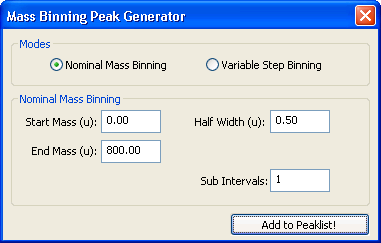
**PCA analysis manual**

1. Export txt data
   1. In the “spectra” window of Surface Lab 6 software → open the set of interesting spectra (either positive or negative)
   2. For every spectrum, follow the following steps to calibrate it:

Double-click a spectrum to select it → press F3 (or Spectrum -> Mass Calibration) to enter the mass calibration menu → default calibration points for negative spectra should be H-, C-, C\_2-, C\_3- and for positive spectra should be H+, H\_2+, CH\_3+, C\_3H\_3+ → if H- is highlighted red, it is oversaturated and needs to be excluded (which is done automatically), so click “Recalibrate” in the lower right-hand corner → exit the mass calibration menu by pressing ESC

* Note for positive spectra calibration: The 4th default ion, C\_3H\_3+, usually causes large deviations (>100ppm) on the calibration; we should be using CH\_3O+ instead. To fix this, on each spectra’s mass calibration menu, you should right click this ion 🡪 select “Delete” 🡪 click on the “Find” search bar on the top-middle section of the screen 🡪 type CH\_3O+ and press enter 🡪 click on the red arrow that pops up on the center panel 🡪 an ID for the species you selected should pop up on the mass calibration menu 🡪 if it is incorrect, tune your result on the zoomed-in spectrum panel below the main panel by clicking the red arrow on that panel 🡪 now CH\_3O+ should show up in the mass calibration panel’s text box 🡪 select “Add” 🡪 select “Recalibrate” 🡪 your calibration should have smaller deviations now!
  1. Activate the 1st spectrum→ “Peak List” → “Create Nominal Mass Binning” → in the new small window, copy the parameters below and click “Add to Peaklist”



* 1. Don’t close the pop-up window → right-click the text box that says “Nominal Mass Binning” in the submenu found in the lower left-hand corner→ select “Add To” -> “All Spectra” → recheck a few sample spectra to make sure all new peak lists have been created → now close the pop-up window
  2. Now right click the dialog box next that says “Nominal Mass Binning” again → select “Replace All Existing Peak Lists Of -> All Spectra to delete the extraneous “List 1” for all other spectra → also make sure to delete “List 1” for the current spectra if it is still there → this list can cause problems later, so MAKE SURE to delete it now!
  3. Select all spectra → click the “Peak Statistics On Selected Spectra” button at the left side of the peak list sub-window → a new window should pop up after 1-2 minutes, showing the results → select “Area normalized by Total Ion Intensity” → save the .txt file

1. Change default working directory: find pca7.py file and its accompanied PCA file folder, e.g., pca.py file is found in the folder of “C:/Anaconda/scripts/” and the accompanied PCA file folder is “C:/Anaconda/scripts/PCA/”🡪 open pca7.py file with notepad 🡪 in line of the default working directory, “pcaDir= “C:/Anaconda/scripts/PCA/” 🡪 if necessary, adjust directory to make it sure these two directories are same.
2. Group column names of spectra file: Open the .txt data file using Excel → remove any irrelevant columns and rows (i.e., we probably don’t need the top- or bottom-most two rows or the stats columns at the end (mean, median, etc.) → make sure the columns are in order: Do this by highlighting all cells and going to “Data” -> “Sort” -> add “Sort by” “Row 1”, then select “Options” -> “Sort left to right” -> click OK → (1) make sure the 1st column’s name is “Mass (u)” 🡪 edit column headers; currently, they might be something like <No Sample Name> (01-N1) \*, <No Sample Name> (01-N2) \*…..; change them to 01-N1, 01-N2, 01-N3, 01-N4, 01-N5, 01-N6, 02-N1, 02-N2, ….. using CTRL+H to find and replace any extraneous character expressions → save the new file as “sims\_data.txt” → paste the new file into the folder: C:/Anaconda/scripts/PCA/OriginalData/ →
3. Now open the file again in notepad (preferably Notepad++ for RegEx editing purposes) → if there are any blank lines, remove them using either RegEx or by writing a quick Python script → also remove any trailing whitespace using Notepad++ by selecting “Edit” -> “Blank Operations” -> “Trim Trailing Space” → once you are certain that no pesky extraneous whitespaces are remaining, save the file and exit (if you have ANY left, you could get an error when doing PCA analysis later; usually, \*\*\*Error! Cannot Recognize Data!\*\*\*
4. Rename group names in final figures: go to C:/Anaconda/scripts/PCA/Group Names/ → open Group Names.txt → after 1, 2, 3, 4 …, rename or input names you want to show in the final figure for group 01, 02, 03, 04….
5. In “Start menu” find out “Anaconda Powershell Prompt (anaconda).exe”, Right click “Anaconda Powershell Prompt (anaconda).exe” 🡪 Run as administrator 🡪 Input “cd C:\Anaconda\scripts” then press Enter 🡪Input command python pca7.py then press Enter
6. Go to the folder C:/Anaconda/scripts/PCA/Exports/ to check PCA results
   1. 10 combinations of 2-d PCA scores plots of PC1-PC5 with confidence circle
   2. 10 combinations of 2-d PCA scores plots of PC1-PC5 without confidence circle
   3. PC1-PC5 single PC scores plots
   4. PC1-PC5 scores tables for all spectra
   5. PC1 to PC5 loadings plots
   6. PC1-PC5 top 20 loadings tables
   7. PC1 to PC10 weight “Percentage of explained variance”
   8. PC1 to PC10 weight “Percentage of explained variance” table

**Note 1:** In 10 combinations of 2-d PCA scores plots of PC1-PC5 with confidence circle**,** 10 combinations of 2-d PCA scores plots of PC1-PC5 without confidence circle, 1-d PC1 to PC5 scores plots, andPC1 to PC5 loadings plots **,** change **coordinate system sizes, figure resolution** 🡪 open pca7.py 🡪 find the following items: line 16-32 🡪 change the parameters 🡪 save it and run again:

'axes.labelsize': '28',

'xtick.labelsize':'28',

'ytick.labelsize':'28',

'lines.linewidth':3,

'lines.markersize':10,

'legend.fontsize': '25',

'dpi = 300'

**Note 2:** The circle is **Confidence Interval** (or “confidence limits”) = 90%. If other values are needed, open pca7.py 🡪 find “nstd = 1.645” 🡪 change “1.645” to the values based on the below table 🡪 save pca7.py and re-run it:

|  |  |
| --- | --- |
| ***Confidence Interval*** | ***Z*** |
| 80% | 1.282 |
| 85% | 1.440 |
| 90% | 1.645 |
| 95% | 1.960 |
| 99% | 2.576 |
| 99.5% | 2.807 |
| 99.9% | 3.291 |

**Note 3:** If more or less **PCs** are required 🡪 open pca7.py 🡪 find the following items 🡪 change the parameters 🡪 save pca7.py and re-run it: (Assume that the number of required PCs is n, modify the red areas highlighted below)

Line 68-69：per\_var=np.round(pca.explained\_variance\_ratio\_\*100,decimals=2)[:n]

labels=['PC'+str(x) for x in range(1,n+1)]

Line 92: pca\_data=pca\_data[:,:n]

Line 111: pca\_df.to\_csv(pcaDir+'Exports/PC1-n\_SCORE\_TABLE.txt')

Line 147-148: for j in range(1, n+1):

for k in range(j+1, n+1):

Line 169-170: for j in range(1, n+1):

for k in range(j+1, n+1):

Line 204: for pc in range(1,n+1):

Line 242-243: loading\_scores=pca.components\_[:n,:]

loadingTable=pd.DataFrame(loading\_scores.T,index=mass,columns=[1,2,3,4,5…,n])

Line 248: for i in range(0,n)

Line 296: loadingTable=pd.DataFrame(loading\_scores.T,indx=mas.columns=['PC1', 'PC1', 'PC1'，'PC1', 'PC1', …, 'PCn']

Line 300: loadingTable.to\_csv(pcaDir+'Exports/PC1-n\_loadingTable.txt')

**Note 4:** In 10 combinations of 2-d PCA scores plots of PC1-PC5 with confidence circle, 10 combinations of 2-d PCA score plots of PC1-PC5 without confidence circle and 1-dPC1 to PC5 score plots, change **symbol sizes, colors** 🡪 open pca7.py 🡪 find the following items: line 136-145 🡪 change the parameters 🡪 save pca7.py and run again

**Note 5:** In10 combinations of 2-d PCA scores plots of PC1-PC5 with confidence circle, change  **figure scale**, **label fontsize, fontname, fontweight and circle linestyle, linewidth, facecolor, alpha**: open pca7.py 🡪 find the corresponding parameters: line 169-201 🡪 change the parameters 🡪 save pca7.py and run again.

**Note 6:** In 10 combinations of 2-d PCA scores plots of PC1-PC5 without confidence circle, change **figure scale, label fontsize, fontname, fontweight**: open pca7.py 🡪 find the corresponding parameters: line 147-167 🡪 change the parameters 🡪 save pca7.py and run again.

**Note 7:** In PC1 to PC5 single PC scores plots, change **figure scale, label fontsize, fontname, fontweight**: open pca7.py 🡪 find the corresponding parameters: line 204-228 🡪 change the parameters 🡪 save pca7.py and run again.

**Note 8:** InPC1 to PC5 loadings plots, change **figure scale,** **fetch number, label fontsize, fontname, fontweight and bar size, color, text fontsize**: open pca7.py 🡪 find the corresponding parameters: line 240-270 🡪 change the parameters 🡪 save pca7.py and run again.

**Note 9:** In PC1 to PC10 weight “Percentage of explained variance” plots, change **figure scale, ticklabel size, label name, fontsize, fontname, fontweight and bar size, color**: open pca7.py 🡪 find the corresponding parameters: line 68-87 🡪 change the parameters 🡪 save pca7.py and run again.

**Note 10:** In PC1-PC5 loadings plots, change **the labeling numbers of m/z**: open pca7.py 🡪 find the corresponding parameters: line 244 🡪 change the parameters 🡪 save pca7.py and run again. (Assume that the number of required labeling numbers is m, modify the red areas highlighted below)

fetchn=m

**Note 11:** In PC1-PC5 loadings tables, change **the** **numbers of +/- loadings**: open pca7.py 🡪 find the corresponding parameters: line 245 🡪 change the parameters 🡪 save pca7.py and run again. (Assume that the number of required numbers of +/- loadings is l, modify the red areas highlighted below)

fetchn\_more=l

**Section 1: Formatting Raw Data File**

The following section provides guidance to the user on how to properly format the raw data txt file. Careful consideration of the formatting is necessary for the code to execute.

**Note 1:** The first column name is ‘Mass (u)’ and the values in the column should be whole number integers of the unit mass.

**Note 2:** The remaining column names should be adjusted as follows:

* The general column name format takes the form ‘XXXX’
* The first two characters are the sample number
  + For the first sample of interest, XX = 01
  + There can be up to 99 samples simultaneously analyzed
* The last two characters consider whether the spectra are in positive or negative ion mode, and the measurement location number
  + For example, if there are six locations analyzed for the first sample, and the data is from negative ion mode for the first location, then XX = N1
* Additional example: Column name for the third sample at the fourth location analyzed in negative mode: 03N4

The following screenshot shows an example of the data formatting:

**Table

Description automatically generated with medium confidence**

**Note 3:** Save the formatted data as a text file in SIM\_PCA/sims-data/OriginalData

**Section 2: Formatting the Metadata File**

To properly correspond the sample name to the number inputted into the txt file from Section 1, the metadata.txt file (found in SIM\_PCA/sims-data/OriginalData) should be appropriately formatted. The following image shows an example; a key consideration is to not have any spaces after the commas in the txt file:

**Graphical user interface, text

Description automatically generated**

**Section 3: Report Generation**

Once the raw data and metadata files have been properly formatted and saved, make the following changes to the main.py file (located in SIMS\_PCA/src) to analyze the data:

**Note 1:** Indicate whether in positive or negative ion mode:

****

**Note 2:** Specify the raw data txt file path and name:

****

**Note 3:** Once the main.py file is updated and saved, execute in Terminal or equivalent location to execute Python code. If successfully run, the complete PCA/SIMS spectra analysis report can be found in SIMS\_PCA/output\_sample/report.docx

**Section 4: Peak Assignment Table in Report**

Once the PCA code is successfully run, a table of peak assignments can be found in the report. The peak assignments are taken from the ‘negative\_doc\_mass\_record.csv’ and ‘positive\_doc\_mass\_record.csv’ documents found in SIMS\_PCA/sims-data. These are the default peak assignments, but the user can manually input different assignments for a given unit mass based on the context of the sample.

For example, in positive ion mode, the signals of C2H3+, Al+, and CNH+ all correspond to a unit mass of 27. Therefore, the peak assignment table can be updated manually from the mass record files to reflect different options of peak assignments for the same unit mass. This enables the user to select peak assignments that are most relevant to the context of the sample analyzed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Default Mass Record Document | | | → | User Updated Mass Record Document | | |
| Unit Mass | Assignment | Document Mass | Unit Mass | Assignment | Document Mass |
| 27 | C2H3+ | 27.02 | 27 | C2H3+  Al+  CNH+ | 27.02  26.98  27.01 |

If the mass record documents are updated, the user must re-run the python script to have the updated peak assignments appear in the exported report document.

**Section 5: Adding Mass Combination Rules for Similar Chemical Species**

The following section provides guidance on how the user can incorporate rules to assign mass values to similar chemical species. For example, for SiOx-based signals, there can be a variety of mass values that are correspondent. Referring to the ‘negative\_doc\_mass\_record.csv’, values related to SiOx are tabulated.

|  |  |  |
| --- | --- | --- |
| Unit Mass | Assignment | Document Mass |
| 60 | SiO2- | 59.9673 |
| 61 | SiO2H- | 60.9751 |
| 76 | SiO3- | 75.9622 |
| 77 | SiO3H- | 76.97 |
| 136 | Si2O5- | 135.9289 |
| 137 | Si2O5H- | 136.9368 |

**Note 1:** To add additional assignments for a particular chemical species, the user must add information regarding unit mass, signal assignment, and document mass into the mass record file

* For positive ion signals, make additions to ‘positive\_doc\_mass\_record.csv’
* For negative ion signals, make additions to ‘negative\_doc\_mass\_record.csv’
* These files can be found in SIMS\_PCA/sims-data

**Note 2:** Once these documents are updated for their respective signals, the user must update the python file which employs the mass combination rules by specifying the relevant unit masses.

* The python file to update is found here: SIMS\_PCA/src/pca\_sims/pca\_sims.py
* The following snapshot is a sample code in the pca\_sims.py file (line 442-460)
* The user should add the unit mass for the signals which correspond to SiOx in line 443

Text, letter

Description automatically generated