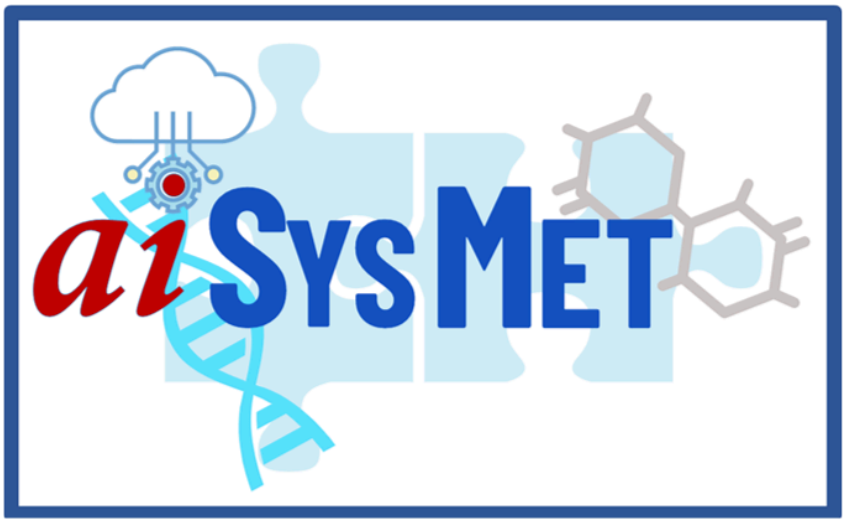


USER MANUAL

Version 0.5

December 2024

OMICSCRAFT

TABLE OF CONTENTS

INTRODUCTION

aiSysMet is an AI-powered platform for analysis of metabolomics data and integration of multii-omics 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 raw and unprocessed omics data
- Facilitates efficient workflow execution by ensuring data accessibility from low-latency storage. 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 (uploaded data or pipeline generated outputs) on a project basis for easy management and retrieval.
- Facilitates collaboration among researchers by sharing projects/pipelines, original data, and results in a transparent manner.

MODULES

aiSysMet's modules are grouped as outlined below. Please note that all modules listed below are available to the user with a full subscription of aiSysMet. In addition to the Data Import modules, users can choose to subscribe to a subset of the three major modules (MetCraft, MetaboQuest, and IntSys) to be included in aiSysMet.

Data Import Modules

This group 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 from local or cloud storage spaces raw metabolomics data for data processing, processed metabolomics data for metabolite annotation, and any processed omics data for biomarker discovery
    - 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, group/sample labels for marker selection, and/or designation of batches if any.
    - 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 omics data along with additional files 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 - Data Processing

This group 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 pre-annotation, i.e., recognizing adducts and isotopes 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.

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

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

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 group 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 MS/MS or EI-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 group 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/PROJECT BUILDER

Steps to Build a Pipeline/Project

1. Click **Project Manager** to create a project. Then, load the project into the pipeline canvas.
2. Starting with one of the Data Import modules, drag modules from the left pane to the pipeline canvas by connecting the modules according to the desired workflow or sequence of analysis.
3. Configure each module by clicking first the module itself and then clicking **Configure Module** in the right pane.
4. Execute individual modules by clicking **Run Module**. You can see the output by clicking **View Result** or download the result table by clicking **Download Result**.
5. Click **Run Project** to run all modules in the project. This requires saving the project first.
6. Click on **Save Project or Save Project As…** to save the project to be able to use it later by accessing it via the Project Manager.

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 canvas
- After placing a module in the pipeline canvas, it can be configured with the appropriate processing settings before execution. To do this, use the **Configure Module** button. If the module is not configured, it will use its default settings up on execution.
- After configuring a module, click the **Run Module** button to execute the module.
- If you choose to run all modules within a project, click the **Run Project** button.
- The progress window below the pipeline canvas 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 (yellow indicates the module is running, pink implies an error has occurred, green means the module completed processing, results can be viewed or downloaded).
- Click **Delete Module** button to remove an unwanted module from the pipeline canvas.
- Click the **Save Project** or **Save Project As …** buttons to the save the entire pipeline on the pipeline canvas
- Click **Clear Pipeline Canvas** to clear any existing pipelines in the pipeline canvas

DEMO DATA

MetCraft - Unprocessed Metabolomics Data

- **Demo1**: a folder consisting of: (1) Demo1a_mzML_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 mzML 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 Demo1a analyzed in the negative mode. Note that both the Peak Detection and Spectral Matching modules work as well with LC-MS/MS data in mzXML format.

MetaboQuest - Processed Data for Metabolite Annotation

- **Demo2**: a folder consisting of: (1) Demo2a_MSMS_pos: a folder of 12 files each consisting of an MS/MS spectrum acquired in the positive mode as well as a file listing the precursor m/z values corresponding to each of the 12 files; and (2) Demo2b_MSMS_pos.txt: all 12 MS/MS spectra from Demo2a listed in one file and a file listing the precursor m/z values corresponding to each MS/MS spectrum. These datasets can be used for metabolite annotation using the Spectral Matching module by uploading the 12 MS/MS spectra or one single file consisting of all MS/MS spectra along with the corresponding precursor files.

- **Demo3**: a folder consisting of: (1) Demo3a 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) Demo3b_MS1.txt that consists of MS spectra for batch processing along with a file listing all the precursor m/z values; and (3) Demo3c_MS2.txt that consists of MS/MS spectra for batch processing along with a file listing all the precursor m/z values.

- **Demo4**: a folder consisting of: (1) Demo4a_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) Demo4b_EI.txt: the same datasets as Demo3a 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.

IntSys - Multi-Omics Data for Integrative Analysis

- **Demo5**: 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 data processing and integrative analysis modules.

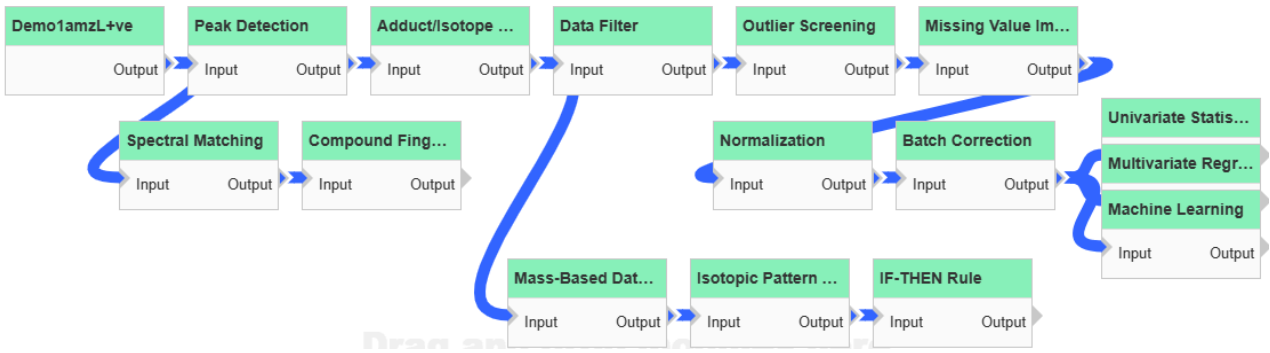
- **Demo6**: a folder consisting of three omics (mRNA expression profile, miRNA expression profile, and metabolomics profile) datasets acquired from the same set of samples. Each dataset, separately or in combination, can be used to test data processing and integrative analysis modules.

- **Demo7**: 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 data processing and integrative analysis modules.

- **Demo8**: 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 data processing and integrative analysis modules.

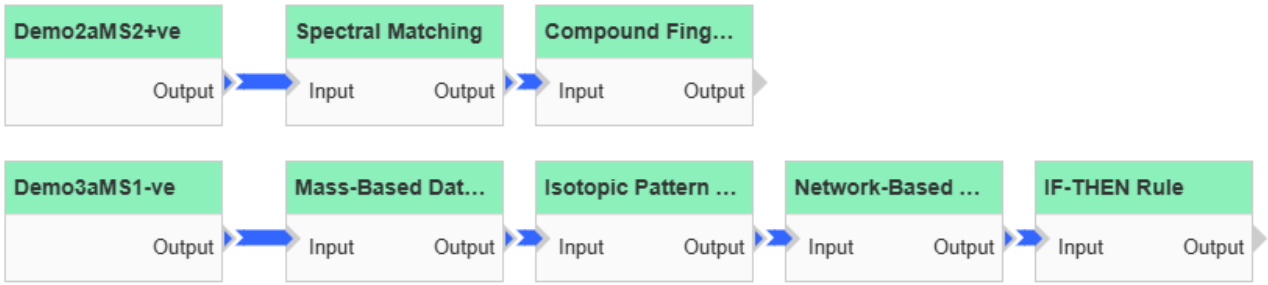
DEMO PROJECT VIDEOS

MetCraft - Data Processing



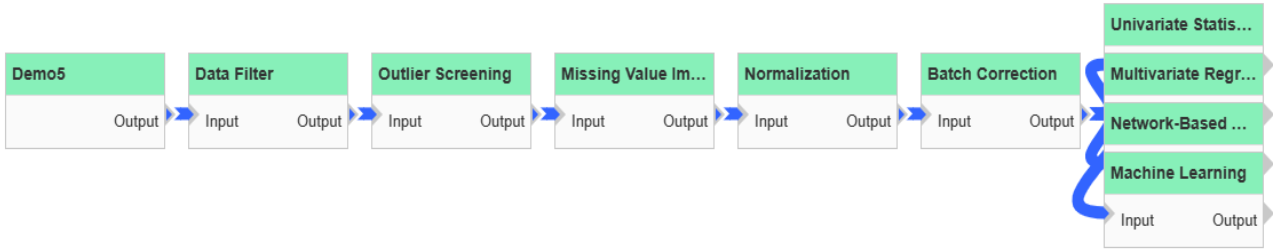
Click here to watch a demo video for this project

MetaboQuest - Metabolite Annotation



Click here to watch a demo video for this project

IntSys - Multi-Omics Data Integration



Click here to watch a demo video for this project