USER MANUAL

March 2024

TABLE OF CONTENTS

INTRODUCTION

OMICSCRAFT software tools consist of management utilities and data analysis modules. The management utilities enable users to create and manage cloud data storage and projects. The data analysis modules can be interconnected to build pipelines via a pipeline builder that allows users to drag and drop, connect, configure, run, etc., various modules. The following describes the management utilities, data analysis modules embedded in different subscriptions (MetaboQuest, MetCraft, and iSysMet), the pipeline builder, and various demo files.

MANAGEMENT UTILITIES

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

Data Manager

This utility allows users to manage cloud storage space to which they can upload their data and store data analysis pipelines along with intermediate outputs and final results. It is designed to improve data access performance, by storing user data closer to the execution environment and hence reducing latency. Repetitive upload of large raw data files from user's local storage degrades performance. The *Data Manager* has a key role in achieving an efficient workflow execution because all data, raw or processed, are accessible from low-latency storage. It offers user interface (UI) for uploading, renaming, deleting, decompressing and downloading of files and directories. These features are implemented via intuitive operating system style interfaces similar to the popular Cloud data management services such as Google Drive.

Project Manager

This utility has two major functions: (1) organize data in the user space (raw uploaded data or pipeline generated outputs) on project basis to support easy management and hassle-free retrieval; and (2) facilitate collaboration among researchers by allowing sharing of project pipelines, raw data, and results in a transparent and intuitive manner.

MODULES

The modules are grouped into categories as outlined below. Please note that all modules listed below are available to the user with MetCraft subscription whereas MetaboQuest and iSysMet subscriptions give the user access to a subset of the modules only.

Data Import Modules

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

Data Upload

This module allows users to upload raw or processed metabolomics or other omics data along with the annotation of the samples from local or cloud storage spaces. The module automatically recognizes the input data type and determines the subsequent modules to be included in the pipeline.

Retrieval from Database

This module searches for preprocessed data deposited in public repositories. Users can use this module to search for preprocessed data deposited in TCGA, CPTAC, and TCIA. The user can specify the program (TCGA, CPTAC, TCIA), primary site (breast, liver, ovarian, lung, brain, etc.), disease type, omics data (mRNA-seq, miRNA-seq, proteomics, phosphoproteomics, etc.) or imaging data, grouping feature based on sample annotation (disease, age, race, days-to-death, etc.). The module displays the available cohorts that meet the selection criteria defined by the user. After importing the data through the ***Retrieval from Database*** module, the user receives a summary of the imported data. This summary includes the demographics of the study subjects that correspond to the imported data.

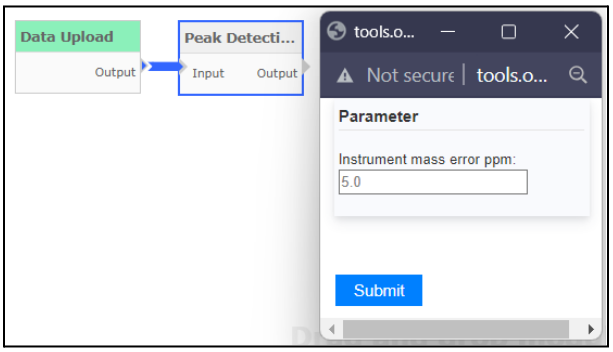
Data Processing Modules

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

Peak Detection

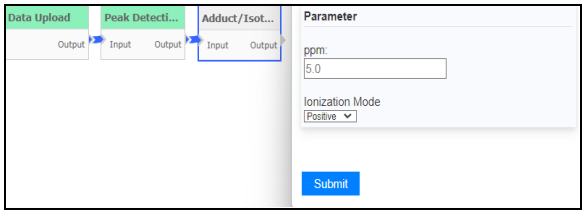
This module allows users to perform peak detection including peak picking, peak integration, and peak alignment. Users are able to upload unprocessed data in the form of MZXML or mzML

files. It detects ion signals based on signal-to-noise-ratio and reconstructs the corresponding peak shape by cubic spline interpolation.



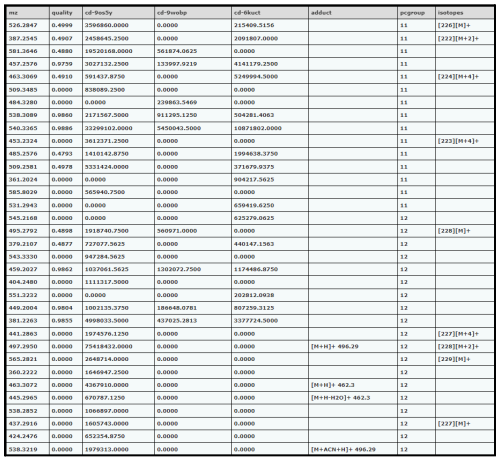
Adduct/Isotope Recognition

This module clusters peaks belonging to the same metabolite into one group by annotating corresponding adduct forms and isotopic peaks. Due to the effects of isotopes, adducts, and neutral-loss fragments, one metabolite often generates multiple peaks with distinct m/z values. Recognition of such clusters of peaks facilitates metabolite annotation.



Outlier Screening

This module applies PCA to visualize samples that look different from the majority of the samples, thereby identifying outliers that should be excluded in subsequent analyses.



mz	quality	cd-9ns5y	cd-9xohp	cd-9kuct	adduct	pcgroup	isotopes
526.2847	0.4999	2596860.0000	0.0000	215409.5156		11	[226][M]+
387.2545	0.4907	2458645.2500	0.0000	2091807.0000		11	[222][M+2]+
581.3646	0.4880	10520168.0000	561874.0625	0.0000		11	
457.2576	0.9739	3027132.2500	133997.9219	4141179.2500		11	
463.3069	0.4910	591437.8750	0.0000	5249994.5000		11	[224][M+4]+
309.3480	0.0000	838089.2500	0.0000	0.0000		11	
484.3280	0.0000	0.0000	239863.5469	0.0000		11	
538.3089	0.9860	2171567.5000	911295.1250	504281.4062		11	
540.3265	0.9886	33299102.0000	5450043.5000	10871802.0000		11	
453.2324	0.0000	3612371.2500	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	371079.9375		11	
361.2024	0.0000	0.0000	0.0000	904217.5625		11	
585.8029	0.0000	565940.7500	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	560971.0000	0.0000		12	[228][M]+
379.2107	0.4877	727077.5625	0.0000	440147.1563		12	
543.3330	0.0000	947284.3625	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	1646947.2500	0.0000	0.0000		12	
463.3072	0.0000	4367910.0000	0.0000	0.0000	[M+H]+ 462.3	12	
443.2965	0.0000	670787.1250	0.0000	0.0000	[M+H-H2O]+ 462.3	12	
538.2852	0.0000	1004897.0000	0.0000	0.0000		12	
437.2916	0.0000	1605743.0000	0.0000	0.0000		12	[227][M]+
424.2476	0.0000	652254.8750	0.0000	0.0000		12	
538.3219	0.0000	1979313.0000	0.0000	0.0000	[M+ACN+H]+ 496.29	12	

Data Filter

This module allows users to select a subset of features for subsequent analyses. If this filter is not applied, all components would need to be analyzed, resulting in a cumbersome computational analysis. The module removes features based on a user-specified threshold for a coefficient of variation across all selected subjects and a threshold for the percentage of missing values.

Missing Value Imputation

This module uses popular missing value imputation methods such as mean value, integer, k-nearest neighbor (KNN), and Random Forest (RF) for imputation of missing values such as a peak missing in a small subset of samples but present in the majority of the samples.

Normalization

This module provides access to various data normalization methods such as quantile normalization, median normalization, mean normalization, cycLoess, global Robust Linear Regression (RLR), and global intensity normalization.

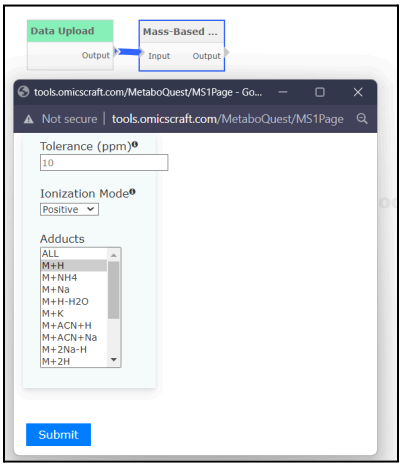
Batch Correction

This module uses empirical Bayes frameworks to adjust data from large scale studies whose measured values are impacted by running order or due to data acquisition in batches.

Metabolite Annotation Modules

This category includes several modules for metabolite annotation including *Spectral Matching* and *Compound Fingerprint Prediction* which use precursor m/sz and MS/MS data provided by the user for annotation. Other modules such as *Mass-Based Search*, *Isotopic Pattern Analysis*, and *Network-Based Annotation* search for putative IDs based on m/z and MS1 spectrum.

Spectral Matching



This module searches for putative metabolite IDs by matching **EI-MS** or **MS/MS** spectra against those in spectral libraries. To input an MS/MS spectrum, users can either enter it as a series of m/z-intensity pairs (with m/z and intensity values separated by a space) or upload data in mzML, mzXML, or plain text formats. The uploaded data may contain single or multiple MS/MS spectra for a batch search. If the data are in mzML or mzXML formats, an accompanying file consisting of the selected precursor m/z is needed for spectral matching.

Compound Fingerprint Prediction

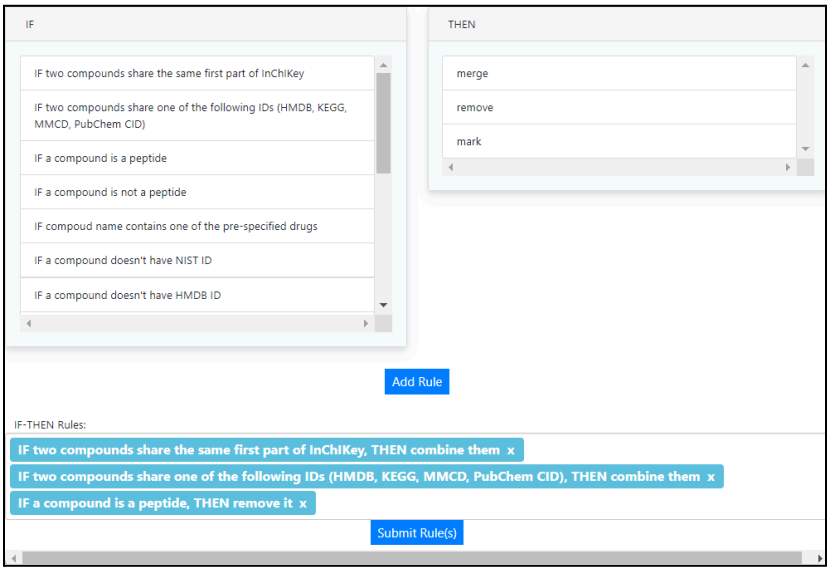
This module uses a deep/machine-learning model to predict compound fingerprints based on MS/MS data and uses the predicted fingerprints to rank candidate metabolites. This is designed for analytes that lack reference measurements in spectral libraries or have low spectral matching scores.

Mass-Based Search

This module enables customers to search for putative metabolite IDs in MetDB based on m/z values. Users can enter m/z values or use uploaded or processed data by a preceding module to search for putative IDs after calculating monoisotopic mass values based on the m/z values and user-specified adducts, ionization mode, mass tolerance in ppm. In addition, the module allows users to select which database IDs to be included in the result table.

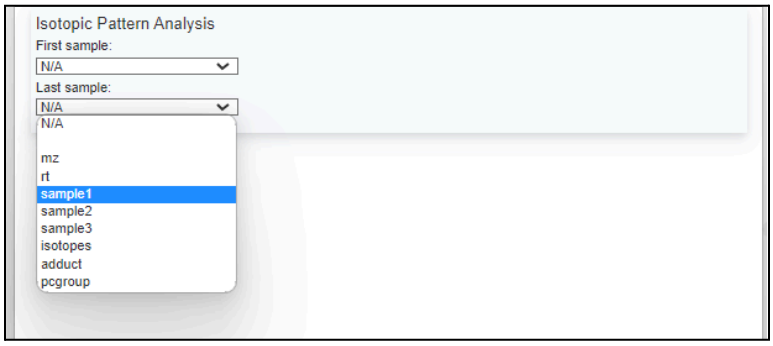
IF-THEN Rule

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



Isotopic Pattern Analysis

This module assigns scores to putative IDs based on their isotopic patterns. When comparing potential IDs with varying elemental formulas, scores are calculated by comparing the observed isotopic patterns from MS spectra with the theoretical isotopic patterns.



Network-Based Annotation

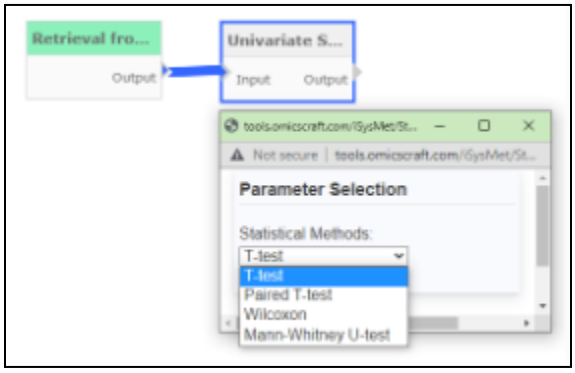
This module assigns scores to putative IDs using a network-based method. Specifically, the module constructs the metabolic network between putative IDs by extracting biochemical pathway information from databases such as MetaCyc and KEGG. This network assigns probability scores to putative IDs, indicating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

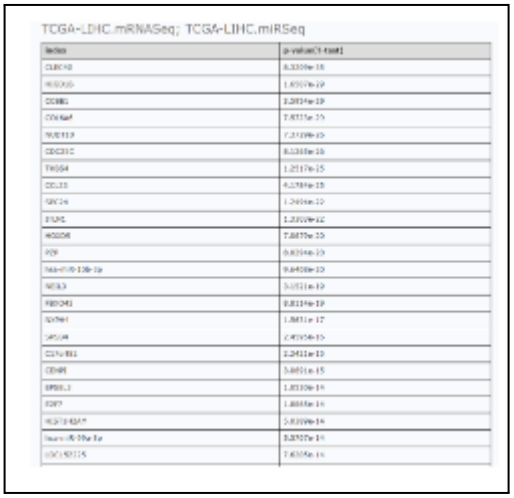
Data/Integrative Analysis

This category consists of modules that allow users to identify significantly altered metabolites or a panel of multi-omics features by integrative analysis. Each module in this category can be used for analysis of single omics or multi-omics data. These modules are linked to a set of tools for visualization including ROC curves, box plots, volcano plots, and heatmaps. Also, tools such as t-SNE or hierarchical clustering are included to visualize the data structure.

Univariate Statistical Analysis

This module performs parametric (Student t-test) or non-parametric (Mann-Whitney U-test) univariate analysis to select analytes/features statistically significantly altered between two independent samples. For matched/paired samples (i.e., tumor and adjacent non-tumor tissue from the same patient), this module allows users to apply parametric (paired t-test) or non-parametric (Wilcoxon signed-rank test) univariate analysis. The module can analyze preprocessed single omics or multi-omics imaging data.





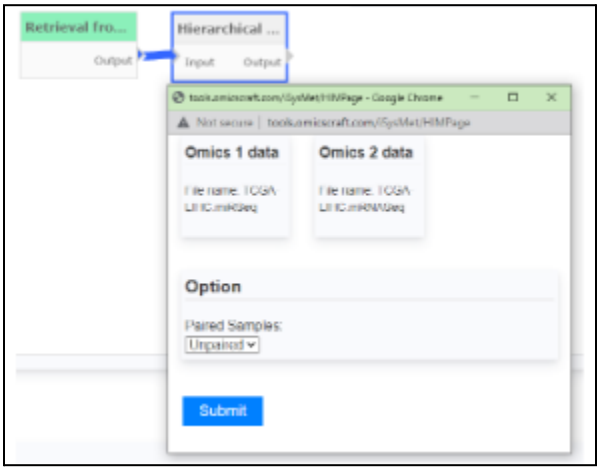
Multivariate Regression Analysis

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

Hierarchical Integrative Analysis

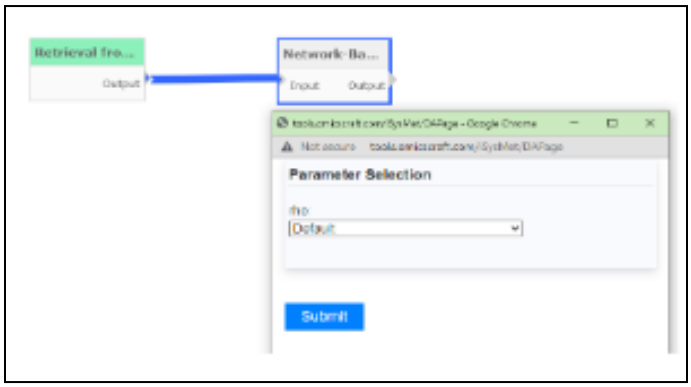
This module offers the opportunity to associate analytes measured in multi-omics studies to uncover novel relationships about disease status. The model allows us to investigate flexible modeling approaches based on penalized likelihood methods and expected maximization (EM)

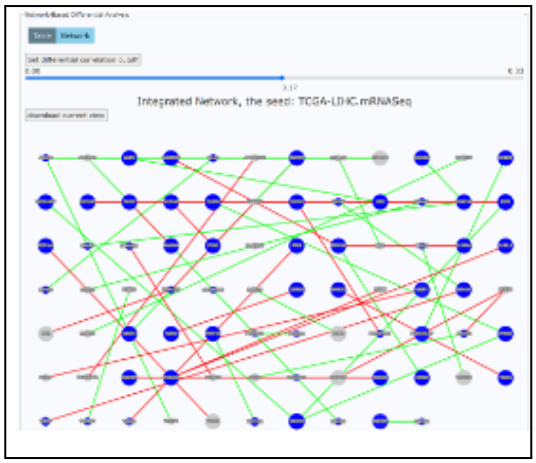
algorithms under various biological relationship scenarios between the different molecular features and their effects on a clinical outcome.



Network-Based Analysis

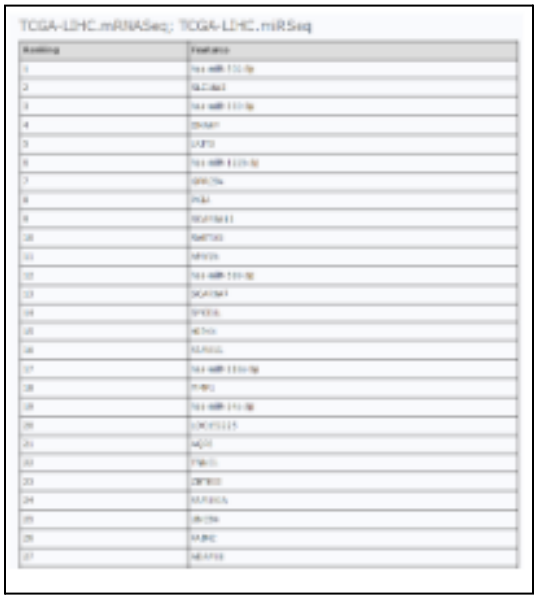
This module offers network-based methods for differential feature analysis of analytes in single omics, multi-omics, or imaging data. To find disease-associated interactions, differential networks are used. These networks show the differences in correlation among analyte pairs within the disease group compared to the control group. By analyzing these networks, users can gain insights into changes in pairwise interactions of analytes in disease versus control groups.

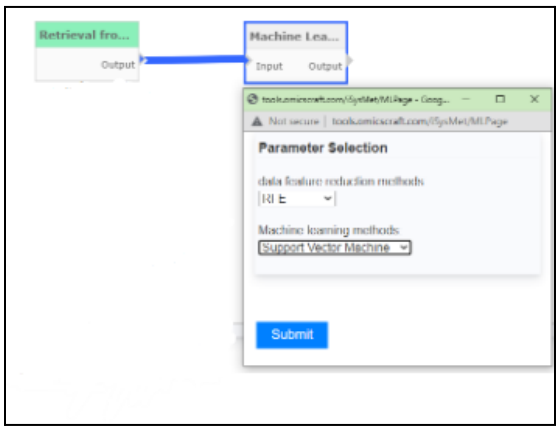




Machine Learning

This module uses two machine learning methods (support vector machine and random forest) and the recursive feature elimination method to select a panel of disease-associated features from single omics or multi-omics data. To achieve the latter, all standardized features are combined to obtain a vector that contains a set of concatenated features for each sample. The integrated, standardized features are then fed into the machine learning methods to find a panel of features that can predict the disease status.





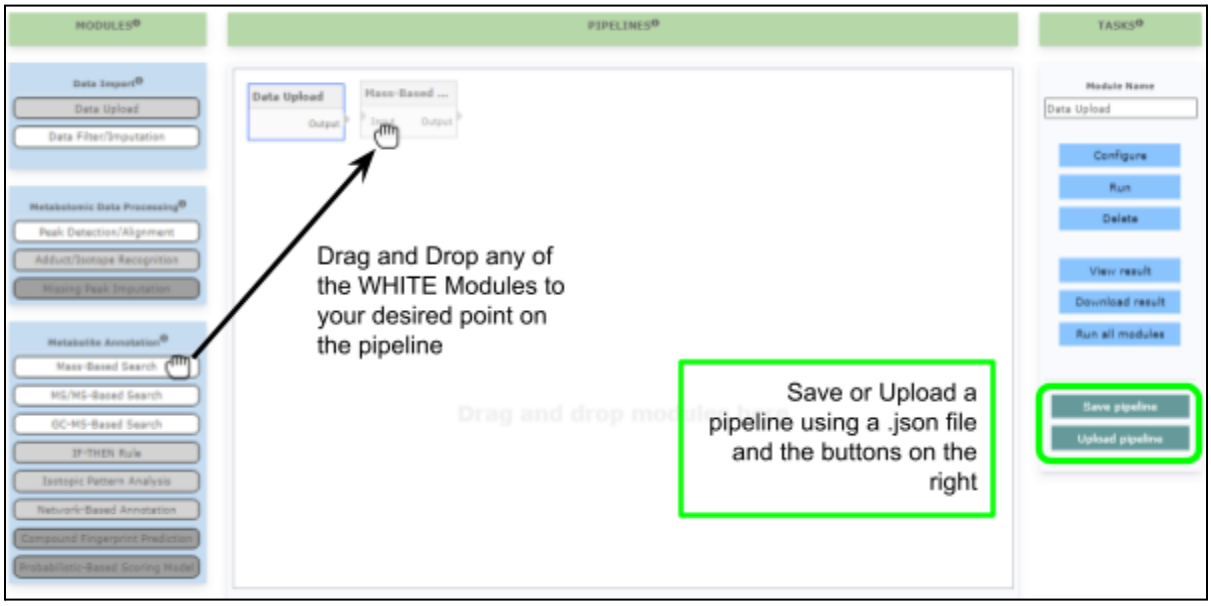
Generative AI

Note: *This module is coming soon*

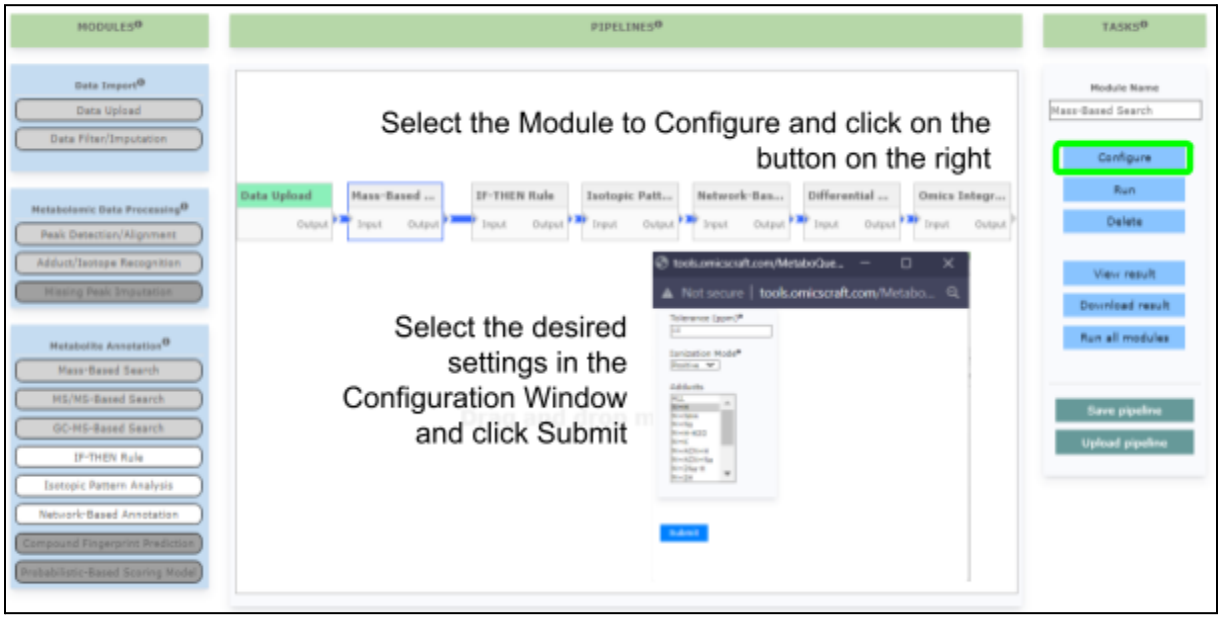
PIPELINE BUILDER

Steps to Build a Pipeline

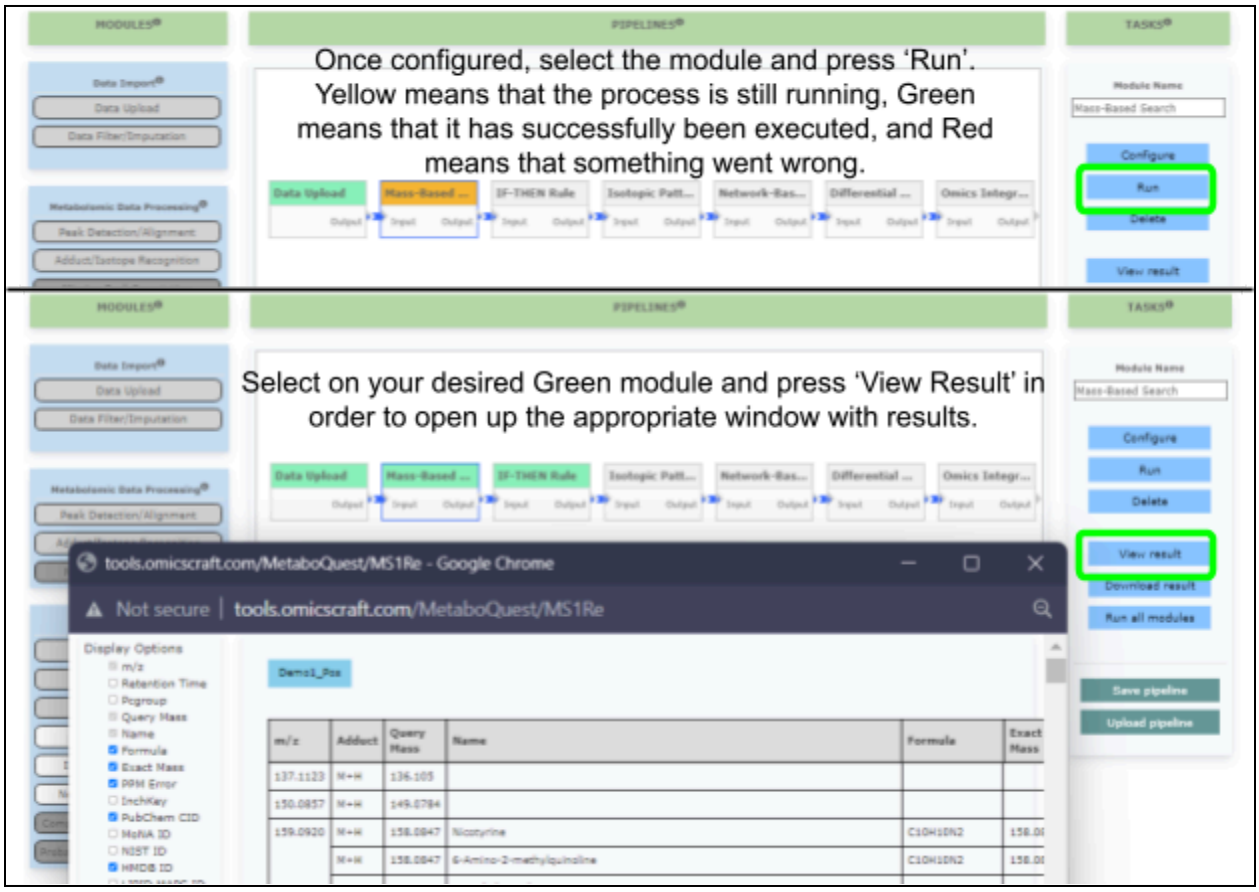
1. Lay down the tools/modules to create or upload an existing pipeline. Users can create a pipeline, save it and upload it later on. Modules in pipelines will require reconfiguration after being uploaded again.



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 AND PIPELINES

Data Processing & Metabolite Annotation

demoMetaboliteAnnotation is a folder consisting of demo data outlined below to test the Data Processing and Metabolite Annotation modules.

LC-MS Processed Data

Demo1a_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.

Demo1b_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.

Demo2a_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.

Demo2b_peaks_pos.csv: the same set of peaks as Demo2a 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.

Demo3a_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.

Demo3b_peaks_neg.csv: the same set of peaks as Demo3a 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.

LC-MS/MS Processed Data

Demo4a_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.

Dem4b_MSMS_pos.txt: all 12 MS/MS spectra from Demo4a 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.

Demo5a_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.

Demo5b_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.

GC-MS Processed Data

Demo6a_EI: 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.

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.

LC-MS/MS Unprocessed Data

Demo7a_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.

Demo7b_mzML_neg: the same samples as Demo7a analyzed in the negative mode.

Demo8a_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.

Demo8b_mzML_neg: the same datasets as Demo8a analyzed in the negative mode.

Demo9a_mzXML_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.

Demo9b_mzML_pos: the same datasets as Demo9a but converted to mzML format.

Data/Integrative Analysis

Metabolomics and Other Omics Data for Annotation and Marker Selection

demoMetaboliteAnnotation&DataIntegrativeAnalysis is a folder comprising processed metabolomics and other omics datasets in the following four folders to test the Metabolite Annotation and Data/Integrative Analysis modules.

Demo1: 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.

Demo2: 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.

Demo3: 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.

Demo4: 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.

Multi-Omics Data for Data/Integrative Analysis

demoDataIntegrativeAnalysis is a folder comprising processed multi-omics datasets and pipelines in the following five folders to test the Data/Integrative Analysis modules.

Demo1: 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.

Demo2: 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.

Demo3: 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.

Demo4: 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.