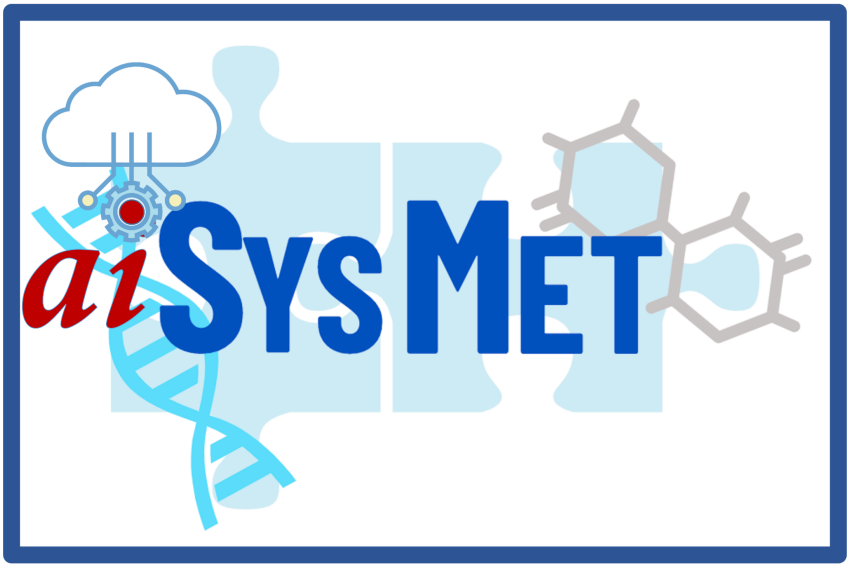


USER MANUAL

August 2024

OMICSCRAFT

TABLE OF CONTENTS

INTRODUCTION

aiSysMet is a powerful AI-powered platform for analysis of metabolomics data and integration of multi-omics and imaging data. It leverages advanced biomedical data analytics to accelerate biomarker discovery. In this user manual, we describe various components of aiSysMet outlined in the following categories: management utilities, data import modules, data processing modules (MetCraft), metabolite annotation modules (MetaboQuest), data/integrative analysis modules (IntSys), and pipeline builder. Furthermore, demo datasets and projects are described.

MANAGEMENT UTILITIES

The following management utilities allow users to create and manage cloud data storage and projects.

Data Manager

- Manages cloud storage space
- Allows users to upload and store data projects
- Facilitates efficient workflow execution by ensuring data accessibility from low-latency storage (both processed and unprocessed data). Repetitive uploading of large raw data files from local storage degrades performance
- Offers user friendly interface for uploading, renaming, deleting, decompressing and downloading files and directories
- Operates with an intuitive, operative, and system-style interface similar to Google Drive

Project Manager

- Organizes data in the user space (**raw uploaded data** or **pipeline generated outputs**) on a project basis for easy management and retrieval.
- Facilitates collaboration among researchers by sharing project pipelines, raw data, and results in a transparent manner.

MODULES

aiSysMet's modules are grouped into categories as outlined below. Please note that all modules listed below are available to the user with a full subscription of aiSysMet. Users can choose to subscribe a partial subscription of aiSysMet by choosing one or more from three modules (MetCraft, MetaboQuest, and IntSys) to be included in aiSysMet.

Data Import Modules

This category consists of two modules for either uploading data from local and cloud storage or retrieval of data from pre-specified databases.

Data Upload

- Allows users to upload raw metabolomics data for data processing, processed metabolomics data for metabolite annotation, and any processed omics data from local or cloud storage spaces
    - Raw/unprocessed metabolomics LC-MS/MS data in mzXML or mzML formats can be uploaded along with a list of precursor m/z values for metabolite annotation and/or group/sample labels for marker selection.
    - Processed metabolomics data including MS/MS, MS, or GC-MS in plain text format along with a list of m/z precursor values in either plain text or csv formats for metabolite annotation. In addition, processed metabolomics data can be directly entered into a window interface provided.
    - Processed any omics data along with additional files corresponding to the omics data including sample labels or group information, designation of batches, precursor m/z list, etc. for marker selection.
- Automatically identifies data types to determine which subsequent modules are allowed to build pipelines

Retrieval from Database

- Searches for preprocessed data in public repositories such as TCGA, CPTAC, and TCIA
- User can specify various search criteria including:
    - Program (TCGA, CPTAC, TCIA)
    - Primary site (breast, liver, ovarian, lung, brain, etc.)
    - Disease type

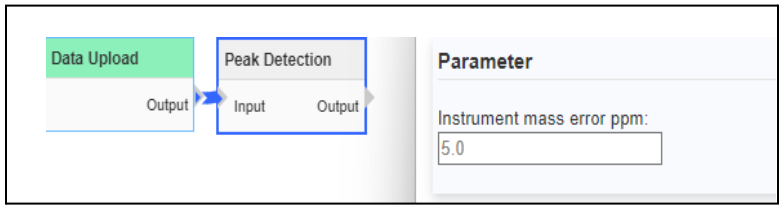
    - Omics data (mRNA-seq, miRNA-seq, proteomics, phosphoproteomics, etc.)
    - Imaging data
    - Grouping features (based on sample annotations such as disease, age, race, days-to-death, etc.)
- Displays available cohorts that meet user-defined selection criteria
- Displays a summary of imported data including demographics of study subjects

MetCraft Modules - Metabolomics Data Processing

This category consists of the following modules to process raw LC-MS/MS metabolomics data or apply various data treatment methods to any processed omics data.

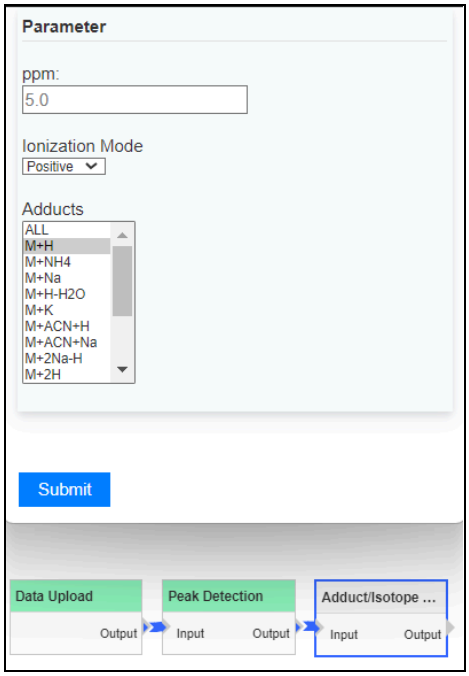
Peak Detection

- Analyzes unprocessed metabolomics data in mzXML or mzML formats to perform peak detection including peak pricing, peak integration and peak alignment
- Detects ion signals based on signal to noise ratio
- Reconstructs peak shapes using cubic spline interpolation



Adduct/Isotope Recognition

- Facilitates subsequent metabolite annotation steps by recognizing adducts and isotope through peak clustering. This is important because one analyte may generate multiple peaks with distinct m/z values due to the effects of isotopes, adducts and neutral-loss fragments.



Outlier Screening

- Applies Principal Component Analysis (PCA) to visualize samples that look different from the majority
- Identifies outliers that should be excluded in subsequent analyses

mz	quality	cd-9oc5y	cd-9vokp	cd-6kuct	adduct	pcgroup	isotopes
526.2847	0.4998	3590860.0000	0.0000	215409.5156		11	[220][M]+
287.2545	0.4907	2458645.2500	0.0000	2091807.0000		11	[223][M+2]+
581.2646	0.4880	19520168.0000	561874.0625	0.0000		11	
457.2576	0.9759	3027132.2300	132997.9219	4141179.2500		11	
463.3069	0.4910	591437.8750	0.0000	5249994.5000		11	[224][M+4]+
509.2485	0.0000	828089.2500	0.0000	0.0000		11	
484.3280	0.0000	0.0000	239863.5469	0.0000		11	
538.3089	0.9860	2171567.5000	911295.1250	504281.4063		11	
540.3265	0.9886	33299102.0000	5450043.5000	10871803.0000		11	
453.2324	0.0000	3612371.2300	0.0000	0.0000		11	[223][M+4]+
485.2576	0.4793	1410142.8750	0.0000	1994638.3750		11	
509.2581	0.4978	5331424.0000	0.0000	371679.9375		11	
361.2024	0.0000	0.0000	0.0000	904217.5625		11	
585.8029	0.0000	365940.7500	0.0000	0.0000		11	
531.2943	0.0000	0.0000	0.0000	659419.6250		11	
545.2168	0.0000	0.0000	0.0000	625279.0625		12	
495.2792	0.4898	1918740.7500	360971.0000	0.0000		12	[228][M]+
279.2107	0.4877	727077.5625	0.0000	440147.1563		12	
543.3330	0.0000	947284.5625	0.0000	0.0000		12	
459.2027	0.9862	1037061.5625	1302072.7500	1174486.8750		12	
404.2480	0.0000	1111317.5000	0.0000	0.0000		12	
551.3232	0.0000	0.0000	0.0000	202812.0938		12	
449.2004	0.9804	1002135.3750	186648.0781	807259.3125		12	
381.2263	0.9855	4998033.5000	437025.2813	3377724.5000		12	
441.2863	0.0000	1974576.1250	0.0000	0.0000		12	[227][M+4]+
497.2950	0.0000	75418432.0000	0.0000	0.0000	[M+H]+ 496.29	12	[228][M+2]+
565.2821	0.0000	2648714.0000	0.0000	0.0000		12	[229][M]+
360.2222	0.0000	1646047.2300	0.0000	0.0000		12	
463.3072	0.0000	4367910.0000	0.0000	0.0000	[M+H]+ 462.3	12	
445.2965	0.0000	670787.1250	0.0000	0.0000	[M+H-H2O]+ 462.3	12	
538.2852	0.0000	1066897.0000	0.0000	0.0000		12	
437.2916	0.0000	1605743.0000	0.0000	0.0000		12	[227][M]+
424.2476	0.0000	652354.8750	0.0000	0.0000		12	
538.3219	0.0000	1979313.0000	0.0000	0.0000	[M+ACN+H]+ 496.29	12	

Data Filter

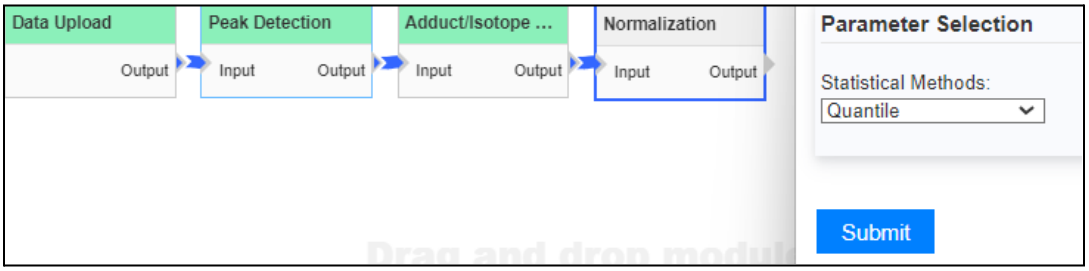
- Allows users to select a subset of features for subsequent analyses
- Features are removed based on: (1) a user-specified threshold for a coefficient of variation across all selected subjects; and (2) a threshold for the percentage of missing values

Missing Value Imputation

- Uses methods such as mean value, integer, k-nearest neighbor (KNN), and Random Forest (RF) to impute missing values, such as a peak missing in a small subset of samples but present in most.

Normalization

- This module provides access to several data normalization methods, including quantile normalization, median normalization, mean normalization, CycLoess, global robust linear regression (RLR), and global intensity normalization.



Batch Correction

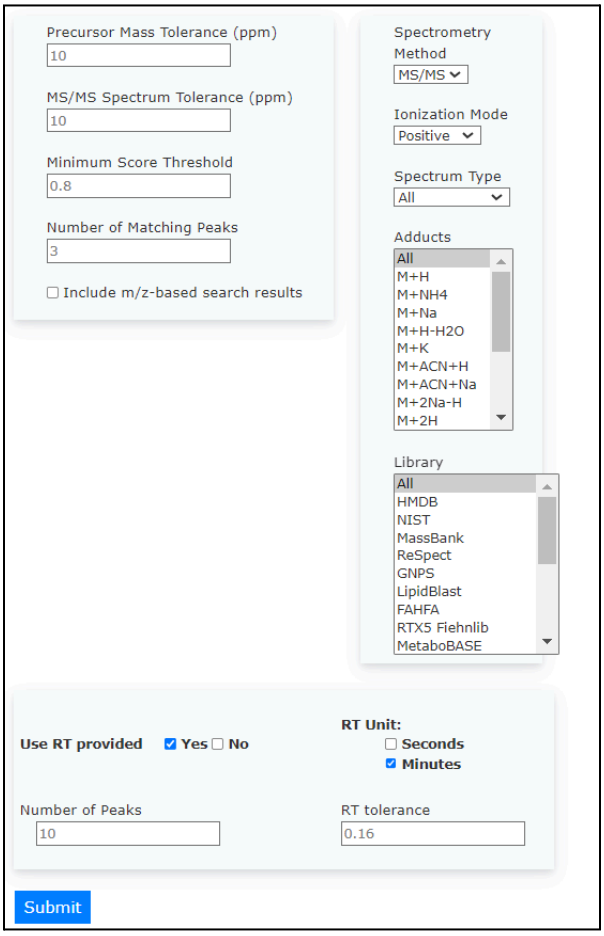
- Uses empirical Bayes frameworks to adjust data from large-scale studies affected by running order or batch acquisition

MetaboQuest Modules - Metabolite Annotation

This category includes several modules that perform mass-based search and spectral matching for metabolite annotation. Other modules that apply Isotopic pattern analysis, network-based annotation, IF-THEN rule, and compound fingerprint prediction help organize and rank putative metabolite IDS.

Spectral Matching

- Searches for putative metabolite IDs by matching EI-MS or MS/MS spectra with our spectral database (SpectDB)



Compound Fingerprint Prediction

- Uses a deep/machine-learning model to predict compound fingerprints based on MS/MS data

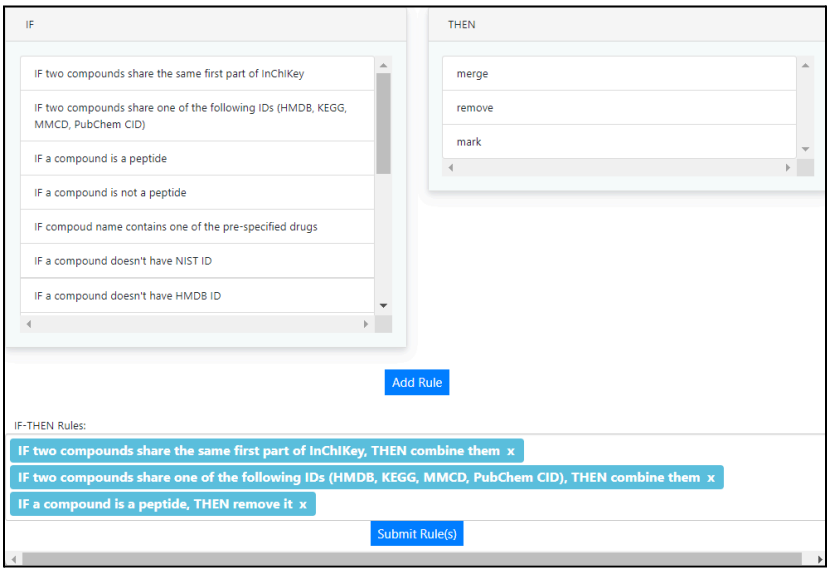
- Utilizes predicted fingerprints to rank candidate metabolites
- Designed for analytes that lack reference measurements in spectral libraries or have low spectral matching scores

Mass-Based Search

- Enables search for putative metabolite IDs in MetDB based on m/z values
- Users are able to enter m/z values or use uploaded or processed data from a preceding module to search for putative IDs
- Calculates monoisotopic mass values based on the m/z values and user-specified adducts, ionization mode, mass tolerance in ppm.

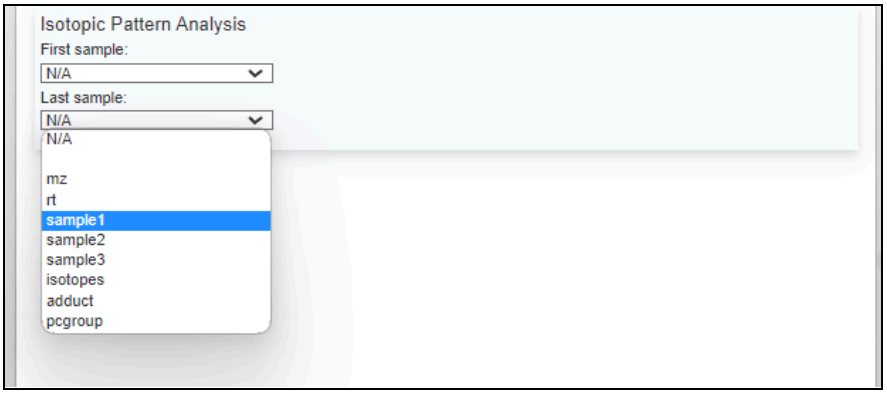
IF-THEN Rule

- Allows users to select IF-THEN rules in order to combine, remove, or mark putative metabolite IDs.



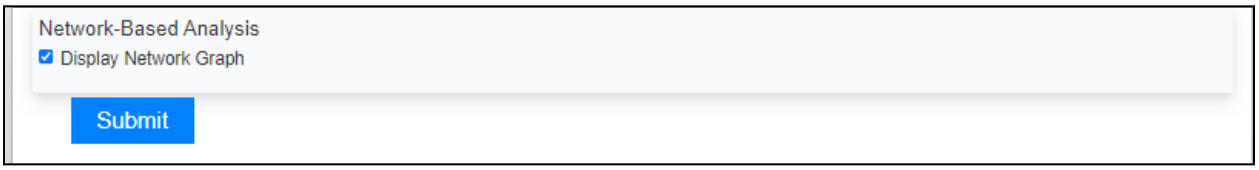
Isotopic Pattern Analysis

- Assigns scores to putative metabolite IDs based on their isotopic patterns
- Compares potential IDs with varying elemental formulas
- Calculates scores by comparing observed isotopic patterns from MS spectra with theoretical isotopic patterns



Network-Based Annotation

- Assigns scores to putative IDs using a network-based method
- Constructs a metabolic network by extracting biochemical pathway information from databases such as MetaCyc and KEGG
- Assigns probability scores to putative IDs, indicating the likelihood of their accuracy for a peak.

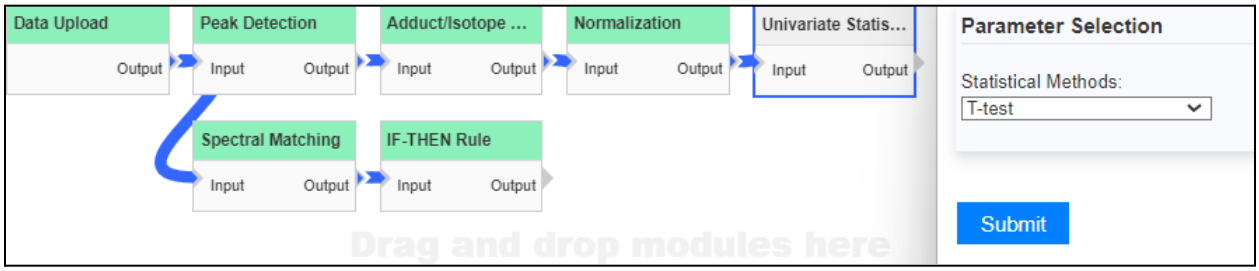


IntSys Modules - Multi-Omics Data Integration

The modules in this category allows users to identify significantly altered metabolites or multi-omics features by integrative analysis. Each module can be used for analysis of single omics or multi-omics data. Modules are linked to tools for visualization including ROC curves, Box plots, Volcano plots, Heatmaps, t-SNE, and Hierarchical clustering.

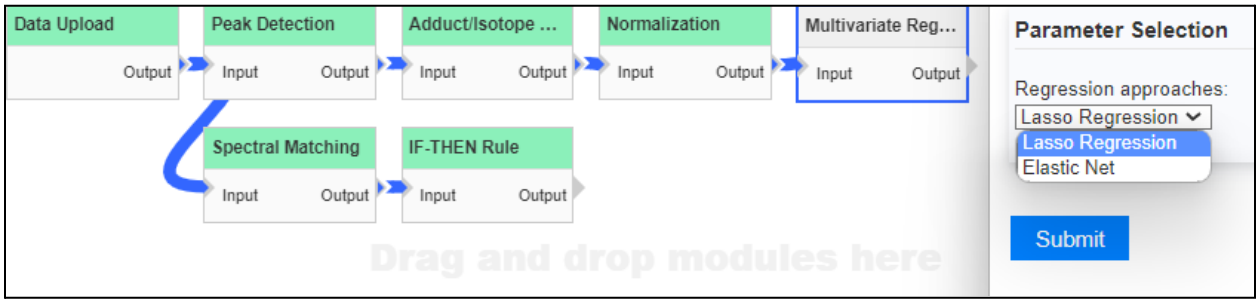
Univariate Statistical Analysis

- Analyzes preprocessed single omics or multi-omics imaging data using parametric (Student t-test) or non-parametric (Mann-Whitney U-test) statistical methods to identify significantly altered features between two independent groups of samples
- Analyzes matched/paired samples (i.e., tumor and adjacent non-tumors) using parametric (paired t-test) or non-parametric (Wilcoxon signed-rank test) to identify significantly altered features between two independent groups of samples
- Multi-omics features are simply concatenated for univariate analysis.



Multivariate Regression Analysis

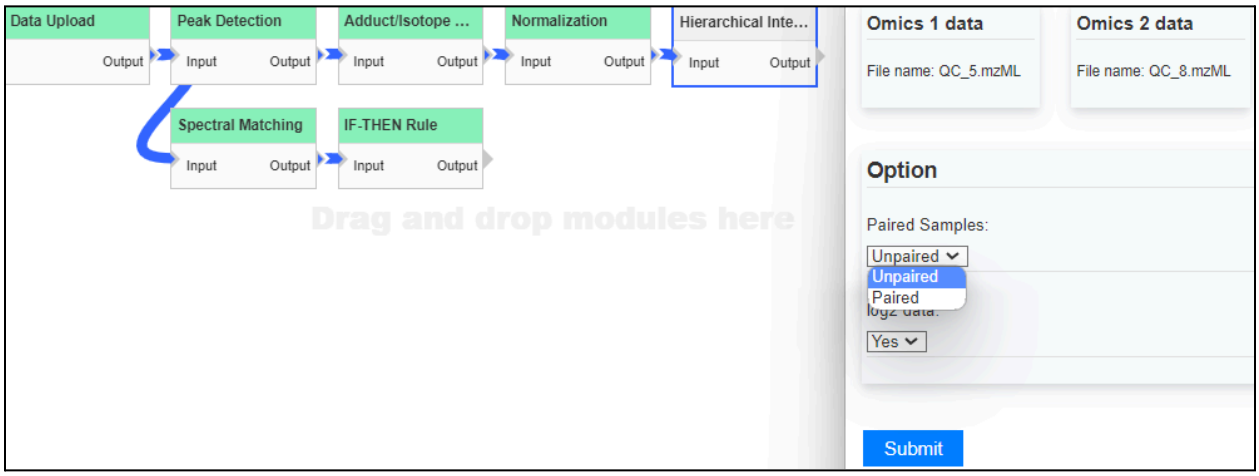
- Allows users to apply multivariate analysis (Lasso Regression and Elastic Net) to select a panel of disease-associated features.



Hierarchical Integrative Analysis

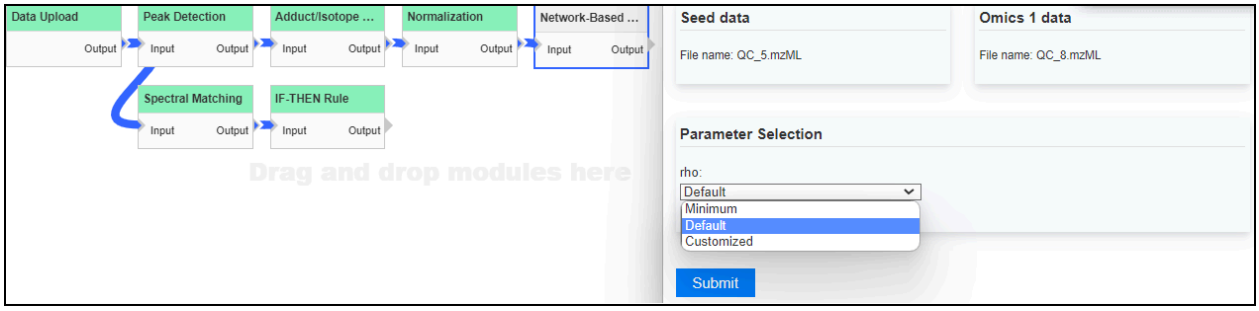
- Associates analytes measured in multi omics studies to discover novel relationships about disease status
- Utilizes modeling approaches with penalized likelihood methods and EM algorithms

- Explores biological relationships between molecular features and their effects on a clinical outcome



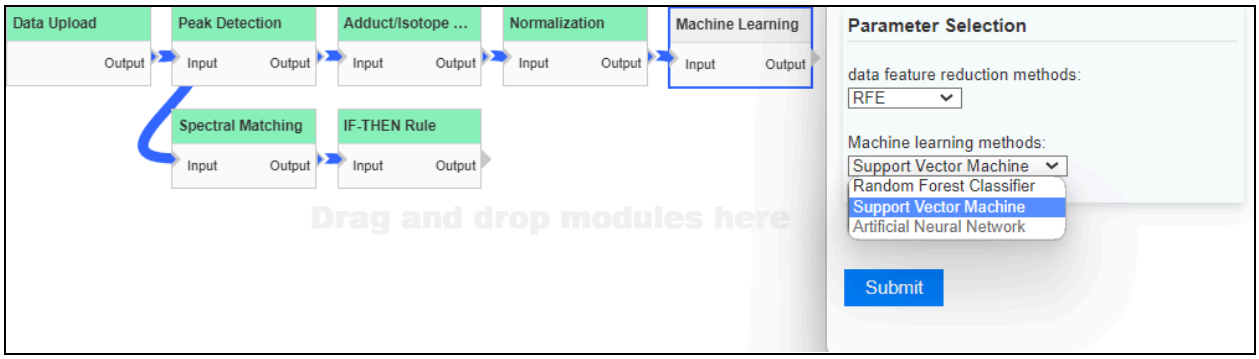
Network-Based Analysis

- Uses network-based methods for differential feature analysis of analytes in single omics, multi-omics, or imaging data
- Uses differential networks to compare correlations between analyte pairs within disease groups vs. control groups
- Helps users understand changes in pairwise interactions of analytes related to disease



Machine Learning

- Uses two machine learning methods: support vector machine and random forest
- Uses the recursive feature elimination method, which selects disease-associated features from single or multi-omics data
- Standardized multi-omics features are combined into vectors for each sample to identify features that predict disease status



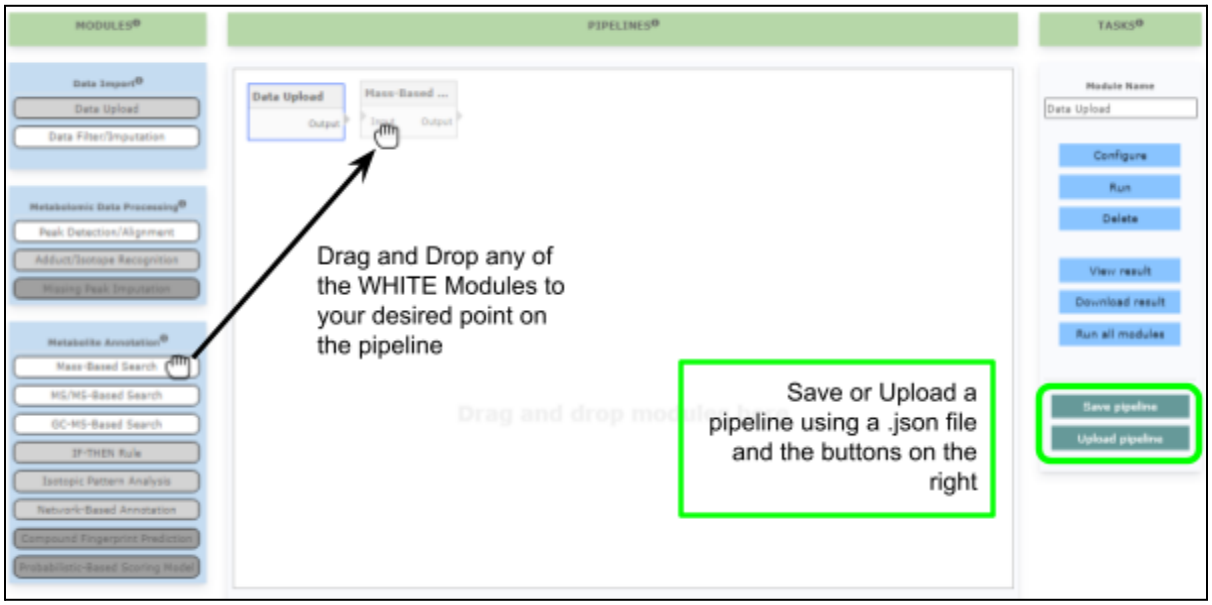
Generative AI

Note: *This module is coming soon*

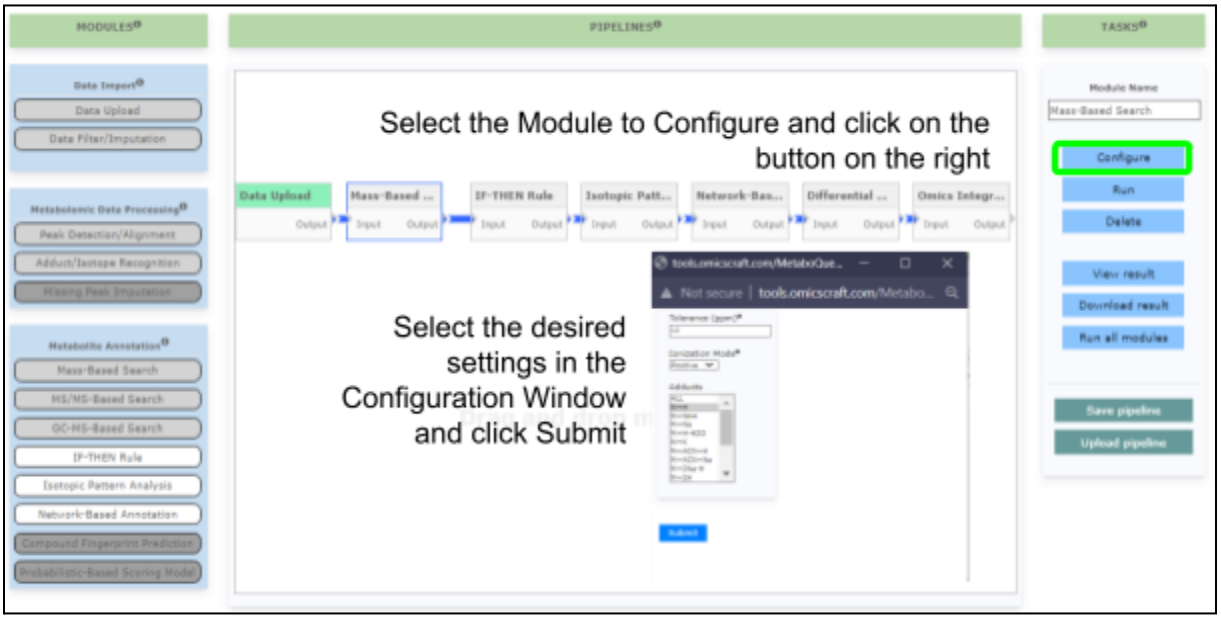
PIPELINE BUILDER

Steps to Build a Pipeline

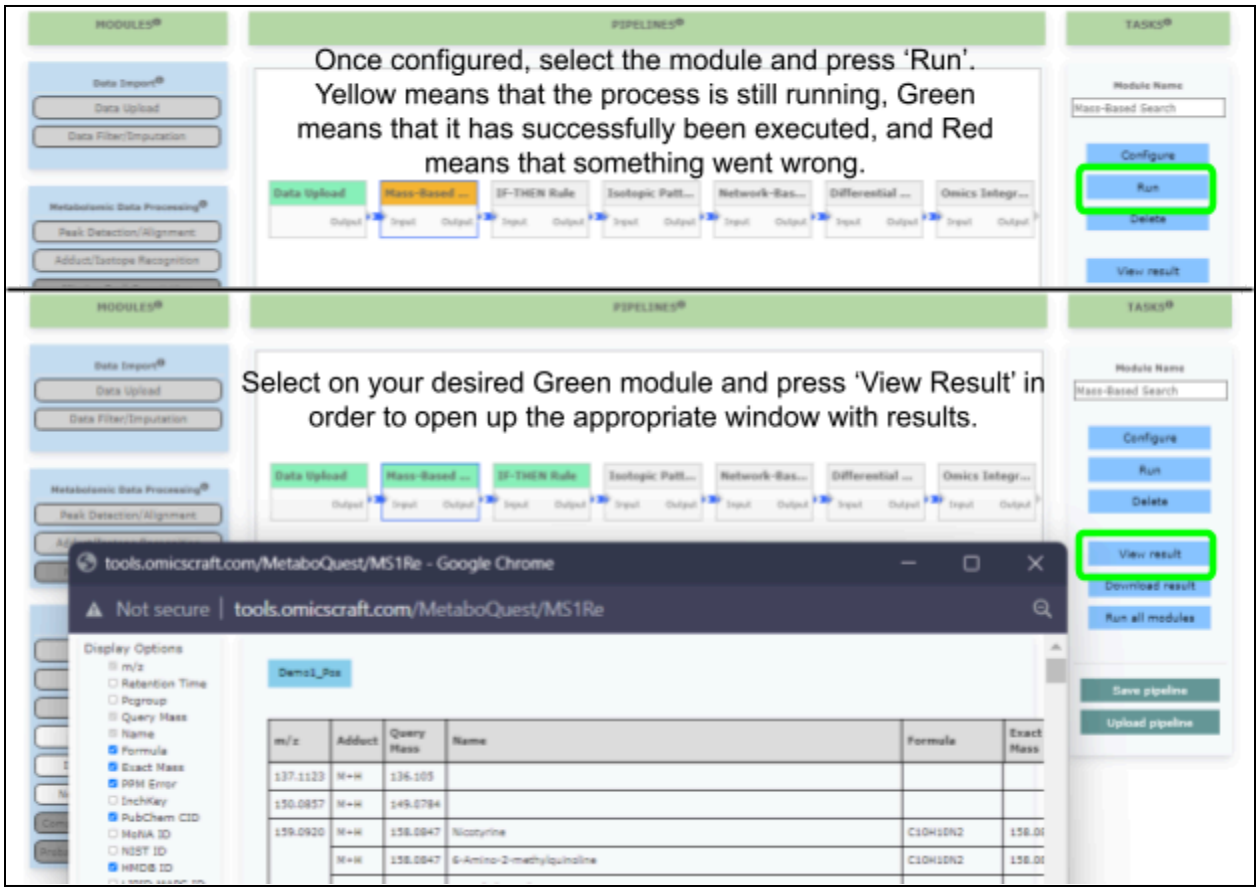
1. Lay down the tools/modules to create or upload an existing pipeline. Users can create a pipeline, save it and upload it later on. Modules in pipelines will require reconfiguration after being uploaded again.



2. Configure modules in order. Click on the individual modules, and then click **Configure**.



3. Execute individual modules by clicking on **Run** in the right pane. You can also see the output by clicking **View Result** if you configured the modules successfully.
4. Click on **Save pipeline** to download the pipeline to your local computer in json format and use it later.



Notes on Pipeline Builder

- The components that cannot be inserted or appended to the current pipeline are grayed. Through this, the ***Pipeline builder*** ensures that the composition of the pipeline follows a logical workflow.
- Therefore, the user should observe the proper sequence for bringing components into the ***Pipeline builder***.
- After placing a module in the ***Pipeline builder***, it must be configured with the appropriate processing settings before execution. To do this, use the **Configure** button in the ***Command Pane***. The module cannot be executed unless it is configured.

- After properly configuring a module, click the **Run** button in the ***Command Pane*** to execute a module.
- The **Progress Window** shows the current operations, selections, and the status of the operations, if available.
- The module execution status can also be determined using a color code, as explained later in this tutorial.
- Click the **Delete** button in the ***Command Pane*** to remove an unwanted module from the ***Pipeline builder*** area.
- Click the **Reset Pipeline** button to clear any existing pipelines in the ***Pipeline builder.***

DEMO DATA

MetCraft Demo Data - Unprocessed Metabolomics Data

Demo1: a folder consisting of: (1) Demo1a_mzXML_pos: a folder of 8 mzML files acquired by metabolomics analysis of 8 QC samples using LC-MS/MS in the positive mode and a precursor file that indicates the m/z and RT values of all precursor ions expected for each mZXML file. This demo dataset can be used for annotation of the analytes indicated in the precursor file using the Spectral Matching module. The module may extract the MS/MS spectra guided by the m/z provided in the precursor file. Users may choose to use the RT values in the precursor file or let the module automatically choose high quality MS/MS spectra across all scans. This demo dataset can also be used for peak detection using the Peak Detection module; and (2) Demo1b_mzML_neg: the same samples as Demo7a analyzed in the negative mode.

Demo2: a folder consisting of: (1) Demo2a_mzML_pos: a folder of 8 mzML files acquired by lipidomics analysis of 8 QC samples using LC-MS/MS in the positive mode and a precursor file that indicates the m/z and RT values of all precursor ions expected for each mZXML file. This demo dataset can be used for annotation of the analytes indicated in the precursor file using the Spectral Matching module. The module may extract the MS/MS spectra guided by the m/z provided in the precursor file. Users may choose to use the RT values in the precursor file or let the module automatically choose high quality MS/MS spectra across all scans. This demo dataset can also be used for peak detection using the Peak Detection module; and (2) Demo2b_mzML_neg: the same datasets as Demo8a analyzed in the negative mode.

Demo3: a folder consisting of: (1) Demo3a_mzML_pos: a folder of 3 mzXML files acquired by LC-MS/MS in the positive mode and a precursor file that indicates the m/z and RT values of all precursor ions expected for each mZXML file. This demo dataset can be used for annotation of the analytes indicated in the precursor file using the Spectral Matching module. The module may extract the MS/MS spectra guided by the m/z provided in the precursor file. Users may choose to use the RT values in the precursor file or let the module automatically choose high quality MS/MS spectra across all scans. This demo dataset can also be used for peak detection using the Peak Detection module; and (2) Demo3b_mzXML_pos: the same datasets as Demo9a but in mzXML format.

MetaboQuest Demo Data - Processed Metabolomics Data

Demo4: a folder consisting of: (1) Demo4a folder with two folders for MS spectra and MS/MS spectra acquired by LC-MS/MS in the negative mode as well as a file listing the precursor m/z values corresponding to the spectra; (2) Demo 4b.txt that consists of MS spectra for batch processing along with a file listing all the precursor m/z values; and (3) Demo 4c.txt that consists of MS/MS spectra for batch processing along with a file listing all the precursor m/z values.

Demo5: a folder consisting of: (1) Demo5a folder with two folders for MS spectra and MS/MS spectra acquired by LC-MS/MS in the positive mode as well as a file listing the precursor m/z values corresponding to the spectra; (2) Demo 5b.txt that consists of MS spectra for batch processing along with a file listing all the precursor m/z values; and (3) Demo 5c.txt that consists of MS/MS spectra for batch processing along with a file listing all the precursor m/z values.

Demo6: a folder consisting of: (1) Demo6a_EI.txt: a set of 5 EI spectra acquired by GC-MS. This demo dataset can be used for batch metabolite annotation using the Spectral Matching module by choosing the GC-MS platform; (2) Demo6b_EI.txt: the same datasets as Demo6a but combined in one file. This demo dataset can be used for metabolite annotation using the Spectral Matching module by choosing the GC-MS platform.

Demo7: a folder consisting of: (1) Demo7a_peaks_pos.csv: a small subset of peaks detected in a metabolomics study using LC-MS in the positive mode. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search module; and (2) Demo7b_peaks_pos.csv: the same set of peaks as Demo1a but with adduct and isotope information provided to some of the peaks. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search followed by the Isotopic Pattern Analysis module.

Demo8: a folder consisting of: (1) Demo8a_peaks_pos.csv: another small subset of peaks detected in a metabolomics study using LC-MS in the positive mode. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search module; and (2) Demo8b_peaks_pos.csv: the same set of peaks as Demo8a with adduct and isotope information provided to some of the peaks. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search followed by the Isotopic Pattern Analysis module.

Demo9: a folder consisting of: (1)Demo9a_peaks_neg.csv: an entire set of peaks detected by analysis of metabolomics data acquired using LC-MS in the negative mode. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search module; (2) Demo9b_peaks_neg.csv: the same set of peaks as Demo9a with adduct and isotope information provided to some of the peaks. This demo dataset can be used to perform metabolite annotation using the Mass-Based Search followed by the Isotopic Pattern Analysis module.

Demo10: a folder consisting of: (1) Demo10a_MSMS_pos: a folder of 12 files each consisting of an MS/MS spectrum acquired in the positive mode. In each file, the first line presents the precursor m/z and retention time (RT) values separated by a comma. The next lines present a list of m/z and intensity pairs separated by space and entered one pair per line. This demo dataset can be used for metabolite annotation using the Spectral Matching module by uploading the 12 MS/MS spectra together; and (2) Dem10b_MSMS_pos.txt: all 12 MS/MS spectra from Demo10a listed in one file. In the file, the first line presents the precursor m/z and retention time (RT) values separated by a comma. The next lines present a list of m/z and intensity pairs

separated by space and entered one pair per line. This format is repeated for all remaining MS/MS spectra, each separated by a blank line. This demo dataset can be used for metabolite annotation using the Spectral Matching module.

Demo11: a folder consisting of: (1) Demo11a_MSMS_neg: a folder of 4 files each consisting of an MS/MS spectrum acquired in the negative. In each file, the first line presents the precursor m/z and retention time (RT) values separated by a comma. The next lines present a list of m/z and intensity pairs separated by space and entered one pair per line. This demo dataset can be used for metabolite annotation using the Spectral Matching module by uploading the 4 MS/MS spectra together; and (2) Demo11b_MSMS_neg.txt: all 4 MS/MS spectra from Demo5a listed in one file. In the file, the first line presents the precursor m/z and retention time (RT) values separated by a comma. The next lines present a list of m/z and intensity pairs separated by space and entered one pair per line. This format is repeated for all remaining MS/MS spectra, each separated by a blank line. This demo dataset can be used for metabolite annotation using the Spectral Matching module.

IntSys Demo Data - Multi-Omics Data for Integrative Analysis

Demo12: a folder consisting of three omics (metabolomics, glycomics, and proteomics) datasets acquired from the same set of samples. Each dataset, separately or in combination, can be used to test the Data/Integrative Analysis modules.

Demo13: a folder consisting of three omics (metabolomics, glycomics, and proteomics) datasets acquired from the same set of samples. Each dataset can be used to test the Differential Analysis module. Each dataset, separately or in combination, can be used to test the Data/Integrative Analysis modules.

Demo14: a folder consisting of three omics (mRNA expression profile, miRNA expression profile, and metabolomics profile) datasets acquired from the same set of samples. Each dataset, separately or in combination, can be used to test the Data/Integrative Analysis modules.

Demo15: a folder consisting of two omics (mRNA expression profile and miRNA expression profile) datasets acquired from the same set of samples comprising tumor and non-tumor pairs. Each dataset, separately or in combination, can be used to test the Data/Integrative Analysis modules.

Demo16: a folder consisting of preprocessed metabolomics data acquired in the positive mode, proteomics data, and glycomics data from an overlapping set of samples and three groups of annotation files. The datasets can be used to test the Data/Integrative Analysis modules.

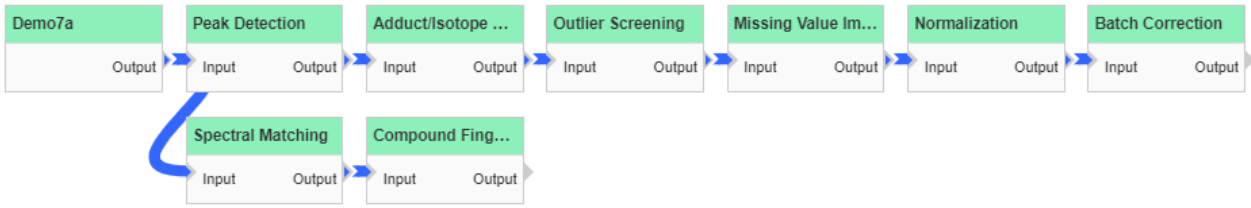
Demo17: a folder consisting of preprocessed metabolomics data acquired in the positive mode and a group annotation file. The datasets can be used to test the Data/Integrative Analysis modules.

Demo18: a folder consisting of preprocessed metabolomics data acquired in the positive mode and a group annotation file. The datasets can be used to test the Data/Integrative Analysis modules.

Demo19: a folder consisting of preprocessed metabolomics data acquired in the negative mode and a group annotation file. The datasets can be used to test the Data/Integrative Analysis modules.

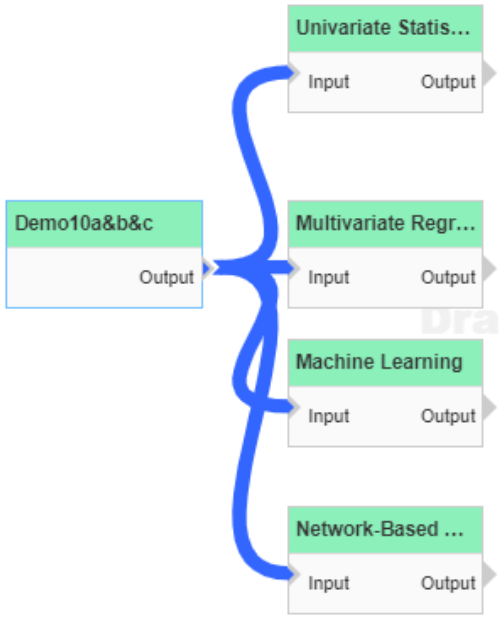
DEMO PROJECTS

MetCraft Demo Project - Metabolomics Data Processing



MetaboQuest Demo Project - Metabolite Annotation

IntSys Demo - Multi-Omics Data Integration



DEMO VIDEOS

MetCraft Demo Video - Metabolomics Data Processing

MetCraft Demo.mp4

MetaboQuest Demo Video - Metabolite Annotation

MetaboQuest Demo.mp4

IntSys Demo Video - Multi-Omics Integration

IntSys Demo.mp4