Apr 28, 2025

Untargeted Metabolomics Data Processing Using FragHub and MS-DIAL Softwares

DOI
https://dx.doi.org/10.17504/protocols.io.j8nlk8jywl5r/v1

Akila Wijerathna Yapa[1], Stanislav Sokolenko[1]

[1]Department of Process Engineering and Applied Science, Dalhousie University, 5273 DaCosta Row, PO Box 15000, Halifax, B3H 4R2, NS, Canada

Akila Wijerathna Yapa
The Hebrew University of Jerusalem

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.j8nlk8jywl5r/v1

Protocol Citation: Akila Wijerathna Yapa, Stanislav Sokolenko 2025. Untargeted Metabolomics Data Processing Using FragHub and MS-DIAL Softwares. **protocols.io** https://dx.doi.org/10.17504/protocols.io.j8nlk8jywl5r/v1

Manuscript citation:

Protocol status: Working
We use this protocol and it's working

Created: September 20, 2024

Last Modified: April 28, 2025

Protocol Integer ID: 108101

Keywords: Metabolomics, Mass Spectrometry, MS/MS, MS-DIAL, FragHub, Untargeted Metabolomics, Lipidomics, untargeted metabolomics data processing, untargeted metabolomics data analysis, comprehensive approach for untargeted metabolomics data analysis, robust framework for untargeted metabolomics data processing, metabolomic, based metabolomic, facilitating enhanced metabolite identification, enhanced metabolite identification, metabolite discovery, metabolite annotation, reproducibility of metabolite discovery, annotation of metabolite, reliability of metabolite annotation, mass spectrometry, metabolite, multiple mass spectral library, various open mass spectral library, data independent analysis software, ms spectral matching, metadata, spectral matching against library, compound identification through retention time, spectral library, chromatogram deconvolution, source software platforms fraghub, various ms vendor

Abstract

This protocol presents a comprehensive approach for untargeted metabolomics data analysis using the open-source software platforms **FragHub** and Mass Spectrometry–Data Independent AnaLysis software (**MS-DIAL**). The workflow is designed for LC-MS/MS-based metabolomics, focusing on optimizing the detection, identification, and annotation of metabolites. **FragHub** integrates multiple mass spectral libraries and formats, including .MSP, .MGF, .JSON, and .CSV, using RDKit to harmonize metadata, which improves the accuracy and reliability of metabolite annotation. It enables users to unify various open mass spectral libraries (OMSLs), facilitating enhanced metabolite identification. **MS-DIAL**, a versatile tool supporting various MS vendors and instruments, enables peak detection, deconvolution, and spectral matching against libraries. MS-DIAL allows efficient precursor ion detection, chromatogram deconvolution, and compound identification through retention time, accurate mass, and MS/MS spectral matching. Together, FragHub and MS-DIAL provide a robust framework for untargeted metabolomics data processing, enhancing the accuracy and reproducibility of metabolite discovery.

Data Acquisition via LC-MS/MS

1 For comprehensive metabolomics analysis, data acquisition via high-resolution mass spectrometry (HRMS) using LC-MS/MS is the gold standard. The following settings are typical for untargeted metabolomics:

Chromatographic Conditions:

- **Column:** Select a column suited for small molecule separation (e.g., C18 or HILIC columns).
- **Mobile Phases:** Typically, water with 0.1% formic acid (A) and acetonitrile with 0.1% formic acid (B).
- **Gradient:** A typical gradient may start with 95% solvent A, gradually decreasing to 5% solvent A over 10-15 minutes.

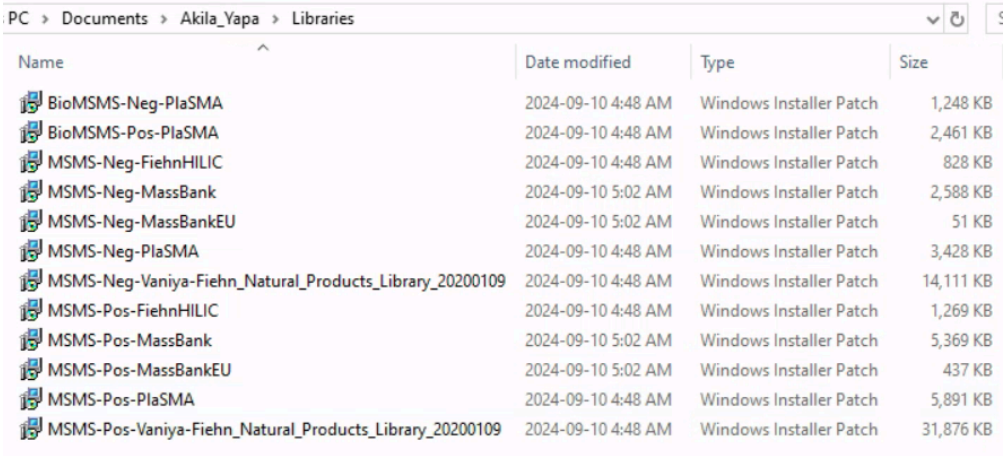
Mass Spectrometry Conditions:

- Use both **positive and negative ion modes** to enhance metabolome coverage.
- **Ion Source**: Electrospray ionization (ESI) is typically used for untargeted metabolomics.
- **Scan Range**: Acquire data over a mass range of 50-1000 Da for MS1, with MS/MS fragmentation covering the same range.
- **Collision Energy**: For data-dependent acquisition (DDA), set collision energies around 30V for positive mode and -25V for negative mode.
- **Data Storage**: Store raw LC-MS/MS data in mzML, WIFF, raw, or other vendor-specific formats.

FragHub: A mass spectral libraries data integration workflow

2 **FragHub** offers an advanced framework to enhance metabolite annotation by integrating multiple mass spectral libraries in different formats, ensuring greater coverage and accuracy **(Dablanc, et al. 2024, Analytical Chemistry)**. FragHub mass spectral libraries data integration workflow is publicly avaible at at the following repository: https://github.com/eMetaboHUB/FragHub. For this current study below steps were followed.

2.1 Publicly available ***.msp*** (mass spectra format) spectral libraries were downloaded from **MS-DIAL** website.

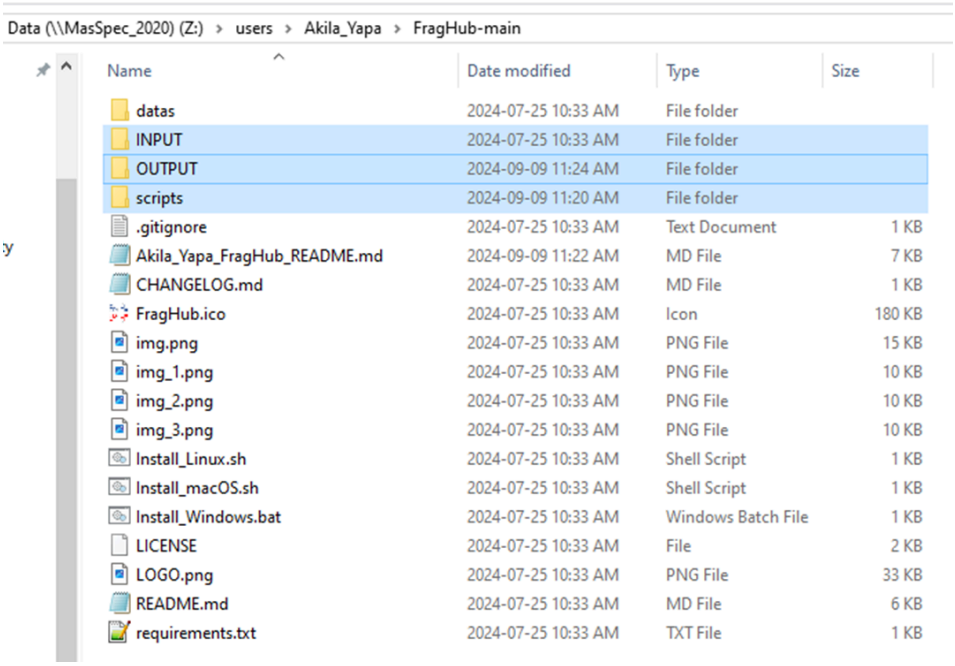


2.2 Python version 3.9 or higher should be downloaded and installed, ensuring that the option 'Install Python to PATH' is selected.
The Python version can be verified using the following command in the command line:
Windows: in command prompt
MacOS/MacOS: in terminal

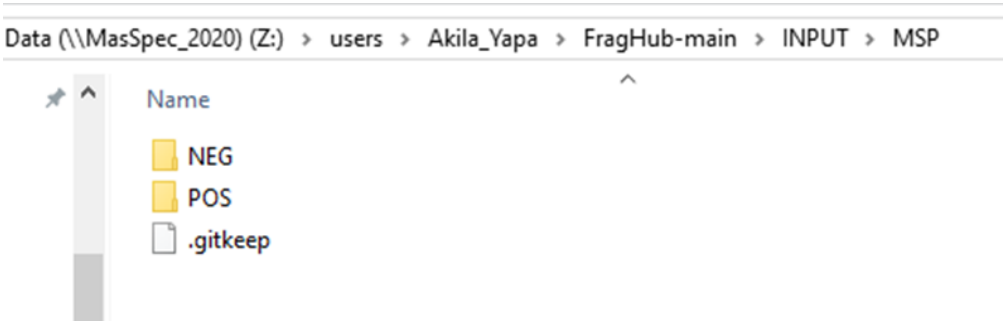
```
python --version
```

Access: **https://www.python.org/downloads/**

2.3 FragHub can be downloaded and installed from **GitHub**. and detailed instructions can be found https://pubs.acs.org/doi/suppl/10.1021/acs.analchem.4c02219/suppl_file/ac4c02219_si_002.pdf

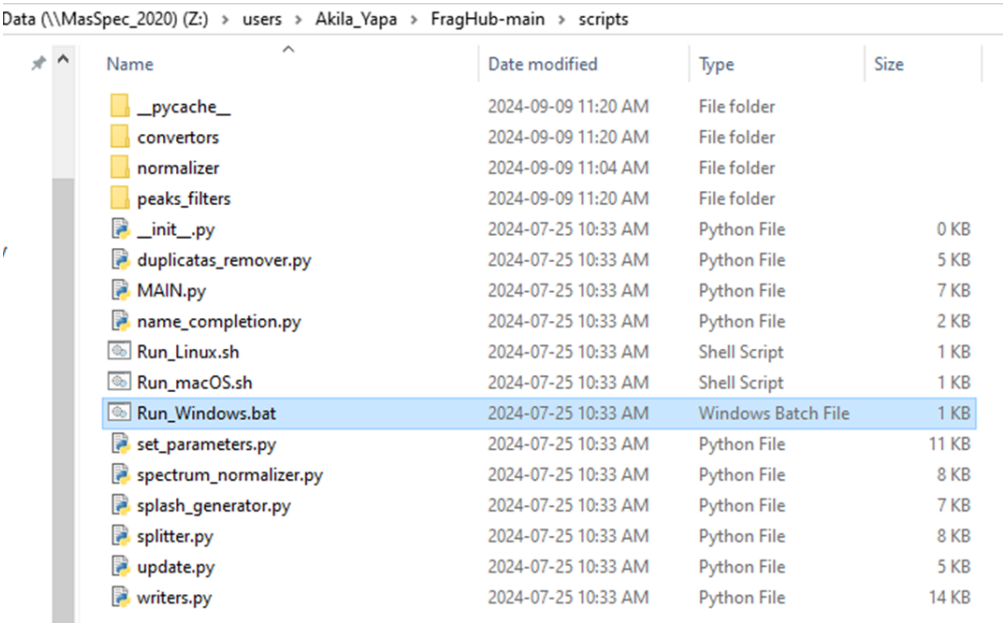


2.4 The downloaded MSP files were saved into the 'INPUT/MSP/<dedicated folder>'.

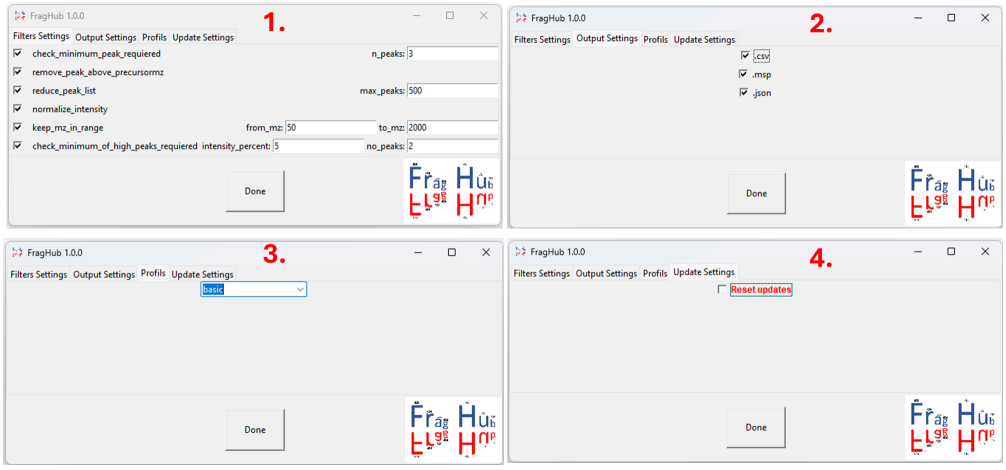


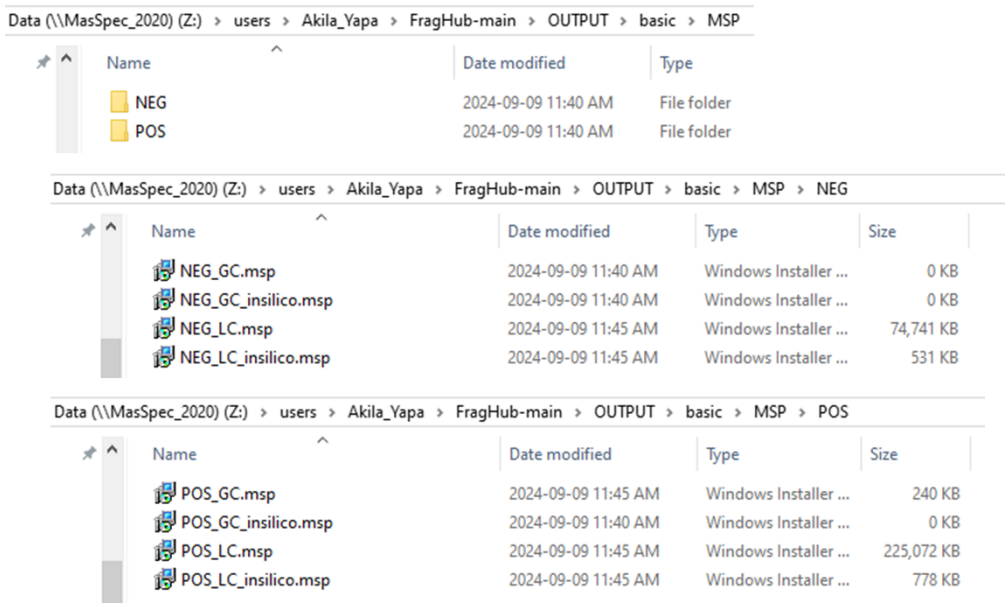
.msp libraries were separated into Positive and Negative mode for easy use.

2.5 The corresponding OS run script in the scripts folder should be double-clicked to start the FragHub GUI.



2.6 The default settings of the FragHub GUI were followed, and once execution was completed, the cleaned files from the OUTPUT folder were copied and placed in a different location.





Data Processing Using MS-DIAL

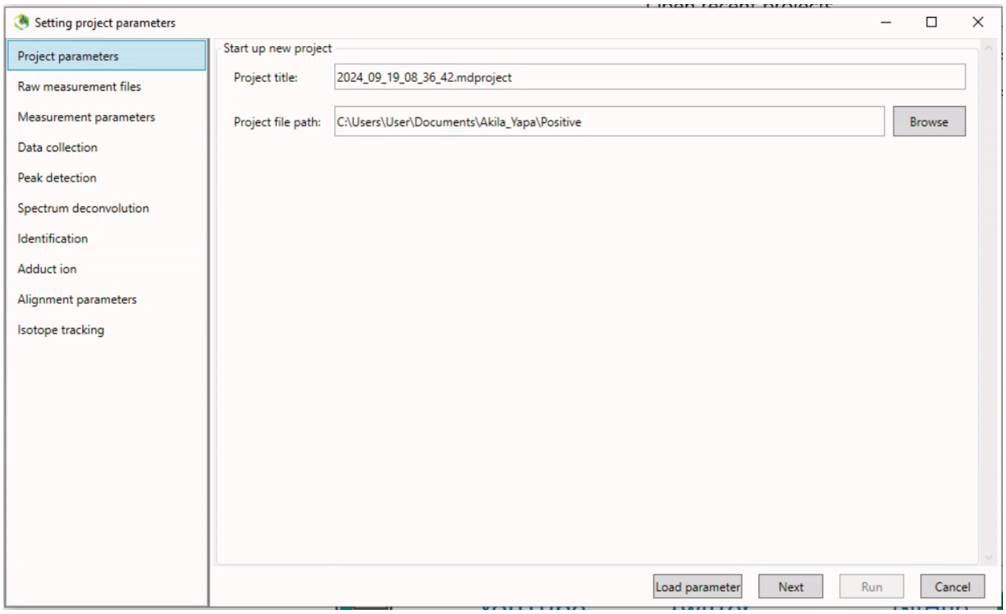
3 MS-DIAL is a flexible software tool for untargeted metabolomics that can handle data from multiple MS vendors and instruments.

3.1 **File Conversion**:

- If necessary, convert raw data to **ABF format** using the Reifycs Analysis Base File Converter **(ABF converter)**, particularly for fast data retrieval and analysis in MS-DIAL.

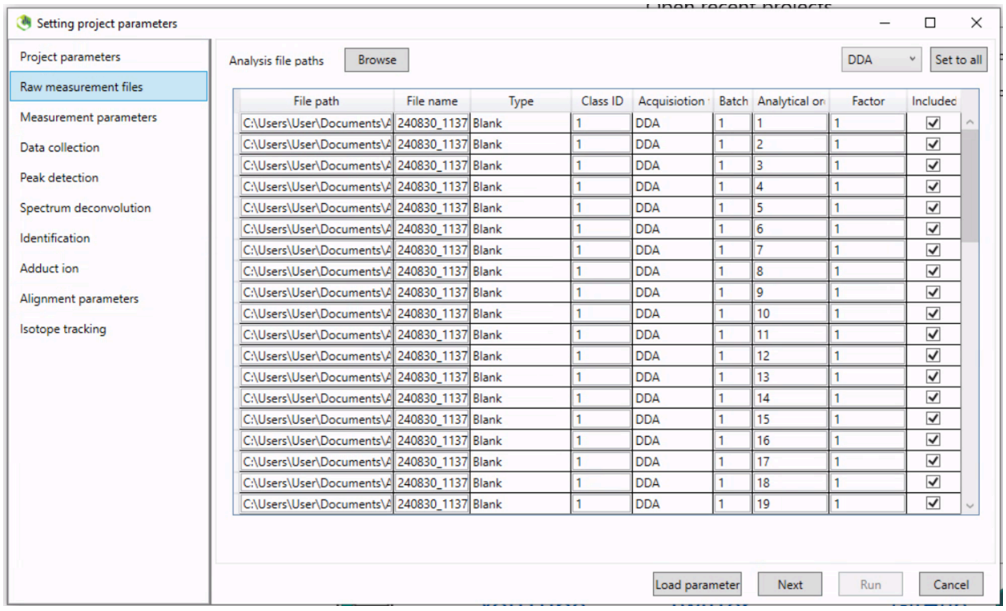
3.2 **Project Setup**:

- Start a new project in MS-DIAL and load the converted raw files. Set appropriate file paths for data storage.



3.3 **Upload the raw files**:

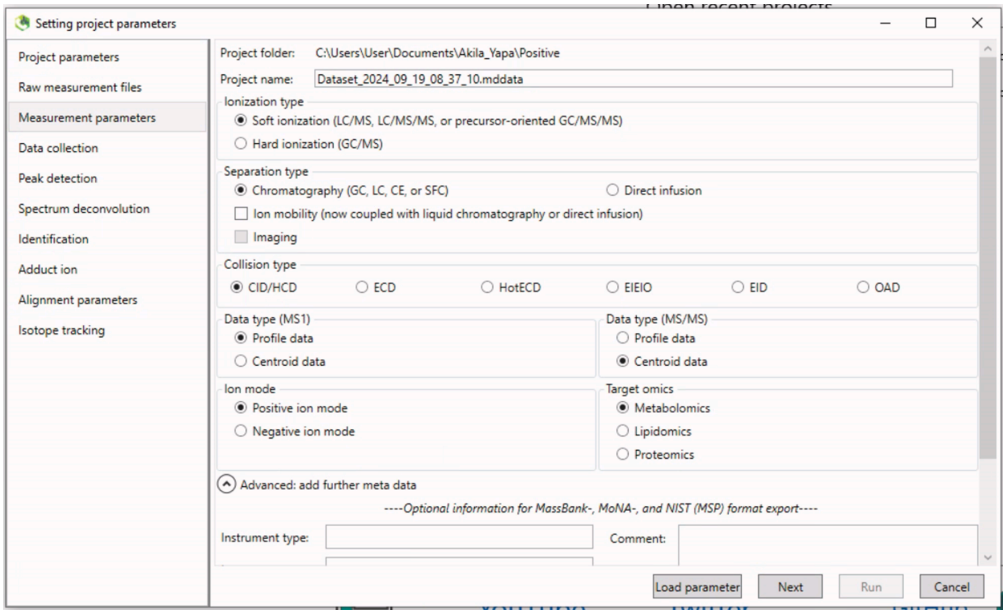
- Categorize raw files in to Blanks, QC, Standard, and Samples based on the study.



3.4 **Data Processing Parameters**:

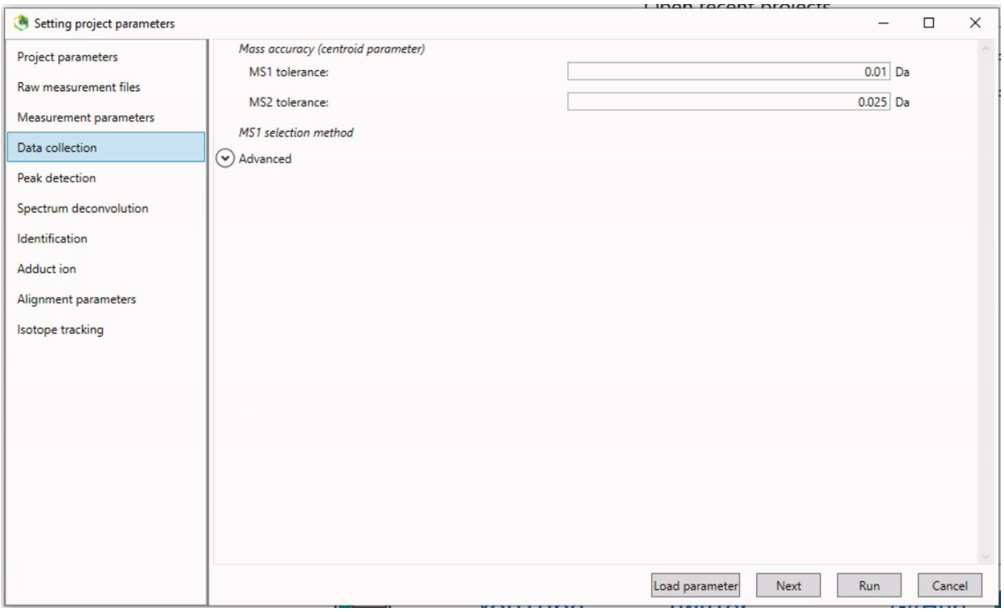
- Ionization type: Soft ionization

- Separation type: LC
- Collision type: CID/HCD
- Data type for MS1 is Profile data and MS/MS is Centroid data for our data, acquired using a Thermo Fisher Scientific Orbitrap Fusion Lumos Tribrid Mass Spectrometer.
- Ion Mode: Either positive or negative ionisation
- Target omics: Metabolomics



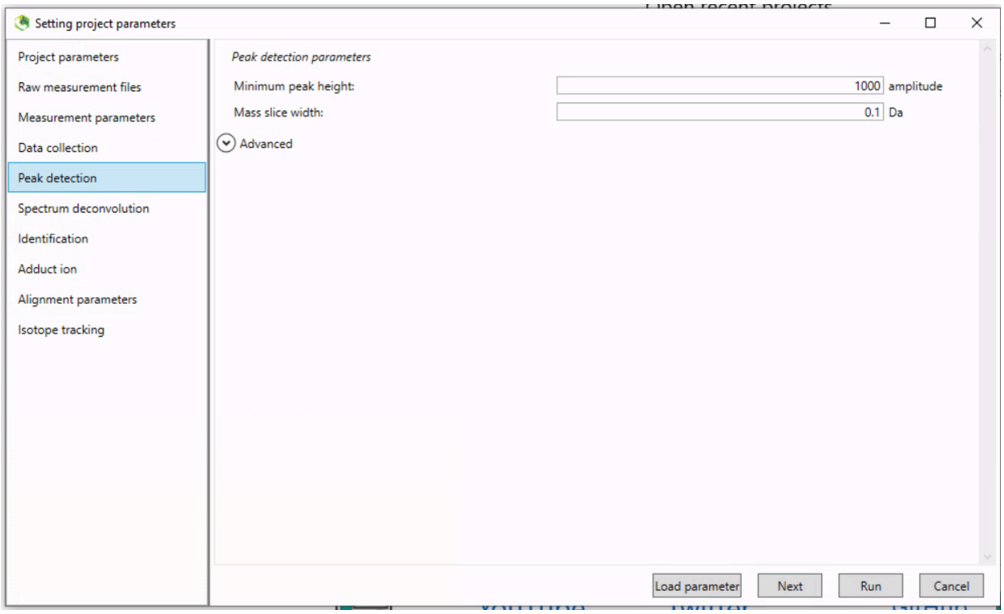
3.5 **Data Collection Parameters**:

- **Mass Accuracy**: Set MS1 tolerances **0.01 Da** and MS2 tolerance to **0.025 Da** for high-resolution data.



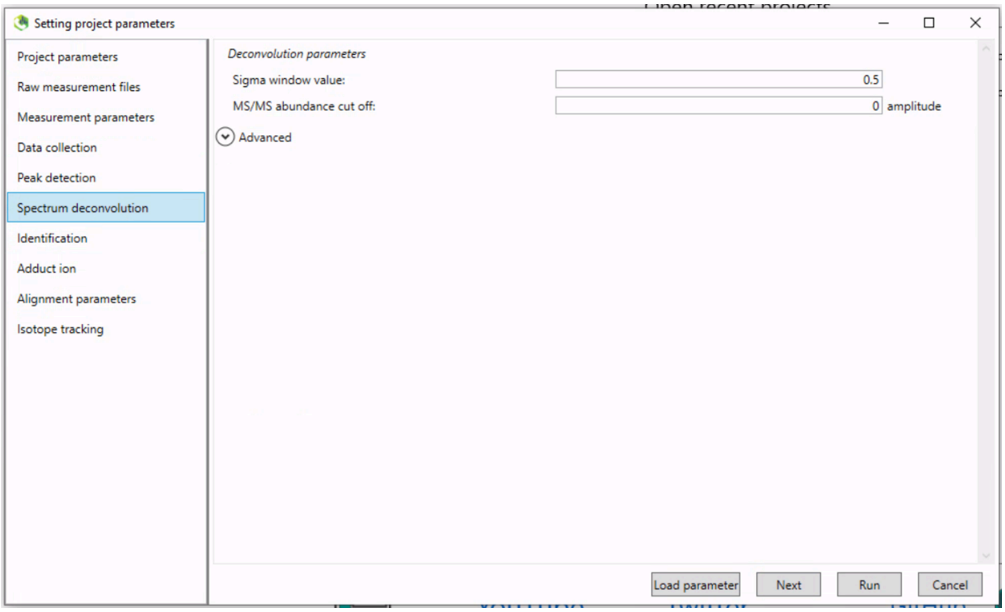
3.6 **Peak Detection Parameters**:

- **Peak Detection**: Adjust the **minimum peak height** based on the signal intensity of your dataset. A common threshold is **1000** for sensitive instruments.
- Leave the **Mass slice width** value to the **default**, along with all options in the drop-down **Advanced** menu.



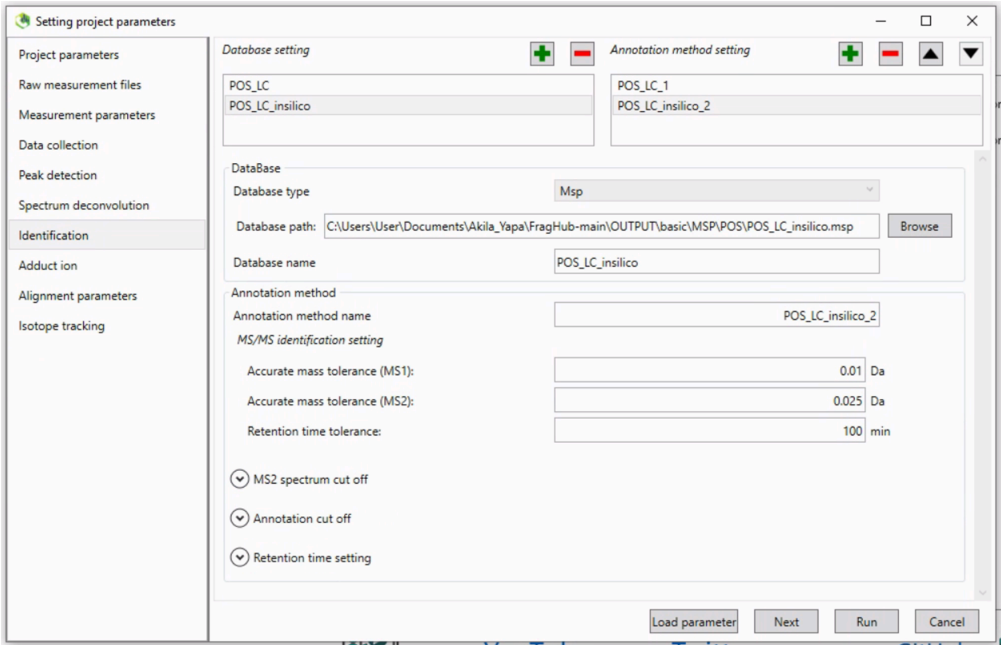
3.7 **Spectrum Deconvolution Parameters**:

- Leave the **default** values.



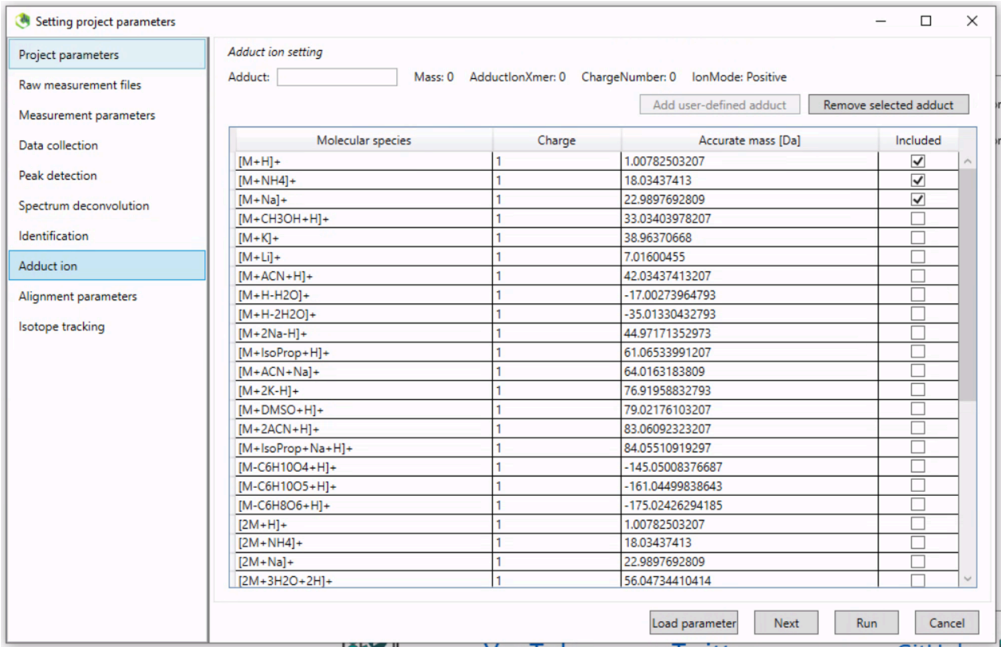
3.8 **Identification Parameters**:

- Load the ***.msp*** library files created using **FragHub.**
- Accurate mass tolerance (MS1)：0.01 Da, Accurate mass tolerance (MS2)：0.025 Da, Retention time tolerance: 100 min.
- Leave other parameters as default.



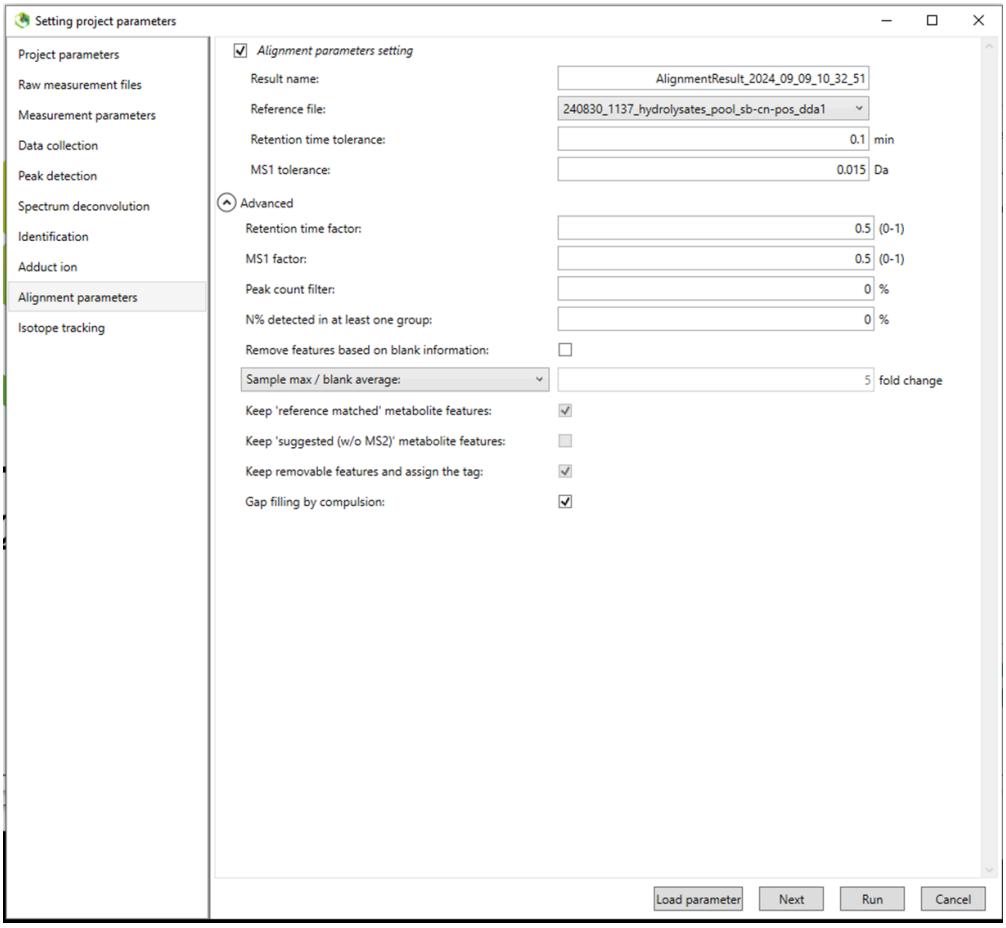
3.9 **Adduct ion Parameters**:

	Positive	Negative
	[M+H]+	[M-H]-
	[M+NH4]+	[M+Na-2H]-
	[M+Na]+	[M+Cl]-



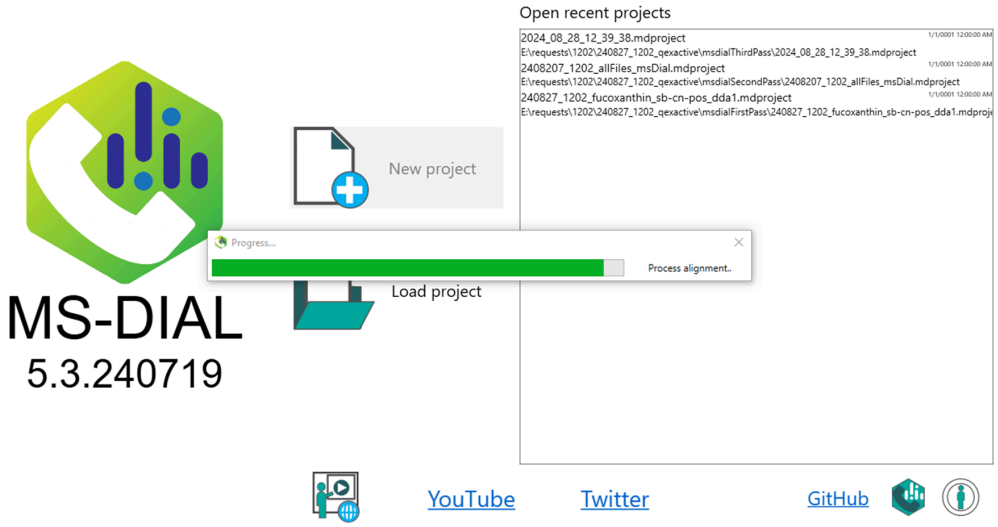
3.10 **Alignment parameters:**

- Set the Reference file to QC sample.
- The 'Retention time tolerance: leave the default.
- Check Gap filling by compulsion, as it improves peak alignment (resulting in fewer highly similar features).



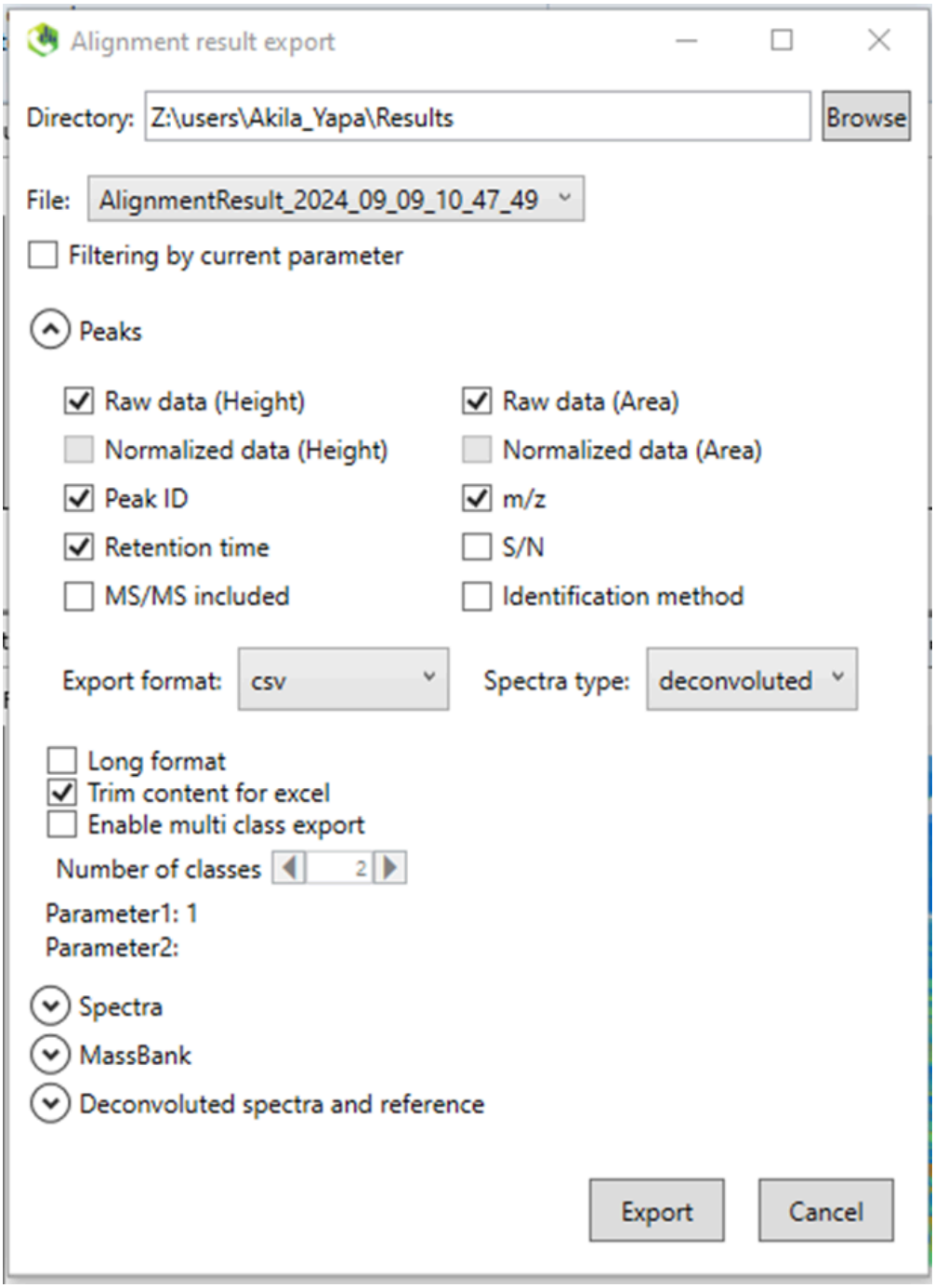
3.11 **Run the pipeline:**

- Once the analysis parameters are set up, click Run to begin the data analysis.



3.12 **Data Export**:

- Export processed data from MS-DIAL in **.CSV** format for statistical and downstream analysis.
- Export the **aligned feature matrix** and annotated compounds for use in statistical programs such as **R** or **Python**.



I. Export Aligned Feature Matrix Select "Export Aligned Results" or a similarly named option. This allows you to export the aligned feature matrix, which contains all the detected peaks across the samples.

- In the export window, select the format as **.CSV**
- Choose the folder or directory where you want to save the file.
- The exported **.CSV** file will include rows representing individual features and columns for sample names, retention times, m/z values, and feature intensities.

II. Export Compound Annotations

- If you have performed compound identification or annotation, you can also export the

identified compounds.
- In the "Export" tab, select "Export Identified Compounds" or a similarly named option.
- Choose the format as .CSV and select the desired directory to save the annotated compound list.
- This export typically contains detailed information about the compounds, including names, molecular formulas, m/z values, retention times, and identification scores.

III. Include Additional Metadata (If Needed)

- Depending on your analysis requirements, you might want to include additional metadata such as sample names, types, or group information. Make sure this metadata is present in your exported files or manually add it to the CSV files using a spreadsheet program.

IV. Verify Exported Data

- Open the exported CSV files in software like Excel, R, or Python to verify that the data has been correctly exported.
- Check the alignment of features and annotations to ensure the integrity of the data.
- Once verified, use the exported feature matrix and annotated compound data in statistical programs such as R or Python for further analysis.

Downstream Data Analysis

4 In a mass spectrometry-based metabolomics analysis, there are several key steps in the workflow that ensure data integrity and accuracy before performing downstream statistical analysis. These steps involve handling blank samples, quality control (QC) data, normalizing the data, and standardizing it for statistical comparison.

4.1 **Deduct Blank Sample Values (Background Subtraction)**

- Calculate the average intensity of each metabolite across the blank samples.
- Subtract this average blank intensity from all other sample intensities to remove background noise and contaminants. If the resulting value for any sample is negative or zero, it can be replaced with a small constant (e.g., half of the minimum non-zero value) to allow for further processing.

4.2 **Coefficient of Variation (CV) for Quality Control (QC) Samples**

- Calculate the Coefficient of Variation (CV) for each metabolite across the QC sample replicates.
- Assess data quality by evaluating the CVs. A low CV (e.g., <20-30%) indicates good technical reproducibility (Zhang et al., 2020, *Analytical chemistry*).

This step helps identify metabolites with high variability, which may be excluded from further analysis if they exceed an acceptable threshold.

4.3 **Ratio of Each Metabolite in Blank vs. Sample**

- For each metabolite, calculate the ratio of the average blank intensity to the sample intensity. This ratio helps identify metabolites that may be dominated by background noise.

- If a metabolite's signal in the sample is close to the blank (e.g., ratio >0.8), consider excluding it from further analysis, as it may represent background noise rather than a true biological signal.

4.4 **Data Normalization**

Normalize the data to correct for technical variability across samples. Common normalization methods include:

- **Total Ion Current (TIC) Normalization**: Adjusts metabolite intensities by the total signal of all metabolites in each sample.
- **Normalization to Internal Standards**: Uses known concentrations of spiked internal standards to correct for differences in sample processing or instrument performance.

- This step ensures that observed differences in metabolite intensities reflect biological variation rather than technical artifacts.

4.5 **Missing Value Imputation**

After normalization, impute missing values to handle gaps in the dataset. The choice of imputation method depends on the nature of the data:

- **K-Nearest Neighbors (KNN) Imputation**: Estimates missing values based on the intensities of similar metabolites or samples.
- **Median/Mean Imputation**: Replaces missing values with the median or mean intensity of that metabolite across samples.
- **Small Constant Value**: For values below the detection limit, replace with a small constant (e.g., half of the minimum observed non-zero intensity) to avoid distorting the data distribution.

4.6 **Data Scaling** (if necessary)

Scale the data to ensure that metabolites with different absolute intensities contribute equally to multivariate analyses. Common scaling methods include:

- **Mean-Centering and Autoscaling**: Subtracts the mean and divides by the standard deviation for each metabolite, bringing all metabolites to a comparable scale.
- **Log Transformation**: Compresses the dynamic range of the data, reducing the impact of high-intensity metabolites and highlighting fold changes.

4.7 **Data Standardization**

- Standardize the data (often using Z-score standardization) to ensure that each metabolite has a mean of 0 and a standard deviation of 1. This step is particularly useful for multivariate analyses, such as PCA or clustering, where you want all features (metabolites) to have an equal contribution.
- Standardization is optional and is applied based on the analysis goals. It is most commonly used when performing multivariate analyses that are sensitive to differences in metabolite scales.

4.8 **Statistical Analysis**

Proceed with downstream statistical analyses using the pre-processed data:

- **Multivariate Analysis**: Use techniques like Principal Component Analysis (PCA) or Partial Least Squares Discriminant Analysis (PLS-DA) to identify patterns, clusters, or differences in metabolite profiles between groups.
- **Univariate Analysis**: Apply t-tests, ANOVA, or other statistical tests to compare individual metabolites between experimental groups.

The choice of analysis depends on the biological questions being addressed and the nature of the dataset.

Protocol references

Dablanc, A., Hennechart, S., Perez, A., Cabanac, G., Guitton, Y., Paulhe, N., Lyan, B., Jamin, E.L., Giacomoni, F. and Marti, G., 2024. FragHub: A mass spectral library data integration workflow. *Analytical Chemistry*, *96*(30), pp.12489-12496.

Takeda, H., Matsuzawa, Y., Takeuchi, M., Takahashi, M., Nishida, K., Harayama, T., Todoroki, Y., Shimizu, K., Sakamoto, N., Oka, T. and Maekawa, M., 2024. MS-DIAL 5 multimodal mass spectrometry data mining unveils lipidome complexities. *bioRxiv*, pp.2024-02.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O. and Arita, M., 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods*, *12*(6), pp.523-526.

Zhang, X., Dong, J. and Raftery, D., 2020. Five easy metrics of data quality for LC–MS-based global metabolomics. *Analytical chemistry*, *92*(19), pp.12925-12933.