Molecule_ID based MSMS fingerprint Similarity

USER MANUAL

Fayçal Hassani

TABLE OF CONTENTS

Intended Use

The Molecule_ID software, based on MSMS fingerprint similarity, aims primarily to calculate the similarity between new MS2 spectra and a pre-recorded reference database.

By comparing these spectra, the software determines the most similar spectrum from the database.

Additionally, Molecule_ID also allows the display of the newly imported spectrum, thus providing a clear and precise visualization of the analyzed spectral data.

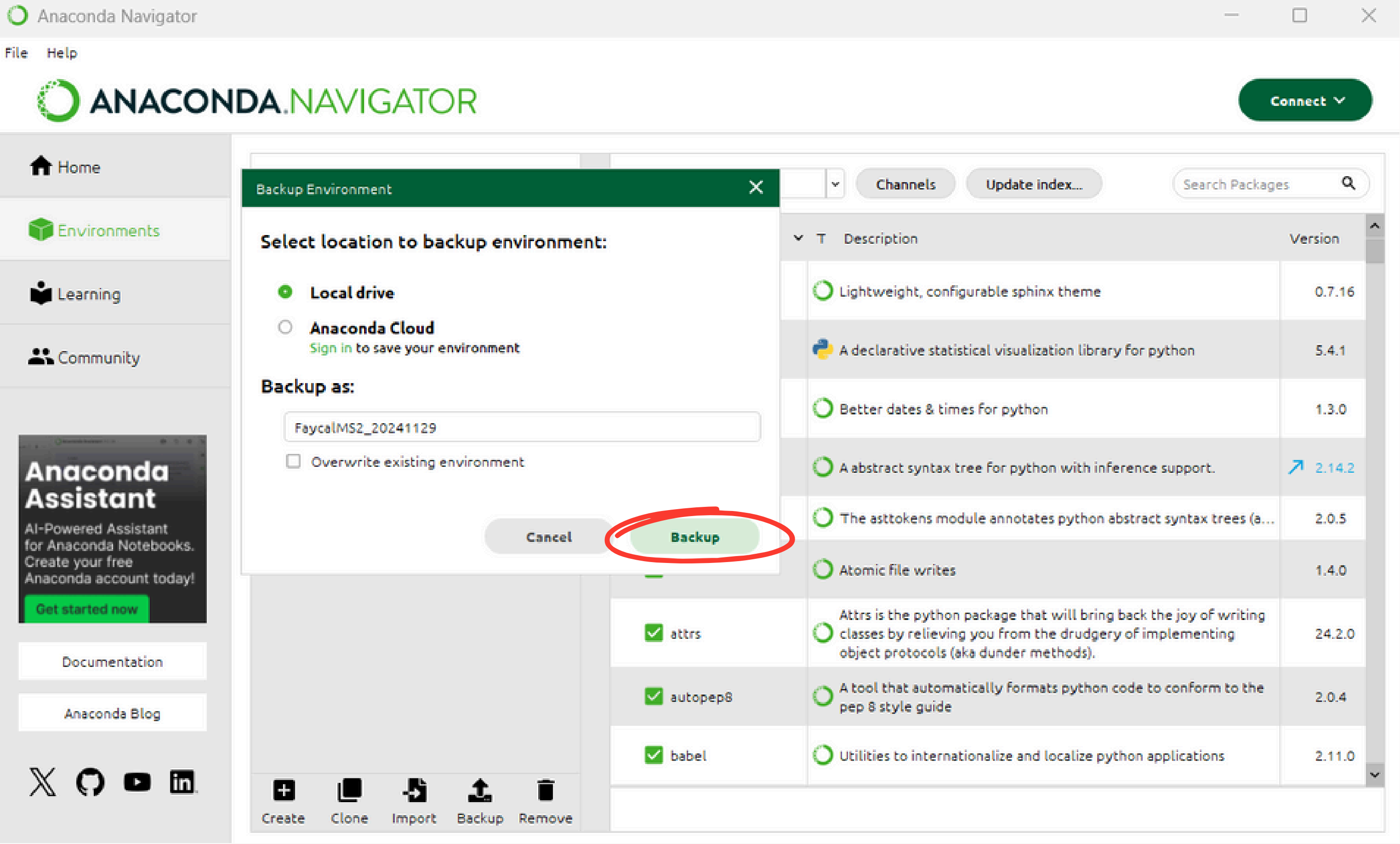
Prerequisites

1.Importing the Anaconda Environment:

Open Anaconda Navigator.

In the main menu, click on "Environments" to access environment management.

Then, go to "Backup Local Drive" to access local backups.



Backup Environment

Select location to backup environment:

● Local drive

○ Anaconda Cloud
Sign in to save your environment

Backup as:

FaycalMS2_20241129

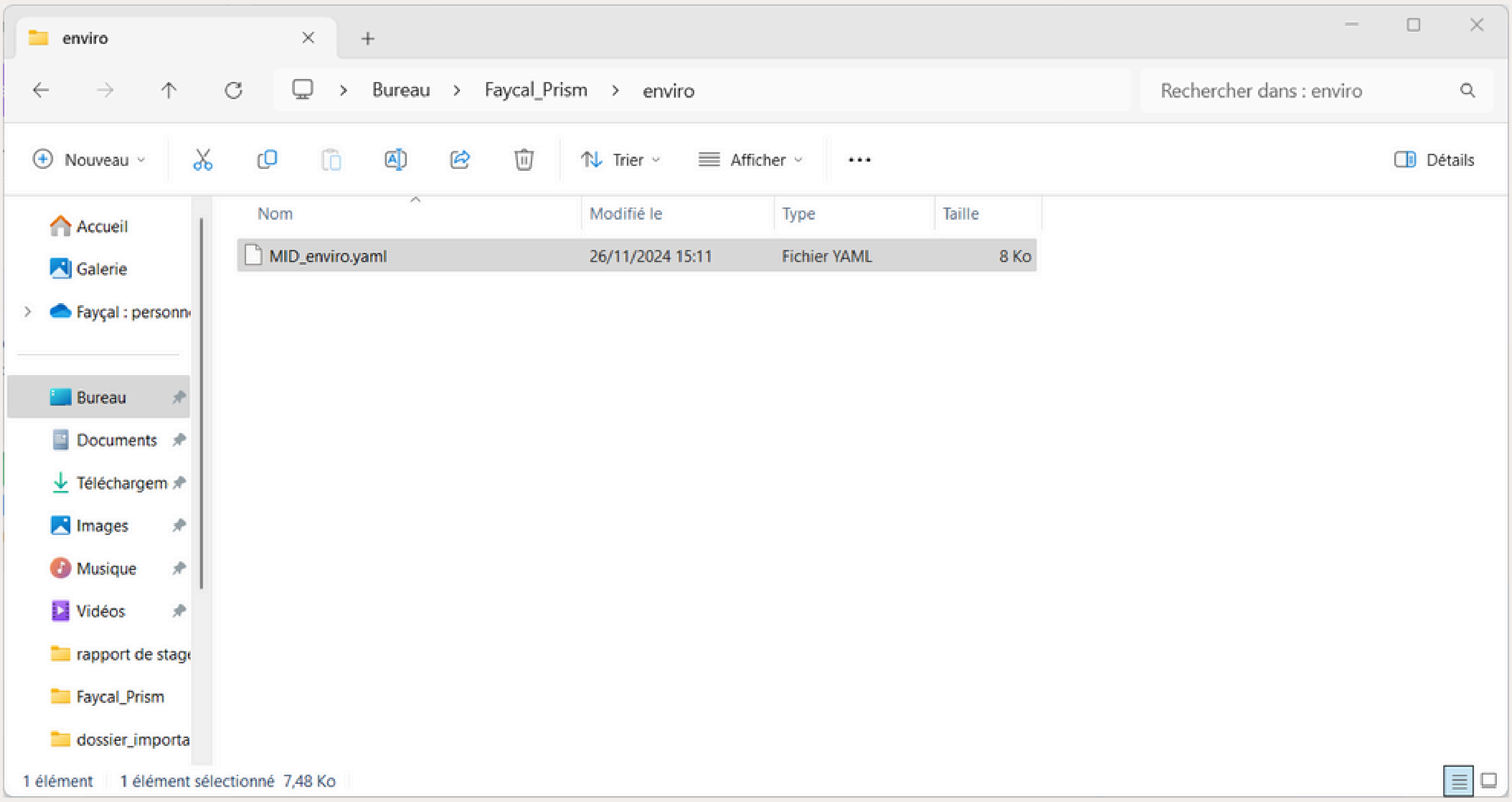
□ Overwrite existing environment

Cancel | Backup

Name	Description	Version
	Lightweight, configurable sphinx theme	0.7.16
	A declarative statistical visualization library for python	5.4.1
	Better dates & times for python	1.3.0
	A abstract syntax tree for python with inference support.	2.14.2
	The asttokens module annotates python abstract syntax trees (a...	2.0.5
	Atomic file writes	1.4.0
☑ attrs	Attrs is the python package that will bring back the joy of writing classes by relieving you from the drudgery of implementing object protocols (aka dunder methods).	24.2.0
☑ autopep8	A tool that automatically formats python code to conform to the pep 8 style guide	2.0.4
☑ babel	Utilities to internationalize and localize python applications	2.11.0

- Locate the folder named "enviro" which contains the saved environments.

- In this folder, find the file "MID_enviro.yaml" and import it by following the on-screen instructions.



Nom	Modifié le	Type	Taille
MID_enviro.yaml	26/11/2024 15:11	Fichier YAML	8 Ko

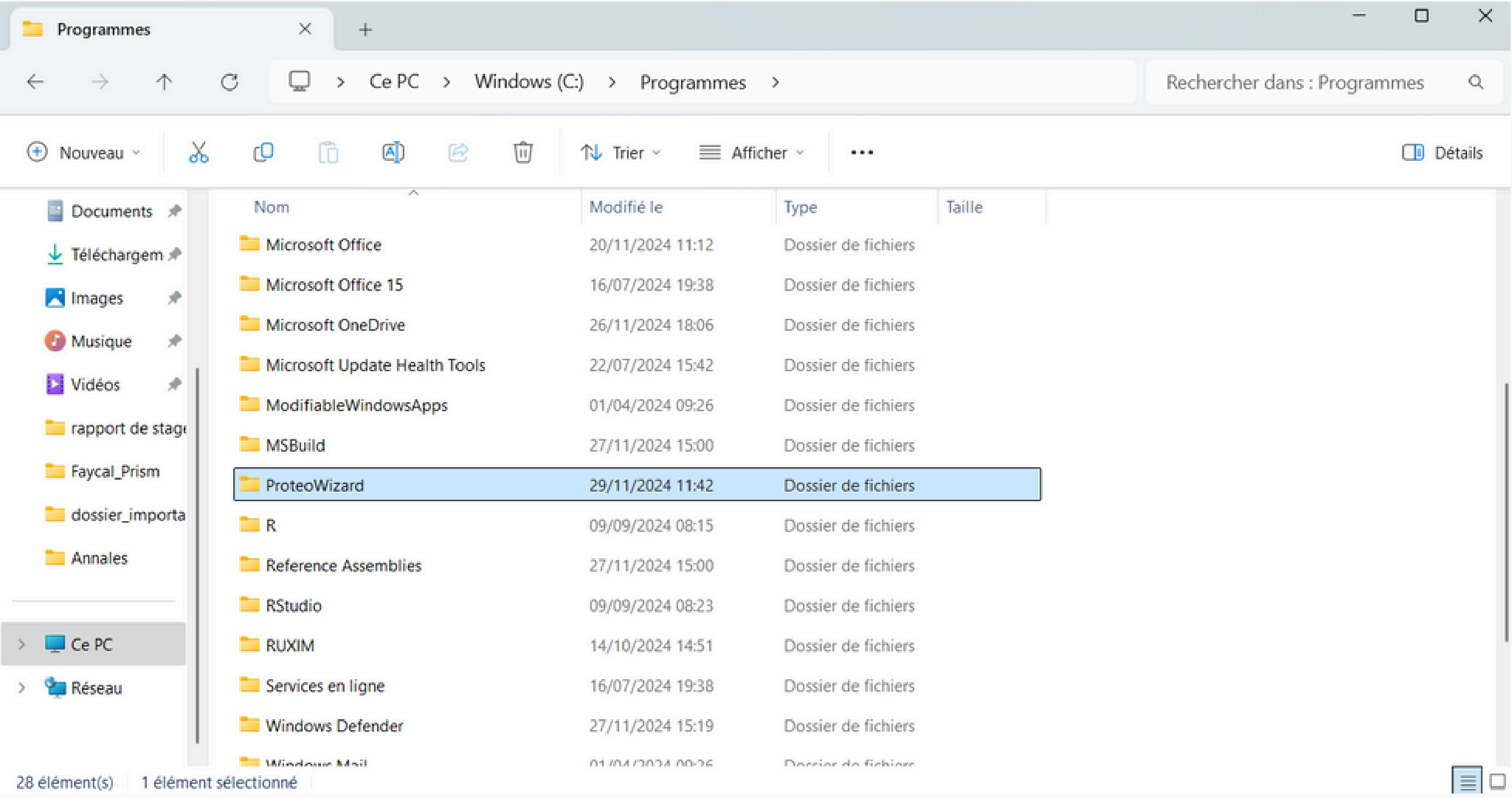
1 élément 1 élément sélectionné 7,48 Ko

- Once the import is complete, the environment will be ready for use and execution.

2.Installation of MSconvert:

- In the provided folder, locate the "ProteoWizard" folder.

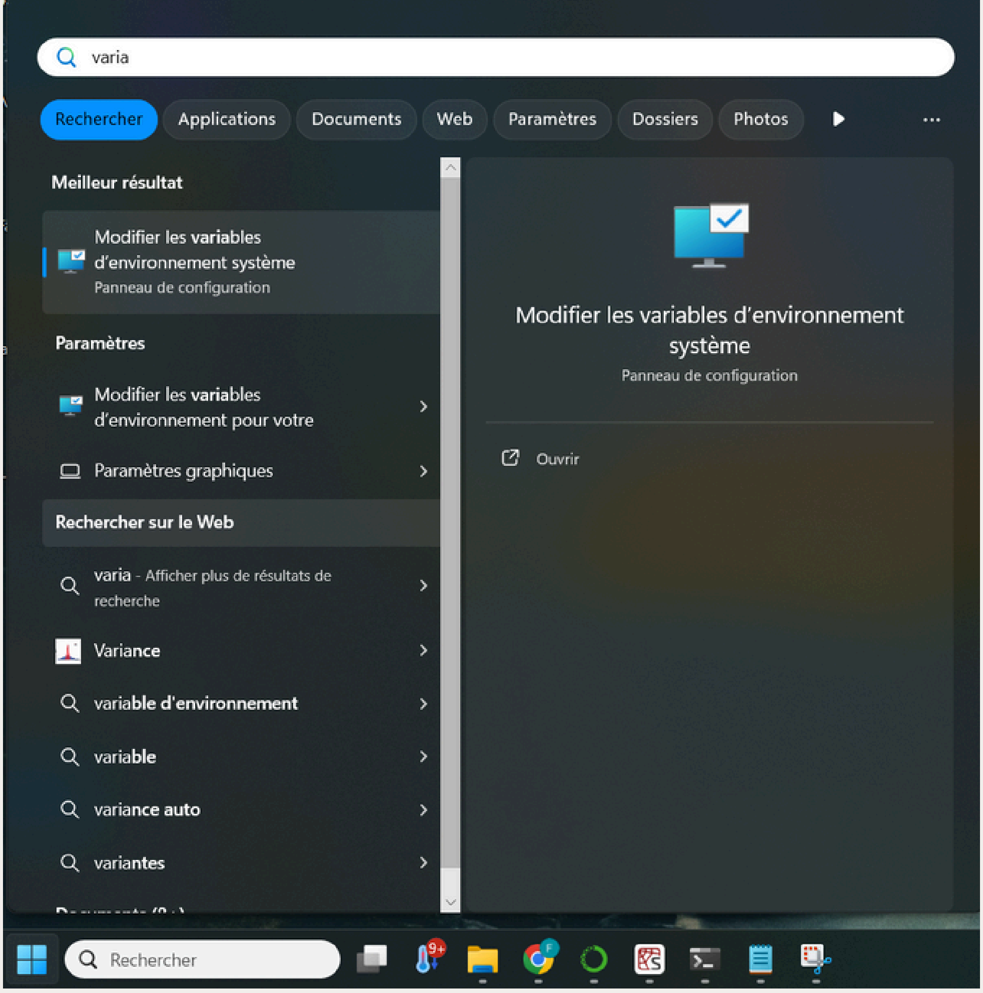
- Copy this folder and paste it into the "Programs" directory on your hard drive (typically located under "C:\Program Files" depending on your system).



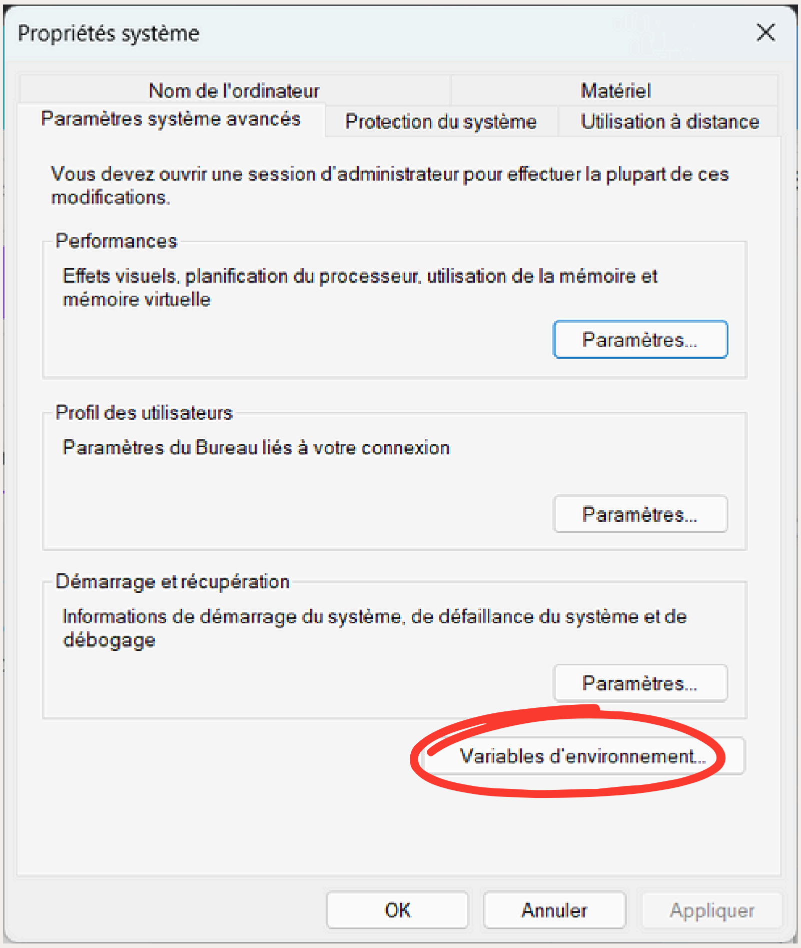
Nom	Modifié le	Type	Taille
Microsoft Office	20/11/2024 11:12	Dossier de fichiers	
Microsoft Office 15	16/07/2024 19:38	Dossier de fichiers	
Microsoft OneDrive	26/11/2024 18:06	Dossier de fichiers	
Microsoft Update Health Tools	22/07/2024 15:42	Dossier de fichiers	
ModifiableWindowsApps	01/04/2024 09:26	Dossier de fichiers	
MSBuild	27/11/2024 15:00	Dossier de fichiers	
ProteoWizard	29/11/2024 11:42	Dossier de fichiers	
R	09/09/2024 08:15	Dossier de fichiers	
Reference Assemblies	27/11/2024 15:00	Dossier de fichiers	
RStudio	09/09/2024 08:23	Dossier de fichiers	
RUXIM	14/10/2024 14:51	Dossier de fichiers	
Services en ligne	16/07/2024 19:38	Dossier de fichiers	
Windows Defender	27/11/2024 15:19	Dossier de fichiers	
Windows Mail	01/04/2024 09:26	Dossier de fichiers	

28 élément(s) 1 élément sélectionné

- Then, open the Start menu and search for "Edit the system environment variables.



- Click on "Environment Variables" to open the Environment Variables configuration window.



Vous devez ouvrir une session d'administrateur pour effectuer la plupart de ces modifications.

Performances

Effets visuels, planification du processeur, utilisation de la mémoire et mémoire virtuelle

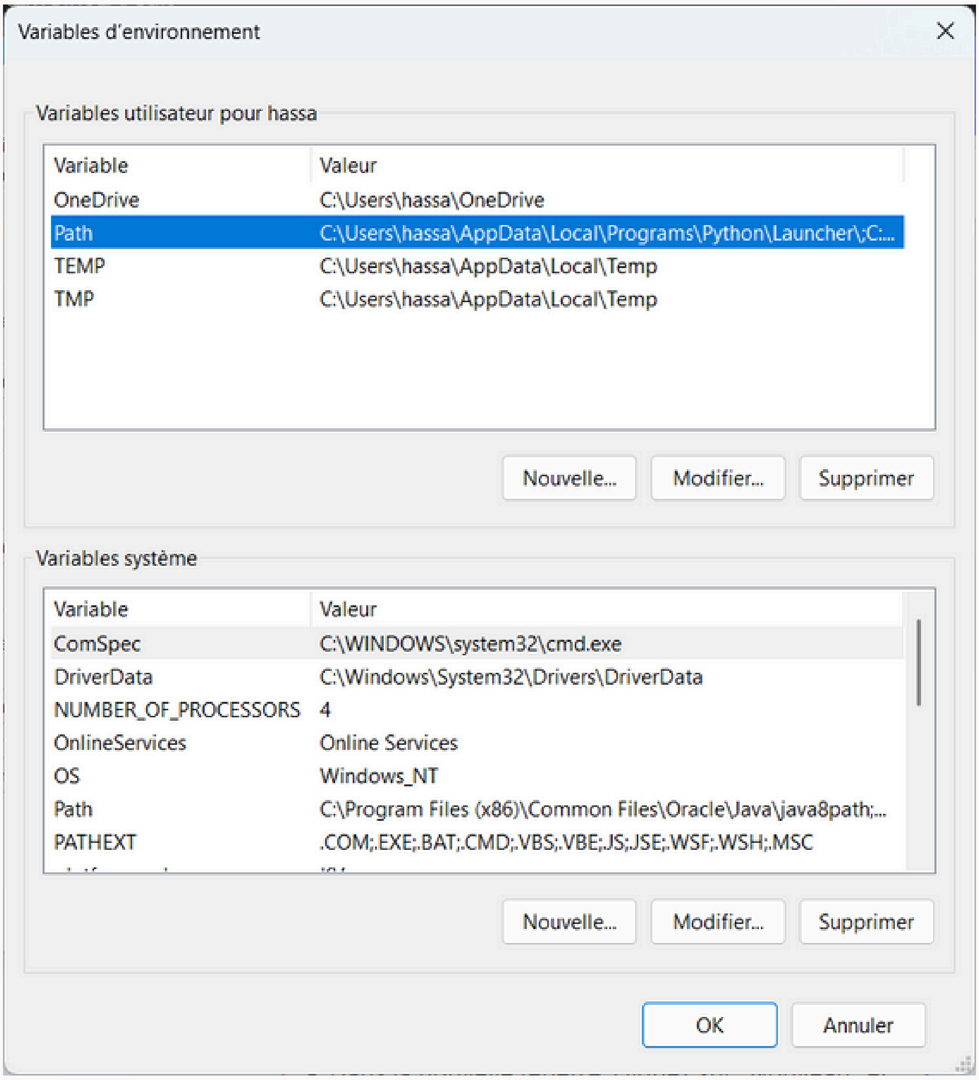
Profil des utilisateurs

Paramètres du Bureau liés à votre connexion

Démarrage et récupération

Informations de démarrage du système, de défaillance du système et de débogage

- In the "System variables" section, find the "Path" variable and select it, then click on "Edit".



Variables d'environnement

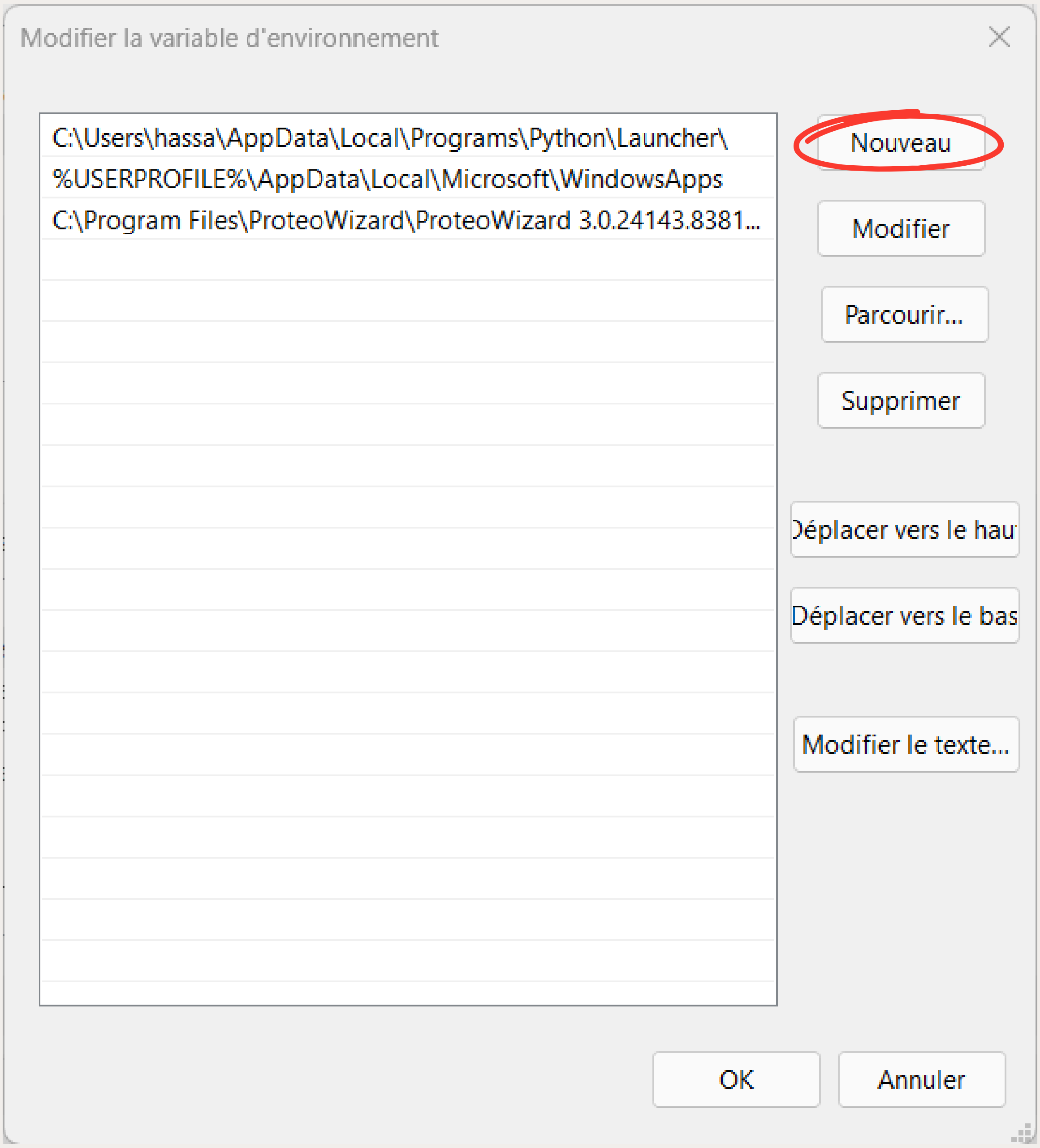
Variables utilisateur pour hassa

Variable	Valeur
OneDrive	C:\Users\hassa\OneDrive
Path	C:\Users\hassa\AppData\Local\Programs\Python\Launcher\;C:...
TEMP	C:\Users\hassa\AppData\Local\Temp
TMP	C:\Users\hassa\AppData\Local\Temp

Variables système

Variable	Valeur
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OnlineServices	Online Services
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\java8path;...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC

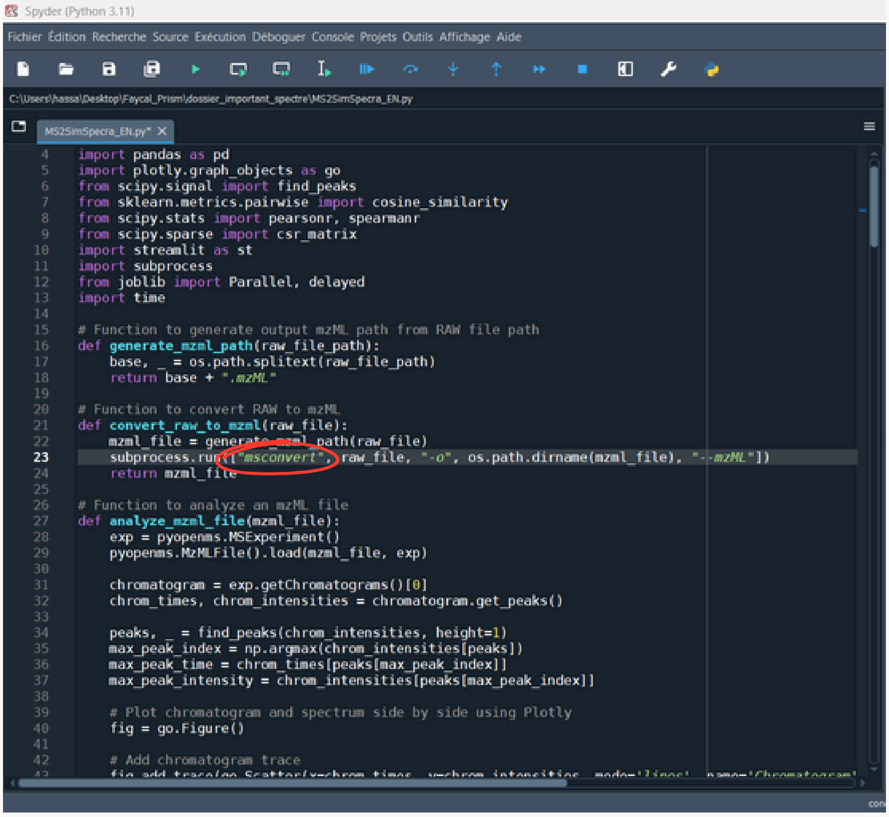
- In the new window, click on "New" and paste the full path of the "ProteoWizard 3.0.24143.8381ed5 64-bit" folder (for example, "C:\Program Files\ProteoWizard 3.0.24143.8381ed5 64-bit").



C:\Users\hassa\AppData\Local\Programs\Python\Launcher\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Program Files\ProteoWizard\ProteoWizard 3.0.24143.8381...

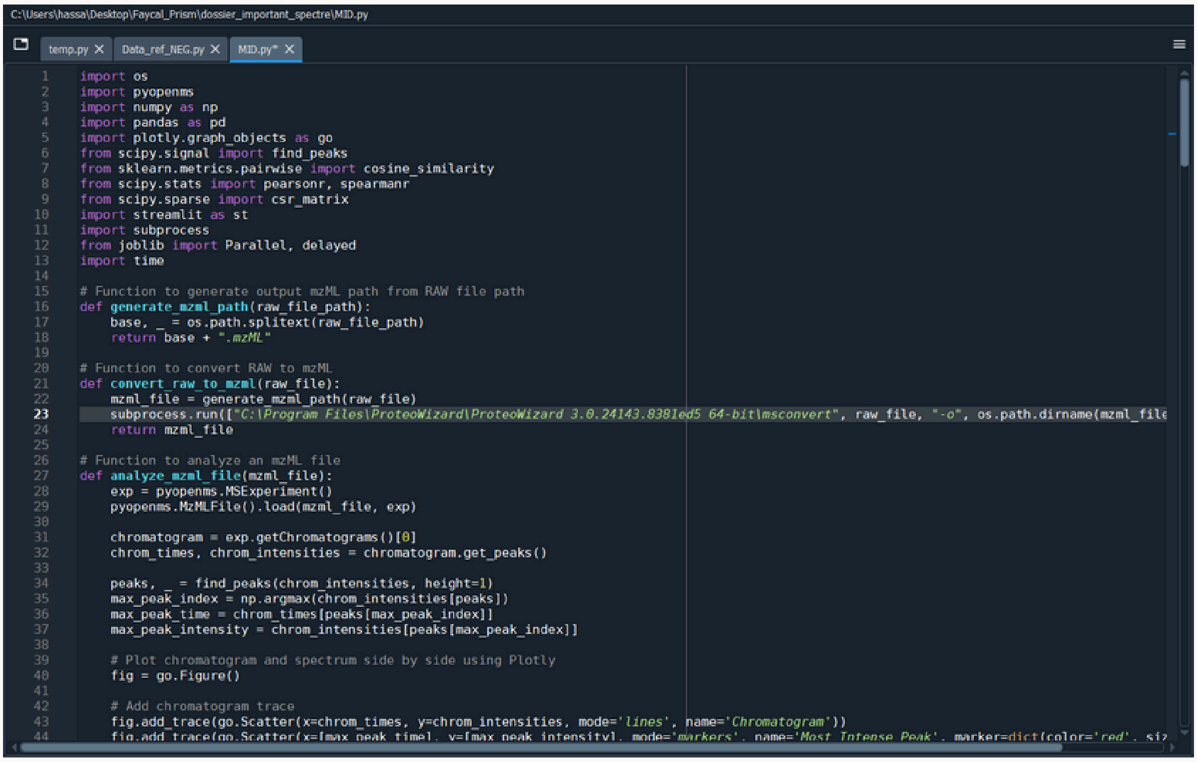
- Confirm the changes and close the configuration windows.

- Or you can go into the script and provide the path to the \msconvert file. (It's up to you to choose which method you prefer, but the first method avoids having to provide the path each time).


Spyder (Python 3.11)
Fichier Édition Recherche Source Exécution Déboguer Console Projets Outils Affichage Aide
C:\Users\hassa\Desktop\Faycal_Prism\dossier_important_spectre\MS2SimSpecra_EN.py
MS2SimSpecra_EN.py*

```python
import pandas as pd
import plotly.graph_objects as go
from scipy.signal import find_peaks
from sklearn.metrics.pairwise import cosine_similarity
from scipy.stats import pearsonr, spearmanr
from scipy.sparse import csr_matrix
import streamlit as st
import subprocess
from joblib import Parallel, delayed
import time

# Function to generate output mzML path from RAW file path
def generate_mzml_path(raw_file_path):
    base, _ = os.path.splitext(raw_file_path)
    return base + ".mzML"

# Function to convert RAW to mzML
def convert_raw_to_mzml(raw_file):
    mzml_file = generate_mzml_path(raw_file)
    subprocess.run(["msconvert", raw_file, "-o", os.path.dirname(mzml_file), "--mzML"])
    return mzml_file

# Function to analyze an mzML file
def analyze_mzml_file(mzml_file):
    exp = pyopenms.MSExperiment()
    pyopenms.MzMLFile().load(mzml_file, exp)

    chromatogram = exp.getChromatograms()[0]
    chrom_times, chrom_intensities = chromatogram.get_peaks()

    peaks, _ = find_peaks(chrom_intensities, height=1)
    max_peak_index = np.argmax(chrom_intensities[peaks])
    max_peak_time = chrom_times[peaks[max_peak_index]]
    max_peak_intensity = chrom_intensities[peaks[max_peak_index]]

    # Plot chromatogram and spectrum side by side using Plotly
    fig = go.Figure()

    # Add chromatogram trace
    fig.add_trace(go.Scatter(x=chrom_times, y=chrom_intensities, mode='lines', name='Chromatogram'
```

- Once these steps are completed, MSconvert will be correctly installed and your software will be operational.


C:\Users\hassa\Desktop\Faycal_Prism\dossier_important_spectre\MID.py
temp.py
Data_ref_NEG.py
MID.py*

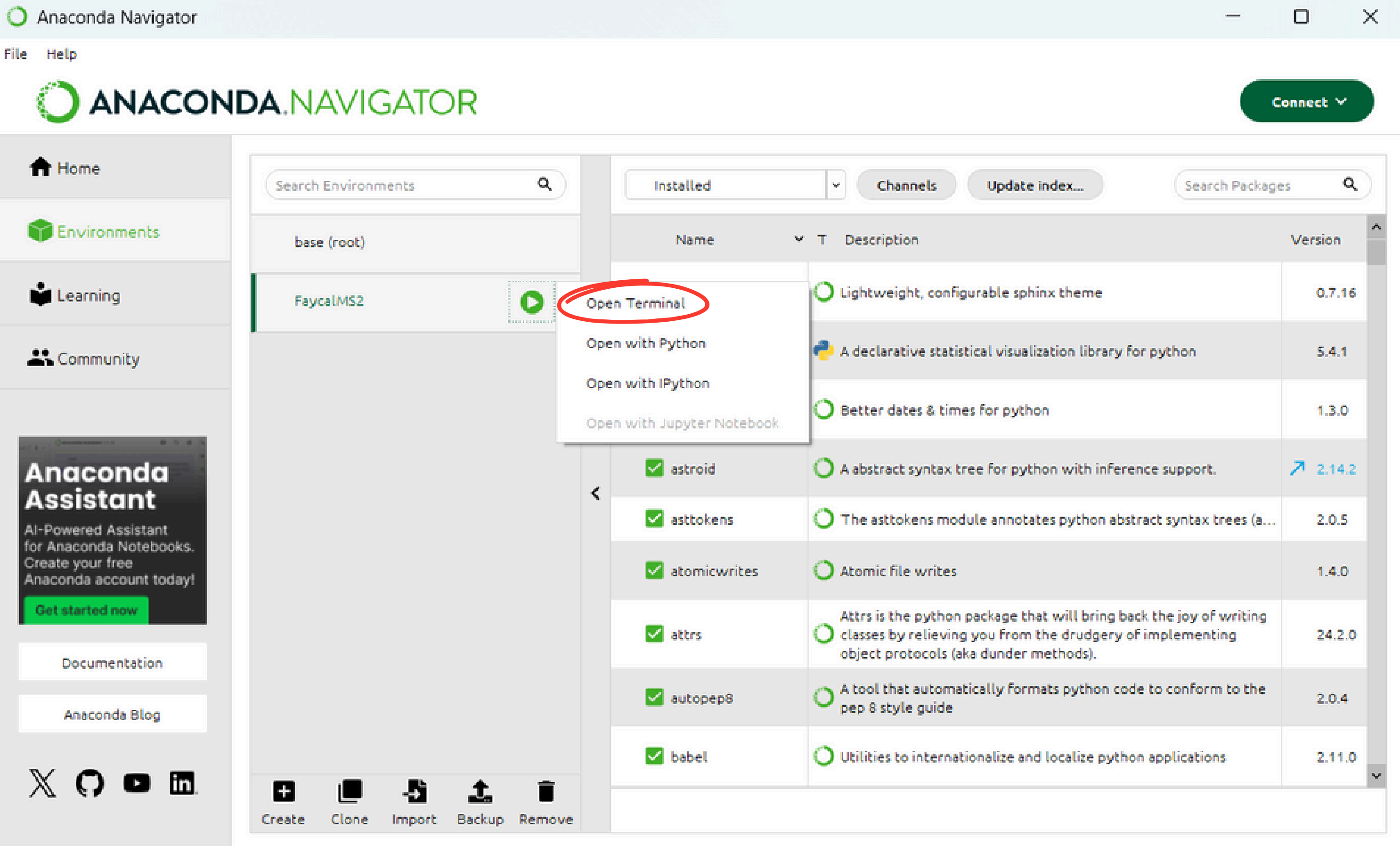
```python
import os
import pyopenms
import numpy as np
import pandas as pd
import plotly.graph_objects as go
from scipy.signal import find_peaks
from sklearn.metrics.pairwise import cosine_similarity
from scipy.stats import pearsonr, spearmanr
from scipy.sparse import csr_matrix
import streamlit as st
import subprocess
from joblib import Parallel, delayed
import time

# Function to generate output mzML path from RAW file path
def generate_mzml_path(raw_file_path):
    base, _ = os.path.splitext(raw_file_path)
    return base + ".mzML"

# Function to convert RAW to mzML
def convert_raw_to_mzml(raw_file):
    mzml_file = generate_mzml_path(raw_file)
    subprocess.run(["C:\Program Files\ProteoWizard\ProteoWizard 3.0.24143.8381ed5 64-bit\msconvert", raw_file, "-o", os.path.dirname(mzml_file
    return mzml_file

# Function to analyze an mzML file
def analyze_mzml_file(mzml_file):
    exp = pyopenms.MSExperiment()
    pyopenms.MzMLFile().load(mzml_file, exp)

    chromatogram = exp.getChromatograms()[0]
    chrom_times, chrom_intensities = chromatogram.get_peaks()

    peaks, _ = find_peaks(chrom_intensities, height=1)
    max_peak_index = np.argmax(chrom_intensities[peaks])
    max_peak_time = chrom_times[peaks[max_peak_index]]
    max_peak_intensity = chrom_intensities[peaks[max_peak_index]]

    # Plot chromatogram and spectrum side by side using Plotly
    fig = go.Figure()

    # Add chromatogram trace
    fig.add_trace(go.Scatter(x=chrom_times, y=chrom_intensities, mode='lines', name='Chromatogram'))
    fig.add_trace(go.Scatter(x=[max_peak_time], y=[max_peak_intensity], mode='markers', name='Most Intense Peak', marker=dict(color='red', siz
```

Software Usage

- To use the software, start by opening Anaconda Navigator.

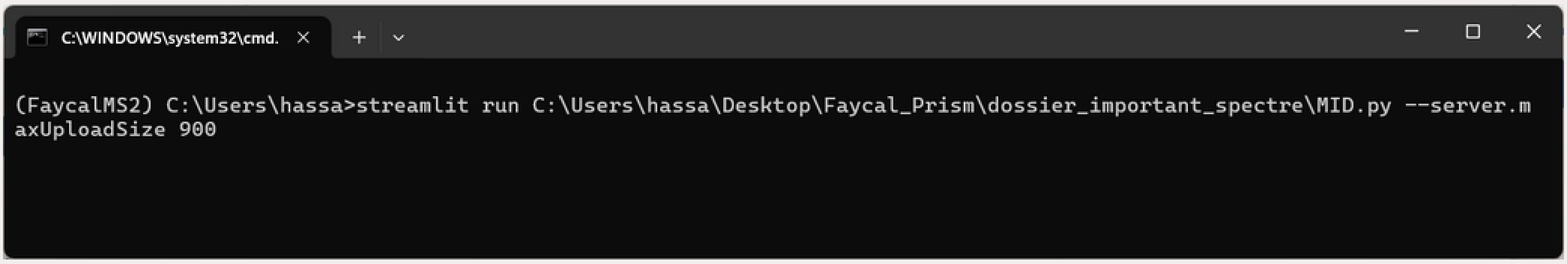
- In Anaconda Navigator, select the environment you previously imported.

- Open the command terminal by clicking on "Open Terminal" in the selected environment.



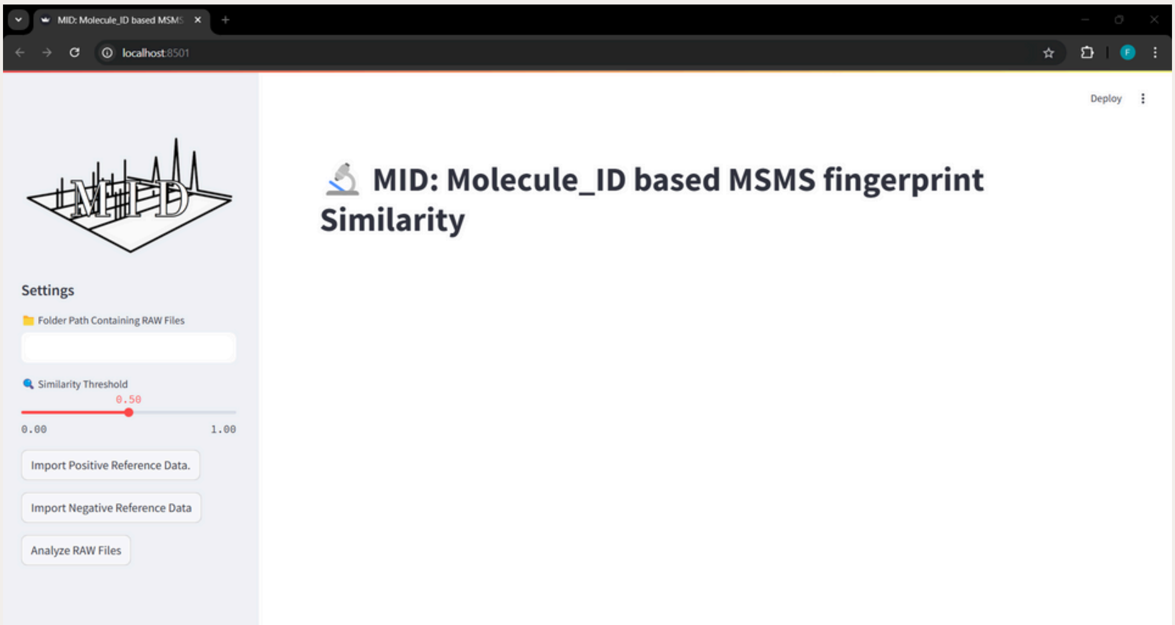
Name	T	Description	Version
	○	Lightweight, configurable sphinx theme	0.7.16
		A declarative statistical visualization library for python	5.4.1
	○	Better dates & times for python	1.3.0
☑ astroid	○	A abstract syntax tree for python with inference support.	↗ 2.14.2
☑ asttokens	○	The asttokens module annotates python abstract syntax trees (a...	2.0.5
☑ atomicwrites	○	Atomic file writes	1.4.0
☑ attrs	○	Attrs is the python package that will bring back the joy of writing classes by relieving you from the drudgery of implementing object protocols (aka dunder methods).	24.2.0
☑ autopep8	○	A tool that automatically formats python code to conform to the pep 8 style guide	2.0.4
☑ babel	○	Utilities to internationalize and localize python applications	2.11.0

- In the terminal, run the following command to launch Streamlit:
- streamlit run path_to_MID.py --server.maxUploadSize 900
- Replace "path_to_MID.py" with the full path to the MID.py file.



```
(FaycalMS2) C:\Users\hassa>streamlit run C:\Users\hassa\Desktop\Faycal_Prism\dossier_important_spectre\MID.py --server.m
axUploadSize 900
```

- Once the command is executed, a new Google page will open with the software interface.


MID: Molecule_ID based MSMS
localhost:8501
Deploy
MID

MID: Molecule_ID based MSMS fingerprint Similarity

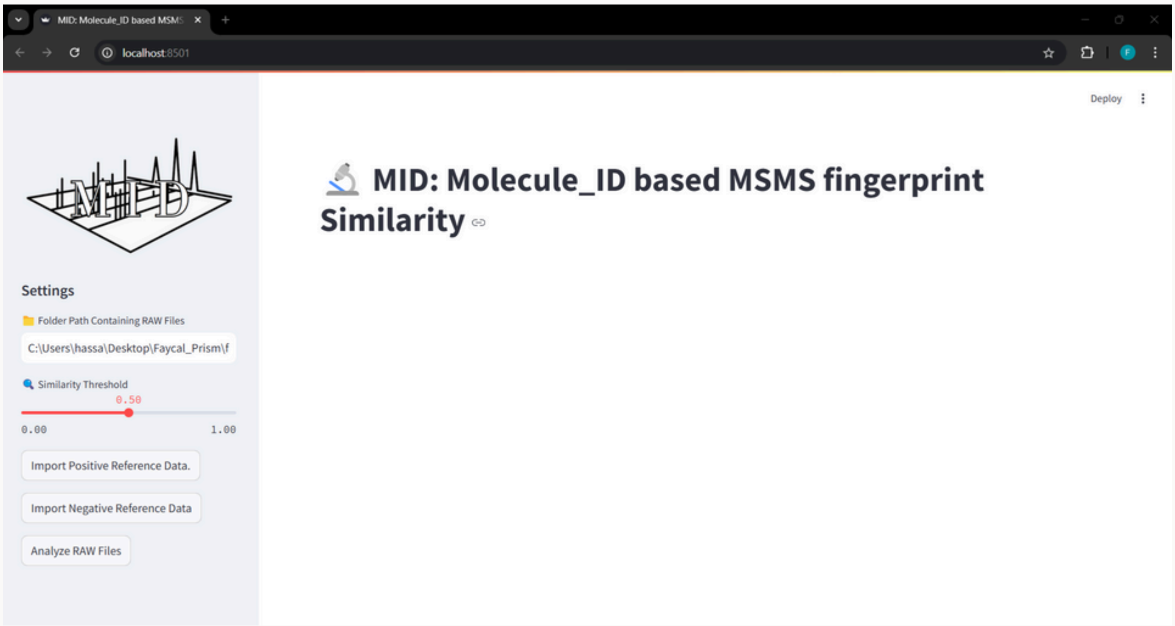
Settings

Folder Path Containing RAW Files

Similarity Threshold 0.50

0.00
1.00
Import Positive Reference Data.
Import Negative Reference Data
Analyze RAW Files

- In the interface, specify the path to the file that contains the raw folder to be imported.


MID: Molecule_ID based MSMS
localhost:8501
Deploy
Import Positive Reference Data.
Import Negative Reference Data
Analyze RAW Files

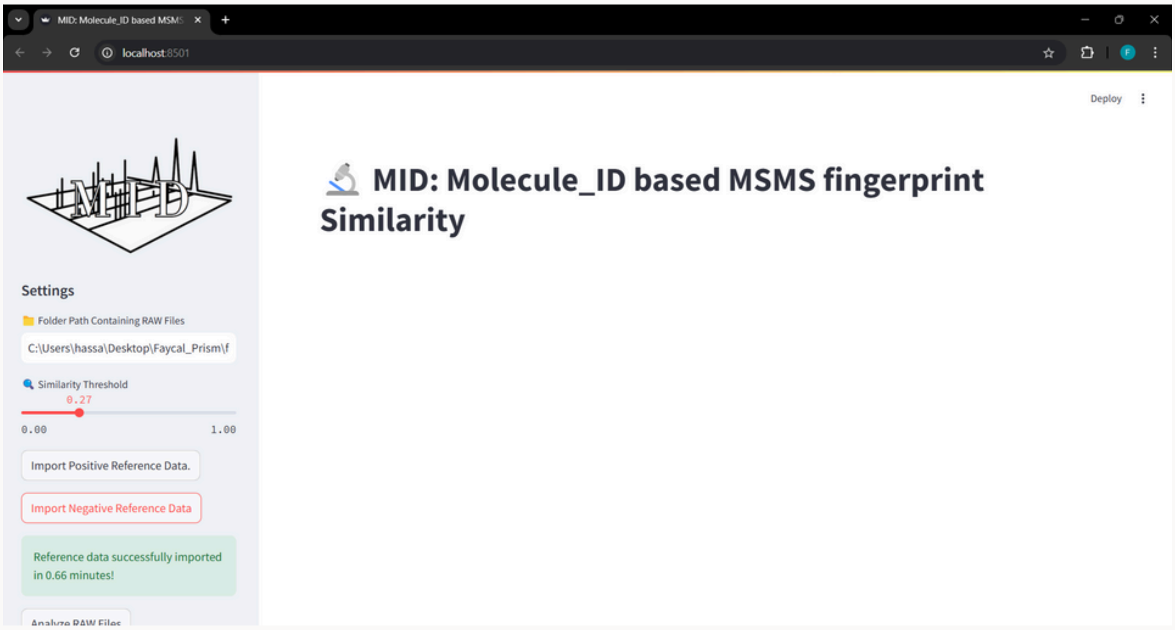
MID: Molecule_ID based MSMS fingerprint Similarity

Settings

Folder Path Containing RAW Files: C:\Users\hassa\Desktop\Faycal_Prism\f

Similarity Threshold: 0.50 (0.00 – 1.00)

- Next, specify the **Similarity Threshold** you wish to use. This threshold should be between 0 and 1.


MID: Molecule_ID based MSMS
localhost:8501
Deploy
Import Positive Reference Data.
Import Negative Reference Data
Analyze RAW Files

MID: Molecule_ID based MSMS fingerprint Similarity

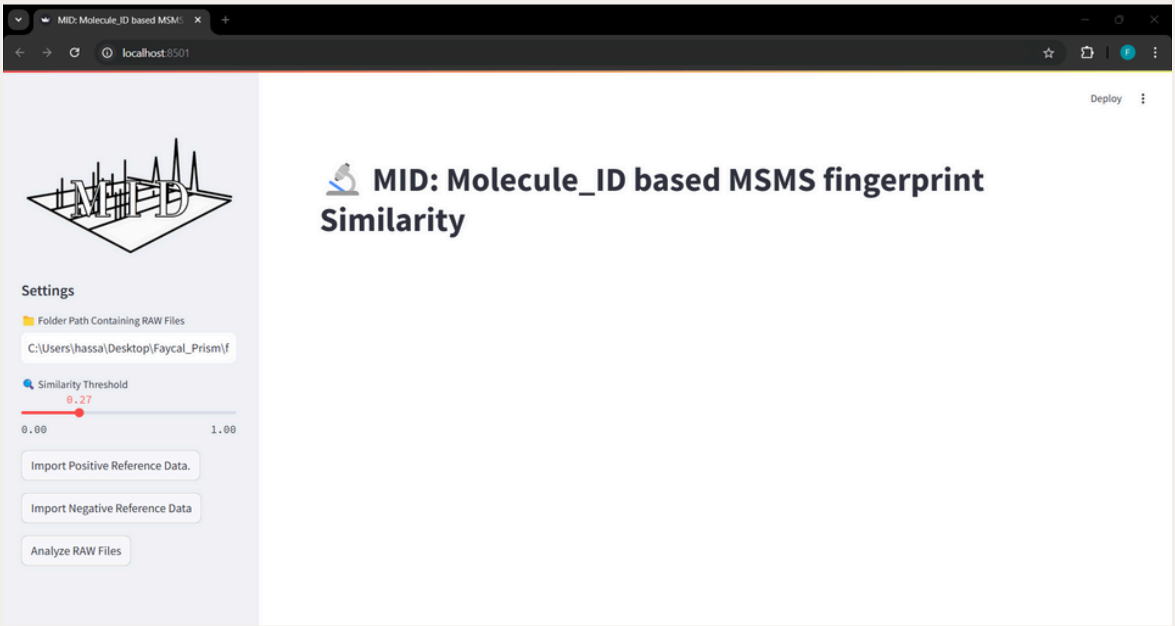
Settings

Folder Path Containing RAW Files: C:\Users\hassa\Desktop\Faycal_Prism\f

Similarity Threshold: 0.27 (0.00 – 1.00)

Reference data successfully imported in 0.66 minutes!

- Import the positive or negative files depending on the file you have imported.


MID: Molecule_ID based MSMS
localhost:8501
Deploy
Import Positive Reference Data.
Import Negative Reference Data
Analyze RAW Files

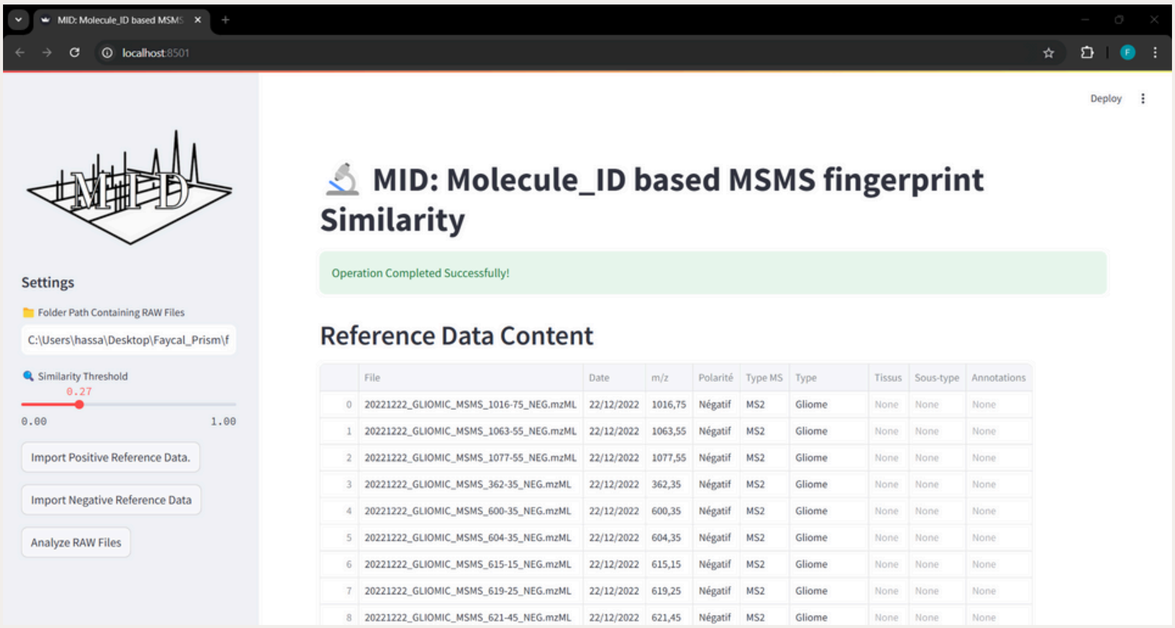
MID: Molecule_ID based MSMS fingerprint Similarity

Settings

Folder Path Containing RAW Files: C:\Users\hassa\Desktop\Faycal_Prism\f

Similarity Threshold: 0.27 (0.00 – 1.00)

- Finally, click on "Analyze Raw files" to start the analysis. The software will then begin processing the files and displaying the results.


MID: Molecule_ID based MSMS
localhost:8501
Deploy
Import Positive Reference Data.
Import Negative Reference Data
Analyze RAW Files

MID: Molecule_ID based MSMS fingerprint Similarity

Settings

Folder Path Containing RAW Files: C:\Users\hassa\Desktop\Faycal_Prism\f

Similarity Threshold: 0.27 (0.00 – 1.00)

Operation Completed Successfully!

Reference Data Content

	File	Date	m/z	Polarité	Type MS	Type	Tissus	Sous-type	Annotations
0	20221222_GLIOMIC_MSMS_1016-75_NEG.mzML	22/12/2022	1016,75	Négatif	MS2	Gliome	None	None	None
1	20221222_GLIOMIC_MSMS_1063-55_NEG.mzML	22/12/2022	1063,55	Négatif	MS2	Gliome	None	None	None
2	20221222_GLIOMIC_MSMS_1077-55_NEG.mzML	22/12/2022	1077,55	Négatif	MS2	Gliome	None	None	None
3	20221222_GLIOMIC_MSMS_362-35_NEG.mzML	22/12/2022	362,35	Négatif	MS2	Gliome	None	None	None
4	20221222_GLIOMIC_MSMS_600-35_NEG.mzML	22/12/2022	600,35	Négatif	MS2	Gliome	None	None	None
5	20221222_GLIOMIC_MSMS_604-35_NEG.mzML	22/12/2022	604,35	Négatif	MS2	Gliome	None	None	None
6	20221222_GLIOMIC_MSMS_615-15_NEG.mzML	22/12/2022	615,15	Négatif	MS2	Gliome	None	None	None
7	20221222_GLIOMIC_MSMS_619-25_NEG.mzML	22/12/2022	619,25	Négatif	MS2	Gliome	None	None	None
8	20221222_GLIOMIC_MSMS_621-45_NEG.mzML	22/12/2022	621,45	Négatif	MS2	Gliome	None	None	None