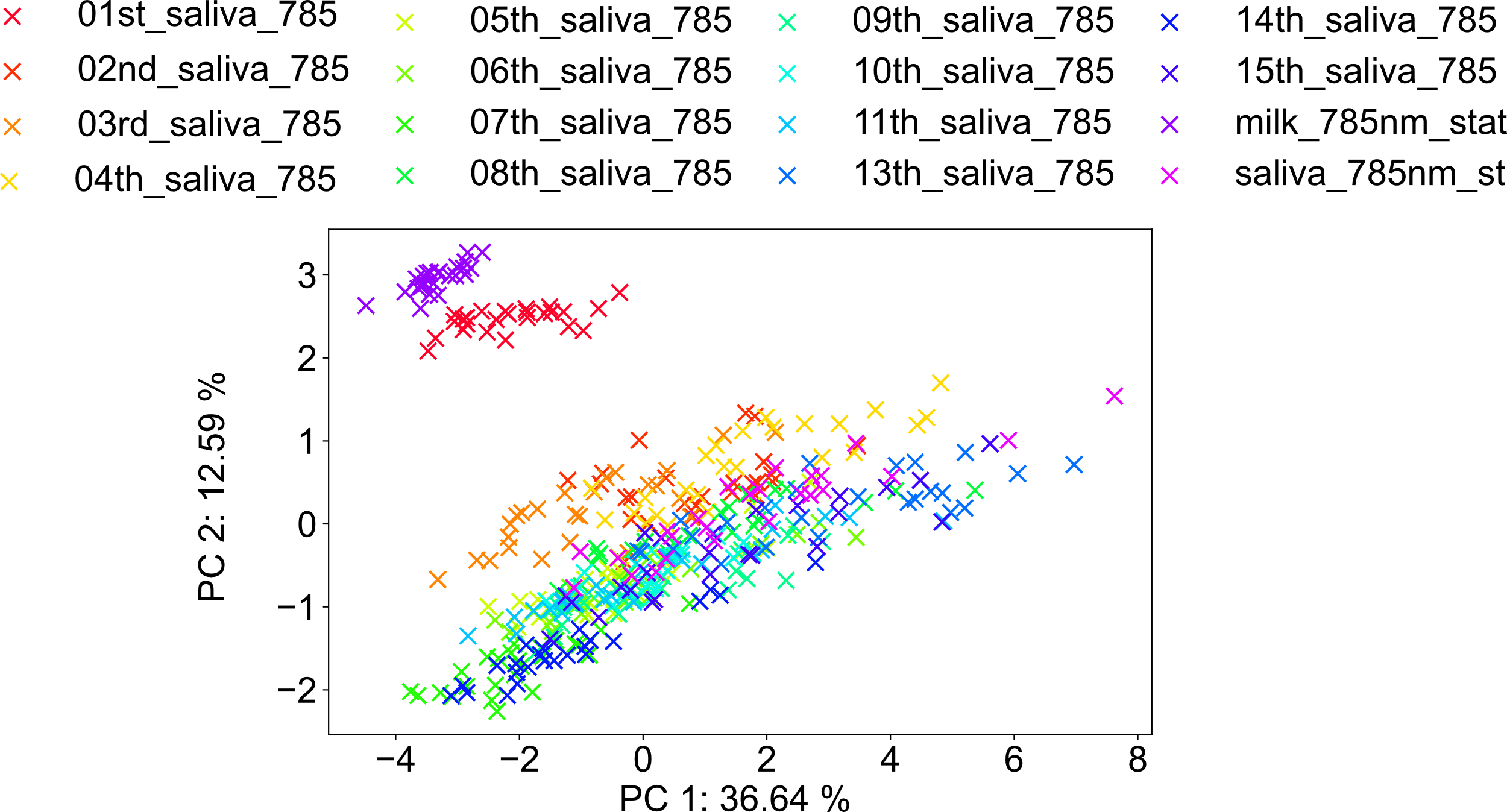
The principal component analysis is a multivariable analysis algorithm that defines the highest variance coordinate system to display the data. From these operations, the eigenvalues and eigenvectors can be operated into defining the Loadings of the PCA. Loadings are similar representations of the original data but seen from the ‘eyes’ of the principal components. Thus, the loadings have much information related to how the separation between different classes occurs.

A couple of commands are used to reduce the dispersion of some data points, for instance, a statistical data analysis like the z-score, changing the plot window and Gaussian mixtures (another statistical analysis tool).

Graphical user interface

Description automatically generated

2D PC plot

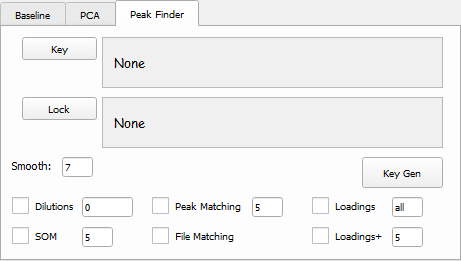


3D PC plots with Gaussian Mixtures

Chart

Description automatically generated

In the Peak finder tab, a plethora of options is displayed.



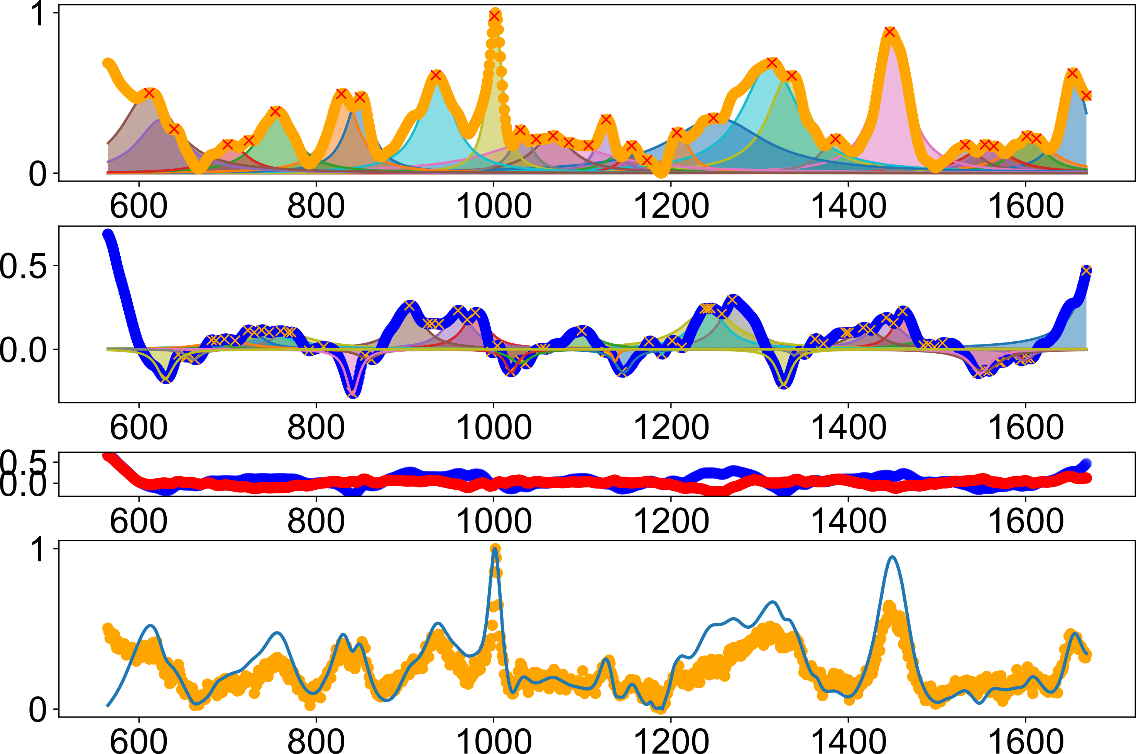
The peak finding method is generated by a key creation and lock solving method. The key option can be used alone, but the lock option always needs a key as input.

If the key option is used on its own, it generates the multipeak fit operation into the data loaded (press key to load the Raman .txt files and then Key Gen button to run the multipeak fit operation). Multiple operations can be ticked like Dilutions, SOM, Peak Matching, File Matching, Loadings and Loadings+.

The Dilution, the SOM, the Loadings and Loadings+ option work only for a loaded key.

The Peak matching, File matching, SOM + File matching, Loadings + File Matching options work if the key and lock are loaded.

If a key is loaded and the Key Gen button is pressed, a Multi Lorentzian fitting/Peak is generated to those files.



A .csv file with all the peaks fitted is generated from the peak finder

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **label** | **height** | **centre** | **width** | **importance** | **err** | **Intensity (a.u.)** |
| saliva\_all. | 0.6 | 1653.5 | 14.5 | major | 2.5 | 564.0 |
| saliva\_all. | 0.2 | 1604.1 | 18.1 | major | 0.9 | 195.0 |
| saliva\_all. | 0.1 | 1556.6 | 19.6 | major | 1.0 | 140.4 |
| saliva\_all. | 0.1 | 1531.1 | 21.2 | major | 0.6 | 131.0 |
| saliva\_all. | 0.9 | 1449.8 | 20.9 | major | 2.6 | 857.0 |
| saliva\_all. | 0.2 | 1386.0 | 15.6 | major | 0.4 | 182.5 |
| saliva\_all. | 0.5 | 1335.1 | 23.9 | major | 4.4 | 514.4 |
| saliva\_all. | 0.6 | 1315.1 | 29.6 | major | 1.3 | 562.6 |
| saliva\_all. | 0.2 | 1209.1 | 11.9 | major | 4.7 | 203.7 |
| saliva\_all. | 0.2 | 1156.4 | 10.2 | major | 0.8 | 162.6 |
| saliva\_all. | 0.3 | 1126.3 | 11.4 | major | 2.9 | 290.8 |
| saliva\_all. | 0.2 | 1080.4 | 35.6 | major | 1.0 | 166.1 |
| saliva\_all. | 0.3 | 1048.9 | 14.8 | major | 2.0 | 286.4 |
| saliva\_all. | 0.2 | 1047.7 | 24.9 | major | 27.8 | 229.8 |
| saliva\_all. | 1.0 | 1001.6 | 9.1 | major | 17.1 | 1003.7 |
| saliva\_all. | 0.5 | 936.9 | 21.4 | major | 0.9 | 470.5 |
| saliva\_all. | 0.3 | 849.7 | 15.4 | major | 5.4 | 341.6 |
| saliva\_all. | 0.3 | 830.7 | 18.8 | major | 2.3 | 320.0 |
| saliva\_all. | 0.3 | 754.2 | 21.5 | major | 0.8 | 292.0 |
| saliva\_all. | 0.2 | 708.6 | 29.6 | major | 1.2 | 154.5 |
| saliva\_all. | 0.1 | 682.2 | 39.7 | major | 0.7 | 83.7 |
| saliva\_all. | 0.2 | 630.5 | 26.7 | major | 6.8 | 221.6 |
| saliva\_all. | 0.3 | 613.9 | 28.1 | major | 2.6 | 325.4 |

We have the option to immediately check where the values between the different samples (key vs lock /X vs X+Y) are. We load the X+Y spectra in the Lock section and the .csv file of the X spectra in the Key section, and we tick the file matching box and press run.

Standard peak matching

Chart, histogram

Description automatically generated

You will have highlighted the peaks that matched your key file and the lock file. Additionally, the key file can have the X and X+Y .csv files to match the X+Y Raman spectra. The peaks highlighted by both peaks can be discarded, and when solely the X+Y peak is highlighted is a good indicator that that is the presence of Y.

The comparison also fits the matching peaks onto the lock sample and find the difference between them (residue), generating a deconvoluted peak matching

A screenshot of a video game

Description automatically generated with medium confidence

And the residue (deconvoluted peak matching – spectra)

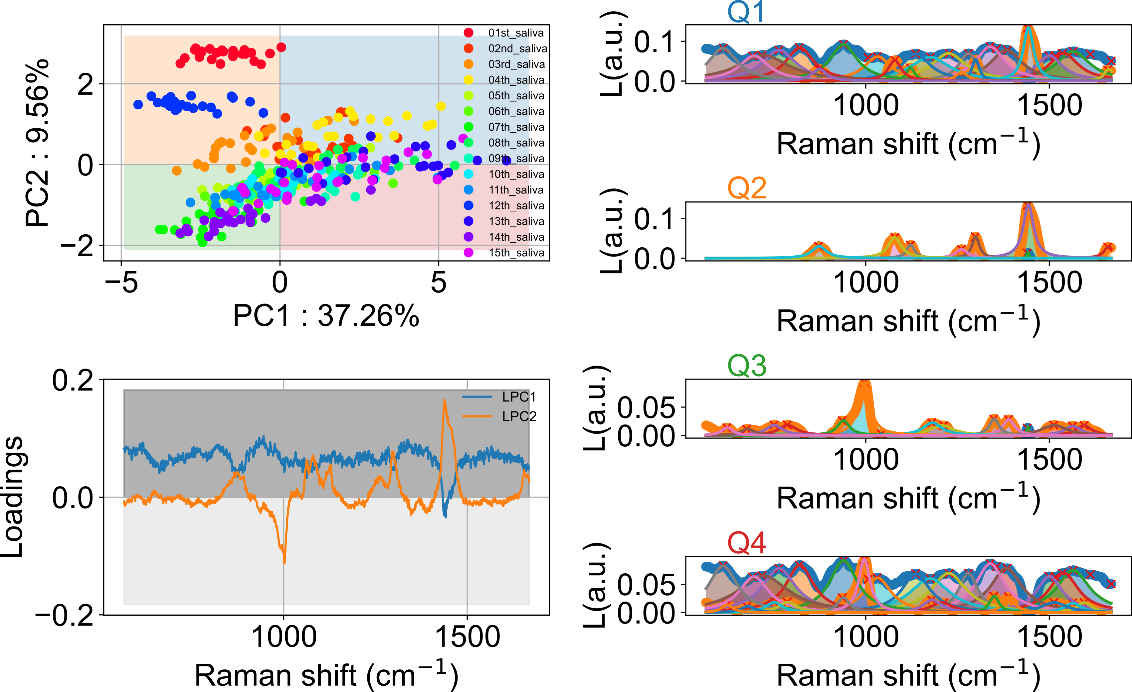
Chart

Description automatically generated

Whichever is above or below zero are important peaks that regard the difference between samples (i.e. the Y).

Since the goal is to obtain the most information related to one sample compared to another, the peak finding options will also work for two different multivariable analysis methods Principal Component analysis (PCA) and Self Organizing Maps (SOM). Information obtained from these multivariable analysis methods is already compartmentalised into optimising the differences between different classes. I.e. knowing X and X+Y, the analysis will highlight the features coming from Y. After loading the X and X+Y .txt files into the key, the user can tick the Loadings box followed by pressing the Key Gen button to generate The PCA + Loadings analysis.

2D Principal component plot with Loadings and Loadings Lorentzian fit per quadrant



From PCA Loadings, involves quadrant positions (Q#)

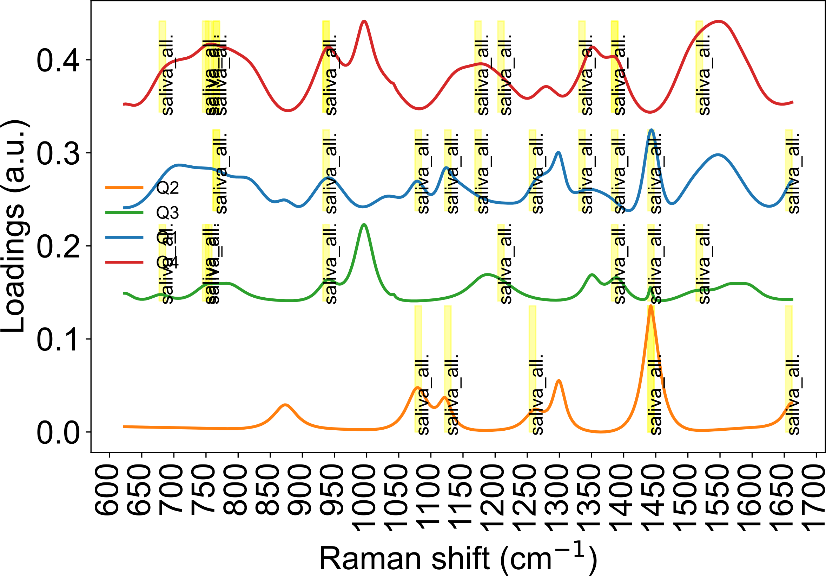
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **label** | **height** | **centre** | **width** | **importance** |
| 01st\_saliva | 0.07 | 1565.02 | 55.01 | Q1 |
| 01st\_saliva | 0.07 | 1539.65 | 48.07 | Q1 |
| 01st\_saliva | 0.06 | 1500.18 | 43.84 | Q1 |
| 01st\_saliva | 0.08 | 1382.95 | 52.29 | Q1 |
| 01st\_saliva | 0.02 | 1441.81 | -3.38 | Q2 |
| 01st\_saliva | 0.03 | 1661.80 | 15.66 | Q2 |
| 01st\_saliva | 0.13 | 1443.17 | 17.09 | Q2 |
| 01st\_saliva | 0.05 | 1299.37 | 12.06 | Q2 |
| 01st\_saliva | 0.02 | 1261.28 | 19.29 | Q2 |
| 01st\_saliva | 0.02 | 1441.81 | -3.38 | Q3 |
| 01st\_saliva | 0.02 | 1594.86 | 24.88 | Q3 |
| 01st\_saliva | 0.02 | 1563.82 | 26.93 | Q3 |
| 01st\_saliva | 0.02 | 1513.45 | 40.57 | Q3 |
| 01st\_saliva | 0.07 | 1565.02 | 55.01 | Q4 |
| 01st\_saliva | 0.07 | 1539.65 | 48.07 | Q4 |
| 01st\_saliva | 0.06 | 1500.18 | 43.84 | Q4 |
| 01st\_saliva | 0.08 | 1382.95 | 52.29 | Q4 |

The peaks present in the .csv file generated from the PCA/Loading analysis can be used to compare against other generated files to identify common peaks and in which Principal Component (PC) most peaks are localised.

To generate, we need to add a lock file. Lock files are the unknown files that we want to find which peaks are present. A practical case would be to generate the PCA/Loadings from the X and X+Y samples. These would outcome a .csv file of the Loadings. Then we would load this .csv file into the lock, and we would load into the key the generated X and X+Y .csv files. To perform the peak matching, we need to tick both Peak matching options with the loading option. The output plot will highlight the location of the X and X+Y peaks into the different PCA/Loadings regions. To identify regions where only Y is present, visually ignore overlapped regions from X and X+Y. Single X+Y highlight would indicate the heavy presence of Y in that region.

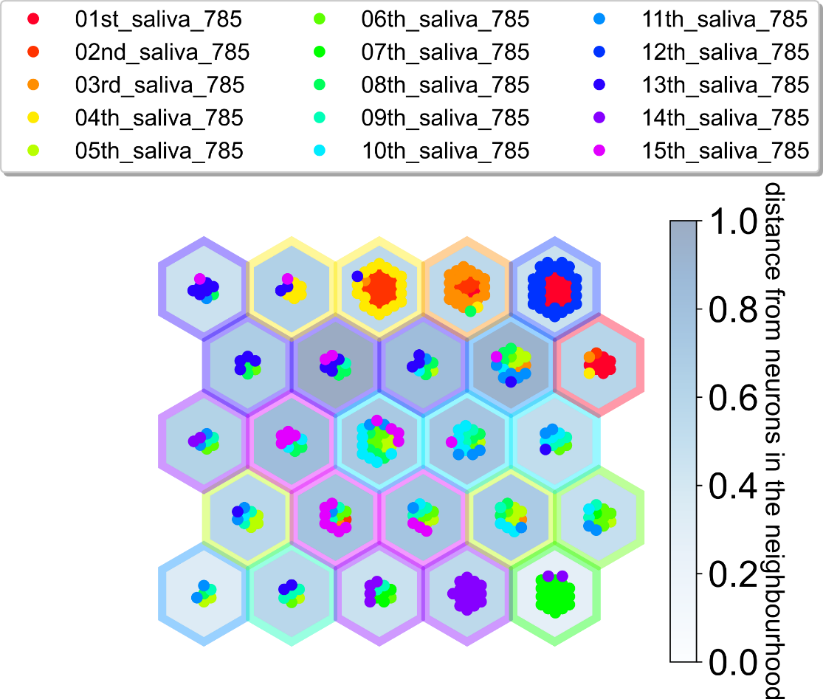
Both the PCA/Loadings plot and the peak fitting of the generated Loadings need to be analysed together to achieve better conclusions into your data.

2D PCA loadings per quadrant peak finder

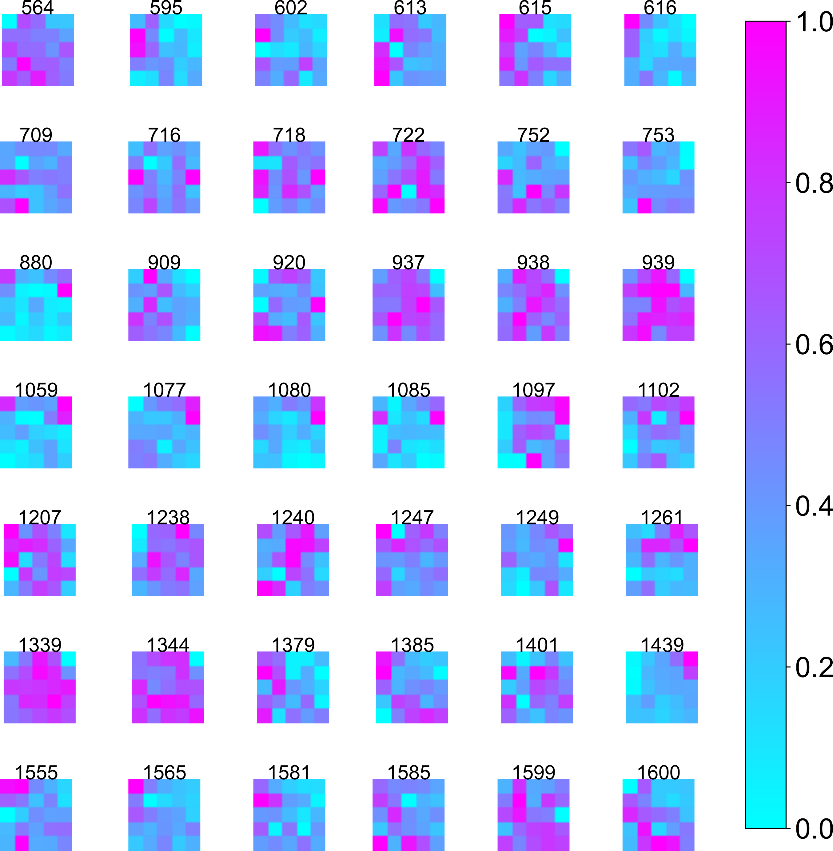


The other multivariable analysis uses artificial neural networks, and its the Self Organizing Maps (SOM). The procedure to use these commands is similar to the PCA, load your data onto the key option, tick the SOM box and press the Key Gen button. The script will organise the Key data into individual dots (each dot representing a spectrum in your Raman data). The computation formulation of the SOM architecture has been implemented in <https://github.com/JustGlowing/minisom>. I have perfected it for the Raman data interpretation. Here SOM only has one parameter, the number of neurons, and depending on that number, sets up a stage of size neuron times neuron to initialise the weights and biases values per neuron. Multiple hexagons represent a neuron, and the location of the dots identifies a determined spectrum-like curve. The hexagonal edge highlights the highest population level present. Also, boundary separations are shown with greyscale on each neuron. Boundaries reflect the high difference in spectra between neighbour neurons.

Self Organizing Maps (SOM)

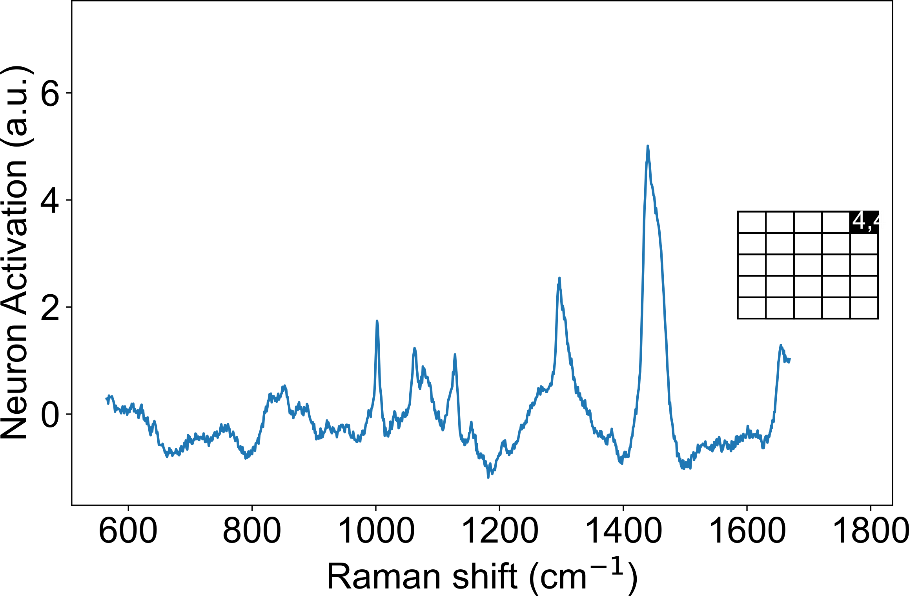


We can inspect each neuron and how its spectral representation is by getting the neuron activation. Here we track multiple peak values in a heatmap formation, with each square representing a similar position on the SOM.



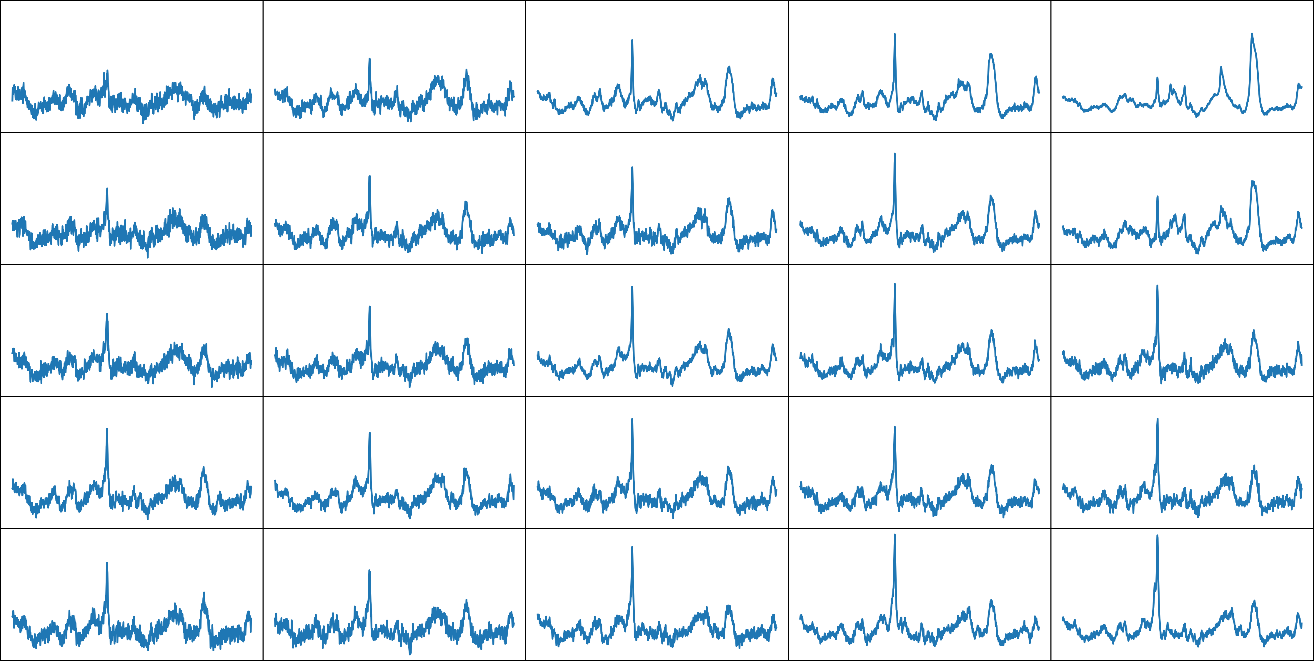
We can also get the full spectra per neuron.

SOM neuron activation localised.



And then the full spectra like for all neurons.

SOM neuron activation all. The spectra position is related to the SOM neuron position.



The same procedure for the PCA Loadings is applied here. Thus all the different peaks in each different neuron spot (x,y) are quantified and written over a .csv file.

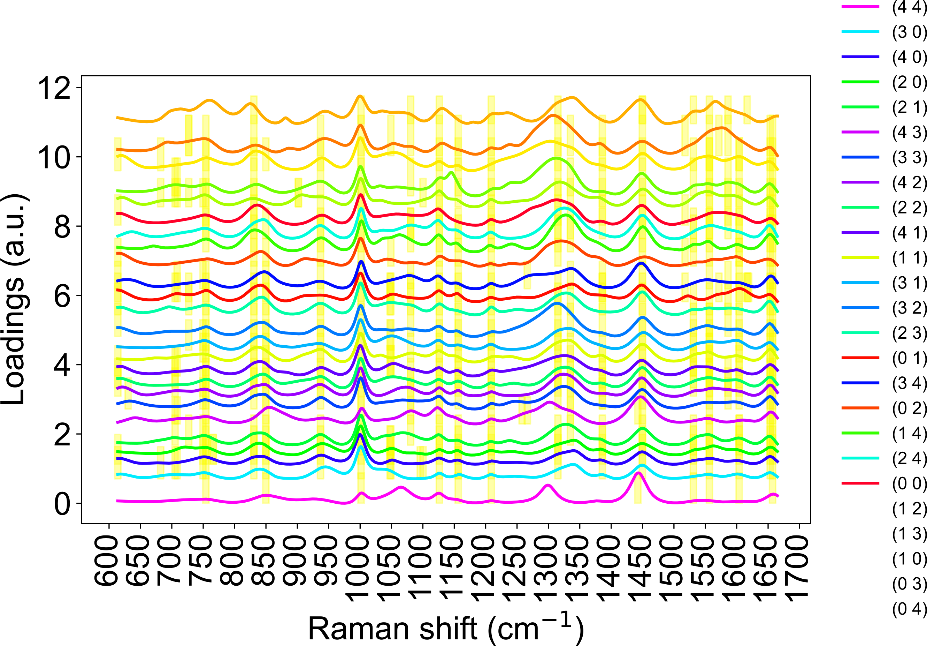
From SOM Activations, involves neuron position (n,n)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **label** | **height** | **center** | **width** | **importance** | **err** | **Intensity (a.u.)** |
| activation | 0.22 | 1609.55 | 22.41 | (0 0) | 0.02 | 1.44 |
| activation | 0.17 | 1592.01 | 33.61 | (0 1) | 0.01 | 1.14 |
| activation | 0.39 | 1657.29 | 25.04 | (0 2) | 0.01 | 2.13 |
| activation | 0.35 | 1582.15 | 18.31 | (0 3) | 0.01 | 1.80 |
| activation | 0.35 | 1613.67 | 22.01 | (0 4) | 0.01 | 1.38 |
| activation | 0.32 | 1640.58 | 79.66 | (1 0) | 1.38 | 2.01 |
| activation | 0.25 | 1601.84 | 22.29 | (1 1) | 0.00 | 1.57 |
| activation | 0.27 | 1550.35 | 44.53 | (1 2) | 0.05 | 1.62 |
| activation | 0.30 | 1603.98 | 18.99 | (1 3) | 0.01 | 1.74 |
| activation | 0.08 | 1496.90 | 13.49 | (1 4) | 0.01 | 0.35 |
| activation | 0.20 | 1600.23 | 20.32 | (2 0) | 0.01 | 1.50 |
| activation | 0.18 | 1601.93 | 20.55 | (2 1) | 0.01 | 1.35 |
| activation | 0.53 | 1653.61 | 15.48 | (2 2) | 0.02 | 3.89 |
| activation | 0.19 | 1583.40 | 25.86 | (2 3) | 0.01 | 1.24 |
| activation | 0.72 | 1652.97 | 15.35 | (2 4) | 0.02 | 4.28 |
| activation | 0.09 | 1532.79 | 34.16 | (3 0) | 0.00 | 0.77 |
| activation | 0.48 | 1653.57 | 16.91 | (3 1) | 0.02 | 3.22 |
| activation | 0.16 | 1556.49 | 37.77 | (3 2) | 0.02 | 1.18 |
| activation | 0.52 | 1653.54 | 15.21 | (3 3) | 0.02 | 3.82 |
| activation | 0.18 | 1546.69 | 59.10 | (3 4) | 0.06 | 1.13 |
| activation | 0.35 | 1653.66 | 14.76 | (4 0) | 0.01 | 2.91 |
| activation | 0.16 | 1602.32 | 19.52 | (4 0) | 0.01 | 1.34 |
| activation | 0.16 | 1561.52 | 53.88 | (4 1) | 0.01 | 1.15 |
| activation | 0.17 | 1556.63 | 18.77 | (4 1) | 0.01 | 1.21 |
| activation | 0.51 | 1653.74 | 14.05 | (4 2) | 0.01 | 3.68 |
| activation | 0.17 | 1606.51 | 17.14 | (4 2) | 0.01 | 1.27 |
| activation | 0.15 | 1585.55 | 16.27 | (4 2) | 0.01 | 1.06 |
| activation | 0.14 | 1537.73 | 41.89 | (4 3) | 0.00 | 0.73 |
| activation | 0.40 | 1659.18 | 18.29 | (4 4) | 0.02 | 2.45 |
| activation | 0.12 | 1604.62 | 30.99 | (4 4) | 0.00 | 0.74 |

Note that in both Loadings and SOM, the ‘height’ values displayed on the .csv file do not have units. Its height value is related to the Loading (PCA) and activation (SOM) space, which do not translate well to the intensity values present in standard Raman data.

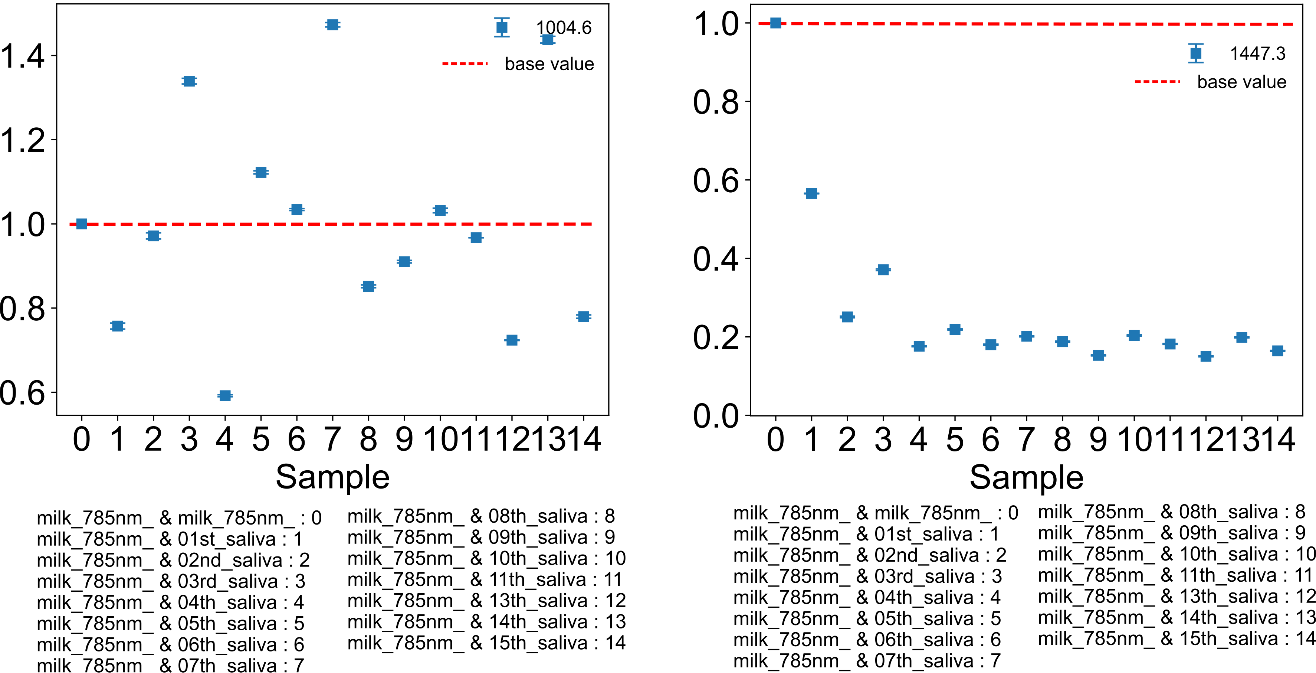
The same peak identification can be made for the SOM; however, the number of spectra can increase rapidly and become difficult to distinguish. It is already enough to take multiple conclusions over your data with the previous plots. A tip for the case that the image gets too crowded is to save the image as .svg and then a posterior process them and have a minute look over the different labelling.

SOM activation peak matching



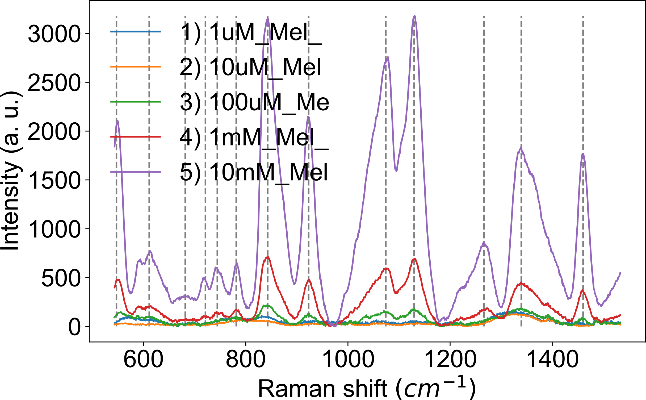
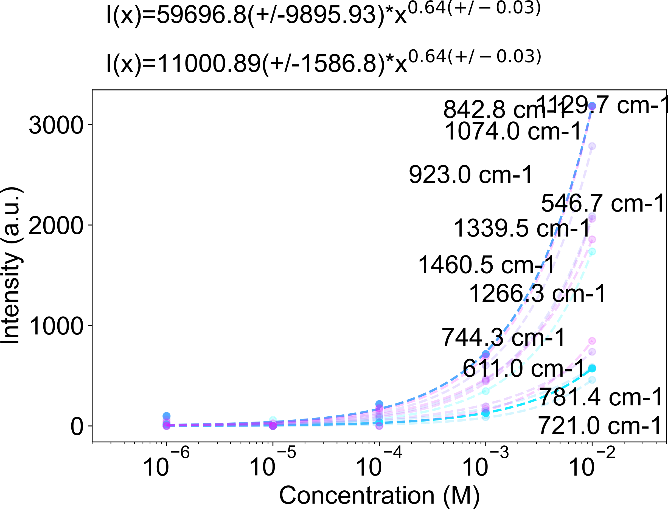
After getting the multiple .csv peak files for different samples, a comparison can be made regarding the ratio of common peaks. This process works better for standard Raman data files. Choose the file you want to compare in the key and lock section. Putting the file on both sides will guarantee that whichever peak the key matches with the lock, there will always be a match with value 1. Then onto the lock, add other files that you want to compare. The matching between the key with the lock will create a list of all the different ratios for the same peak to get important information regarding band increase or decrease. This option can be selected by ticking the file matching option.

Comparative peak matching

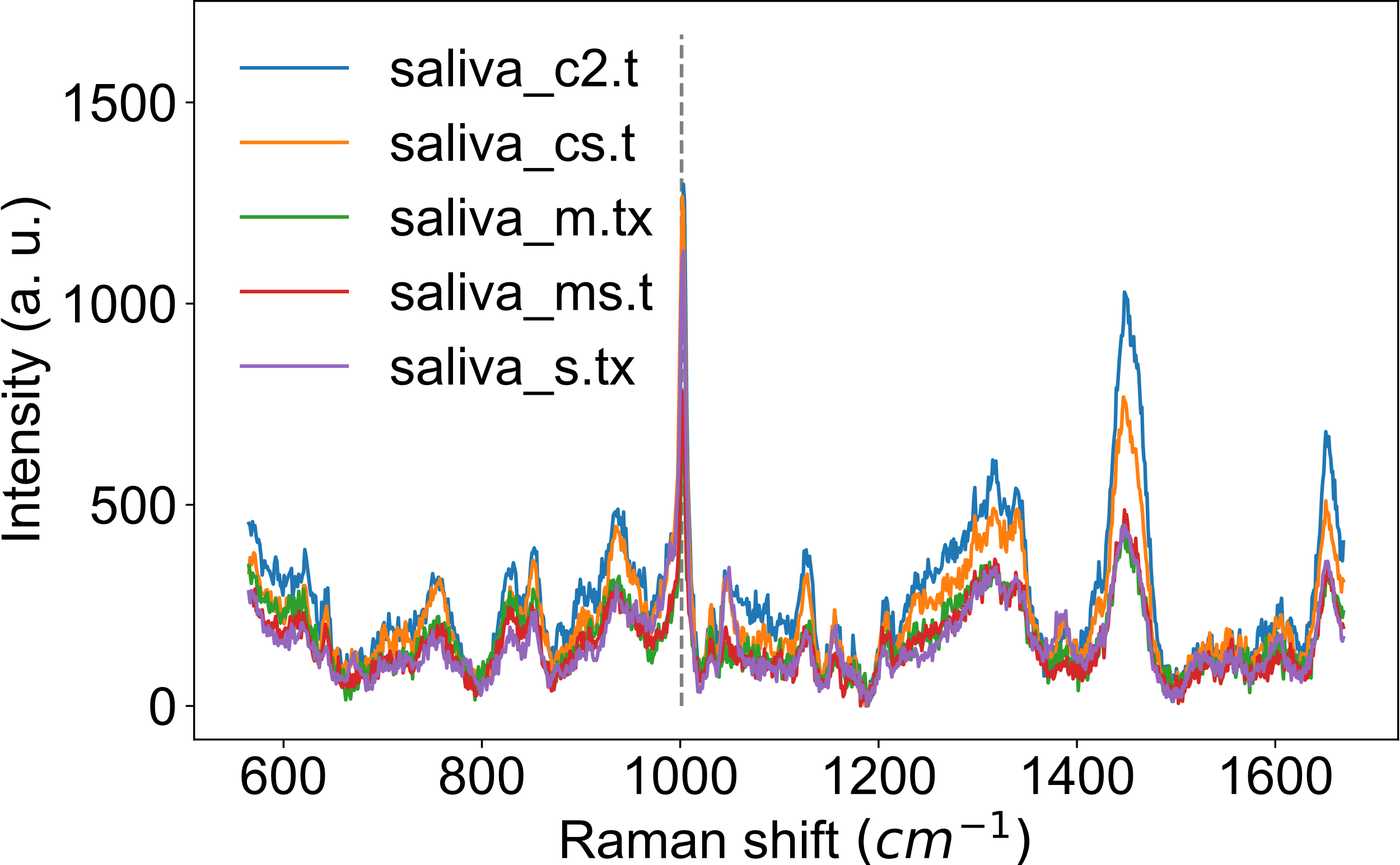
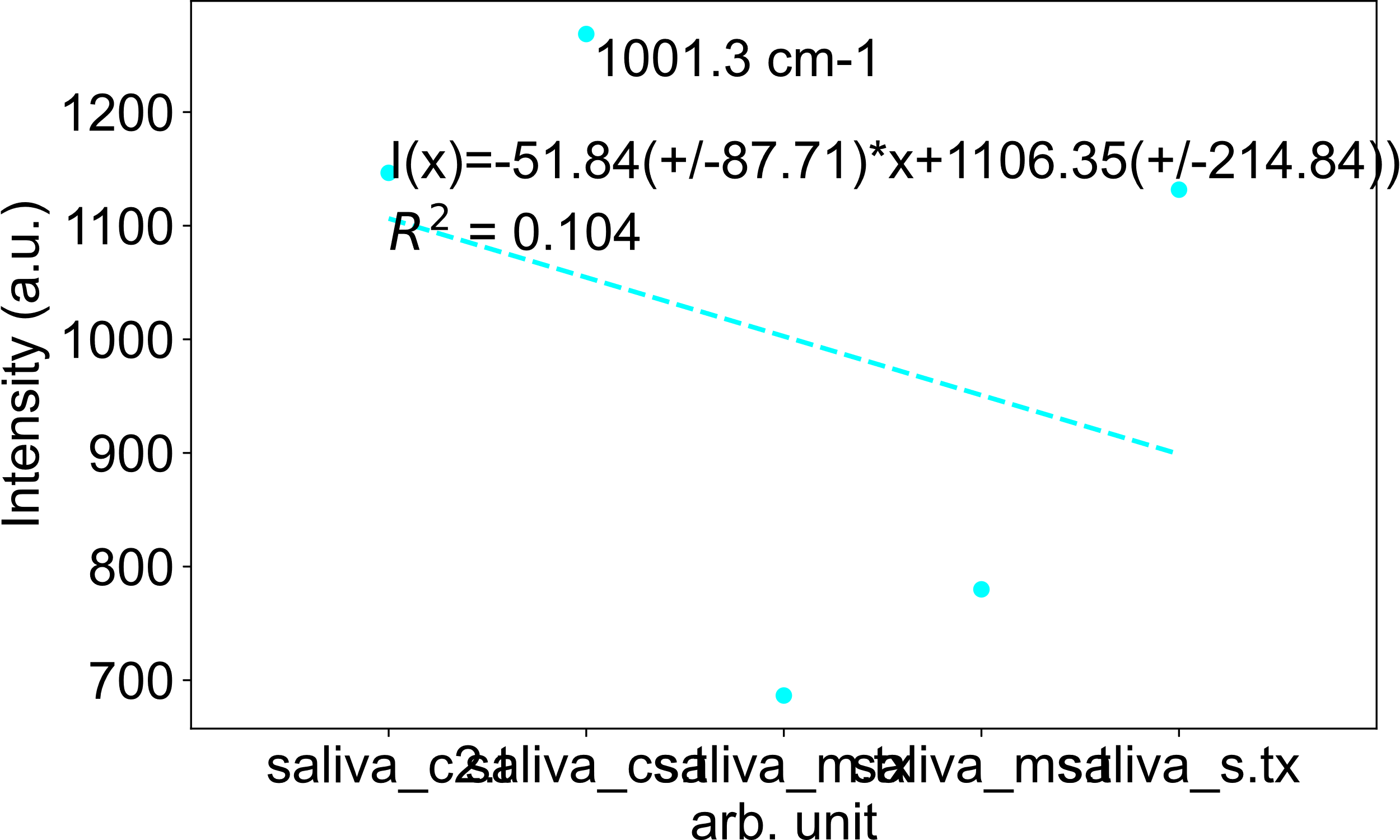


The script also has a dilution curve for matching peaks for different samples. The script can be selected by ticking the Dilutions part. The image shows how different peaks differ per sample and the intensity/concentration fit equation related to the highest and lower curve.

Dilution curves/ Dilution plots with peak assignment



There is also the option of tracking a single peak and comparing it among different samples. To run this, write the region where that peak appears and press the Key gen button.



Finally, this operation also generates a .csv file to inspect the selected values for the fit in more detail.

Dilution peak values table

|  |  |  |  |
| --- | --- | --- | --- |
| **Center (cm-1)** | **Concentration [M]** | **Intensity (a.u.)** | **label** |
| 721 | 1.00E-06 | 0 | 1) 1uM\_Mel\_ |
| 721 | 1.00E-05 | 0 | 2) 10uM\_Mel |
| 721 | 1.00E-04 | 0 | 3) 100uM\_Me |
| 721 | 0.001 | 88 | 4) 1mM\_Mel\_ |
| 721 | 0.01 | 458 | 5) 10mM\_Mel |
| 1074 | 1.00E-06 | 0 | 1) 1uM\_Mel\_ |
| 1074 | 1.00E-05 | 0 | 2) 10uM\_Mel |
| 1074 | 1.00E-04 | 147 | 3) 100uM\_Me |
| 1074 | 0.001 | 603 | 4) 1mM\_Mel\_ |
| 1074 | 0.01 | 2785 | 5) 10mM\_Mel |
| 1340 | 1.00E-06 | 0 | 1) 1uM\_Mel\_ |
| 1340 | 1.00E-05 | 0 | 2) 10uM\_Mel |
| 1340 | 1.00E-04 | 181 | 3) 100uM\_Me |
| 1340 | 0.001 | 443 | 4) 1mM\_Mel\_ |
| 1340 | 0.01 | 1857 | 5) 10mM\_Mel |
| 1461 | 1.00E-06 | 0 | 1) 1uM\_Mel\_ |
| 1461 | 1.00E-05 | 59 | 2) 10uM\_Mel |
| 1461 | 1.00E-04 | 71 | 3) 100uM\_Me |
| 1461 | 0.001 | 346 | 4) 1mM\_Mel\_ |
| 1461 | 0.01 | 1736 | 5) 10mM\_Mel |

* 1. Provide a detailed description of one or more examples, including, where possible:
     + drawings, schematics or flow charts of equipment or processes
     + compositions
     + process parameters
     + operating parameters
     + results or test data

As an illustration of the capabilities of the script, we have saliva samples (X), milk samples(Y) and saliva samples spiked with milk (X+Y).

We start with the visual representation of the spectra from the data, with the baseline subtraction and standard deviation highlighted. It is important to avoid normalisation when making comparisons since normalisations will reduce concentration differences proportional to the intensity values. From an initial inspection, we can already see that milk has more very active peaks than saliva.

Figure 5.2.1A shows the Raman spectra from saliva, milk, and saliva spiked with milk. Baseline subtraction was performed on the data and highlighted with the standard deviation. Following a PCA plot (5.2.1Bi) of the data, a significant separation exists between milk and saliva spiked with milk (populates Quadrant (Q)2 and Q3) with saliva (populates Q1 and Q4). Although some points of the saliva spiked with milk populate the saliva quadrant, representing that some spots were similar to saliva, but most were milk-related. The PCA loadings in 5.2.1Bii and 5.2.1Biii reflect this separation. Q1 is referenced to the principal components (PC)1 and PC2 positive. Thus, we select the LPC1 positive and LPC2 positive from the loading plot and perform a Lorentzian fit to these bands. This process generates the 5.2.1BiiiQ1 plot. Repeating the idea for the other quadrants and respecting the negative and positive signs, we end up with 5.2.1Biii. Connecting all 5.2.1B images, we can observe that in quadrants 1 and 4 its mostly saliva peaks, with the highest peak being the 1000cm-1, and in quadrants 2 and 3 are milk peaks, the 1445cm-1 peak that most contributes to this separation from its high loading value.

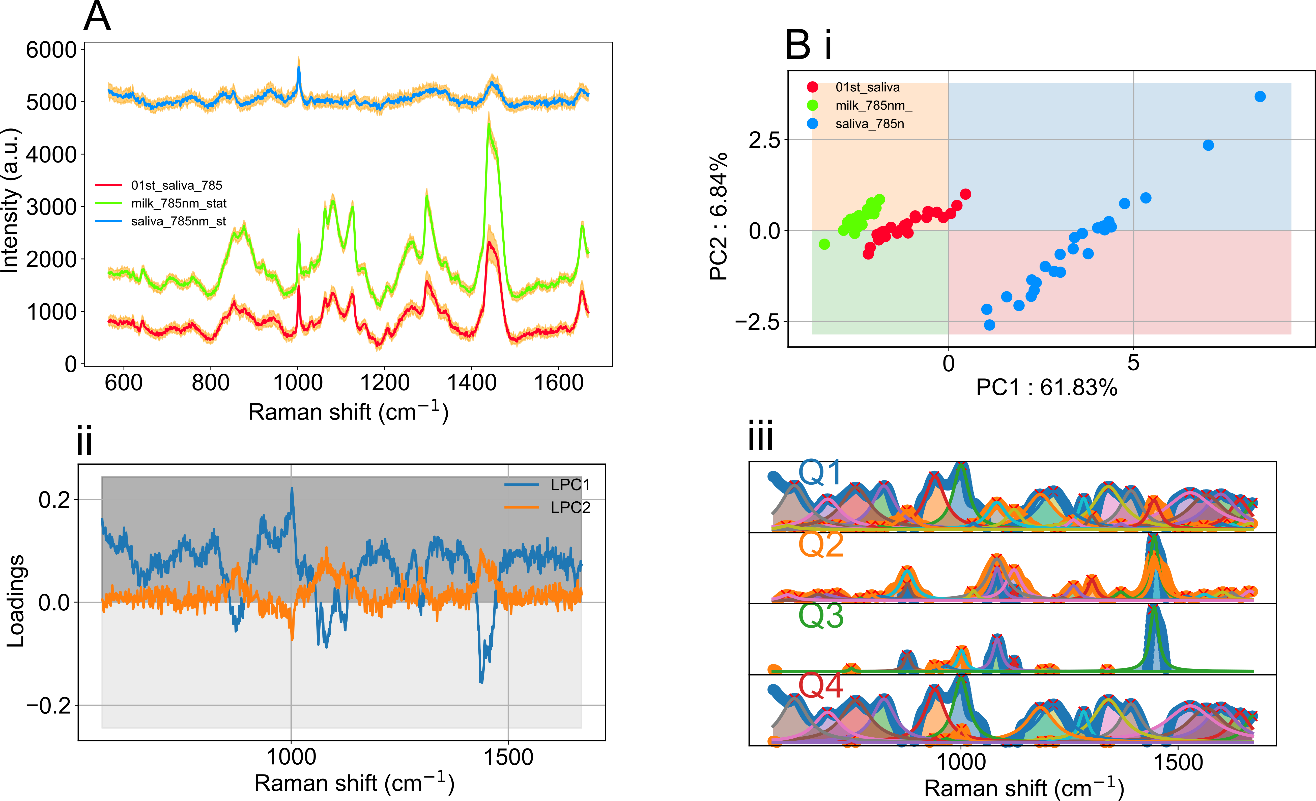


Figure 5.2.1: A) Raman spectra from saliva, milk, and saliva spiked with milk, baseline subtraction and highlighted with the standard deviation. Bi) PCA analysis of the data. ii) PCA loadings. iii) Lorentzian peak fitting of the different Loading components.

Although the separation exists, PCA lacks some fine details that we can obtain from advanced data analysis algorithms involving neural networks.

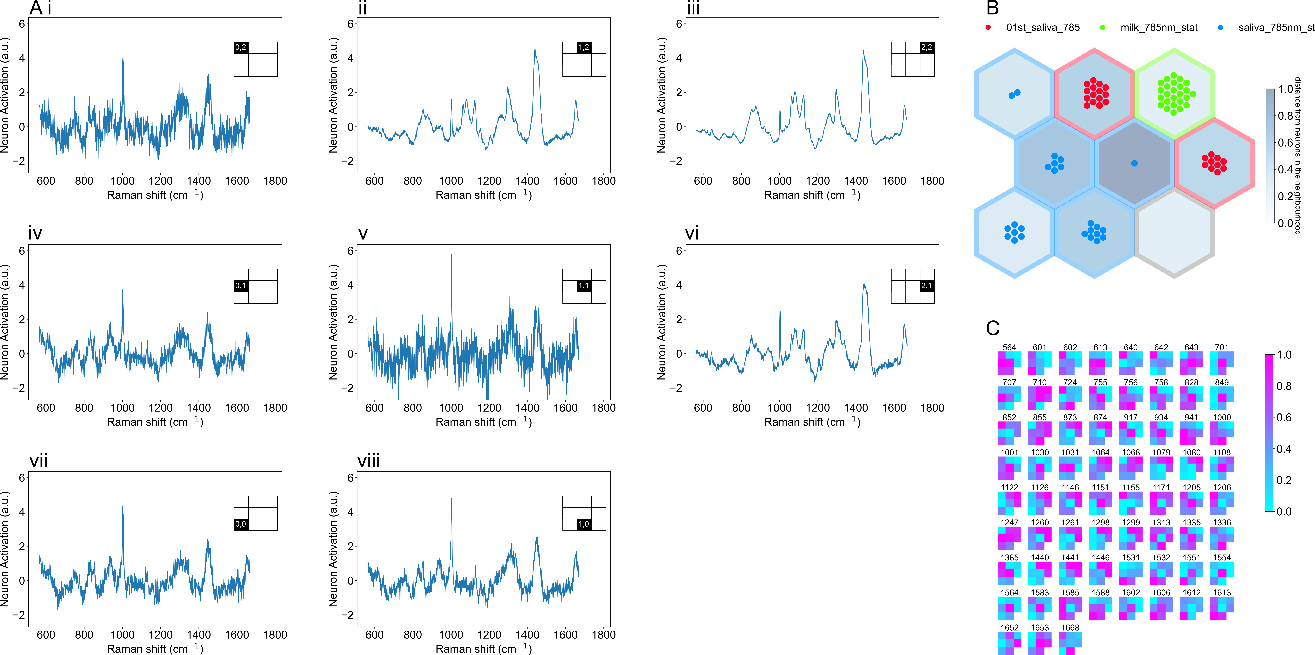


Figure 5.2.2: A)i-viii) Neuron activation spectra from the different neuron spots concerning B. B) Self Organizing Map of milk, saliva and saliva spiked with milk. C) Raman peak Activation heatmap.

From the SOM (Figure 5.2.2B), we can observe a clear distinction between different samples. A clear separation occurs between milk, saliva and saliva spiked with milk. The separation between the milk and saliva has the milk in-between these two samples. Confirming the relation between milk (Y) and saliva(X) and saliva spiked with milk (X+Y)

From each neuron, we can get the activation regarding individual neurons. The activation spectra are related to the populated samples in that neuron. Also, the heatmap activation plot helps in tracking specific peaks and measuring their activation.

The observations were done with the SOM analysis perfect the observation done over the PCA loadings. Although the essence is the same, SOM provides easier angles to interpret the data.

We observe that the peak around 1445cm-1 is highly active for the milk, and saliva spiked with milk compared to the only saliva. Not only that peak but 1079cm-1, 1122cm-1, 1260cm-1, 1300cm-1 are very active over milk and saliva spiked with milk. There could be subtle changes between the spiked milk and milk, for instance, at the 643cm-1, 828cm-1, 874cm-1, 1000cm-1 and 1380cm-1, although subtle. For the saliva, the dominating peak is the 564cm-1, 613cm-1, 755cm-1, 828cm-1, 1000cm-1, 1588cm-1 and 1613cm-1 which are more related to saliva.

Most of the time in the literature, any kind of multivariable data analysis like PCA or neural nets like SOM have normalised data as input datasets. It is still discussed if this is the accepted approach. On the one hand, it makes all datasets comparable in a certain range (0-1). The loss of information related to different bonds’ intensity can prejudicate the analysis and skew the results. For instance, comparing the 1000cm-1 from figure 5.2.1A we can visually notice that the 1000cm-1 peak is the same height for the saliva, saliva spiked with milk and milk samples. However, in the normalised SOM analysis, the 1000cm-1 peak was more active for saliva than the other. Since when normalising, if there exist other higher peaks, this peak value is reduced.

Thus, we have implemented the option to run SOM without data normalisation. This way, we get a better understanding of how the intensity values influence the clustering.

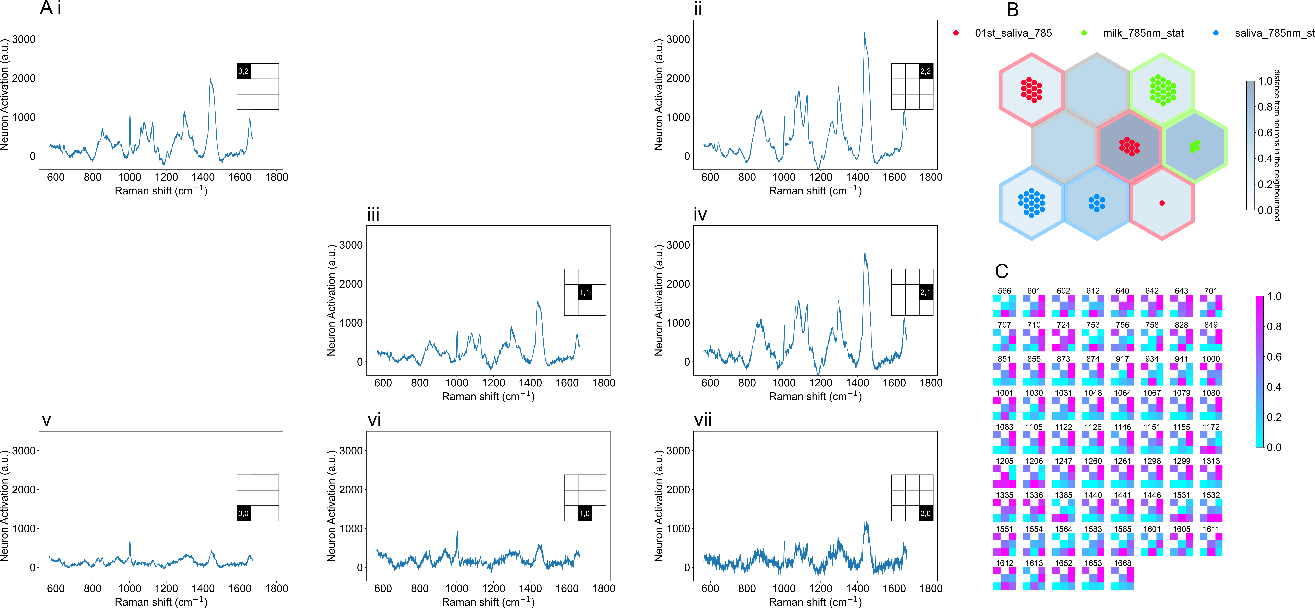


Figure 5.2.3: Data analysis without data normalisation. A)i-vii) Neuron activation spectra from the different neuron spots concerning B. B) Self Organizing Map of milk, saliva and saliva spiked with milk. C) Raman peak Activation heatmap.

Without normalised data, the separation and peak activations are very different. The two most dominant peaks are the 1000cm-1 and the 1445cm-1. The 1000cm-1 peak activation went from mostly saliva-dominated to now showing higher activation for milk and saliva spiked with milk. 1445cm-1 before showed higher activation for both milk and saliva spiked with milk, whereas now it is mostly milk dominated.

Just two examples, but the same logic can be applied to the remaining peaks. It has been pretty much established that normalisation is the way to go, but occasionally it is always better to have the real spectra speak for themselves.

Finally, the peak matching feature (Peak matching, File matching, SOM + File matching, Loadings + File Matching) shows relevant information related to where X sample is shown in X+Y measurement or where Y sample is shown in X+Y measurement (Figure 5.2.4.A B top). Also, the peak subtraction is performed, resulting in the residue sample (Figure 5.2.4.A B bottom). The residue spectra are Y’=(X+Y)-X or X’=(X+Y)-Y. where Y’ is the milk residue when subtracting the saliva (X) from the saliva spiked with milk (X+Y) and X’ is the saliva residue when subtracting the milk (Y) spectra from the saliva spiked with milk (X+Y).

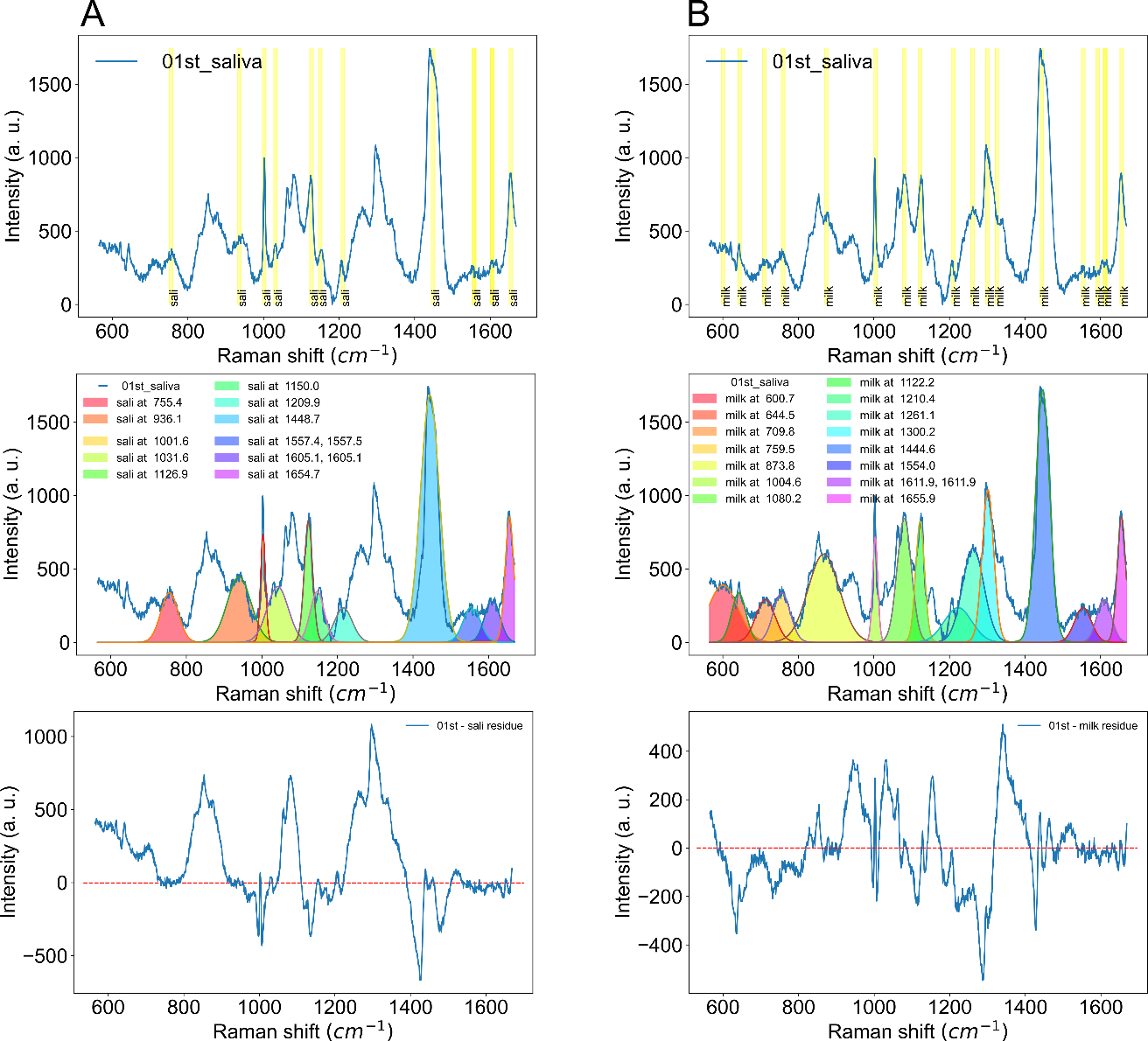


Figure 5.2.4: A)Match saliva peaks with the saliva spiked with milk spectra. Additional results show fitting to the peaks and residue from subtracting the peaks with the spectra. B) Match the milk peaks with the saliva spiked with milk spectra. Additional results show fitting to the peaks and residue from subtracting the peaks with the spectra.

The final step in this demonstration is to determine which chemicals bonds are responsible for the peaks. Also, the ratio between the matching peaks between different samples can be tracked. Here we show a couple of these peaks.

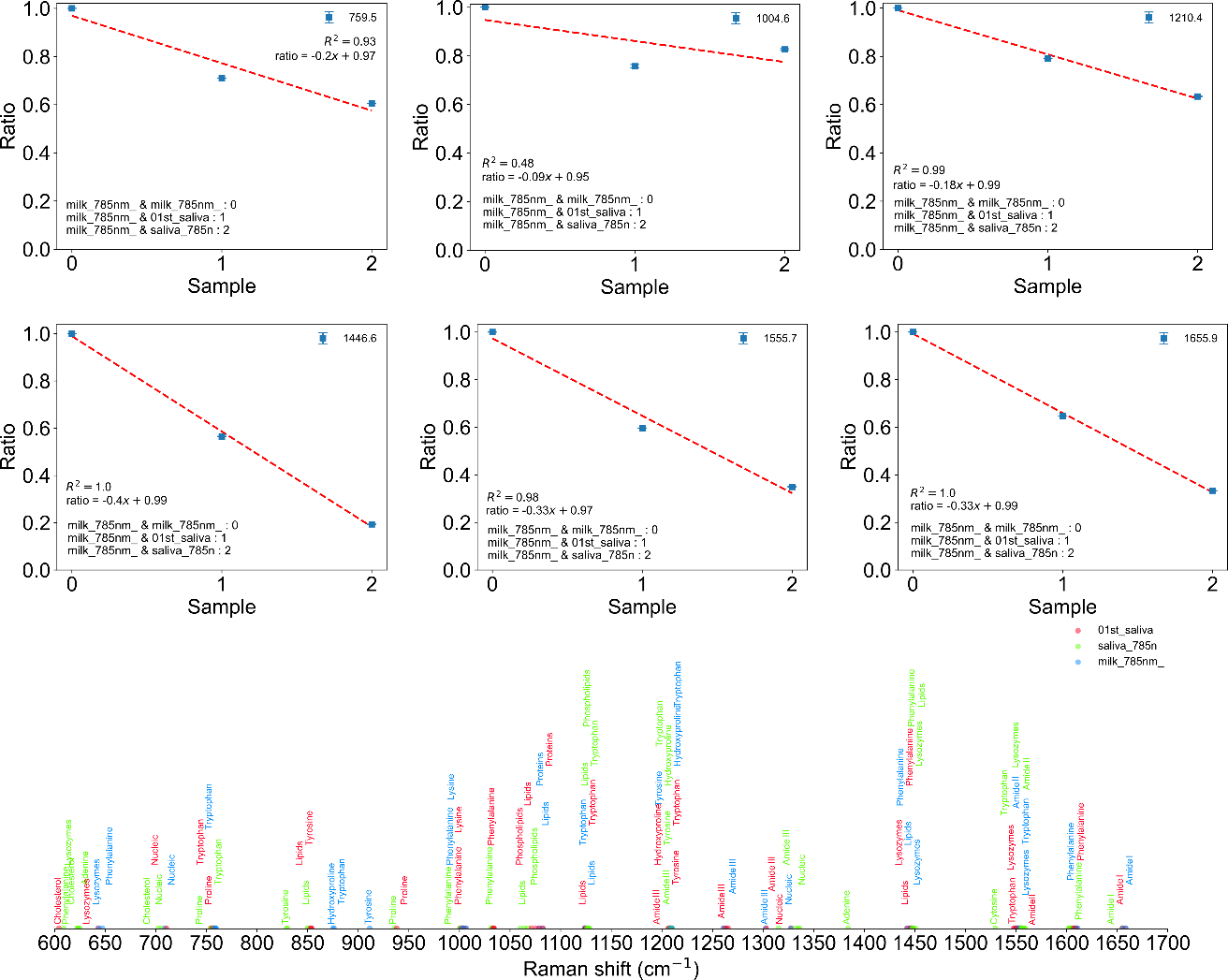


Figure 5.2.5: Ratio between different matching peaks, between milk vs milk, saliva and saliva spiked with milk. Additional information related to chemical bonds related to saliva literature is displayed.

Here we show the difference in the most important peaks between milk saliva and saliva spiked with milk.760cm-1, 1004cm-1, 1210 cm-1, 1445 cm-1, 1555 cm-1 and 1655 cm-1. Overall, all these peaks decrease in intensity (without normalisation) compared to the milk, which validates the not normalised SOM data. Finally, the peaks obtained from the literature can be used (in the appropriate format) and compared with your data, where the matching between bond and band can be made. Multiple peaks are labelled where they appear in the Raman shift values when they match the sample Raman spectra. It is a simple visual representation of the bond location respective to the match of your peaks.

These are some of the plots and data analyses that can be done over the code presented here. We did not go through other features involving the SOM peak analysis of the activations and the PCA Loadings peak analysis, which would only confirm what we discussed. It serves as additional confirmations to our assertions, but it is unnecessary to have them only if a very fine analysis is needed.

1. **Commercial context**
   1. Please briefly describe any products that are currently on the market that do or purport to do what your invention does. If there is no direct competitive product, please provide details of what you think is the closest technology or competing product.

The only product to my knowledge that slightly approximates what we demonstrated here is WIRE from Renishaw. It has peak analysis and fit and a peak library that can be used to compare. However, it misses most data analysis features like SOM and PCA, as discussed here. Also, the freedom to building your peak library is not present in Wire. Furthermore, as minor upgrades, the plotting features are improved with the stack and waterfall plot options.

* 1. Please briefly describe what it is you are trying to protect in a commercial rather than scientific sense.

Major:

Multiple Raman peak fitting method, which generates a peak library;

Use the .csv file generated as your own peak library to compare with other Raman samples;

Facilitated PCA Loadings interpreter with its own peak library;

SOM Activation interpreter with its own peak library;

All output libraries are compatible, and peak matching can be preform between them;

Create your own peak library and compare them with generated libraries;

Peak matching ratio comparison for an easy quantitative Raman analysis.

Minor:

Stack plot and Waterfall plot with optimal spacing;

PCA (2D and 3D) Euclidean cluster distance analyser.