**Automated Stable Isotope Tracer Program**

Developed by Evan Larson, Clinton Curry, Cohen/Hegeman Labs, Department of Horticultural Science, University of Minnesota

**Quick Start:**

- Install Python 3.11

- Install Git

- Run the install\_requirements file to install needed dependencies

- Move mzML data files into the Data/RawData folder

- Run the run\_workflow file

- Select Single Run in the first dropdown and fill in the other relevant parameters

- Press run and track progress on the open terminal until finished

- Open the output folder to see all analyzed data!

**Description:**

Data processing workflow that takes raw liquid chromatography mass spectrometry data and searches scan by scan for any mass pairs that differ by a multiple of the given isotope label mass, for example two pairs of masses that differ by 2.0067 are likely an unlabeled compound and its 13C2 labeled isotopologue. As each scan in the full chromatogram is analyzed, a full list of all possible pairs of unlabeled labeled features above a minimum intensity is compiled. This list is then trimmed down based on: the number of scans in which the feature pairs are present, a mass bin which groups masses within the same range, and a retention time bin which groups retention time within the same range. As this is looking for labeled pairs within a single data file, the data must contain both labeled and unlabeled peaks. If you have separate runs for each, consider rerunning a pool of the labeled and unlabeled sample or using mzmine to extract features from the labeled and unlabeled datasets (see below).

Once this initial feature list is filtered, an extracted ion chromatogram is extracted for each feature pair from each datafile provided. This essentially generates an unlabeled and labeled peak chromatogram. Linear regression is then used to plot the intensity of the unlabeled vs the labeled which should result in a high R2 score if the two peaks are well aligned, as should be the case for 13C and 15N feature pairs. R2 can then be used as an subsequent filtering step as features with high R2 are more likely to be isotopically related pairs. Additionally, the linear regression slope provides the ratio between the intensities for the labeled and unlabeled peaks. , The width of the scan window is optimized using the R2 and 20% R2 bootstrap (described below). Finally, an ideal Gaussian distribution is generated for the labeled and unlabeled based on the RT range and peak maximum and the Pearson correlation of unlabeled and labeled EICs are taken to score the peak shape. This regression data is averaged across all datafiles for each feature and then output with some additional statistics.

Once features are filtered by a given R2, the unlabeled masses are searched against a provided database, which is based on a snapshot of the Pubchem database (Feb 13th, 2024, filtered for CHNOPSCl) (more databases may be added in the future). This provides any formulas which match within a ppm error range while also considering the number of isotopes present in the labeled isotopologue(for example if two pairs differ by 5 13C1, the formula must have at least 5 carbons). It then calculates ppm error for the unlabeled and labeled masses based on the known labels. Additionally, a list of formulas and target compound names can be provided to match up any formulas that are targeted. This does not confirm the identification as the ID is based only on accurate mass, additional orthogonal confirmation based on MS/MS or RT matching with a standard is needed for confirmation.

Finally, feature clustering is done to identify peaks from in-source fragmentation or multiply counted isotopologues. To accomplish this, the labeled and unlabeled EICs from each replicate are summed to a single EIC and a correlation matrix between each feature is done using Pearson correlation. Beginning with the highest intensity feature in the RTbin, features are grouped if they have > 0.6 correlation with the highest intensity feature. All features grouped this way are removed from the pool and the process is repeated until no features remain. These groupings are included in the final output.

**Installing Dependencies**

**If you do not have python installed:**

Download python 3.11 from the link below (scroll down for release 3.11).

**Download python -** <https://www.python.org/downloads/>

**If you already have python installed:**

This workflow requires python 3.11 to run properly. If you have another version installed, either uninstall that version and download 3.11 to keep things simple, or download 3.11 alongside your current python version, just be aware of potential issues. The workflow is designed to work as long as 3.11 is installed and on PATH.

Next make sure Git is installed. Go to <https://git-scm.com/downloads> and pick your specific operating system installer and follow the instructions, making sure that when installed it is added to your PATH.

Now windows users can double-click the install\_requirements.bat file which will create a virtual environment in the Stable Isotope Tracer folder called “venv” that contains all needed dependencies for the workflow and will be used to run the workflow from this environment. This only needs to be done once, after the venv folder is created it will remain and can be used to run the workflow multiple times.

Mac users will need to either manually run the install\_requirements\_mac.command file by opening a terminal and inputting ./install\_requirements\_mac.command

*Or*

Navigating to the directory of the file by typing cd “path/to/Stable Isotope Tracer” and then inputting chmod +x install\_requirements\_mac.command which will make the file “double-clickable” from that point forward.

**Running the Program**

When running the program, make sure to leave the file/folder structure of the StableIsotopeTracer folder! Any alteration of the Database, Scripts, Data, and Sequenced folders will break the program, make sure to just copy/paste datafiles/ouptut files from the folders.

It is recommended to run the program through the command prompt as it will leave any errors in the terminal if anything goes wrong. This can be done by opening a command prompt/terminal, typing cd “path/to/Stable Isotope Tracer”, and entering run\_workflow.bat for windows users. For mac users the same process as before with the install\_requirements file will need to be completed. In the terminal either use ./run\_workflow\_mac.command to run the file manually or chmod +x run\_workflow\_mac.command to make the file run by double-clicking the file.

All datafiles must be converted from a proprietary data format (.Raw, .D, etc.) into .mzML, and centroided The msconvert software from the proteowizard package is suggested for conversion, add the peak picking (vendor) and remove the title maker filter for centroided and best results.

There are two main methods of running the program: Single Dataset and Sequenced.

**Single Dataset**

This method should be used for a single set of files with the same conditions (i.e. labeling, experimental conditions) with either analytical or biological replicates although the program will also run on a single datafile.

Copy all mzML files for this dataset into the folder StableIsotopeTracer/Data/RawFiles

Run the run\_workflow file as described above

Running the program will open the GUI, make sure to select Single Dataset from the "Single Dataset or Sequenced" dropdown

Enter all the necessary run parameters in each of the boxes/dropdowns (described in detail below)

Once all parameters are entered, you can save the config file using the save config button. Additionally, you can load this same file for future runs with the same parameters

Press the run button and follow instructions and the program will start analysis.

Keep an eye on the terminal that opens with the GUI for updates on the program.

Once finished all the output datafiles will be exported to the output folder (described in detail below)

**Sequenced**

This method should be used for running multiple datasets (i.e. labeling, experimental conditions) which are separated into different folders. It works by running the Single Dataset program on individual folders.

Folder/file structure is very important to running a sequenced dataset. An example of the file structure is provided within the Sequenced Example folder in the program folder

Within the sequence folder, there should be a folder for each of the datasets you plan to run which is name accordingly. Within that folder, there should be a folder named Data which contains a folder named Raw Files. All datafiles for that dataset should be placed in the Raw Files folder.

Instead of using the main GUI for the run parameters the parameters are loaded from a csv so within the main folder for each dataset there should be a config file named "Config\_(the name of that folder exactly).csv" which contains all the parameters for running the program. I would suggest using the main GUI to create this config file by using the save config function so that the exact syntax is correct. For the program to work, the first column of the config file must not be edited from the saved version. This also means that the GUI should not be.

You can also mix generate feature lists, using premade feature lists and using MZmine feature lists but relevant files must be placed in the right directories (see below).

Once all folders and datasets are arranged, open the program and select Sequenced from the "Single Dataset or Sequenced" dropdown.

Press the run button, follow the prompts and the program analysis will start.

Keep an eye on the terminal that opens with the GUI for updates on the program.

Once finished all the output datafiles will be exported to the output folder (described in detail below)

**Program Parameters/GUI Buttons**

There are various settings and parameters that can be changed based on the datasets which are all described below:

**General Program Settings**

**Single Dataset or Sequenced** - Described above but chooses the single run or sequenced run type

**Select Program Function** - Chooses the way that the program operates on each run.

**Options:**

**Generate New Feature List** - Works from the start to identify any feature label pairs, performs regression and identifies formula matches

**Work From Premade Feature List** - Works from a feature list already generated from the program. This will be called GeneratedFeatureList.csv in the output folder after the program run. This can be used if you want to search for the same sets of features in a different set of data or if you want to add a new set of features to search for. Make sure to edit a premade list from the program so that the syntax is exactly what the program expects, any changes to the header will break the program. Also all columns must be filled.

**Select Isotope Label** - Select which isotope label to search for within the data. A user chosen mass can be used and the program will search for multiples of that value.

**Feature Lists**

**Premade Feature List File** - If working from the premade feature list selected above, submit the filename for the feature list you want to use. Make sure to place it in the main Stable Isotope Tracer folder (if running Single Dataset) or in each individual dataset folder with the correct name saved in each config file, (if running sequenced mode)

**Feature Detection Settings**

**Minimum Number of Labels** - Sets the minimum number of labels you would like to search for within your dataset.

**Maximum Number of Labels** - Sets the maximum number of labels you would like to search for within your dataset. If the experiment is fully labeled, consider these settings the maximum and minimum number of the labeled atom to look for.

**Mass Window for Label** - Sets the range for finding multiples of feature (i.e. for 13C m/z labeled = m/z unlabeled + n \* (1.003355 +/- this value).For a QExactive Orbitrap, we typically use +/- 0.0002 but it would be better to find the ideal range empirically based on your instrument which can be done by observing the natural C13 - C12 m/zs for a known compound within individual scans over a peak (or preferably a few peaks of varying intensity).

**Retention Time Bin Size/Mass Bin Size** - Sets the "bin" for retention time and mass, only the most abundant feature within the same mass and retention time window will be kept. Smaller windows will result in more features. For our HPLC-Orbitrap system, we typically use 0.5 for RT bin and 0.002 for m/z bin

**Minimum Intensity for Features** - Sets the minimum raw intensity for a feature to be kept

**Minimum Binned Scans for Features** - Sets the minimum number of scans a feature must be in to be kept

**Regression Settings**

**Retention Time Window to Search for Peak Height** - Sets the +/- RT window to search for the peak maximum for a particular feature, this is used to align the peaks and can typically be fairly small (0.1 or smaller)

**Mass Window for Regression** - Sets the mass window to extract for a particular peak from the raw data, should encompass the expected mass error for your instrument. For our HPLC-Orbitrap system, we typically use 0.002. All peaks identified within this range are summed.

**Minimum R2 to Keep Feature** - Sets the minimum R2 to keep a feature after linear regression, if you want to include all features use a very small number like 0.0000001 rather than 0. If this value is set to 0, features on which regression was not performed (due to limited number of datapoints or other issues) will be kept.

**Remove Plots if R2 is below threshold** - Keep or Remove, if "remove" all plots made for the features that fall below the minimum R2 are deleted otherwise they are kept.

**Database Search Settings**

**PPM Error for Formula Matching** - Sets +/- PPM window to search for database matches, only potential formulas within the given range will be considered

**Fully Labeled or Partially Labeled Experiment** - Fully or Partially, this is chosen based on the type of labeling experiment that was performed. If all atoms are expected to be labeled (i.e. samples grown on 100% of a label) choose Fully. If a tracer is used (i.e. samples grown with a metabolite containing a set number of labels) choose Partially. This determines how the number of labels identified is used, if "Fully" is chosen the number of labels is considered the number of that respective atom in potential formulas, if "Partially" is chosen the number of labels identified is considered the minimum number of that respective atom in potential formulas.

**Use Target Formula List** - Yes or No, choose whether to match up a target list of formulas/names with the data output

**Formula List File** - Enter the full name of the Target Formula List file, should be an excel file so enter "filename".xlsx. The file should be placed within the main Stable Isotope Tracer folder of the program (for either single or sequenced) and must contain a column titled "Name" and a column titled "Formula". Make sure that your formulas follow typical IUPAC rules (i.e. CHNOPS) as it matches the formula exactly. This will match any output features with that formula with the given name. Consider that matches are based solely on formulas so the first isomer on a list will be picked.

**Prefer CHNO in Main Output** - Yes or No, choose whether to only put CHNO containing formulas in the main output (see below). Personal preference based on the target metabolic pathway for which this program was developed only contained these atoms, useful for clearing up the data output.

**Buttons**

**Run Button** - Will open the dialogue to start the run

**Save Config** - Will save currently entered parameters into a csv file to keep config settings

**Load Config** - Will load parameters from a previously saved config csv file, make sure matches the syntax of a saved config file or issues will arise

**Exit** - Quits the Program (redundant with exit button on the window)

**Data Output**

As the program runs, an Output folder will be generated either in the main folder (single dataset) or the specific dataset folder (sequenced) which will collect all data outputs.

A key concept for the outputs data is that each feature is given a feature name, UnlabeledBinMass\_UnlabeledBinRT\_LabeledBinMass\_LabeledBinRT, this is used to differentiate between features and unifies the formula, plot and csv outputs. The formula output is named by just UnlabeledBinMass\_LabeledBinMass as isomers at different retention times will have the same tentative formulas

There are 3 folders contained within the Output folder after the program finishes running as well as the output excel files which will be explained below:

**Folders**

**Peaks –** This folder contains the subsequent folders as well as the excel files for each peak rt bin. Within this peak RT bin excel are the clustered features based on pearson correlation between all compounds in that bin. These peak clusters are assigned a code, RT\_PeakGroupNumber. For example, all peaks within the RT bin at 2.0 minutes within the 3rd group would have a code, 2\_3. All other exported information will match to the information in the other outputs.

**Formulas** - This folder contains a separate excel file for each feature. Within each excel file are the databases matches for this feature sorted by the absolute value of the ppm. The top match is put into the main output unless Prefer CHNO is selected (see above). These files are also linked from the final output.

**OtherOutputFiles** – This folder contains all the intermediate output files and can be used to check for errors or artifacts. See below for an explanation of each file’s structure and contents.

**Plots** - This folder contains a separate folder for each feature. Within the folder for each feature, plots of the data are exported from each datafile for that feature.

Half will be plotting the labeled and unlabeled extracted ion chromatograms which will be named Peak(Filename) (Feature), the blue line will be unlabeled peak (intensity on the left axis) and the red line will be the labeled peak (intensity on the right axis).

Half will be plots of the regression of the labeled intensity vs. the unlabeled intensity including the slope and R2 for that feature. These files are also linked from the final output.

**JoinedCSV** - This folder contains the extracted data for each feature extracted from each datafile. The name is featurename\_(Filename)joined.csv. Within is contained the data used for plotting the regression and the extracted ion chromatograms

**Intermediate Output Files**

**AllExtractedFeatures\_File.**csv - Contains all the features before binning/size reduction extracted from each file

**GeneratedFeatureList.csv** - Contains all the features after binning/size reduction compiled over all files within the dataset. This is also the file that can be edited or used again to extract these features, see above in the premade feature list section.

**FullRegressionDataOutput** - Contains all the regression data for each feature in each file

**MeanStandardDevOutput** - Average and Standard Deviation of all the regression data for all of the features provided across each file

**TrimmedFeatureList** - A trimmed version of the feature list used for regression which includes all of the features which have regression above the minimum R2 and the regression data for each

**PeakInfo\_RegressResults** - Contains all the information from the trimmedfeaturelist file but also includes the information on formula matching from the database

**CleanedDataLabelingCalc** - The cleaning step applied here drops any features that have duplicate mass features (binned labeled and unlabeled bin mass) and the same height ratio which could happen if the same mass pair is picked up with RT binning. Will rarely happen with larger RT bins. Also applies links to plots/formula files

**AnnotatedCleanedDataLabelingCalc** - This step applies any annotation from the provided feature lists which is matched based solely on formula. Only present if annotation is used.

**Final Output Files**

There are two versions of the final output files, and they will be described below, and all data columns will also be explained.

**FinalOutput** – This output contains all the features above the user-provided scoring threshold and all associated information.

**FinalOutput\_PeakReduced** – This output contains only the highest scoring and highest intensity in each RT bin, more useful if small RT bins are used.

**Output File Column Explanations**

**Unlabeled\_Labeled Features** - Provides the feature key for this particular feature as described above, used to define all data, plot and formula outputs

**y-intercept** - Average of the y-intercept for the linear regression of this feature pair across all provided data files

**slope** - Average of the slope for the linear regression of this feature pair across all provided data files

**R2** - Average of the R2 for the linear regression of this feature pair across all provided data files

**UnlabeledHeight** - Average of the unlabeled mass intensity at the peak height for this feature pair across all provided data files.

**LabeledHeight** - Average of the labeled mass intensity at the peak height for this feature pair across all provided data files.

**Height Ratio** - Average of the labeled mass intensity/unlabeled mass intensity at the peak height for this feature pair across all provided data files.

**20% R2** - Average of the R2 for the linear regression removing 20% of the highest intensity points over the peak, useful to double check how good the regression is. If the peaks overlap well, removing a percentage of the peaks will not greatly affect the R2

**RegressPoints** - Average number of points used in the linear regression for this feature

**OptimizedScanWidth** - Number of scans on each side of the peak maximum which is optimized using the scan width that provides the highest average R2 and 20% R2. That scan width is then used for the linear regression

**SlopeDerivedLabelIncorporation** – Label incorporation calculated by the slope of the labeled intensities vs the sum of the labeled and unlabeled intensities, this calculation provides a measurement of the percent labeling of the total pool of that metabolite

**20%R2\_R2 Average** – Average of the R2 and 20% R2 scores for this feature

**12CalPearR2, 13CalPearR2, SimPeakAverage** – Scores of the correlation between the simulated gaussian distribution vs the real peak for labeled and unlabeled feature, provides a metric to measure the peak shape. SimPeakAverage is the average of these two metrics

**CompleteScore** – Average of the SimPeakAverage, R2 and 20% R2 to give an overall score for the feature

**m/zUnlabeled, rtUnlabeled** - mass and retention time for the unlabeled feature pulled from a single file during the feature finding

**m/zLabeled, rtLabeled** – mass and retention time for the labeled feature pulled from a single file during the feature finding

**Labeled/UnlabeledHeight** - Height ratio pulled from a single file during the feature finding (this is the output from the feature finding section, so intensity/ratio are only a single file, use the regression version for better results which is above)

**pol** - Polarity in which the feature was identified

**NumLabels** - Number of labeled atoms present in the labeled mass rounded to the nearest integer (i.e. m/z13 = m/z12 + n \* 1.003355, where n = number of labels)

**exact** - exact value of n from previous value

**m/zUnlabeledround, m/zLabeledround, rtUnlabeledround, rtLabeledround** - Binned mass and retention time based on the provided bin size

**spectralcount** - number of spectra within the bin in which the feature is found for that particular file

**feature, massfeature** - feature designation based on binned mass/rt as described above, massfeature includes only the binned masses

**filecount** - number of files in which the feature is found, only keeps features present in n-1 files

**Formula** - Top formula match pulled from the database

**Adduct** - Adduct for the top formula match

**ppm** - ppm error for the top formula match for the unlabeled mass

**Nat13C** - calculated natural abundance of 13C, only relevant if a single label and 13C

**Labeledppm** - ppm error for the top formula match for the labeled mass

**MassUnlabeled** - theoretical mass for the unlabeled compound

**Name** (only present in the annotated data output) - Name of any compounds matched from the provided target list

**Plot Hyperlink, Formula Hyperlink** - Hyperlink to the plots folder or the formula list file. The data output file must be in the folder that contains the plot/formula folders or else the hyperlinks will not function. If you get an error that the file is not found, it is likely due files/folders being moved

**PeakCorrelationScore** – Pearson correlation score for the correlation between the feature and the most abundant feature within the correlated peak group

**HighestIntensityFeature** – Feature within the correlated peak group to which all other features are correlated

**PeakFeatureGroup** – Identifier for the correlated feature grouping within the associated retention time bin. The number given signifies the grouping of compounds correlated with each other above a score of 0.6. The format is RTBin\_FeatureGroup

**FeaturesInPeak** – Number of features present in that PeakFeatureGroup

**Peak Hyperlink** – Hyperlink to output containing all features found within an RTbin, includes all peak groupings and scores

**Optional SIRIUS Workflow Extension**

As an additional extension to this workflow, an option has been provided to run a subsequent analysis step of the provided .mzML data through the SIRIUS software API to get additional structural information from MS/MS fragmentation data. This option requires having SIRIUS 6.0.7 installed and added to your PATH as well as an internet connection and an authenticated SIRIUS account for login. Provided from this extension will be an additional matched\_output.xlsx file that will aggregate data from both datasets (sheet 1 & 2) well as show matches between the regression analysis output and the SIRIUS analysis output (sheet 3) as determined by mass ppm error and retention time difference thresholds. The matched\_output.xlsx file will provide structural information for the found matches, achieving an additional layer of annotation and insight. Also provided is the .sirius output file which will contain the raw SIRIUS analysis results and may be used with the SIRIUS GUI for further inspection.

\*Note: This extension only works with Single Dataset mode. It will detect multiple .mzML files for alignment and analysis through SIRIUS, but does not have the sequenced-mode capabilities of the regression workflow to run multiple experiments within separate folders sequentially.

**SIRIUS 6.0.7 Download link** – <https://github.com/sirius-ms/sirius/releases> (scroll down for 6.0.7)

**Directions:**

1. Make sure the proper SIRIUS version is installed and added to PATH and you have an authenticated user account.
2. Copy/Move the MS/MS .mzML data files into the Data/RawData\_SIRIUS folder.
3. Enter parameter info into SIRIUS\_config.csv.
4. Make sure you are completely logged out of SIRIUS and it is shutdown (open the GUI and log out, then shutdown the program. This is to clear old port files).
5. For Windows users: double-click the run\_extension.bat file.

For Mac/Linux users: as done above with the other .command files, this will need to be made executable in order to allow it to run by double-clicking the file. First open the terminal and either enter ./run\_extension\_mac.command to run the file manually or navigate to the Stable Isotope Tracer folder, then enter the command chmod +x run\_extension\_mac.command. Once this is done once the file can then be double-clicked to run as desired. Double-click the file to run after adding the execution permission.

1. Look for results in the Output folder.

**SIRIUS Config File Parameters Explained:**

**Login credentials –** These are needed to run the structural prediction modules from SIRIUS. This information will not be stored or used in any other way except for API authentication with SIRIUS.

* **username** – your email used for your SIRIUS account.
* **password** – password used for your SIRIUS account.

**adduct\_choices** – A list of the feature adducts found in the SIRIUS GUI is supplied at the bottom of the csv file as well as directions for custom adduct input. Choose desired adducts and add to “Adducts” line separated by a comma and a space.

**zodiac\_option** – Runs the optional Zodiac module which builds a similarity network to improve formula prediction ranking. Will increase run-time and should only be used with full-run biological samples as enough datapoints are needed to achieve precision.

**project\_name** – Name identifier used for your .sirius project file.

**ppm\_error** – Set threshold for desired mass PPM error for matching algorithm between regression and SIRIUS data. This will be the maximum calculated ppm difference allowed between the m/z of the unlabeled regression features and SIRIUS features to qualify for a match.

**mass\_accuracy** – Threshold used for mass accuracy (ppm) used when running the SIRIUS analysis.

**Additional Matched Output Page Explained:**

In the matched\_output.xlsx file, the first two sheets correspond to the input datasets(Regression & SIRIUS) while the third sheet will include a selection of columns from the regression analysis data with the new information from SIRIUS which are explained below.

Important to note is the SIRIUS output is organized by structure. It will provide the top 5 structure candidates that include one of the user-supplied adduct choices for each matched feature, this is why you will see multiples of the same matches (they will have different structure ranks).

**Columns included from Regression Output File:**

Unlabeled\_Labeled Features, CompleteScore, mzUnlabeled, mzLabeled, rtUnlabeled, NumLabels, Tentative Matches, Tentative Matches PPM Error, Plot Hyperlink, Formula Hyperlink, Peak Hyperlink

**Additional Columns from SIRIUS output:**

**SIRIUS\_Feature ID** - Unique identifier assigned to each feature detected by SIRIUS. Can be used with the SIRIUS GUI.

**SIRIUS\_Feature Mass** - The mass-to-charge ratio (m/z) of the detected feature.

**SIRIUS\_Retention Time** - The retention time of the feature in minutes.

**SIRIUS\_Quality** – Overall quality assessment of the feature spectra.

**SIRIUS\_Structure Rank** - The ranking of the predicted structure based on confidence scores (1 = highest confidence).

**SIRIUS\_SMILES** - The SMILES for the chemical structure.

**SIRIUS\_CSI Score** - The confidence score from the CSI:FingerID module indicating how well the structure matches the spectral data.

**SIRIUS\_Structure Adduct** - The adduct form detected for the structure.

**SIRIUS\_Molecular Formula** - The predicted molecular formula from SIRIUS for the feature.

**SIRIUS\_SIRIUS Score** - The score assigned by the SIRIUS module indicating confidence in the molecular formula prediction.

**SIRIUS\_Zodiac Score** - The score from the ZODIAC module assessing the likelihood of the predicted molecular formula.

**SIRIUS\_Structure DB Info** - Database info (e.g., PubChem) for additional information about the predicted structure.

**Mass Error (ppm)** – Calculated mass ppm error between the regression analysis feature and SIRIUS feature.

**RT Difference** – Calculated retention time difference between the regression analysis feature and SIRIUS feature.