**Targeted Regression Workflow**

**Installing dependencies:**

Download and install Python 3.11.9 from this link: <https://www.python.org/downloads/windows/>

During install make sure to check: Install python with Admin privileges and Add Python to PATH

Windows users run **install\_requirements.bat** file to create virtual environment to run the program, once created it will not need to be repeated unless venv folder is deleted

Mac users navigate to the working directory in the terminal and run chmod +x install\_requirements\_mac.command. Then run the **install\_requirements\_mac.command** file either from the terminal or the folder itself.

**Program operation**

**MS1 is explained below, MS2 is very similar but the caveats will be explained in a different section.**

1. Fill Out masslist based on the example (**aMasslist\_Height\_Example.xlsx**) for your dataset,
   1. Fill in the Std (Unlabeled) and SIL (Labeled) for all the target isotopologues, for example 13C8 Trp Std would be 205.09 and SIL would be 213.126. The program simply does linear regression between the two target masses so it does not have to be an unlabeled and labeled mass, just keep track of which mass was used in what column. The Std2 and SIL2 are used if there is significant in-source fragmentation, as they will be summed with the other respective masses when the EIC is extracted. Using a mass window and masses that align with your dataset rather than theoretical data will improve the output significantly.
   2. For retention time, Std/SIL RT will be the same unless deuterium is used as a label. This retention time is used to identify the peak maximum for the target masses, so make sure that the RT range contains that maximum but does not contain any other peaks that fall within the mass range.
   3. RTadj is a simple y/n. “y” will try to line up the peak maxima identified by the provided range, this is necessary for deuterated masses, MS/MS, or other applications where there may be peak shift.
   4. Mass window is the window around the target mass where the data will be extracted.
2. Place all MS1 data in the MS1 folder and any MS2 data in the MS2 folder. The data files must be in .mzML format.
3. Click on the **run\_SILExtract.bat** script in the folder (for Windows users. Max users will need to repeat the same process as done above with the install\_requirements\_mac file. From the terminal run chmod +x run\_SILExtract\_mac.command and then run the file from the terminal or folder), then navigate to the correct mass list you filled out, select it and click Run. Also select if your dataset has replicates (i.e. A\_1, A\_2, A\_3, B\_1,B\_2,B\_3), if yes, this will take the mean and standard deviation for each sample group (i.e. A, B). Replicate files must have ”\_#” where # is the replicate number at the end of the file names for this to function.
4. As the data is being analyzed, leave the window open. Once finished, text will pop up indicating how long analysis took.
5. Data will be exported into an excel file in the output folder called **replicatestats.xlsx** if the dataset contained replicates or **RegressionDataOutput.xlsx** if not. This contains all the output data. The plot folders contain the EICs and regression plots for each feature in each datafile. The joinedCSV folder contains the data for the overlapping scans in the “joined” csv and the entire extracted EIC in the “FullEIC” csv.
6. Within the replicate stats file, each compound's regression data will be sorted by datafile. **I highly recommend double-checking the EICs in the raw data to confirm regressions, especially for unfamiliar datasets. With any automatic extraction, artifacts may happen, like overfitting or extracting an unwanted mass.**
7. If the output does not look as expected, changing the mass or RT window may help, it may require a few iterations to nail the extraction.

**MS2 caveats**

1. In the masslist, the SIL and Std are now the fragments you are extracting and doing linear regression on. The precursor masses must also be included, **Make sure you match the exact precursor mass from the raw data.**
2. Make sure to put a “y” in the RTAdj column for any MS2 as they will be collected in separate scans