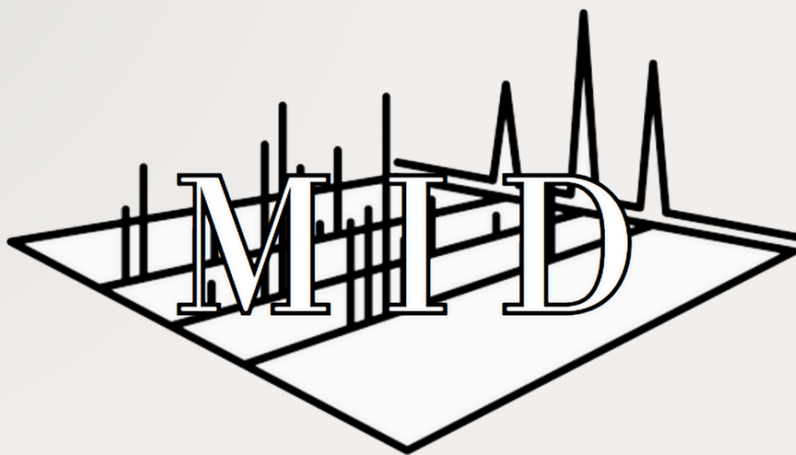


Molecule_ID based MSMS
fingerprint Similarity



ADD A
SPECTRUM TO
THE REFERENCE
DATABASE.

Fayçal Hassani

TABLE OF CONTENTS

01. EXPLANATION

02. SPECTRUM ADDITION

Explanation

To obtain a model that is as reliable as possible, it is crucial to have a large reference database.

To achieve this, it is necessary to convert a RAW file to mzML format.

Once this conversion is completed, the spectrum must then be added to the dictionary.

This step enriches the database and improves the accuracy and reliability of the model.

Spectrum addition

To convert raw data files to mzML format, it is necessary to open the command prompt and use the following command:

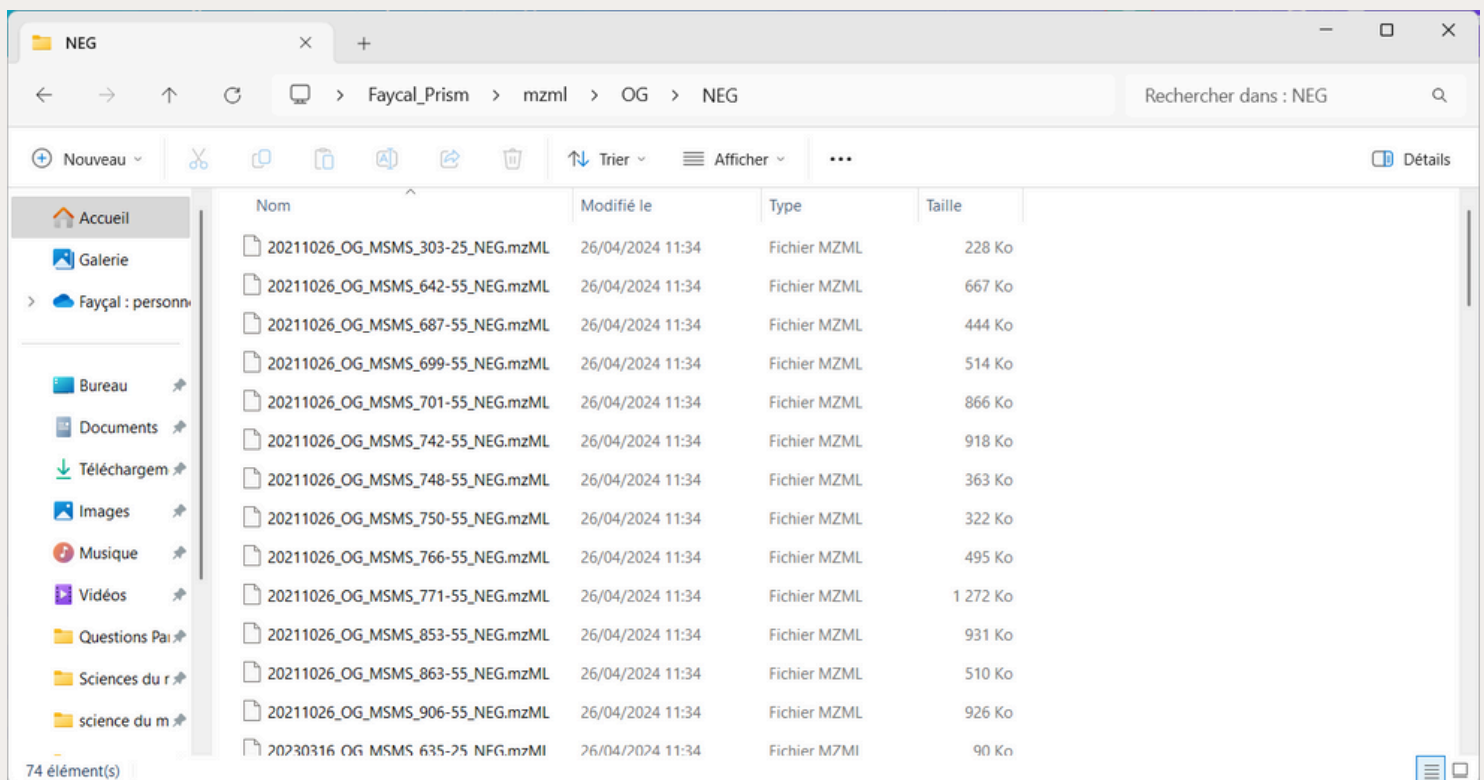
```
msconvert "$raw_file" --mzML -o  
"$output_directory".
```

However, before being able to execute this command, it is imperative to install ProteoWizard.

To do this, it is recommended to consult the user manual to ensure that the installation is correctly performed and that all necessary dependencies are met.

The next step is to move this newly created file to the directory where all other mzML files are stored.

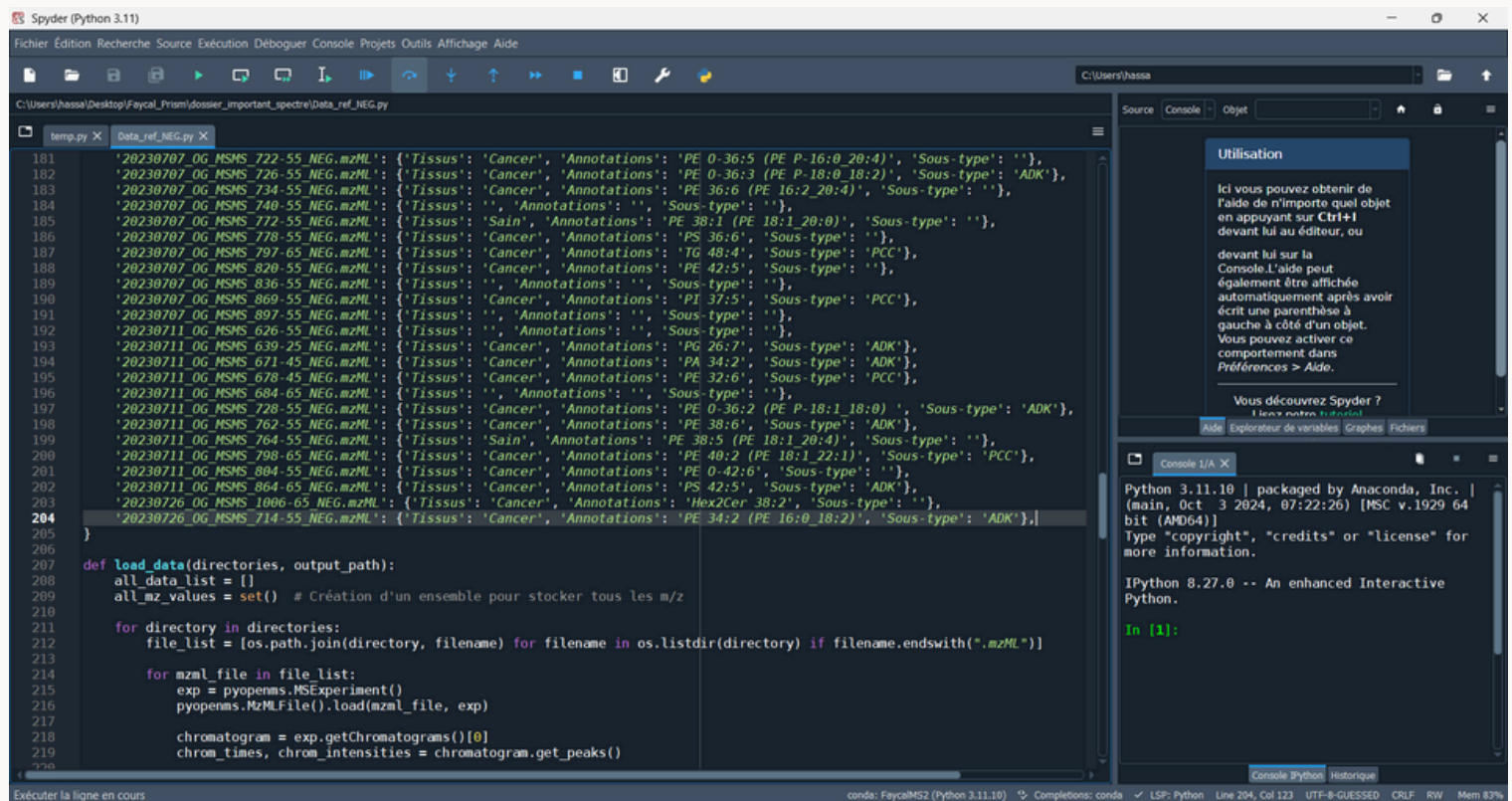
This can be done by simply dragging the converted file into the appropriate folder.



This organization allows for the centralization of all mzML files in one location, thus facilitating their management and subsequent analysis.

Now, you need to add an entry to the dictionary following this format:

```
"filename.mzML": {'Tissue': '', 'Annotations': '',  
'Subtype': ''}.
```



```
181 '20230707_06_MSMS_722-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 0-36:5 (PE P-16:0_20:4)', 'Subtype': ''},  
182 '20230707_06_MSMS_726-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 0-36:3 (PE P-18:0_18:2)', 'Subtype': 'ADK'},  
183 '20230707_06_MSMS_734-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 36:6 (PE 16:2_20:4)', 'Subtype': ''},  
184 '20230707_06_MSMS_740-55_NEG.mzML': {'Tissue': '', 'Annotations': '', 'Subtype': ''},  
185 '20230707_06_MSMS_772-55_NEG.mzML': {'Tissue': 'Sain', 'Annotations': 'PE 38:1 (PE 18:1_20:0)', 'Subtype': ''},  
186 '20230707_06_MSMS_778-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PS 36:6', 'Subtype': ''},  
187 '20230707_06_MSMS_797-65_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'TG 48:4', 'Subtype': 'PCC'},  
188 '20230707_06_MSMS_820-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 42:5', 'Subtype': ''},  
189 '20230707_06_MSMS_836-55_NEG.mzML': {'Tissue': '', 'Annotations': '', 'Subtype': ''},  
190 '20230707_06_MSMS_869-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PI 37:5', 'Subtype': 'PCC'},  
191 '20230707_06_MSMS_897-55_NEG.mzML': {'Tissue': '', 'Annotations': '', 'Subtype': ''},  
192 '20230711_06_MSMS_626-55_NEG.mzML': {'Tissue': '', 'Annotations': '', 'Subtype': ''},  
193 '20230711_06_MSMS_639-25_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PG 26:7', 'Subtype': 'ADK'},  
194 '20230711_06_MSMS_671-45_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PA 34:2', 'Subtype': 'ADK'},  
195 '20230711_06_MSMS_678-45_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 32:6', 'Subtype': 'PCC'},  
196 '20230711_06_MSMS_684-65_NEG.mzML': {'Tissue': '', 'Annotations': '', 'Subtype': ''},  
197 '20230711_06_MSMS_728-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 0-36:2 (PE P-18:1_18:0)', 'Subtype': 'ADK'},  
198 '20230711_06_MSMS_762-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 38:6', 'Subtype': 'ADK'},  
199 '20230711_06_MSMS_764-55_NEG.mzML': {'Tissue': 'Sain', 'Annotations': 'PE 38:5 (PE 18:1_20:4)', 'Subtype': ''},  
200 '20230711_06_MSMS_798-65_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 40:2 (PE 18:1_22:1)', 'Subtype': 'PCC'},  
201 '20230711_06_MSMS_804-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 0-42:6', 'Subtype': ''},  
202 '20230711_06_MSMS_864-65_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PS 42:5', 'Subtype': 'ADK'},  
203 '20230726_06_MSMS_1006-65_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'Hex2Cer 38:2', 'Subtype': ''},  
204 '20230726_06_MSMS_714-55_NEG.mzML': {'Tissue': 'Cancer', 'Annotations': 'PE 34:2 (PE 16:0_18:2)', 'Subtype': 'ADK'},  
205 }  
206  
207 def load_data(directories, output_path):  
208     all_data_list = []  
209     all_mz_values = set() # Cr ation d'un ensemble pour stocker tous les m/z  
210  
211     for directory in directories:  
212         file_list = [os.path.join(directory, filename) for filename in os.listdir(directory) if filename.endswith(".mzML")]  
213  
214         for mzml_file in file_list:  
215             exp = pyopenms.MSExperiment()  
216             pyopenms.MzMLFile().load(mzml_file, exp)  
217  
218             chromatogram = exp.getChromatograms()[0]  
219             chrom_times, chrom_intensities = chromatogram.get_peaks()  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230
```

Here, you can specify and provide additional information such as the type of tissue, relevant annotations, and the specific subtype of the mzML file.

Now, simply run the program to obtain the Parquet file, which will serve as the reference base for all subsequent analyses.