GC-MS Data Processing with PipMet

Authors: Tatiani Brenelli Lima, Anna Clara Freitas Couto and Juliana Aricetti

The PipMet is a package with functions wrappers for xcms and CAMERA GC-MS data processing workflow with additional packages to generate automatic images for data and algorithm evaluation, as well as the evaluation of results. It can work entirely based on pop-ups windows after initiating the main function workData() if the information is not previously provided by arguments. It also provides a quantification method which relies on a representative ion (most intense ion) for each spectrum proposed, and it is confirmed by the presence of the second most intense ion. In the case of derivatization, the software can remove ions introduced by the derivatization reactions from the quantification process, such as *m/z* 73 for the trimethylsilyl groups.

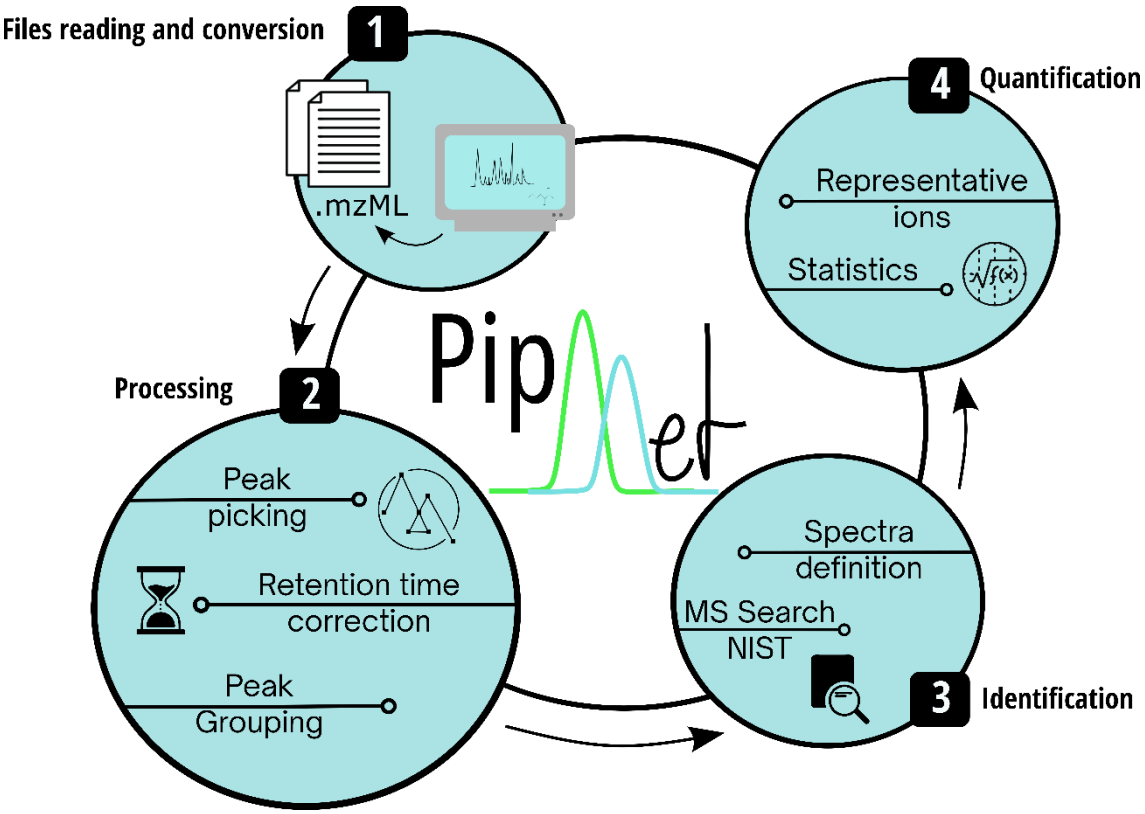


Figure 1. Overview of PipMet workflow.

This vignette describes the installation process and usage of the global function workData() from the PipMet R package to process GC-MS acquired data. The workData() is a wrapper for files readings, preprocessing raw data into viable spectra and a quantification table and normalization. The user may provide the information required through parameters to the workData() function or through pop-up windows.

For PipMet installation, follow the following step:

1.1. Prerequisite Environment Install the following software in this exact order:

- **R (v4.2 or higher):** https://cran.r-project.org/
- **RStudio:** https://posit.co/download/rstudio-desktop/
- **Rtools 4.5 (Critical for Windows users):** Download here. *Without Rtools, PipMet cannot be compiled.*

1.2. Install Dependencies Open RStudio and paste the following code into the **Console**. This script is optimized for R 4.5 and will automatically align with the correct Bioconductor version (3.22).

```
# 1. Install BiocManager and align with Bioconductor 3.22 (Required for R 4.5)
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(version = "3.22", ask = FALSE, update = TRUE)

# 2. Install Devtools
if (!requireNamespace("devtools", quietly = TRUE)) install.packages("devtools")

# 3. Install Bioconductor specific dependencies
bioc_pkgs <- c("xcms", "MSnbase", "CluMSID", "metaMS", "BiocParallel",
        "Biobase", "ProtGenerics", "CAMERA", "NormalyzerDE")
BiocManager::install(bioc_pkgs, ask = FALSE, update = TRUE)

# 4. Install CRAN dependencies
cran_pkgs <- c("svDialogs", "pheatmap", "ddpcr", "webchem", "fritools", "pracma")
new_pkgs <- cran_pkgs[!(cran_pkgs %in% installed.packages()[,"Package"])]
if(length(new_pkgs)) install.packages(new_pkgs)
# Install devtools
devtools::install_github("AnnafCouto/PipMet", force = TRUE)
```

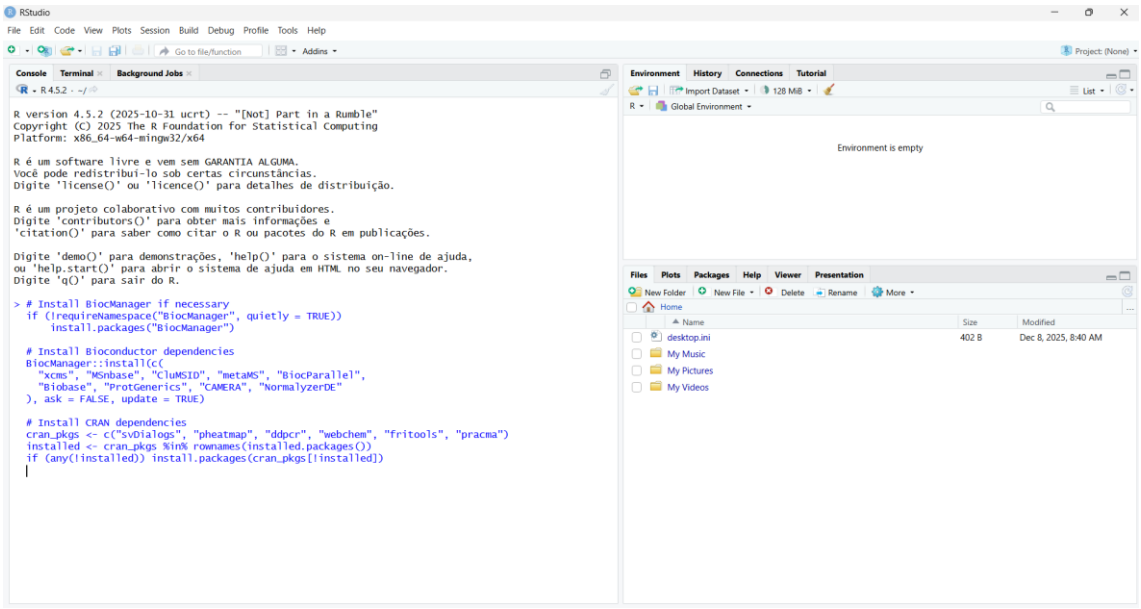


Figure 2. Example of how to copy and paste for dependencies installation.

2. The Quick Start (Full Workflow Walkthrough)

The PipMet package includes a built-in dataset of eight GC-MS files from sugarcane and energycane gems. This block is designed to walk you through the **entire logic** of the software. By the end of this block, you will have processed, audited, and analyzed a complete metabolomics experiment.

2.1 Initialization

To begin, load the library and execute the main wrapper function with the example flag enabled.

```
library(PipMet)

# This triggers the interactive engine using built-in data

result <- workData(example = TRUE)
```

- **Internal Logic:** Setting example = TRUE by passes manual folder selection and maps the tool to the internal inst/ext directory, reading 8 .mzXML files and a pre-configured metadata.csv.

2.2 Interactive Configuration & Preprocessing

Before the heavy processing starts, PipMet will request technical definitions through interactive pop-ups. This is where you define the computational power and sensitivity of the analysis:

A. iRT & Monitor EICs

- **The Concept:** Used for tracking **Internal Standards** (e.g., Ribitol).

- **Action:** Select **Yes**. Enter an (m/z) (e.g., 217) and an RT (e.g., 645s). PipMet creates a specific EIC for this ion to monitor machine stability across all injections.

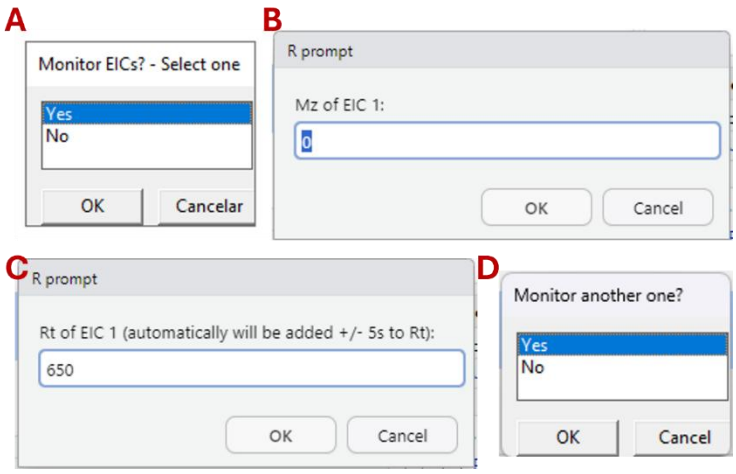


*Figure 3. **Monitor EICs** Pop-ups. (A) Select whether you want to monitor any specific ion or not. (B) Add the m/z. (C) Add expected retention time for this ion. (D) Select whether you want to monitor another specific ion or not.*

B. Parallelization (Computing Strategy)

- **The Concept:** Tells R how many "brains" (CPU cores) to use.
- **Action:** * **Windows:** Select **SnowParam** (use 2 or 4 workers).
    - **Linux/Mac:** Select **MulticoreParam**.

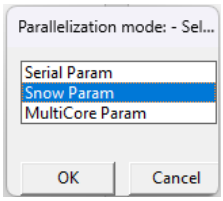


Figure 4. Parallelization Menu. Select the parallelization strategy for the data analysis based on our computer conditions.

C. Folder creation

- **The Concept:** PipMet automates the organization by creating a dedicated "Parent Folder." This ensures that all outputs—from the initial chromatograms to the final statistical models—are stored in a structured, reproducible hierarchy.
- **Action:** When the prompt appears, type a clear, descriptive name for your study (e.g., PipMet_Example). PipMet will then generate the internal sub-folder structure (e.g., Visualization_results, Statistics, etc.) within your working directory. For this dataset, it recommended to use: PipMet_example

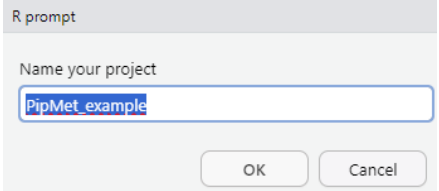


Figure 5. Folder creation pop-up.

2.2 Initial Audit: The Visualization Results

As soon as the files are read, PipMet creates the **Visualization_results** folder. **Stop and check these files** before moving to peak processing:

- **TIC & BPC (Chromatograms):** * **Folder:** Visualization_results/...
    - **What to check:** Ensure the runs overlap consistently. If one graph is significantly lower or "empty," that file may have a failed injection.
- **Correlation Heatmap:** * **Folder:** Visualization_results/...
    - **What to check:** Biological replicates must show high correlation (darker red colors). This confirms your experiment is reproducible before you spend time on complex math.
- **Monitoring ions:** * **File:** Visualization_results/ Monitoring ions/...
    - **Concept:** This step generates an **Extracted Ion Chromatogram (EIC)** specifically for the target ion you selected in Step 2.2A (e.g., your Internal Standard like Ribitol at *m/z* 217). It allows you to "zoom in" on a single molecule across all samples to check for technical drift.
    - **What to check:** 1. **Peak Presence:** Confirm that a clear, defined peak exists for every single sample. 2. **Retention Time (RT) Shift:** Check if the peak appears at the same time in all runs. If the peaks are shifting left or right significantly, your GC column might be degrading or the pressure is inconsistent. 3. **Peak Shape (Morphology):** The peaks should be **Gaussian** (bell-shaped). If you see "shoulders," jagged edges, or flat tops, it indicates a problem with the injection or detector saturation. 4. **Intensity Consistency:** While biological samples vary, your Internal Standard should have a relatively stable height. If one sample has a tiny peak compared to others, you likely had a partial injection failure.

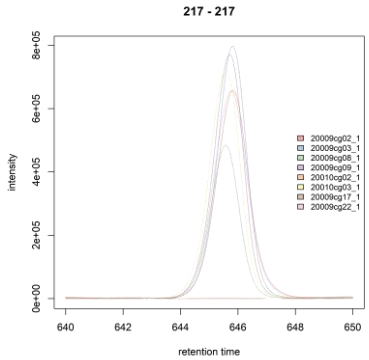


Figure 6. Example of EIC of our dataset.

2.3 Initial Data-processing setup

A. Intensity Threshold (The Noise Filter)

- **The Concept:** GC-MS detectors pick up background "chemical noise."

- **Action:** Enter **150**. The MatchedFilter algorithm will ignore any signal below this threshold, cleaning your quantification table of baseline "grass". Others equipment may need another threshold.

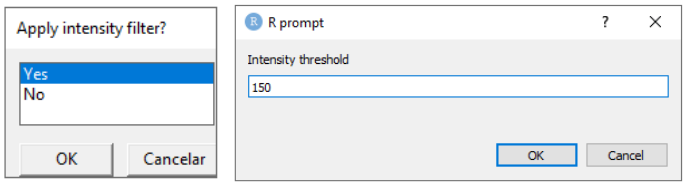


Figure 7. Folder creation pop-up.

B. Condition to group

- **The Concept:** For PipMet to calculate Fold-Change and P-values, it needs to know which experimental factor (variable) you want to compare. By selecting a "Condition," you are telling the software how to split the data for the **Volcano Plots**, **PCA**, and **Heatmaps**.
- **Action:** 1. Look at the list of columns from your metadata.csv provided in the pop-up. 2. Select the column that represents your primary comparison (e.g., Treatment, Genotype, or Timepoint). 3. Note: Ensure this column has at least two different levels (groups) so that a statistical comparison can be made. For this dataset, select **group**.

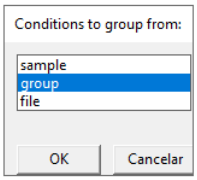


Figure 8. Example of EIC of our dataset.

C. Audit: Peak Processing Results

After filtering, check the **peakProcessing_results** folder:

- **RT Alignment:** plotAdjustedRtime.png shows how PipMet "stretched" the samples to line up. The lines should be smooth, not chaotic.
- **Peak Distribution:** peakDistribution.png shows where most metabolites were detected during the run.

D. Acquisition mode

- **The Concept:** This defines charge of the acquisition.
- **Action:** For the example, select **Positive**.
- **Why it matters:** This is essential for calculating the **Retention Index (RI)**, a standardized ID for molecules that allows your data to be compared with other labs worldwide.

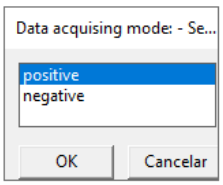


Figure 9. Acquisition mode pop-up: Select the acquisition mode (positive or negative) that your data has been acquired.

E. Chemistry: Column & Temperature Setup

- **The Concept:** This defines the physical chemistry of your run.
- **Action:** For this example, select **Non-Polar** (e.g., DB-5), and **Ramp**.
- **Why it matters:** This is essential for calculating the **Retention Index (RI)**, a standardized ID for molecules that allows your data to be compared with other labs worldwide.

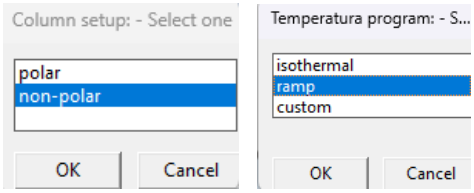


Figure 10. Column configuration and chromatographic program pop-up: Select the acquisition mode (positive or negative) that your data has been acquired.

F. Monitoring spectra EIC of the 6 most intense *m/z*

- **The Concept:** PipMet automatically identifies the 6 most intense ions across your entire dataset and extracts their individual chromatograms (EICs). Since these are your strongest signals, they represent the best-case scenario for your run.
- **Action:** For this example, select **Yes.** When prompted, PipMet will automatically generate a PDF report. You don't need to select ions manually; the software finds the most intense ones to provide an unbiased view of your peak quality.
- **EIC extracted file:** * **File:** Visualization_results/EIC_XIC.pdf

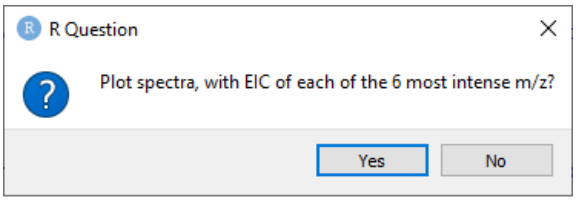


Figure 11. Select if you want to monitor the most intense ions.

2.3 The Identification Break & NIST (Optional)

A message will appear: *"pre_anno.csv and spectra.msp were created".*

* **Strategic Choice:** You have two options:

1. **Identify Now:** Open **NIST MS Search** (for more information and download: https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nistlibs), import the .msp, identify the peaks, and fill the pre_anno.csv.

2. **Identify Later:** Click **OK** without filling the CSV. PipMet will label peaks as "Unknowns". This allows for a results-driven identification strategy, where you prioritize identifying only the statistically significant features (the 'winners'), drastically reducing manual annotation time.

- **What to check:** * The Files: Ensure that pre_anno.csv and spectra.msp have appeared in your project folder.
- **The Format:** If you choose to identify now, ensure the names in the CSV match the NIST suggestions exactly to avoid errors in the final library creation.
- **Action:** For the example run, click **OK** without adding annotations to proceed quickly to the statistical results.

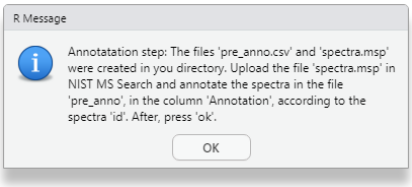


Figure 12. The Identification and Annotation Gateway. Press "OK" after annotating the compounds or if you want to procced without annotation.

2.4 Retention index information

- **The Concept**: The Retention Index (RI) is like a "GPS coordinate" for molecules. While Retention Time (RT) changes depending on your column's age or length, the RI is standardized. By providing RI information, you increase the confidence of your identifications, moving from "putative" to "confirmed" compounds.
- **Action:** For this example, click **No**.

Why it matters: If you had a list of standard alkanes or a specific RI library, selecting "Yes" would allow PipMet to calibrate your samples against international databases, making your data comparable to any other GC-MS lab in the world. Additional information about retention index can be found in **3.3 Creating the RI calibration file.**

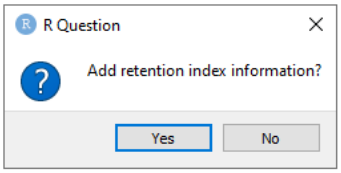


Figure 13. Additional retention index information pop-up.

2.6 Normalization & Derivatization Handling

A. The TMS/Derivatization Filter

- **The Concept:** Derivatization (TMS) creates a huge, useless peak at (m/z) 73.

- **Action:** Select **Yes**. PipMet will skip *m/z* 73 and use the *second* most intense ion as the **Representative Ion** for quantification.

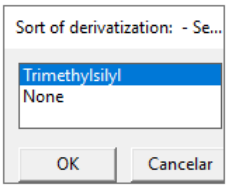


Figure 14. Pop-up for selecting the derivatization application and skip or not the ion 73.

B. Merging and removing compounds

- **The Concept:** During peak picking, the algorithm sometimes detects the same metabolite as two separate peaks due to slight shifts or complex fragmentation. This step allows you to "clean" your data by merging these redundant signals into a single representative entry. This prevents the final quantification table from being cluttered with duplicate information for the same molecule.
- **Action:** Select **No** for this example.

  Strategic Note:** You should only select Yes (Merge) if you have already identified your compounds in the pre_anno.csv and noticed that two different peaks were assigned the same name. In that case, PipMet will combine their intensities. Moreover, you should only select No (Remove), if you have already peaks or compounds **undesired.

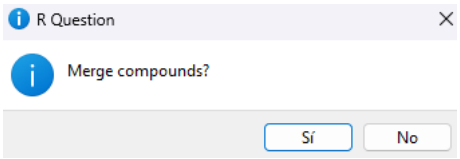


Figure 15. Pop-up for merging compounds.

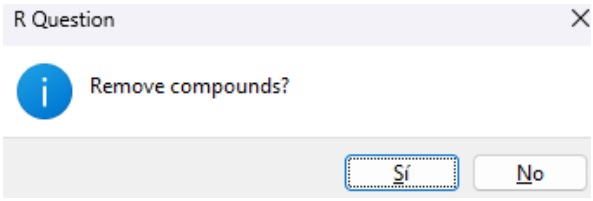


Figure 16. Pop-up for removing compounds.

C. Assisted Normalization by mass or number of cells/ etc. (OPTIONAL)

- **The Concept (Normalization by Statistics):** Sometimes, technical normalization (such as CycLoess) is not enough. If your samples have variations in fresh weight, cell counts, or volume recorded during the experiment, you must account for them. **PipMet** allows you to perform an initial normalization based on these external factors to ensure that final concentrations are biologically comparable. To do this, you can open and edit the non-normalized data file (.../Statistics/NotNormalized_quantification.csv), divide the values by your specific factors, and save the document. During this process, you will encounter two confirmations: **first, click "Yes" in the pop-up**, and **then click "OK"** to tell PipMet

that the file is ready. This file contains raw peak areas before any scaling. Once saved and confirmed, you can proceed with the *normalizeDE* function for further statistical refinement.

- **Action:**
    - **Normalize by mass/number of cells?** Select **No** for this example. (Select Yes only if you have a specific column in your metadata with these numeric values).

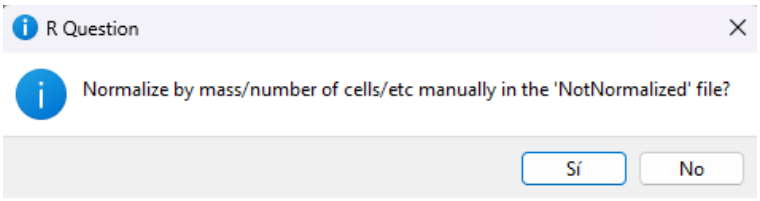


Figure 17. Normalization Menu. Select the normalization strategy for the data analysis based on the best results from computer conditions.

D. Assisted Normalization

As soon as you select merging and removing compounds, assisted normalization will occur and you can select the best strategy. PipMet creates the **Statistics** folder. **Stop and check these files into Normalyzer_results** folder before moving foward:

- **The Concept:** Minor differences in injection volume or sample concentration can occur.
- **Action:** Open the PDF report located in the **Normalyzer_results** folder (Norm-report-Normalyzer_results.pdf).
    1. Look at the Boxplots: The best method is the one where the boxes are most horizontally aligned (centered).
    2. Look at the Density Curves: The best method is the one where the colored curves overlap most perfectly.
    3. Select: Usually, **CycLoess** is the most robust choice. Type the name of your chosen method exactly as it appears in the pop-up list.

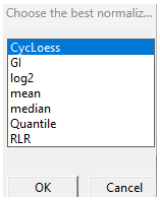


Figure 18. Normalization Menu. Select the normalization strategy for the data analysis based on the best results from computer conditions.

E. Additional Normalization steps (if necessary)

- **The Concept:** Sometimes, normalization (like CycLoess) didn't go good. You can perform another data normalization or conclude.
- **Action (The Final Decision):**

- **Conclude or Re-do?** This is your final quality check.
- **Conclude normalization:** Select this if you are happy with the boxplots and density curves (Check: Statistics/Normalized_quantification.xlsx).
- **Re-do normalization:** Select this if the first method you chose didn't look right and you want to try another strategy (e.g., switching from CycLoess to Median, for example).

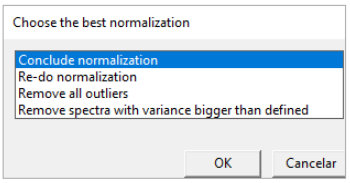


Figure 19. Normalization Menu. Select the normalization strategy for the data analysis based on the best results from computer conditions.

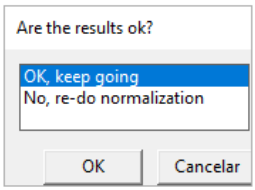


Figure 20. Normalization Menu. Select the normalization strategy for the data analysis based on the best results from computer conditions.

D. Normalization Results

After the normalization step is completed, PipMet exports the quantification tables. These results can be observed and analyzed by the user in the **Statistics/** folder within the working directory.

- **Not-normalized data:** Look for NotNormalized_quantification.csv. This file contains the raw peak areas for each compound (or mass feature) before any scaling or normalization algorithms were applied.
- **Normalized data:** Look for Normalized_quantification.csv. This file contains the final areas after normalization. The values in this table are ready for statistical comparisons and can be used to perform **absolute quantification** if standards were used.
- **Additional information:** The **Statistics/** folder will also contain diagnostic images and supplementary .csv files generated during the normalization process, which help in verifying the quality of the data before proceeding to the final analysis.

2.7 Statistical Analysis and Data Visualization

After data normalization, PipMet enters the biological interpretation and comparison phase. In this stage, the software utilizes the unidentified or identifications from NIST and the corrected intensities to generate high-quality plots that validate your experimental

findings. For example, unidentified strategies can be applied to just identify significantly altered metabolites before spending time on a full NIST library search.

- **Action:**
    - **All:** For this example, select this option. It will ask and perform all the statistical comparisons. If you select the other options, it will perform only the selected analysis, and the data analysis will be finished. We recommend selecting them all.

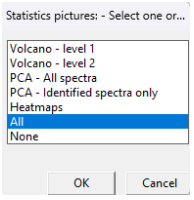
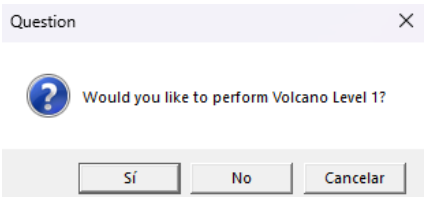


Figure 21. Selecting statistical analysis and visualization parameters. Volcano – level 1: Performs a primary univariate comparison between total groups (e.g., Group A vs. Group B). Volcano – level 2: Performs a refined comparison of specific conditions or retention time variations within a single group. PCA – All spectra: Executes a Principal Component Analysis using the entire dataset, including both identified and unidentified features. This is the recommended option if a full compound identification has not been completed. PCA – Identified spectra: Performs a PCA focused strictly on compounds named in the NIST library. ***Note: A.*** *Do not select this option if identification is not performed, as it will result in a processing error. Heatmaps: Automatically generates a hierarchical clustering heatmap to visualize abundance patterns and technical replicate consistency. All: Sequentially executes all the analyses listed above.* ***B.*** *If you select* ***"All"*** *but have not identified in the compounds in pre_anno.csv file, the software will prompt you with a confirmation menu before attempting the* ***PCA – Identified spectra****. This prevents the script from crashing and allows you to skip that specific analysis while still generating the others.*

2.7.1 Univariate Analysis: Volcano Plots

The software identifies significant changes between groups using *t*-tests. These results are visualized through Volcano Plots, plotting statistical significance (*p*-value) against the magnitude of change (Fold Change).



- **Level 1 (Standard Comparison):** This is the mandatory first screen. It makes a direct comparison between two specific groups (e.g., **Group A vs. Group B**). It is highly effective for "unidentified strategies," allowing you to see which mass features are statistically relevant before naming them.

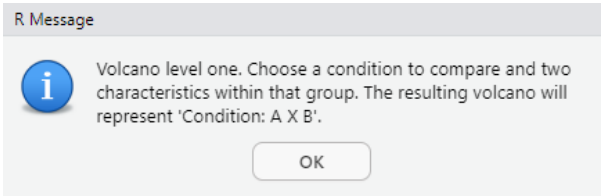


Figure 22. Dialogue box to further define the groups for the Level 1 analysis. The message header confirms that the software is in the ***"Volcano Level 1"*** *stage.*

- **Action:** For this example, select **group**. Further, select Energy and Sugar, or any other comparison. The results can be observed in folder: Statistics/Volcanos. The table document: Significance_compounds_vol_lvl1.csv will be created containing the information Fold Change, p-value and up or down regulated.
    - After this comparison, select **Next Step**.

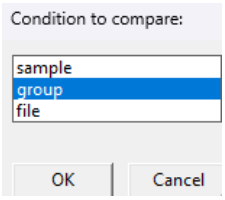


Figure 23. ***Selecting the metadata column for comparison.*** *Before defining the groups, a dialogue box asks which metadata category should be used for the analysis (e.g., files, group, or sample). This step determines which column the software will use to aggregate the data and perform the statistical tests.*

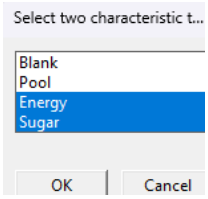


Figure 24. Users must select the two specific conditions from the metadata to be compared (e.g., Control vs. Treatment). This selection determines the direction of the Fold Change and the statistical significance displayed in the resulting plot.

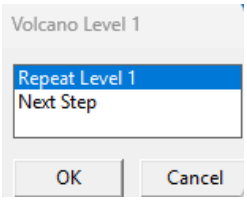


Figure 25. Users must select if want to repeat Volcano – level 1 or follow for the next step.

- **Level 2 (Contrast Comparison - OPTIONAL):** After Level 1, PipMet allows for a higher-level meta-analysis. This level is used to **compare the results of two different Volcano Plots** (e.g., comparing the hits of **A vs. B** against the hits of **A vs. C**). This is a powerful tool to identify metabolites that are specifically altered in one condition but not in another, or to find common biomarkers across multiple experimental contrasts.
    - **Action:** For this example, select **group for both comparisons**. Further, select **Energy** as Target and **Blank** as Baseline in Pair 1; and **Sugar** as Target and **Blank** as Baseline in Pair 1, or any other comparison. The results can be observed in folder: Statistics/Volcanos. The table document: Significance_Level2_Table.csv will be created containing the information Fold Change, *p*-value and up or down regulated in a specific comparison. Further, select **Next Step**.

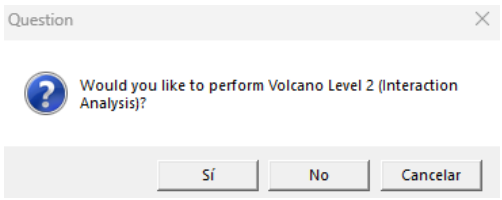


Figure 26. Users must select if want to perform Volcano – level 2 or follow for the next step (PCA).

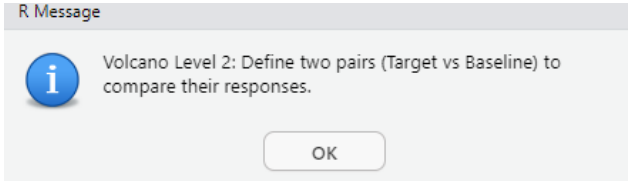


*Figure 27. Dialogue box to further define the groups for the Volcano Level 2 analysis. The message header confirms that the software is in the **"Volcano Level 2"** stage.*

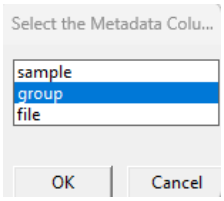


Figure 28. Users must select the metadata column to select the comparison conditions.

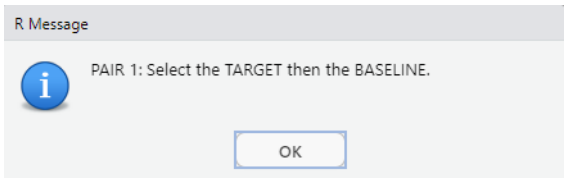


Figure 29. Dialogue box to further define the Target and Baseline for the first comparison for the Volcano Level 2 analysis.

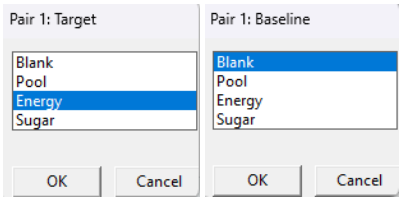


Figure 30. Users must define the Target and Baseline for the first comparison for the Volcano Level 2 analysis.

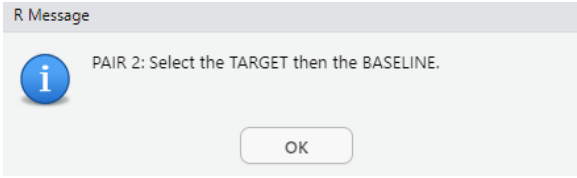


Figure 31. Dialogue box to further define the Target and Baseline for the second comparison for the Volcano Level 2 analysis.

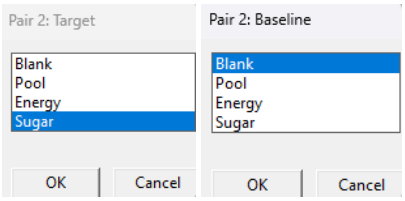


Figure 32. Users must define Target0020and Baseline for the second comparison for the Volcano Level 2 analysis. Further PipMet will compare the difference between Volcano 1 and Volcano 2.

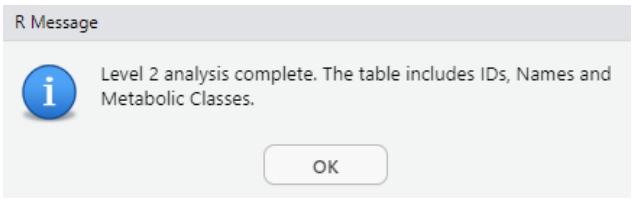


Figure 33. Dialogue box showing that the Volcano level 2 comparison is done.

Navigation Controls:

- **Next Step:** Use this to finalize the univariate phase and move forward multivariate analysis (PCA).

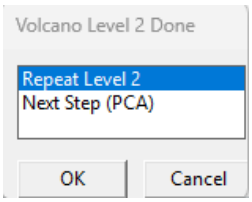


Figure 34. Dialogue box for workflow control during statistical analysis. After performing Volcano Level 2, the system prompts the user to either repeat the analysis with different group interactions or proceed to the Principal Component Analysis (PCA) step.

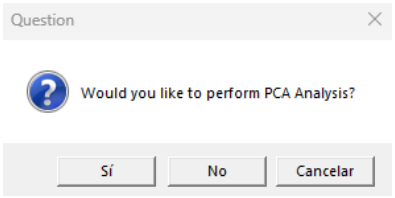


Figure 35. Dialogue box asking if the user wants to perform the PCA analysis.

2.7.2 Multivariate Analysis: Principal Component Analysis (PCA)

PCA is essential for understanding the global variance of your samples. PipMet separates this into two distinct approaches:

A. PCA General (All Spectra)

This plot utilizes every detected signal in your batch. It is the most reliable way to:

- Identify sample outliers.
- Observe the natural clustering of your biological groups.
- Validate the overall quality of the extraction and injection process.
- A specific group comparison can also be made.
    - **Action:** For this example, select **General PCA**. Further, select **Yes** for specific groups/conditions. Then, **group** for select column metadata and Energy and Sugar as groups in conditions to compare. The results can be observed in folder: PCA_general. Further, select Next Step.

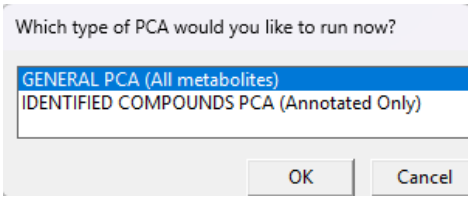


Figure 36. Dialogue box asking if the user wants to perform the PCA analysis. Just select Identified compounds PCA if has identified the compounds before.

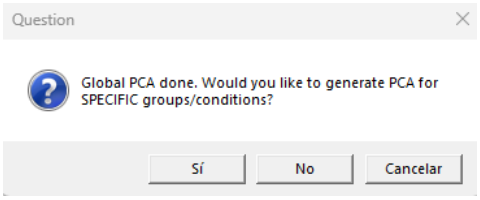


Figure 37. Dialogue box asking if the user wants to perform the PCA analysis for specific subgroups.

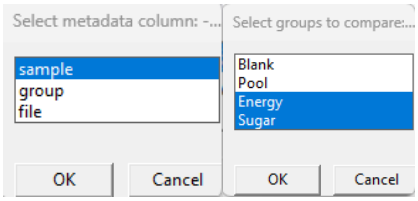


Figure 38. Dialogue box is asking the user to select the column from metadata to perform the specific PCA analysis and further select the groups that want to compare.

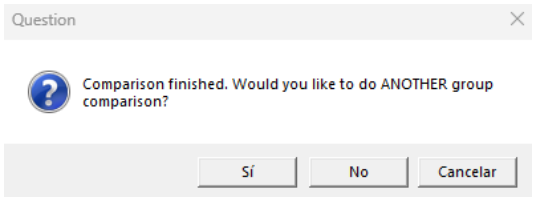


Figure 39. Dialogue box informing that the PCA analysis was done and asking if the user wants to perform another comparison.

B. PCA Identified

Once you have finished your NIST identification, you can generate a PCA based **only** on named metabolites.

- **Why it is optional:** This plot requires a well-populated and anotated pre_anno.csv. Since identifying metabolites is time-consuming, PipMet asks if you want to generate this plot now or skip it.
- **Outcome:** It provides a "cleaner" biological story, focusing strictly on the known metabolic pathways affected by your experiment.
    - **Action:** For this example, select **Next Step**. Note: Just repeat the PCA if you want to perform or another specific group comparison; or the PCA analysis with the identified compounds.

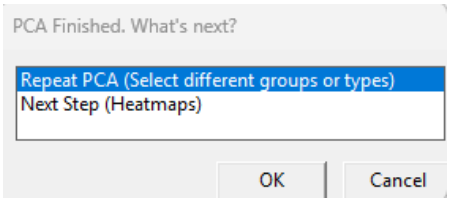


Figure 40. Dialogue box asking if the user wants to perform again the PCA analysis. If the user wants to perform PCA from identified compounds only (not all the spectra), select repeat.

2.7.3 Heatmaps: The Metabolic Signature

The final visualization tool is the **Heatmap**, which performs hierarchical clustering on both samples and metabolites to reveal the "metabolic fingerprint" of your experiment.

- **Clustering Logic:** PipMet automatically groups metabolites and samples with similar abundance patterns. This highlights which compounds are up-regulated or down-regulated across different experimental conditions.
- **Color Scale:** The map uses a standard color gradient (typically **Red** for high abundance and **Blue** for low abundance). This allows for immediate visual identification of biomarkers that define a specific group or treatment.
- **Technical Validation:** The Heatmap is an excellent tool to verify **Technical Replicates**. If replicates do not cluster together, it may indicate a need to revisit the normalization or injection steps.
    - **Action:** For this example, select **Yes**. Further, select in name the samples as: **sample** and **Finish Statistics**.

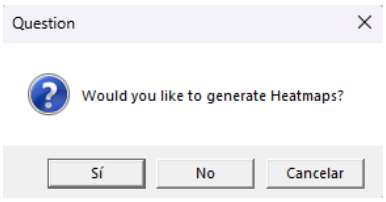


Figure 41. Dialogue box asking if the user wants to generate Heatmaps.

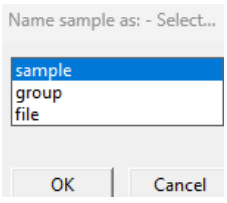


Figure 42. Pop-up asking to user select the name of the samples for the heatmap.

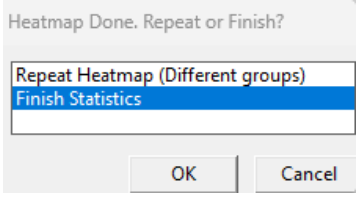


Figure 43. Repeat Heatmap or Finish statistics pop-up.

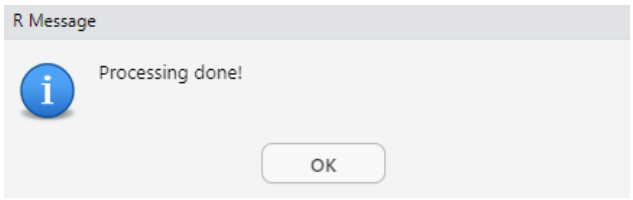


Figure 44. Dialog box demonstrating finishes in data processing.

3. Processing your own data

To analyze your own data, the user must follow these steps:

I. **Workspace:** Create a main folder where PipMet will generate all processing and result folders.

II. **File Conversion:** Convert your raw files to **.mzML** or **.mzXML** using software like ProteoWizard (MSConvert).

III. **Metadata Creation:** Create a .csv file with three essential columns: sample, group, and file.

- **Note on File Path:** The file column must contain the **full pathway** + file name (e.g., C:/Users/Project/data/sample01.mzXML). Use forward slashes /.

IV. **Additional Metadata:** Beyond the three essential columns, the user is encouraged to add as many additional columns as needed to describe the experimental design. Examples include:

- **Time points** (e.g., 0h, 24h, 48h)
- **Treatment doses** (e.g., Control, Low, High)
- **Biological replicates** (e.g., BioRep1, BioRep2)
- **Technical factors** (e.g., Extraction_Batch, Run_Order)

These additional columns allow you to easily switch the grouping variable during statistical analysis.

3.1 Running the Analysis

Once your folder is set and the metadata.csv is ready, initiate the pipeline with:

```
library(PipMet)
# This triggers the interactive engine using built-in data
result <- workData( )
```

PipMet will open interactive windows for you to select your folders and the metadata file (Figures 46-48).

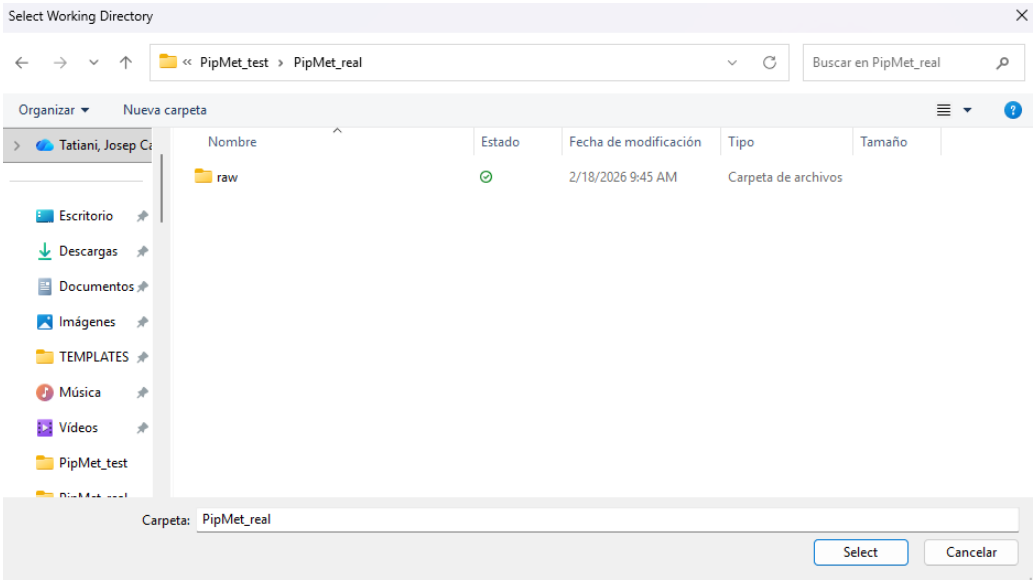


Figure 45. Pop-up for main folder selection.

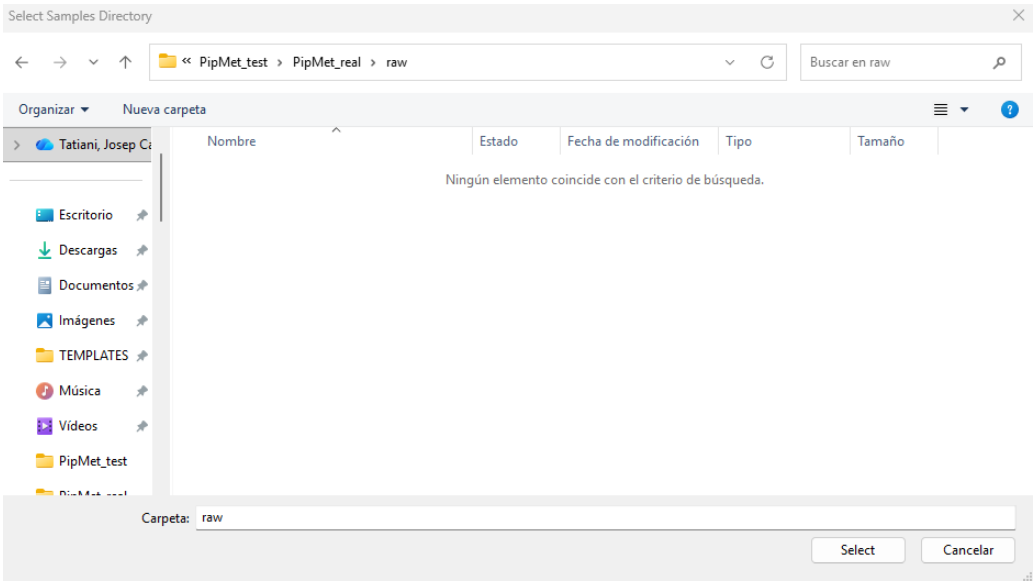


Figure 46. Pop-up for raw file folder selection.

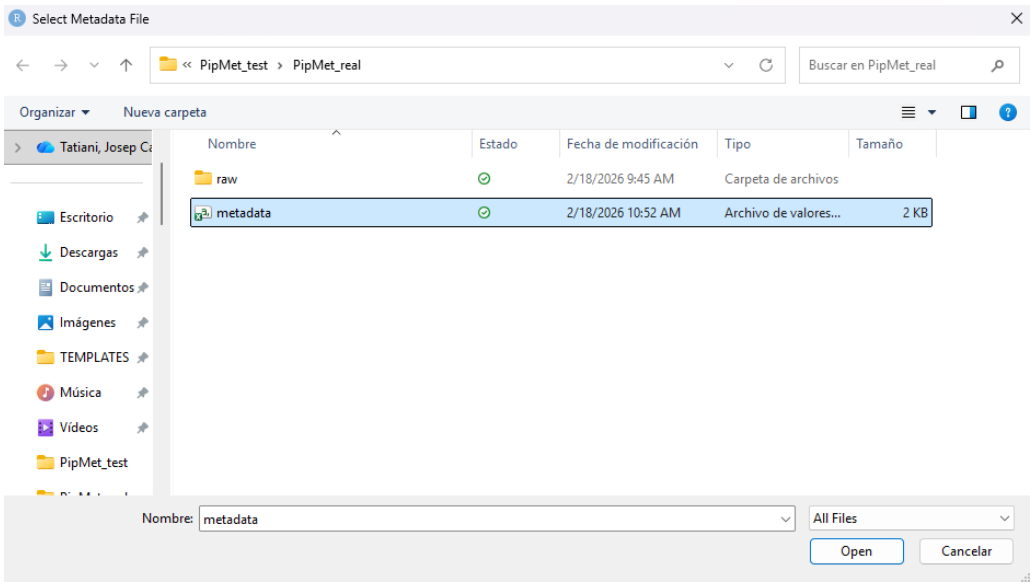


Figure 47. Pop-up for metadata archive selection.

3.2 Automated Workflow

After the initial setup, you should follow the interactive steps as described in the **Data Processing Example** (Section 2). The logic remains identical:

- **Pre-processing:** Peak picking and grouping.
- **Annotation (optional):** Confirmation of compounds via NIST/Library. An .msp file is automatically saved in the main working directory for compound identification by the user.
- **RI adjustment (optional):** retention index information can be added by the user .
- **Mass Normalization (optional):** When prompted, you can normalize your data by weight or volume. **Edit** the NotNormalized_quantification.csv in the *Statistics* folder, save it, and press OK.
- **Statistical Normalization:** Choose the best method (e.g., Median, TIC, Loess) via the NormalyzerDE results. If you need further manual adjustments, you can edit the resulting Normalized_quantification.csv.
- **Final Statistics:** PipMet will automatically generate PCAs, Volcano Plots, and Heatmaps based on the columns you provided in your metadata (e.g., group, treatment, or time).

3.3 Creating the RI calibration file

The RI file must be a **CSV** (Comma Separated Values) containing two mandatory columns:

1. **'rt'**: The retention time of your standards (e.g., alkanes). Ensure the unit (seconds or minutes) matches your data.
2. **'RI'**: The theoretical Retention Index for each standard (e.g., 1000 for C10, 1100 for C11).

Tip: You can create this in Excel and "Save As" CSV (Comma delimited). Do not use spaces or special characters in the column headers. An example can be observed in https://github.com/PipMet/PipMet/inst/extdata/RI_example.csv

- **Action:** click Yes, when asked for adding retention index information. Further, select the file. Once the retention index (RI) information is added, a new .msp file (e.g., xxx_RI.msp) is automatically generated in the working directory.

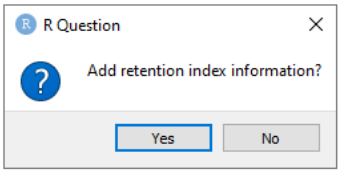


Figure 48. Additional retention index information pop-up.

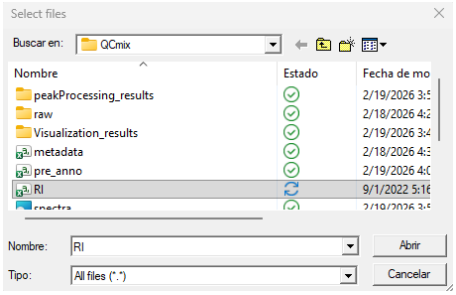


Figure 49. Retention index file selection dialog box.

3.3 Additional information

After data processing, PipMet generates a standard .msp file containing the processed spectra. This file can be directly imported into the NIST MS Search program for compound identification. Users can then use these identified spectra to build and manage their own in-house libraries within the NIST environment. The .msp file is automatically saved in the main working directory.

4. References

1. Abreu LGF, Silva N, Ferrari A, Carvalho L, Fiamenghi M, Carazolle M, Fill T, Pilau E, Pereira G, Grassi M: Metabolite profiles of energy cane and sugarcane reveal different strategies during the axillary bud outgrowth. Plant Physiology and Biochemistry, 2021, 167:504-516.

2. NIST. **NIST Standard Reference Database. Available in: https://chemdata.nist.gov/mass-spc/ms-search/docs/Ver20Man_11.pdf.

3. TANAKA S, FUJITA Y, PARRY HE, YOSHIZAWA AC, MORIMOTO K, MURASE M, YAMADA Y, UTSUNOMIYA S, KAJIHARA K, FUKUDA M, IKAWA M, TABATA T, KATAHASHI K, AOSHIMA K, NIHEI Y, NISHIOKA T, ODA Y, TANAKA K: Mass++: a visualization and analysis tool for mass spectrometry. *Journal Of Proteome Research, 2014, 13:3846-3853.

4. KESSNER D, CHAMBERS M, BURKE R, AGUS D, MALLICK P: ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics, 2008, 24(21):2534-2536.

5. DEPKE T, FRANKE R, BRÖNSTRUP M: CluMSID: an R package for similarity-based clustering of tandem mass spectra to aid feature annotation in metabolomics. Bioinformatics, 2019, 35(17): 3196-3198.