# Analysis Parameters Explained

This document describes the practical effect of each option in the **analysis** section of the integration workflow configuration file (/input\_data/config/project\_config.yml), including the analysis tag, feature selection, network analysis, and MOFA modeling. Each section lists available methods, their options, and default values, and at the end there is an example of a full analysis workflow.

# Setup Parameters

## 1. Tagging

* **data\_analysis\_tag**

Allows user to create a new analysis folder (Analysis--**TAG**) underneath the data processing folder to store analysis results. This is useful if analysis settings are changed (see below) and a new set of outputs should be produced that is separate from previous runs. Default: “0”.

*Note*: if an analysis folder already exists with the supplied tag and overwriting is disabled, the workflow will return an error message indicating that you should change the tag

## 1. Feature Selection

### Options:

* **variance** [default]

Selects the features with the highest variance across samples. Useful for keeping only the most variable and potentially informative features.

* + *max\_features*

Maximum number of features to retain (integer > 0; default: 5,000)

* **glm**

Generates a Generalized Linear Model (GLM) to identify features significantly associated with a metadata category. Filters by FDR-corrected p-value and minimum log2 fold change to keep only significant features.

* + *metadata\_category*

Metadata column to use for group comparison (must match a variable in the user\_settings->variable\_list in the configuration file).

* + *metadata\_category\_reference*

Reference group for the GLM - a specific group within the selected metadata category.

* + *significance\_level*

FDR-corrected p-value cutoff (real number between 0 and 1; default: 0.05).

* + *log\_fold\_level*

Minimum absolute log2 fold change to consider significant (real number > 0; default: 0.5).

* + *max\_features*

Maximum number of features to retain (integer > 0; default: 5,000).

* **kruskalwallis**

Use the Kruskal-Wallis test to identify features significantly associated with a metadata category. Filters by FDR-corrected p-value and minimum log2 fold change.

* + *metadata\_category*

Metadata column to use for group comparison (must match a variable in the user\_settings->variable\_list in the configuration file).

* + *significance\_level*

FDR-corrected p-value cutoff (real number between 0 and 1; default: 0.05).

* + *log\_fold\_level*

Minimum absolute log2 fold change to consider significant (real number > 0; default: 0.5).

* + *max\_features*

Maximum number of features to retain (integer > 0; default: 5,000)

* **feature\_list**

Selects features from a user-provided list (one feature ID per line in a file, must match the feature IDs in the analysis.integrated\_data table).

* + *feature\_list\_file*

Filename containing the list of features to keep. This file must be saved into the correct analysis output directory (e.g., /output\_data/project\_name/Data\_Processing--TAG/Analysis--TAG/). You can drop this file directly into the folder via the JupyterLab interface.

* + *max\_features*

Maximum number of features to retain (integer > 0; default: 5,000)

* **none**

No feature selection is performed; all features are retained.

* + *max\_features*

Maximum number of features to retain (integer > 0; default: 5,000). *Warning*: currently, a limit still needs to be imposed even when no feature selection is performed due to memory constraints when calculating large correlation matrices/networks. The top features are selected in the order they appear in the data table.

*Note*: the max\_features option during feature selection, which shows up in almost all modes, restricts the number of features that go into downstream correlation analysis and networking - this is designed to reduce the size and scale of calculations and should be set to a value lower than 10,000 when possible.

## 2. Feature Correlation

### Options:

* **corr\_method** 
  + *pearson* [default]

Calculates Pearson correlation between features.

* + *spearman*

Calculates Spearman rank correlation.

* + *kendall*

Calculates Kendall rank correlation.

* **corr\_cutoff**

Correlation threshold for including edges in the network (real number between 0 and 1; default: 0.5). Only feature pairs with correlation above this value are included.

* **keep\_negative**

If true, includes both positive and negative correlations above the absolute threshold. If false, only positive correlations are included.

## 3. Network Analysis

### Options:

* **network\_mode**
  + *bipartite* [default]

Constructs a network only between features from different datasets (e.g., transcript and metabolite node edges)

* + *full*

Constructs a network including all feature-feature correlations, regardless of dataset. Not currently recommended.

*Note*: Functionally, this **network\_mode** option is also passed to the feature correlation step above to keep the cached correlation matrix as small as possible.

* **submodule\_mode**
  + *community* [default]

Extracts submodules using community detection algorithms (Louvain method).

* + *subgraphs*

Extracts submodules as connected components.

* + *none*

No submodules are extracted from the main graph.

## 4. MOFA Analysis

### Options:

* **num\_mofa\_factors**

Number of latent factors to compute in the MOFA model (integer > 0; default: 5). Controls model complexity.

* **num\_mofa\_iterations**

Number of training iterations for MOFA (integer > 0; default: 1,000). Higher values may improve convergence.

* **seed\_for\_training**

Random seed for reproducibility (integer > 0; default: 555). Ensures consistent results across runs – set a different random seed to produce a different (non-deterministic) result.

## Example

Suppose you run an analysis with the following configuration (removed some unused feature selection settings for this example):

analysis:

data\_analysis\_tag: VARIANCE

analysis\_parameters:

feature\_selection:

selected\_method: kruskalwallis

…

kruskalwallis:

metadata\_category: temperature

significance\_level: 0.01

log\_fold\_level: 0.5

max\_features: 10000

…

correlation:

corr\_method: pearson

corr\_cutoff: 0.75

keep\_negative: false

networking:

network\_mode: bipartite

submodule\_mode: community

mofa:

num\_mofa\_factors: 3

num\_mofa\_iterations: 1000

seed\_for\_training: 555

**Result:**

* Only features with normalized abundance that was significantly different (FDR<0.01 and LFC>0.5) between samples of different "temperature" categories (e.g., samples with low vs. medium vs. high) by a Kruskal-Wallis test by ranks are kept.
* The correlation is performed with the Pearson rho value and pairs of features are only kept if they have a positive correlation ≥ 0.75.
* The network includes only bipartite edges (between different data types)
* Multi-omics factor analysis is run with 3 factors, 1000 iterations, and a fixed random seed of 555 for reproducibility.

**Final Output:** After these analysis steps, your results will include a subset of the integrated, QC-ed, and normalized features from the data processing step, a correlation network focused on strong cross-omics relationships, and a MOFA model summarizing features that are a major sources of variation across samples and datasets.