MetFlow tutorial

Xiaotao Shen [*1] ***and Dr. Zheng-Jiang Zhu*** [†1]

[1]Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences.

*shenxt@sioc.ac.cn †jiangzhu@sioc.ac.cn

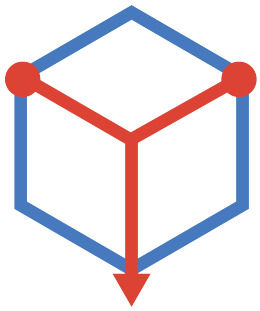
12 December 2018

Contents

1 Data Preparation . . . 3
1.1 Prepare MS1 peak tables . . . 3
1.2 Prepare a sample information file . . . 3
1.3 Important notes for data preparation . . . 4
2 Log In or Sign Up . . . 5
2.1 Sign up . . . 5
2.2 Log in . . . 7
3 Data Cleaning . . . 7
3.1 Upload Data Files . . . 7
3.2 Check Data Files . . . 7
3.3 Batch Alignment . . . 7
3.4 Data Quality Check . . . 9
3.5 Missing Value Processing . . . 9
3.6 Zero Value Processing . . . 10
3.7 Data Normalization . . . 11
3.8 Data Integration . . . 13
3.9 Outlier Removal . . . 14
3.10 Data Quality Visualization . . . 16
3.11 Result Download . . . 16
4 Differential Metabolite Discovery . . . 16
4.1 Upload Data Files . . . 16
4.2 Check Data Files . . . 16
4.3 Univariate Analysis . . . 16
4.4 Multivariate Analysis . . . 18

4.5 Differential Metabolite Selection 20
4.6 Performance Validation 20
4.7 Result Download 22

5 Pathway Enrichment Analysis 23
5.1 Paste Differential metabolites/Peaks 23
5.2 Check Data Files 23
5.3 Pathway Enrichment 23
5.4 Result Download 24



MetFlow

Metabolomics Data Cleaning
and Differential Metabolite Discovery

1 Data Preparation

MetFlow requires the import of the following files, including:

- One or multiple MS1 peak tables (.csv format),
- A table for sample information (.csv format).

1.1 Prepare MS1 peak tables

The MS1 peak table is a list of metabolic peaks with annotated m/z, retention times (RTs) and peak abundances.

1.1.1 Untargeted metabolomics data

LC-MS or GC-MS untargeted metabolomics data can be generated using processing sortware, such as XCMS or MS-DIAL. The peak table from software must be modified for `MetFlow`.

- The format of MS1 peak table must be csv;
- The first column is the peak name ("name");
- The second column is the mass-to-charge ratio ("mz");
- The third column is the retention time ("rt");
- The unit of retention time must be second (not minute);
- Other columns are peak abundances of MS1 peaks in each sample.

IMPORTANT: the order and names of the first three columns must be "name", "mz", and "rt".

Note: If you use `fillPeaks` function in XCMS to process data, there will be no missing values (MV) in the MS1 peak table.

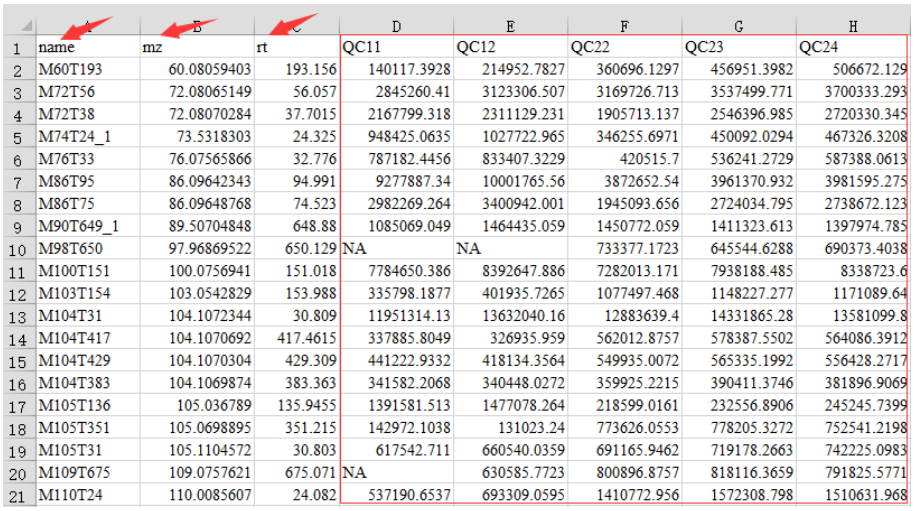
The final generated MS1 peak table should look like:

1.1.2 Targeted metabolomics data

For targeted metabolomics data, such as MRM, because there is no accurate m/z, so users must assign pseudo m/z values for each peak.

1.2 Prepare a sample information file

The sample information file (.csv format) is designed to describe the sample injection order, class, batch and group information. The first column is named as "sample.name", while the second column is named as "injection.order", the third column is "class", the fouth column



	A	B	C	D	E	F	G	H
1	name	mz	rt	QC11	QC12	QC22	QC23	QC24
2	M60T193	60.08059403	193.156	140117.3928	214952.7827	360696.1297	456951.3982	506672.129
3	M72T56	72.08065149	56.057	2845260.41	3123306.507	3169726.713	3537499.771	3700333.293
4	M72T38	72.08070284	37.7015	2167799.318	2311129.231	1905713.137	2546396.985	2720330.345
5	M74T24_1	73.5318303	24.325	948425.0635	1027722.965	346255.6971	450092.0294	467326.3208
6	M76T33	76.07565866	32.776	787182.4456	833407.3229	420515.7	536241.2729	587388.0613
7	M86T95	86.09642343	94.991	9277887.34	10001765.56	3872652.54	3961370.932	3981595.275
8	M86T75	86.09648768	74.523	2982269.264	3400942.001	1945093.656	2724034.795	2738672.123
9	M90T649_1	89.50704848	648.88	1085069.049	1464435.059	1450772.059	1411323.613	1397974.785
10	M98T650	97.96869522	650.129	NA	NA	733377.1723	645544.6288	690373.4038
11	M100T151	100.0756941	151.018	7784650.386	8392647.886	7282013.171	7938188.485	8338723.6
12	M103T154	103.0542829	153.988	335798.1877	401935.7265	1077497.468	1148227.277	1171089.64
13	M104T31	104.1072344	30.809	11951314.13	13632040.16	12883639.4	14331865.28	13581099.8
14	M104T417	104.1070692	417.4615	337885.8049	326935.959	562012.8757	578387.5502	564086.3912
15	M104T429	104.1070304	429.309	441222.9332	418134.3564	549935.0072	565335.1992	556428.2717
16	M104T383	104.1069874	383.363	341582.2068	340448.0272	359925.2215	390411.3746	381896.9069
17	M105T136	105.036789	135.9455	1391581.513	1477078.264	218599.0161	232556.8906	245245.7399
18	M105T351	105.0698895	351.215	142972.1038	131023.24	773626.0553	778205.3272	752541.2198
19	M105T31	105.1104572	30.803	617542.711	660540.0359	691165.9462	719178.2663	742225.0983
20	M109T675	109.0757621	675.071	NA	630585.7723	800896.8757	818116.3659	791825.5771
21	M110T24	110.0085607	24.082	537190.6537	693309.0595	1410772.956	1572308.798	1510631.968

Figure 1:

is "batch" and the fifth column is "group". "class" is used to descibe the class of samples: subject sample ("Subject") or QC sample ("QC"). The "group" is used to describe the group information of samples, and QC samples should be names as "QC". The sample information file should look like:

NOTE: The "sample.name" column in sample information file must be the **EXACTLY** same as the sample names in the MS1 peak table.

1.3 Important notes for data preparation

- In the MS1 peak table, make sure that no "-" or blank appears in the peak name or sample name. If there are some symbols that cannot be recognized by our program, the data processing may be failed.
- The "sample.name" column in sample information file must be the **EXACTLY** same as the sample names in the MS1 peak table.
- Please make sure that sample information (.csv format) and MS1 peak table (.csv format) are separated by comma. Because in some countries or regions (European and some French-speaking regions), the default separator is semicolon. You can open the sample information or MS1 peak table with notepad or other text editors to check whether they are separated by comma.

	A	B	C	D	E
1	sample.name	injection.order	class	batch	group
2	QC11	1	QC	1	QC
3	EC6225	2	Subject	1	Case
4	EC567	3	Subject	1	Control
5	EC5A1395	4	Subject	1	Case
6	EC4604	5	Subject	1	Case
7	EC7542	6	Subject	1	Case
8	EC7528	7	Subject	1	Case
9	EC6345	8	Subject	1	Case
10	EC6108	9	Subject	1	Case
11	QC12	10	QC	1	QC
12	EC34A1771	11	Subject	1	Case
13	ECA1469	12	Subject	1	Case
14	EC24A1581	13	Subject	1	Case
15	ECA558	14	Subject	1	Control
16	EC6513	15	Subject	1	Case
17	EC4385	16	Subject	1	Case
18	EC6305	17	Subject	1	Case
19	EC6893	18	Subject	1	Case
20	QC13	19	QC	1	QC
21	EC8289	20	Subject	1	Case
22	ECFA123	21	Subject	1	Case
23	EC6894	22	Subject	1	Case
24	EC6659	23	Subject	1	Case
25	EC3768	24	Subject	1	Case

Figure 2:

2 Log In or Sign Up

2.1 Sign up

If you are using `MetFlow` for the first time, please sign up first.

1. Click “Sign up” tab;
2. Enter your information;
3. Click “Sign up” button.

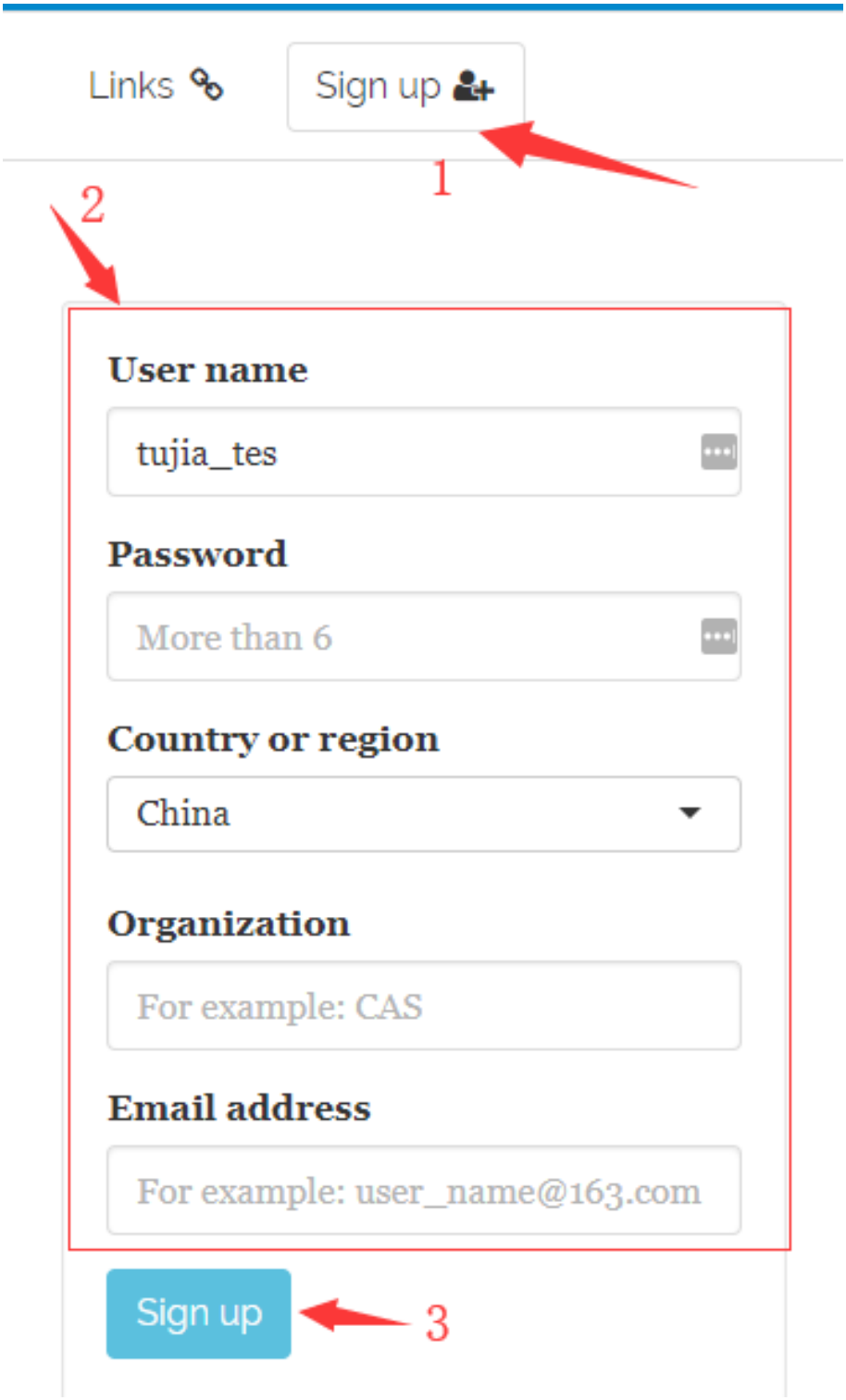

Links
Sign up
1
2
User name
tujia_tes
Password
More than 6
Country or region
China
Organization
For example: CAS
Email address
For example: user_name@163.com
Sign up
3

Figure 3:

2.2 Log in

1. Click "Log in & Account" tab;
2. Enter your user name and password;
3. Click "Log in" button.

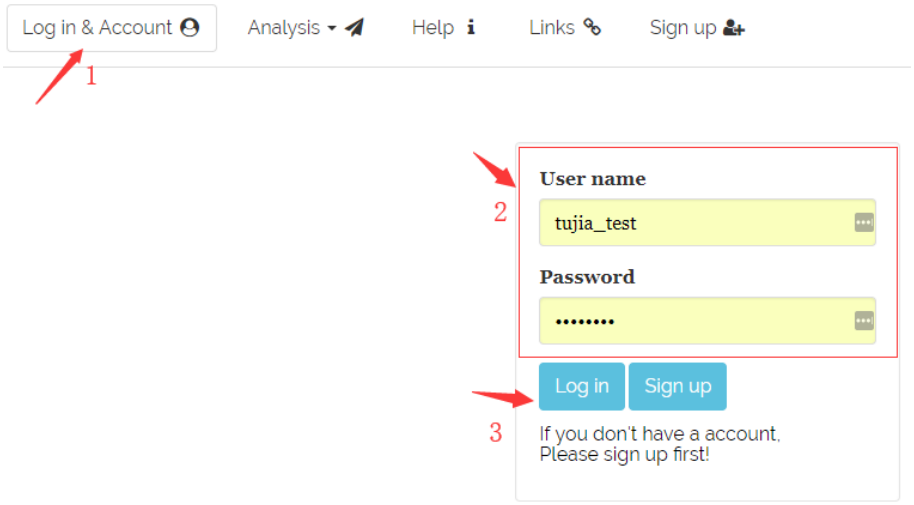


Figure 4:

3 Data Cleaning

Data cleaning is implemented as a step-wised and standardized workflow under "Data Cleaning" tab. Users should process data step by step.

3.1 Upload Data Files

1. Enter the project name;
2. Select the MS1 peak tables (.csv format) and Sample information (.csv format);
3. Or you can use demo data;
4. Click "Submit" button to upload data.

3.2 Check Data Files

Then `MetFlow` check the data format of MS1 peak tablss and sample information. If there are error in you data, please click `Previous` to check your data and upload again. If there is no error, you can click `Next` for the next step.

3.3 Batch Alignment

3.3.1 Parameter setting

1. Set parameters for rough alignment;
2. Click `Submit` for batch alignment.

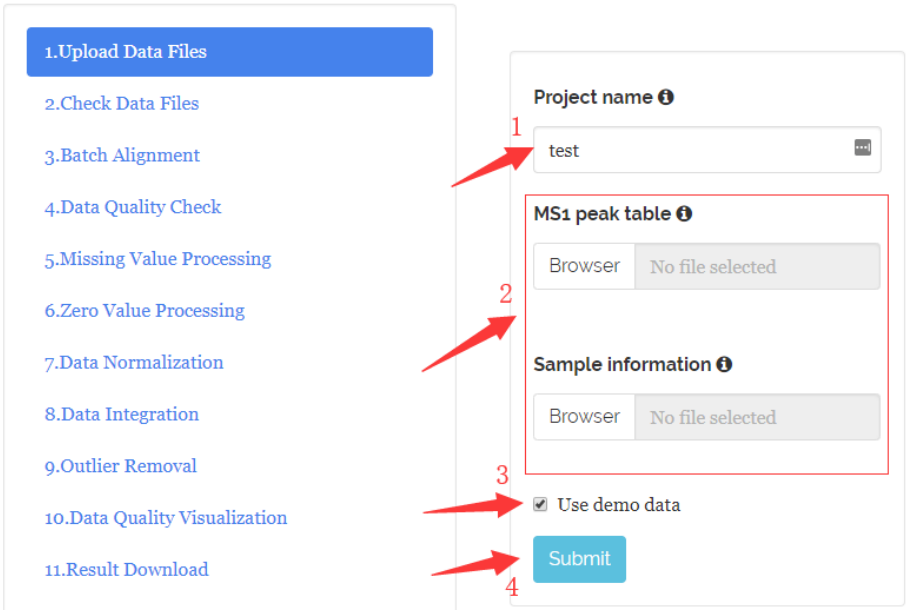


Figure 5:

Table 1: Parameters of batch alignment

Paramter	Meaning
m/z tolerance (ppm)	m/z tolerance (ppm) for rough alignment.
Retention time tolerance (second)	Retention time tolerance (ppm) for rough alignment.

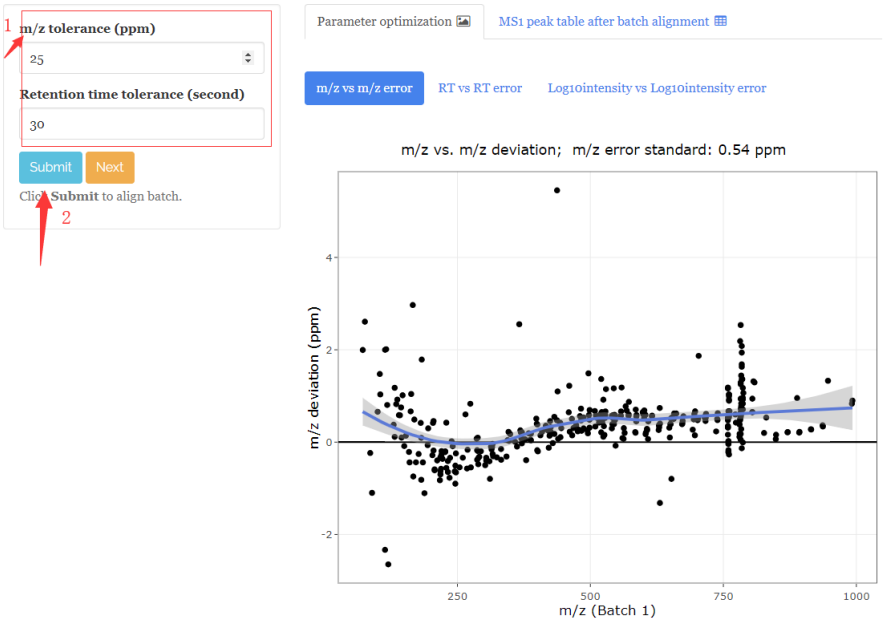


Figure 6:

3.3.2 Result

1. The "Parameter optimization" tab shows the m/z error, RT error and log10intensity error deviation in rough alignment.
2. The "MS1 peak table after batch alignment" tab shows the aligned MS1 peak table, users can click "Download" to download it.
3. Then click "Next" for the next step.

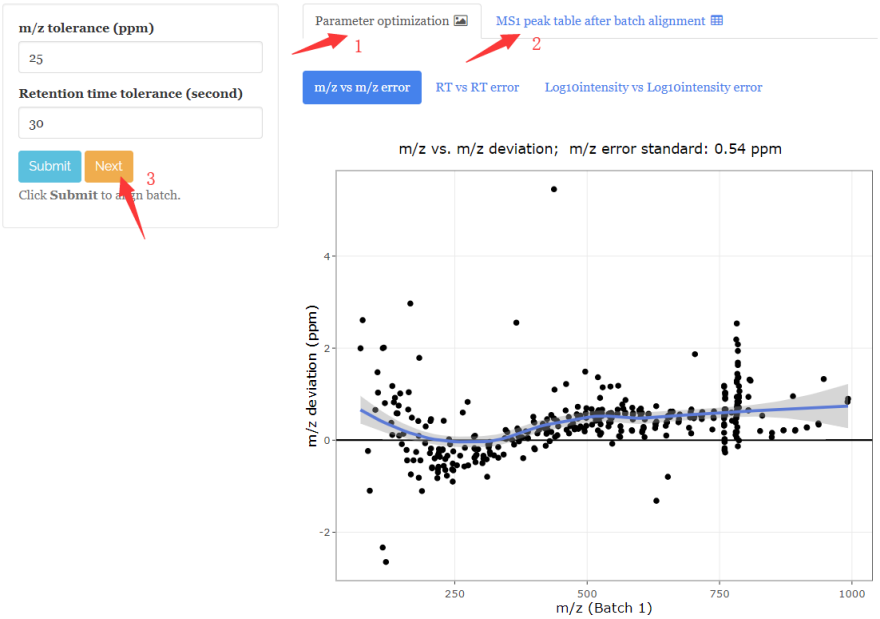


Figure 7:

3.4 Data Quality Check

Then the data quality is visually checked using 7 different creteria:

1. Data profile: m/z vs RT vs log10(intensity);
2. Missing value distribution: Missing value ratios in peaks and samples;
3. Zero value distribution: Zero value ratios in peaks and samples;
4. RSD distribution: RSD distribution in QC samples, you can also use different group to calculate RSD;
5. PCA score plot: PCA score plot of different batches;
6. QC intensity boxplot: QC auto-scaled intensity boxplot in different batches;
7. QC correlation: The correlations of QC samples;
8. All the figures can be downloaded. Then click "Next" for the next step.

3.5 Missing Value Processing

3.5.1 Parameter setting

1. Set parameters for missing value processing;
2. Click `Submit`.

Figure 8:

Table 2: Parameters of missing value processing

Paramter	Meaning
Remove peaks with MV ratio > (%)	It means that if the MV ratio larger than the threshold you set, the peaks will be removed from the dataset. For example, the default of this parameter is 50, it means that for each peak, if its MV ratio > 50%, this peak will be removed.
Imputation method	'MetFlow' has 9 methods for missing value imputation: 1) Zero value, 2) Mean, 3) Median, 4) Minumun, 5) KNN, 6) missForest and 7) BPCA. The default is KNN.

3.5.2 Results

1. Summary: Show the peaks which are removed from the dataset;
2. MS1 peak table (after MV processing): You can download the MS1 peak table after MV processing;
3. Click `Next`.

3.6 Zero Value Processing

1. Set parameters for zero value processing;
2. Click `Submit`.
3. Summary: Show the peaks which are removed from the dataset;
4. MS1 peak table (after zero processing): You can download the MS1 peak table after zero processing;
5. Click `Next`.

NOTE: If there are no missing values in your data, you can select any imputation method.

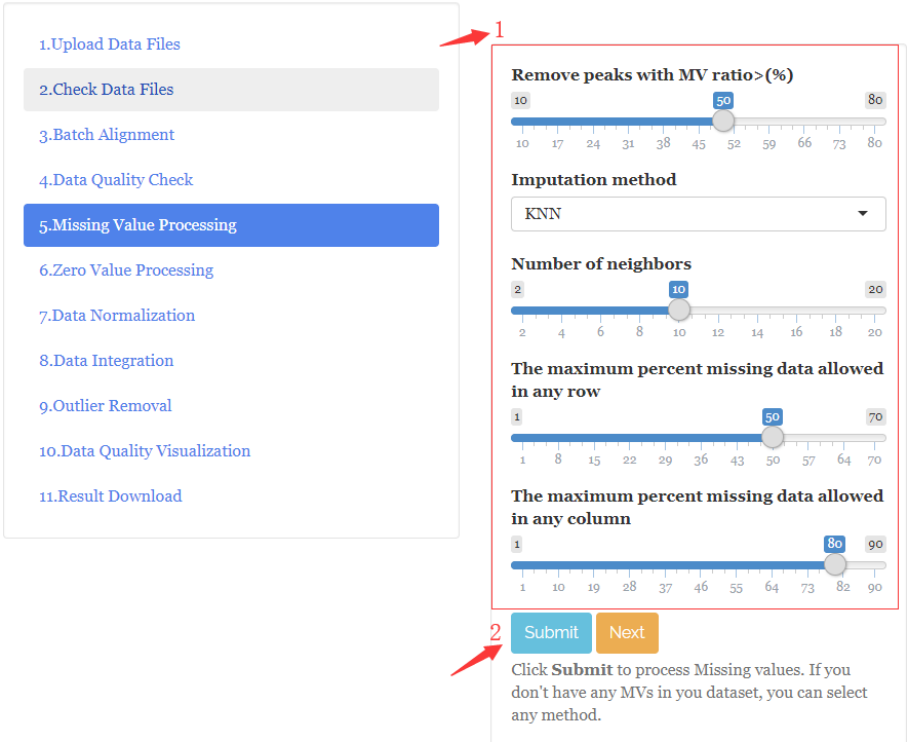


Figure 9:

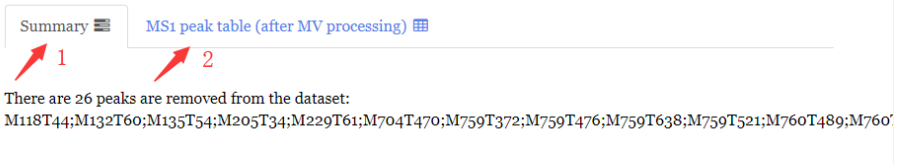


Figure 10:

3.7 Data Normalization

3.7.1 Parameter setting

1. Set parameters for data normalization;
2. Click `Submit`.

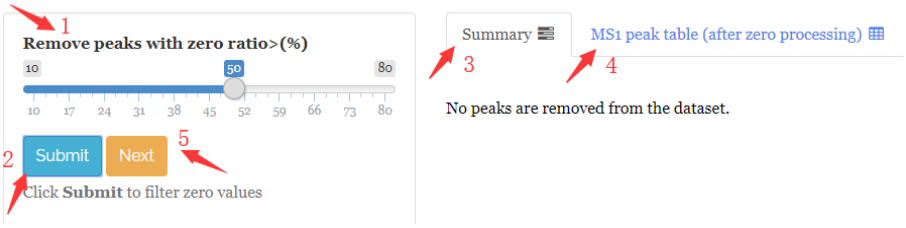


Figure 11:

Table 3: Parameters of zero value processing

Paramter	Meaning
Remove peaks with zero ratio > (%)	It means that if the zero ratio larger than the threshold you set, the peaks will be removed from the dataset. For example, the default of this parameter is 50, it means that for each peak, if its zero ratio > 50%, this peak will be removed.

Table 4: Parameters of data normalization

Paramter	Meaning
QC sample-based methods	You can check the methods based QC sample or not.
Normalization method	There are 3 common used non-QC sample-based methods: 'Mean', 'Median' and 'Total'. And there are two common used QC sample-based methods: 'QC SVR (MetNormalizer)' and 'QC LOESS'.

3.7.2 Results

3.7.2.1 Summary

1. QC intensity box plot before normalization;
2. QC intensity box plot after normalization;
3. RSD comparison;

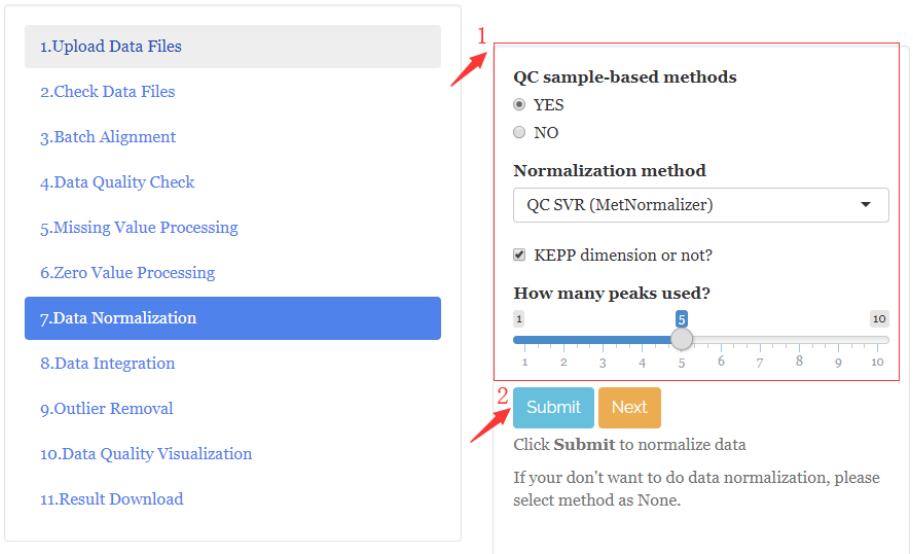


Figure 12:

4. The distribution of RSDs of peaks.

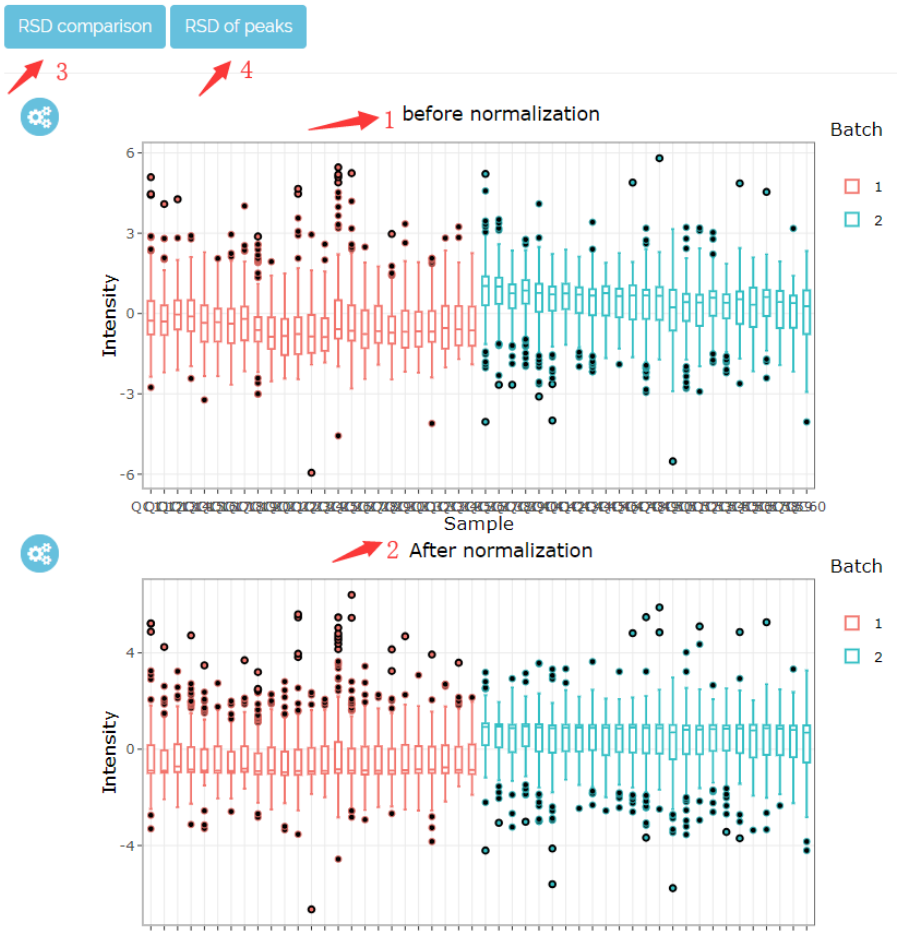


Figure 13:

3.7.2.2 MS1 peak table (after data normalization)

1. For each peak, you can select it, then click "Before normalization" or "After normalization" to show it's intensity drift.
2. Click `Next`.

3.8 Data Integration

3.8.1 Parameter setting

1. Set parameters for data integration;
2. Click `Submit`.

Like data normalization, you can also see the single peak intensity plot, QC auto-intensity boxplot, RSD comparison plot and RSD of peaks. Then click "next" for next step.

Table 5: Parameters of data integration

Paramter	Meaning
QC sample-based methods	You can check the methods based QC sample or not.
Integration method	There are 2 common used non-QC sample-based methods: 'Subject mean' and 'Subject median'. And there are two common used QC sample-based methods: 'QC mean' and 'QC median'.

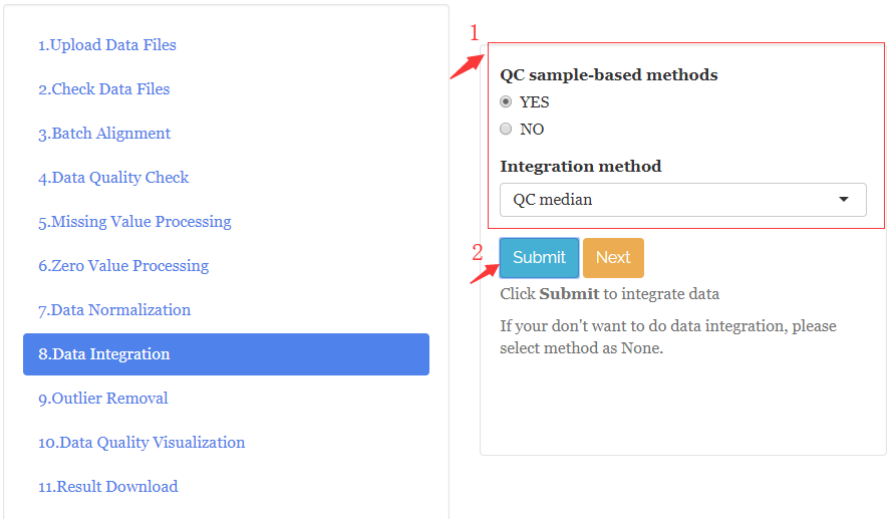


Figure 14:

3.9 Outlier Removal

3.9.1 Parameter setting

Table 6: Parameters of outlier removal

Paramter	Meaning
Logarithm method	default is 'Log 10'.
Scale method	default is 'Auto scale'.
Samples will be considered as outliers outside % CI	It means that if one sample is outside % confidence interval, the sample will be considered as outlier samples. The default is 95%.
Samples will be considered as outliers with zero value ratio > %	It means that it one sample with zero value ratio bigger than %, the sample will be considered as outliers. The default is 50%.

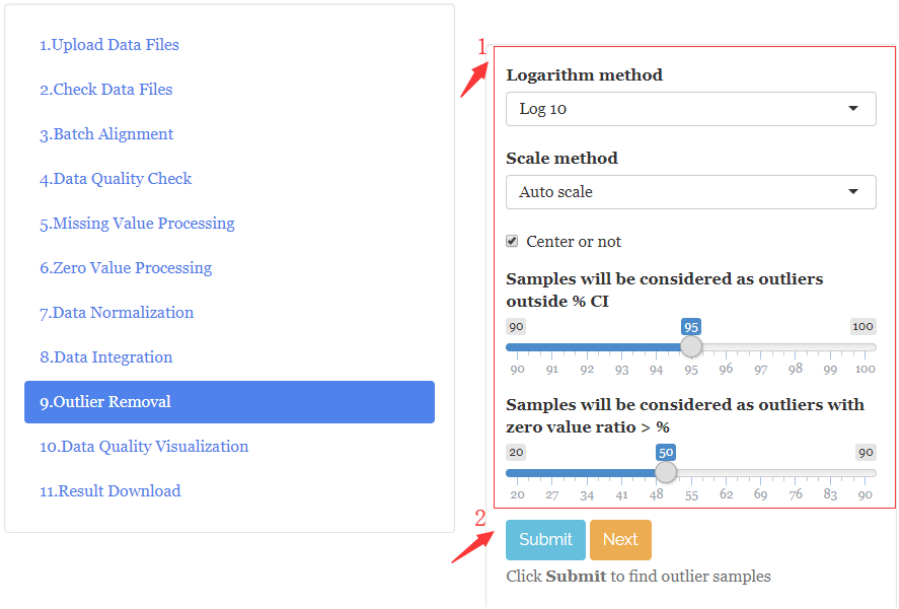


Figure 15:

3.9.2 Delete outlier samples

1. The information outlier samples;
2. Select outlier samples which you want to remove;
3. Click `Delete`;
4. Click `Submit` again.

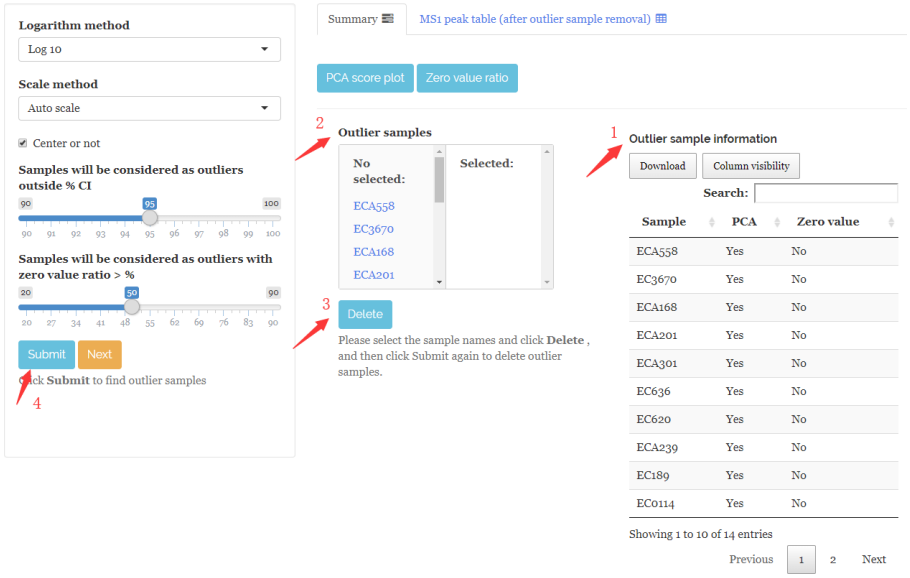


Figure 16:

3.10 Data Quality Visualization

`MetFlow` also visually assesses the data quality after data cleaning.

3.11 Result Download

1. Click "Generate HTML Summary" to generate analysis report (html format);
2. Then click "Download HTML Summary" to download the analysis report;
3. Click "Generate Analysis Result" to generate analysis result (zip foramt);
4. Then click "Download Analysis Result" to download the analysis result.

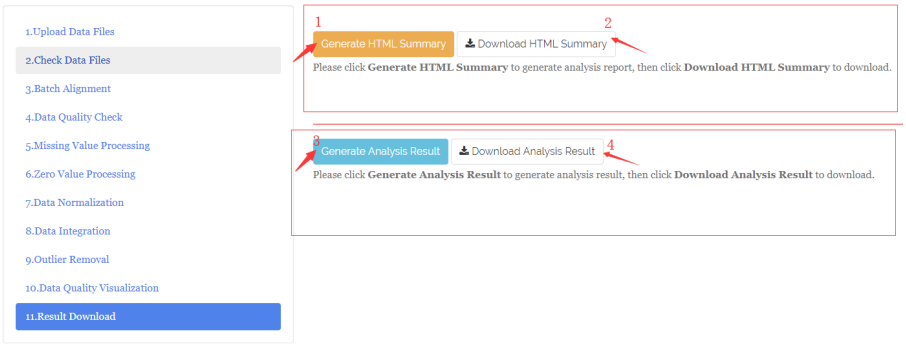


Figure 17:

4 Differential Metabolite Discovery

Differential metabolite discovery analysis is implemented as a step-wised and standardized workflow under "Differential Metabolite Discovery" tab. Users should process data step by step.

4.1 Upload Data Files

This step is same as "Data Cleanning".

4.2 Check Data Files

This step is same as "Data Cleanning".

4.3 Univariate Analysis

4.3.1 Parameter setting

1. Set parameters for univariate analysis;
2. Click `Submit`.

4.3.2 Results

1. Volcano plot: The volcanplot is utilized to visualized the differential metabolites.;

Table 7: Parameters of univariate Analysis

Paramter	Meaning
Control group	Select the control group.
Case group	Select the case group.
Logarithm method	Select logarith method, default is 'No log'.
Use what to calculate fold change	Use mean or median value of samples in one group to calcualte fold-change, default is 'Mean'.
Hypothesis testing method	'Student's t test' or 'Wilcoxon test'.
Alternative	'Two sided', 'Less' or 'Greater'.
Paired t-test	Paired or not.
Correction method	Select Correction method, default is 'False discovery ratio (FDR)'.
P-value cutoff	Default is 0.05.
Fold change cutoff	Default is 2, it means fold change (case/control) > 2 or < 0.5.

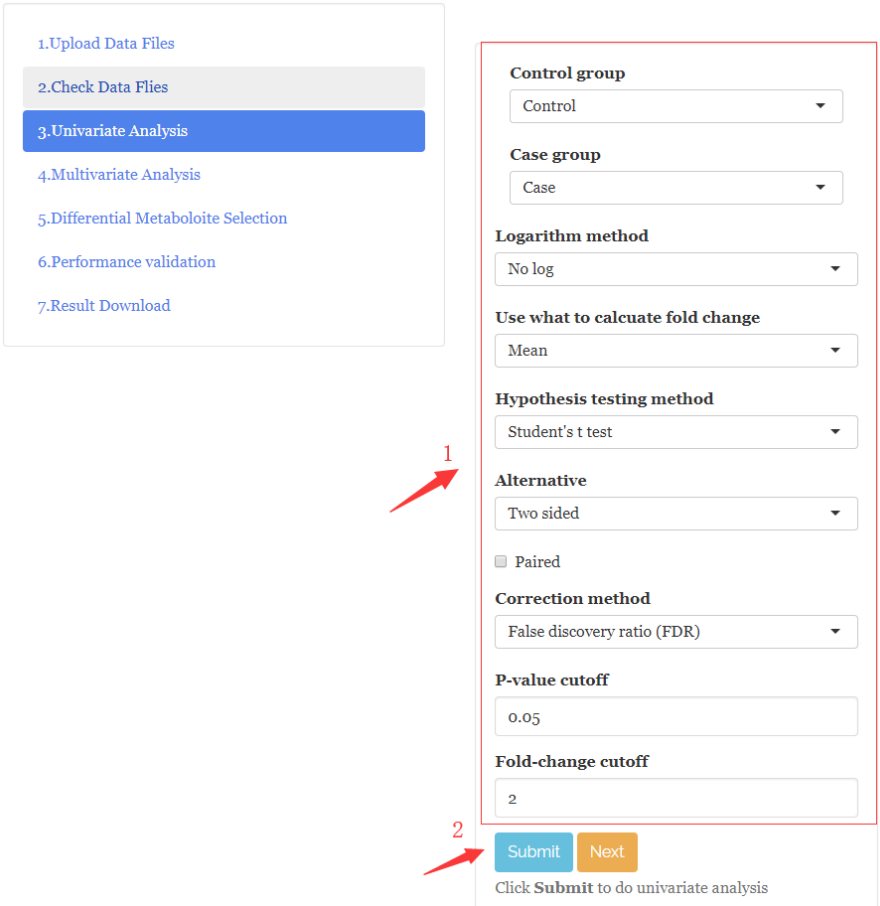


Figure 18:

2. Fold change and P-value: Fold-changes and P-values for all peaks.

4.4 Multivariate Analysis

4.4.1 Set parameters

1. Set parameters for multivaraite analysis;
2. Click "Submit".

Table 8: Parameters of multivariate analysis

Paramter	Meaning
Logarithm method	Select logarith method, default is 'Log 10'.
Scale method	Select scale method, default is 'Auto scale'.
Center or not	Default is checked.

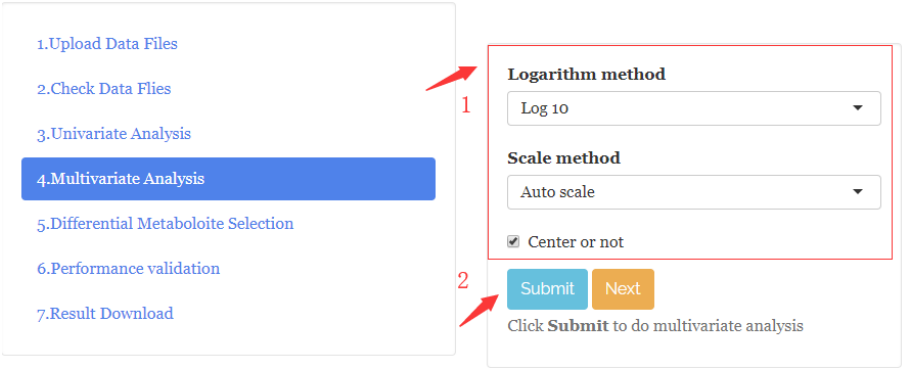


Figure 19:

4.4.2 Results

4.4.2.1 PCA analysis

The PCA score plot.

4.4.2.2 PLS analysis

1. Click "Q2cum" and select the ncomp with the biggest Q2cum, and then click "Submit";
2. Click "Q2cum&R2cum" to see the final Q2cum and R2 cum of the PLS model.

4.4.2.3 HCA analysis

1. Click "Parameter setting" to set parameters for HCA analysis;
2. Click "Download" to download heatmap.

4.4.2.4 Fold-change&P-value&VIP

Fold-changes, P-values and VIP values for all peaks.

Figure 20:

Table 9: Parameters of HCA analysis

Paramter	Meaning
Distance measure used in clustering rows	Distance measure used in clustering rows. Default is 'Euclidean'.
Distance measure used in clustering columns	Distance measure used in clustering columns. Default is 'Euclidean'.
Clustering method	Clustering method used. Default is 'Ward.D'.
Cluster rows	Cluster rows or not.
Cluster columns	Cluster column or not.
Show row names	Show row names or not.
Show column names	Show column names or not.
Control group color	Color for control group.
Case group color	Color for case group.
Low color	Color used in heatmap for low intensity.
Middle color	Color used in heatmap for middle intensity.
High color	Color used in heatmap for high intensity.

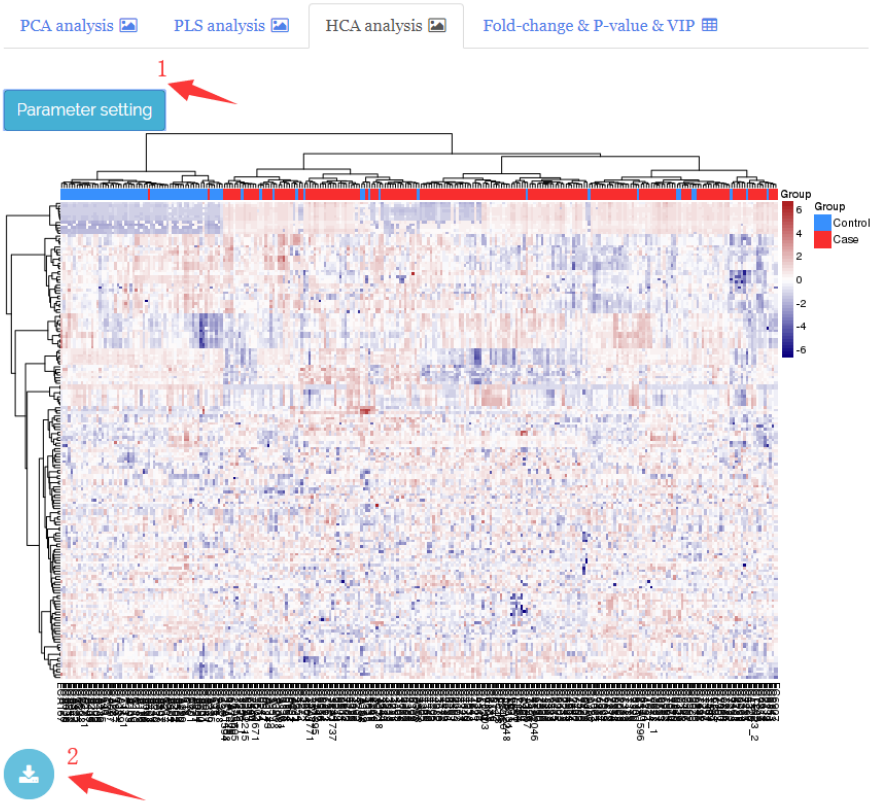


Figure 21:

4.5 Differential Metabolite Selection

1. Set parameters;
2. Click `Submit`;
3. 3D plot for visualization of differential metabolite selection;
4. Differential metabolite table.

Table 10: Parameters of differential metabolite delection

Paramter	Meaning
P-value cutoff	The cutoff of P-values.
Fold-change cutoff	The cutoff of fold-changes.
VIP cutoff	The cutoff of VIP.

4.6 Performance Validation

4.6.1 Upload validation dataset

1. If you have validation dataset, please select them and click “Upload”;
2. Click “Submit”.

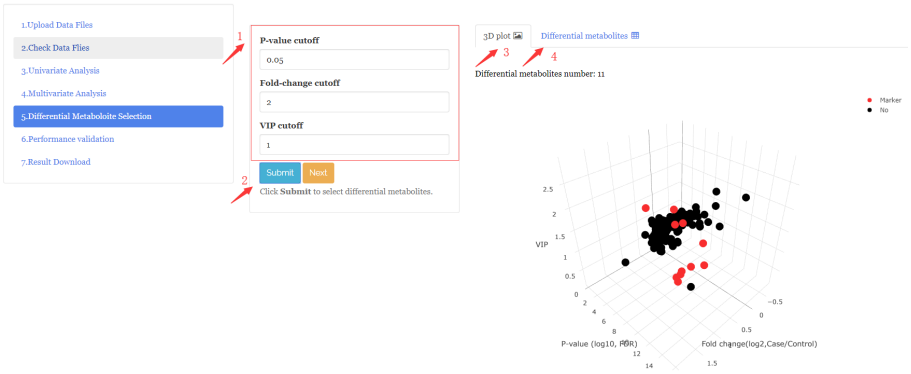


Figure 22:

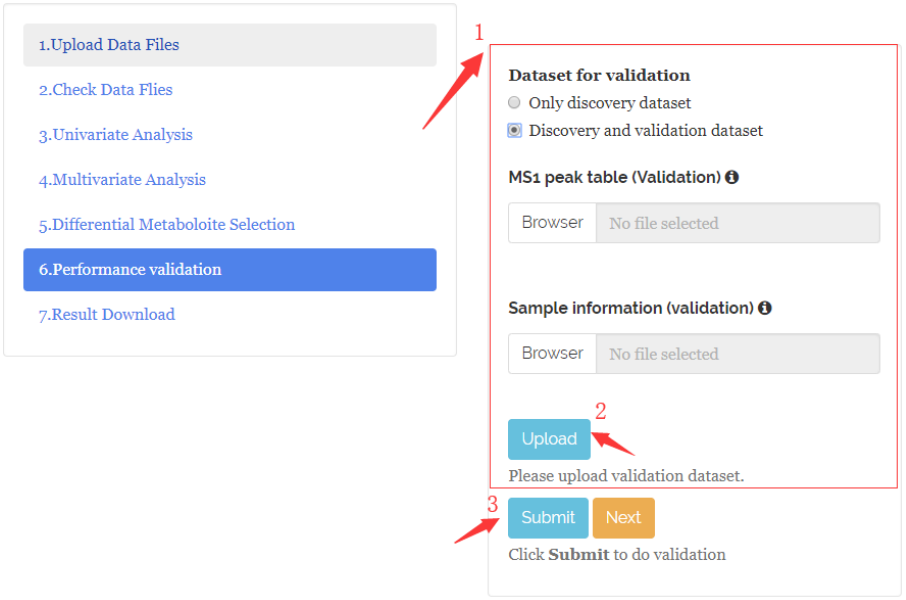


Figure 23:

4.6.2 Results

4.6.2.1 PCA, PLS and HCA analysis

"PCA analysis", "PLS analysis" and "HCA analysis" are performed using the differential metabolites in your discovery dataset and validation dataset.

4.6.2.2 ROC analysis

1. Select prediction model you want to use. There are four models, PLS, random forest, support vector machine and logistic regression;
2. Click "Submit".

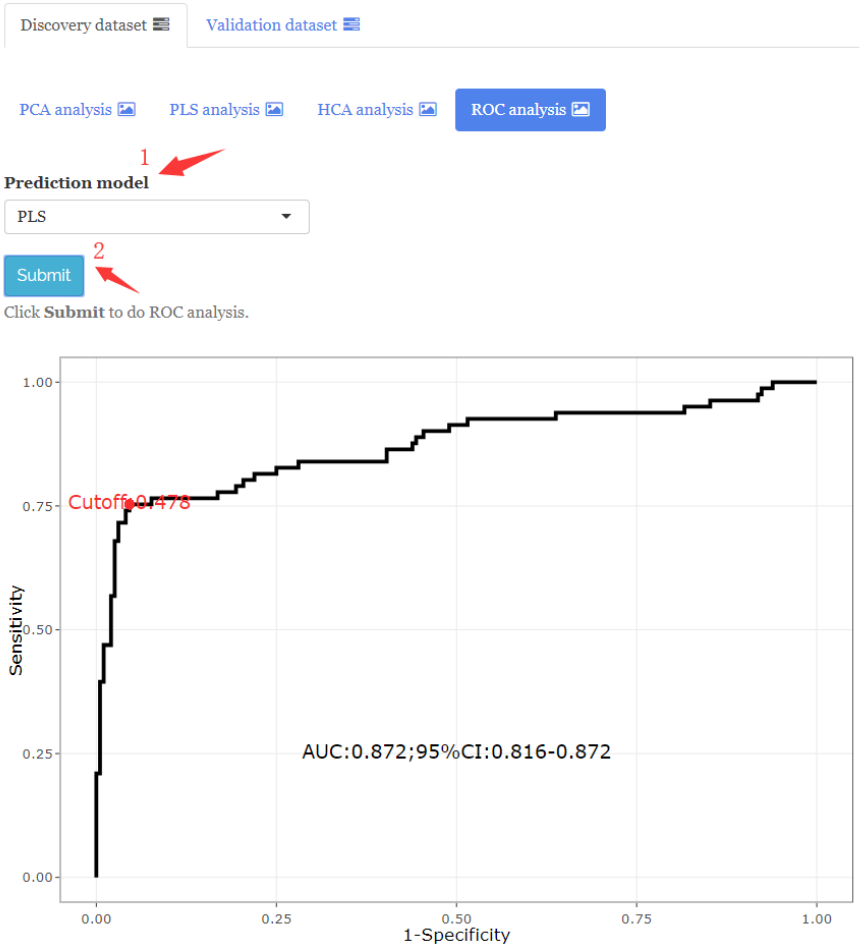


Figure 24:

4.7 Result Download

1. Click "Generate HTML Summary" to generate analysis report (html format);
2. Then click "Download HTML Summary" to download the analysis report;
3. Click "Generate Analysis Result" to generate analysis result (zip foramt);
4. Then click "Download Analysis Result" to download the analysis result.

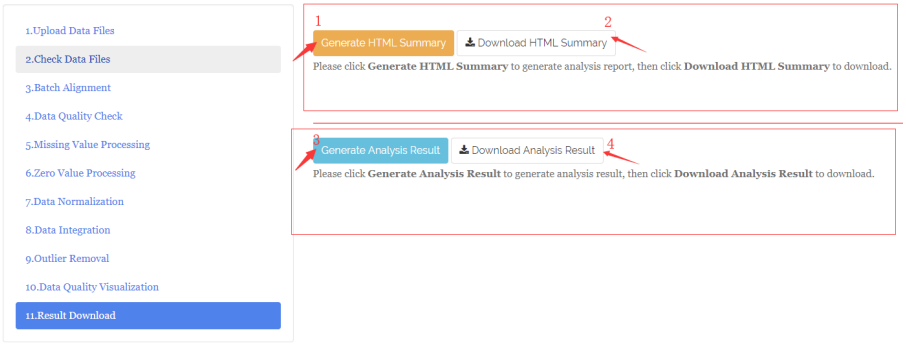


Figure 25:

5 Pathway Enrichment Analysis

Pathway enrichment analysis is implemented as a step-wised and standardized workflow under "Pathway Enrichment Discovery" tab. Users should process data step by step.

5.1 Paste Differential metabolites/Peaks

1. Enter the project name;
2. Then click `Metabolite type` to select which type of metabolites you want to provide, `KEGG ID` is KEGG ID of metabolites, `Peak (m/z)` is the m/z values of peaks;
3. Paste KEGG ID or m/z values, one row one ID or m/z. If you don't paste your values, the demo data will be used;
4. Then click "Submit".

5.1.1 Paste differential peaks (m/z)

If you paste m/z values of differential peaks, you must set the parameters for metabolite identification.

- Ionization polarity: `Positive` or `Negative`;
- Adduct type;
- m/z match tolerance;
- Database for metabolite identification.

5.2 Check Data Files

For `KEGG ID`, MetFlow match them in KEGG database, and shows the match result. For `Peak (m/z)`, the identification rsult is shown.

5.3 Pathway Enrichment

Please set the parameters for pathway enrichment first, and then click `Submit`.

1. Pathway library;
2. Pathway analysis algorithm: `Hypergeometric test` or `Fisher's Exact Test`;
3. P.value cutoff;

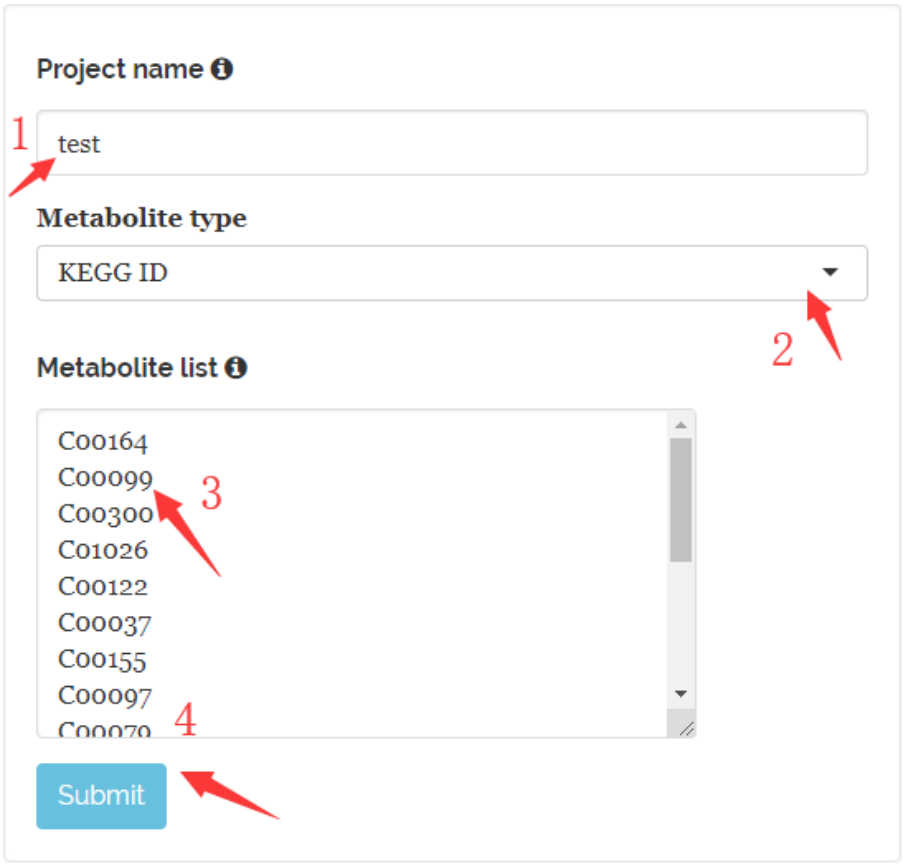


Figure 26:

4. Overlap cutoff.

The pathway enrichment result (plot and table) are shown on the right panels.

5.4 Result Download

The HTML summary and analysis results are generated and downloaded as data cleaning and differential metabolite analysis.

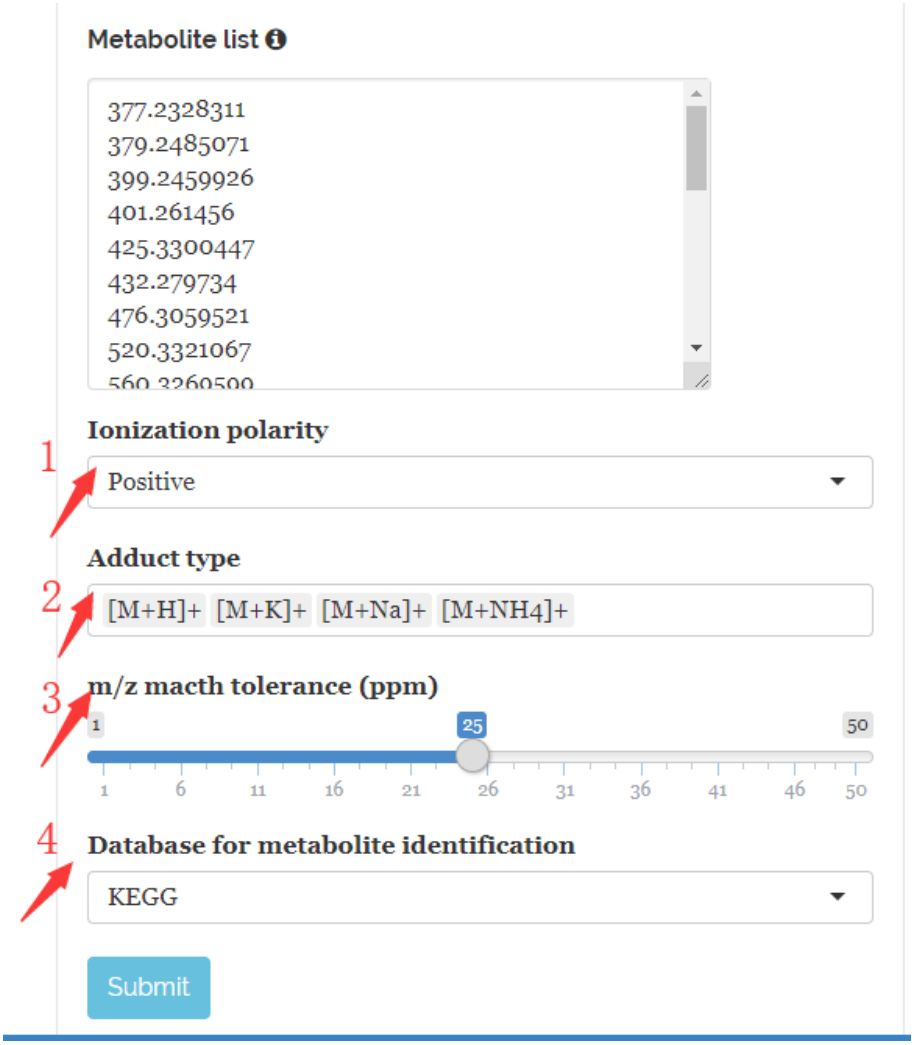
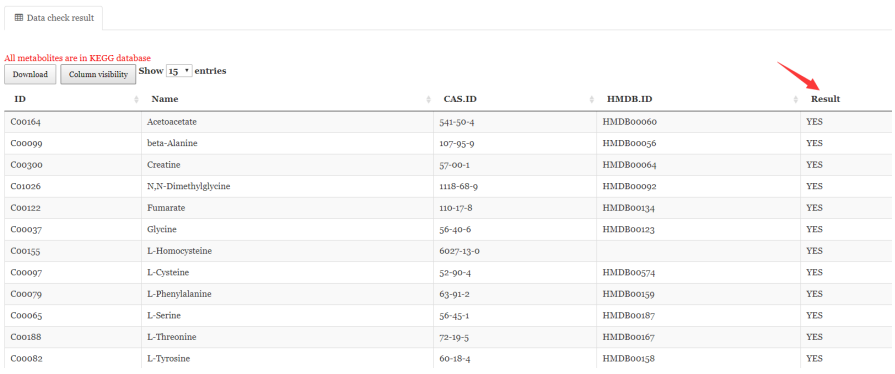


Figure 27:



ID	Name	CAS.ID	HMDB.ID	Result
C00164	Acetoacetate	541-50-4	HMDB00060	YES
C00099	beta-Alanine	107-95-9	HMDB00056	YES
C00300	Creatine	57-00-1	HMDB00064	YES
C01026	N,N-Dimethylglycine	1118-68-9	HMDB00092	YES
C00122	Fumarate	110-17-8	HMDB00134	YES
C00037	Glycine	56-40-6	HMDB00123	YES
C00155	L-Homocysteine	6027-13-0		YES
C00097	L-Cysteine	52-90-4	HMDB00574	YES
C00079	L-Phenylalanine	63-91-2	HMDB00159	YES
C00065	L-Serine	56-45-1	HMDB00187	YES
C00188	L-Threonine	72-19-5	HMDB00167	YES
C00082	L-Tyrosine	60-18-4	HMDB00158	YES

Figure 28:

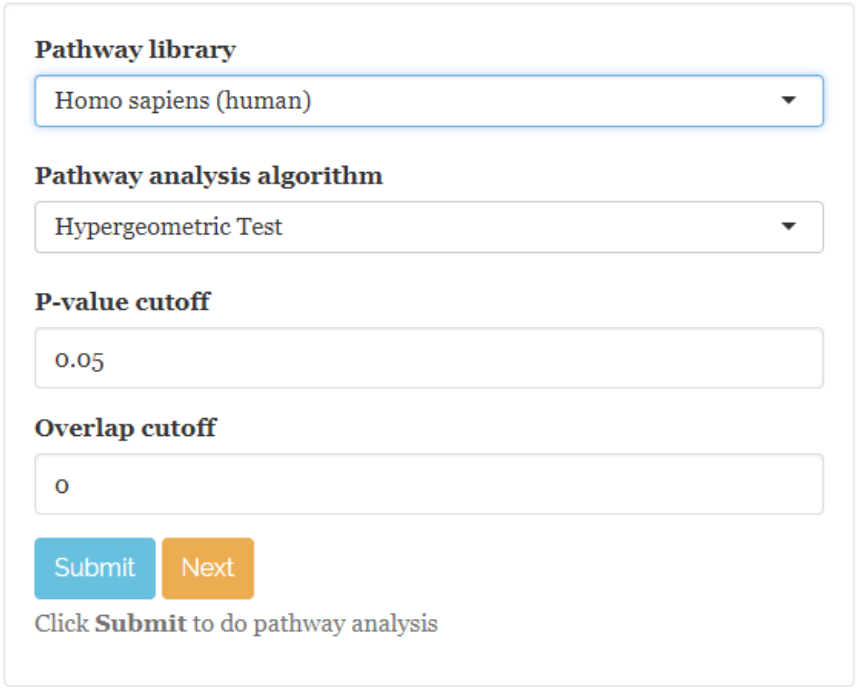


Figure 29:

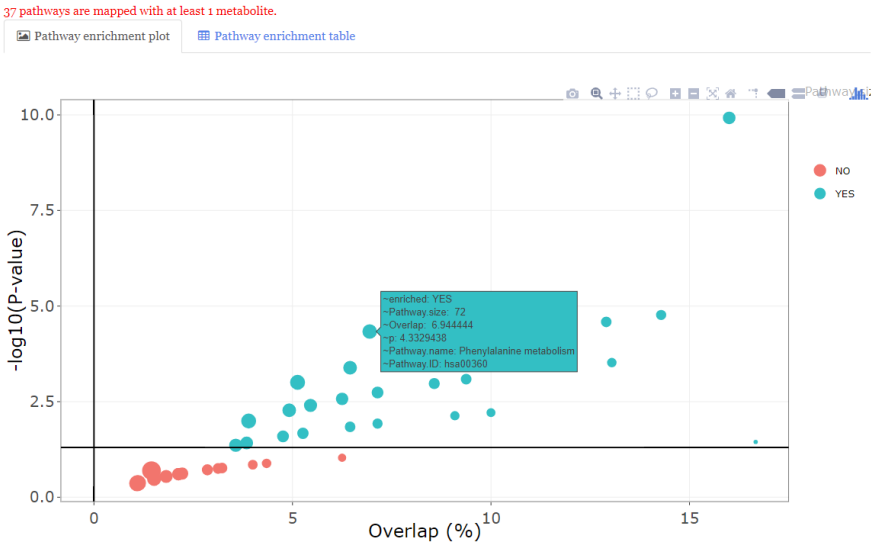


Figure 30: