

# Pathway Analysis of Untargeted Metabolomics Data using the MS Peaks to Pathways Module

---

By: Jasmine Chong, Jeff Xia

Date: 14/02/2018

The aim of this tutorial is to demonstrate how the MS Peaks to Pathways module of MetaboAnalyst can be used to directly infer biological activity from mass peaks, bypassing the bottleneck of metabolite identification. This module implements the *mummichog* algorithm (Version 1.0.10) from Li et al. 2013 (<https://doi.org/10.1371/journal.pcbi.1003123>). The example data used in this tutorial comes from the original Python implementation, which are human samples that were collected using an Orbitrap LC-MS.

## Introduction to MS Peaks to Pathways

High-throughput analysis and functional interpretation of untargeted MS-based (mass spectrometry-based) metabolomics data continues to be a major bottleneck in metabolomics. Conventional MS-based procedures typically include peak identification and annotation prior to functional interpretation, which are prone to human bias and error. Further, previous versions of MetaboAnalyst encompassed two modules for functional analysis, metabolic pathway analysis (MetPA) and metabolite set enrichment analysis (MSEA), which required metabolite identifications prior to use. However, if users were uncertain of the validity of their metabolite identifications, they would be even more uncertain about the functional interpretation of metabolic features. One promising approach to reduce problems associated with compound misidentification and thereby pathway misinterpretation is to shift the unit of analysis from individual compounds to individual pathways. In particular, the *mummichog* algorithm bypasses the bottleneck of metabolite identification prior to pathway analysis by leveraging *a priori* pathway and network knowledge to directly infer biological activity based on MS peaks. Due of its popularity and repeated user requests, we have implemented the *mummichog* (version 1.0.10) algorithm in R to be consistent with MetaboAnalyst workflow in a new, user-friendly interface. The main steps for MS Peaks to Pathways are as follows:

- i. Upload your data as a table containing three columns, m/z features, p-values, and statistical scores (T-score, fold-change values).
- ii. Specify the mass accuracy and the ion mode of your mass-spec instrument, as well as the p-value cut-off.
- iii. Select the organism's library from which to perform pathway analysis.
- iv. View the pathway analysis results.
- v. Visualize the results in a global KEGG metabolic network.

## Data Upload Preparation

Upload your data in either a tab-delimited (.txt) or a comma separated value file (.csv) format. Make sure that the uploaded table contains three columns with these exact names: **m.z**, **p.value**, and **t.score**. An example dataset is shown below:

	A	B	C
1	m.z	p.value	t.score
2	304.2979	1.02E-10	14.7179316
3	177.1024	1.62E-10	14.2666
4	345.0277	1.72E-10	-14.209195
5	491.0325	1.83E-10	-14.146348
6	258.0048	2.17E-10	-13.987636
7	483.1205	2.22E-10	-13.967634
8	694.9937	2.81E-10	-13.745172
9	270.9767	3.27E-10	13.6060705
10	371.604	3.53E-10	-13.534483
11	316.5773	3.71E-10	13.4893333
12	451.0505	4.04E-10	-13.412347

## MS Peaks to Pathways Step-by-Step

**Step 1:** On the MetaboAnalyst home page, press “**click here to start**” to enter the module overview.



**MetaboAnalyst 4.0** -- a comprehensive tool for metabolomics analysis and interpretation

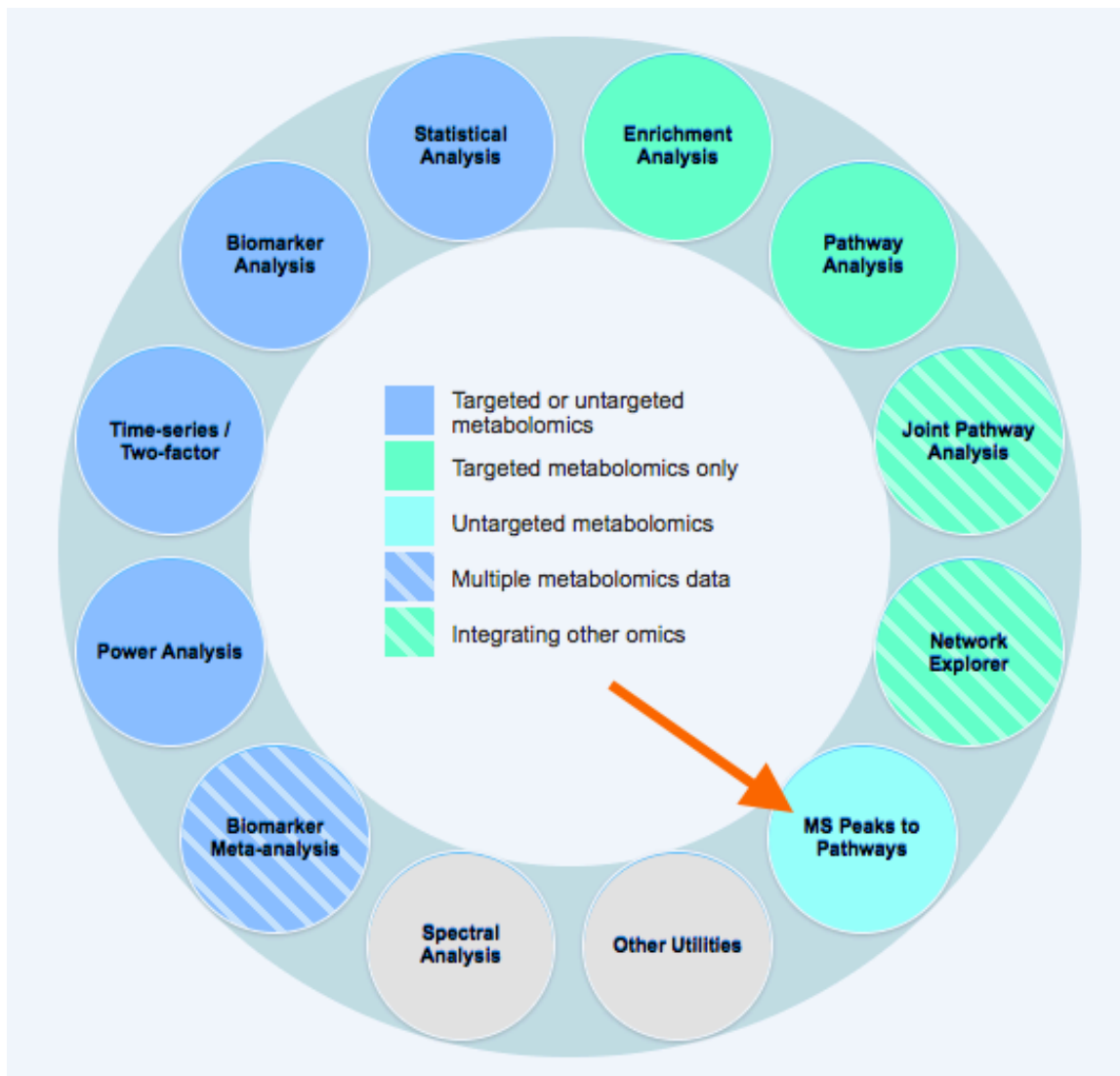
Home  
Overview  
Data Formats

Welcome >> [click here to start](#) <<

News & Updates

- Added a new tutorial for the three new modules (02/09/2018); **NEW**
- Minor bug fixes and feature enhancements based on user feedback (02/06/2018); **NEW**
- Release of **MetaboAnalyst 4.0** together with a companion R package **MetaboAnalystR**. You can still access [version 3.0 here](#) (01/29/2018); **NEW**
- Updated the interface for module selection (01/22/2018); **NEW**

**Step 2:** On the Module View page, click the “MS Peaks to Pathways” circle to begin.



**Step 3:** In the MS Peaks to Pathways upload page, first specify the “Mass Accuracy” and the “Analytical Mode” of your mass-spectrometry instrument (see screenshot below). These are both drop-down menus; select the option that best matches your instrument. Also select the “P-value cutoff” for downstream analysis, which is also a drop-down menu. Finally, upload your data using the “Choose File” button. Click “[Submit](#)” to continue. If p-values have not yet been calculated for their data, users can use the exploratory statistical analysis module to upload their raw peak tables,

process the data, perform t-tests or fold-change analysis, and then upload these results into the module.

The screenshot shows the 'Upload a peak list profile' section of the MetaboAnalyst 4.0 web interface. The page title is 'MetaboAnalyst-- a comprehensive tool for metabolomics analysis and interpretation'. On the left is a navigation menu with options: Upload, Data check, Set parameter, View result, Metabolic network, Download, and Exit. The main content area contains the following settings: 'Mass accuracy' set to 10 ppm, 'Analytical Mode' with 'Positive Mode' selected, 'P-value Cutoff' set to 1.0E-4 (marked as editable), and 'Choose Data File' with 'Choose File' and 'No file chosen' buttons. Below these settings is a checkbox for 'Use the example data', which is currently unchecked. A 'Dataset' link is provided, with a description: 'An example peak list data obtained from untargeted metabolomics using Orbitrap LC-MS (positive mode, human samples, p.value cutoff: 0.0001)'. A 'Submit' button is located at the bottom of this section. On the right side, there is an 'R Command History' panel with a 'Keep collapsed' checkbox and a 'Save' button.

**Use Case:** To use the example data, click the box next to “Use the example data” on the bottom of the upload page. The data will be uploaded automatically, which is a peak list dataset collected using Orbitrap LC-MS in positive mode. Click “Submit” to continue.

**Step 4:** After uploading the data, a data integrity check is performed to check the quality of your data, the number of features to be uploaded, and the total number of significant features based on your selected p-value cutoff. Click “Skip” to continue. Please note that on the right-hand side is the R Command history, which reveals the step-by-step of your analysis in real-time. This R Command History can be used to reproduce your analysis locally in R with the MetaboAnalystR package. Also, on the left-hand side is the navigation panel, which can directly navigate you through different steps of the MS Peaks to Pathways module.

The screenshot shows the 'Data Integrity Check' section of the MetaboAnalyst 4.0 web interface. The page title remains 'MetaboAnalyst-- a comprehensive tool for metabolomics analysis and interpretation'. The navigation menu on the left is updated, with 'Data check' highlighted. The main content area displays a list of four check items: 1. Checking the class labels - at least three replicates are required in each class. 2. If the samples are paired, the pair labels must conform to the specified format. 3. The data (except class labels) must not contain non-numeric values. 4. The presence of missing values or features with constant values (i.e. all zeros). Below this list is a dashed box containing 'Data processing information': 'Checking data content ...passed', 'A total of 3934 input mz features were retained for further analysis', 'The optimal number of significant features ~300', and 'A total of 261 significant mz features were found based on the selected p-value cutoff: 1e-04'. At the bottom of this box are two buttons: 'Missing value estimation' and 'Skip'. On the right side, the 'R Command History' panel is expanded, showing a list of R commands: 1. InitDataObjects("mass\_all", "numichog", FALSE); 2. mSet<-read.PeakListData(mSet, "ReplACING\_with\_your\_file\_path"); 3. mSet<-UpdateNumichogParameters(mSet, "ten", "positive", 1.0E-4); 4. mSet<-SanityCheckNumichogData(mSet)

**Step 5:** Following the data integrity check, select a pathway library that best fits your organism. Here, the knowledge-base consists of five genome-scale metabolic models obtained from the original Python implementation which have either been manually curated or downloaded from BioCyc, as well as an expanded library of 21 organisms derived from KEGG metabolic pathways. Note that there can exist several libraries for the same organism. For instance, one human genome-scale metabolic model has been manually curated and originates from a number of sources (Human MFN model - KEGG, BiGG, and Edinburgh Model), while the other genome-scale metabolic models are directly derived from BioCyc. As the example data are human samples, we will use the *Homo sapiens* (MFN) library. Click “Submit” at the bottom of the page to continue.

**Please select a pathway library:**

<b>Mammals</b>	<input checked="" type="radio"/> Homo sapiens (human) [MFN] <input type="radio"/> Homo sapiens (human) [BioCyc] <input type="radio"/> Homo sapiens (human) [KEGG] <input type="radio"/> Mus musculus (mouse) [BioCyc] <input type="radio"/> Mus musculus (mouse) [KEGG] <input type="radio"/> Rattus norvegicus (rat) [KEGG] <input type="radio"/> Bos taurus (cow) [KEGG]
<b>Birds</b>	<input type="radio"/> Gallus gallus (chicken) [KEGG]
<b>Fish</b>	<input type="radio"/> Danio rerio (zebrafish) [KEGG] <input type="radio"/> Danio rerio (zebrafish) [MTF]
<b>Insects</b>	<input type="radio"/> Drosophila melanogaster (fruit fly) [KEGG] <input type="radio"/> Drosophila melanogaster (fruit fly) [BioCyc]
<b>Nematodes</b>	<input type="radio"/> Caenorhabditis elegans (nematode) [KEGG]
<b>Fungi</b>	<input type="radio"/> Saccharomyces cerevisiae (yeast) [KEGG] <input type="radio"/> Saccharomyces cerevisiae (yeast) [BioCyc]
<b>Plants</b>	<input type="radio"/> Oryza sativa japonica (Japanese rice) [KEGG] <input type="radio"/> Arabidopsis thaliana (thale cress) [KEGG]
<b>Parasites</b>	<input type="radio"/> Schistosoma mansoni [KEGG] <input type="radio"/> Plasmodium falciparum 3D7 (Malaria) [KEGG] <input type="radio"/> Trypanosoma brucei [KEGG]

**Step 6:** The objective of the MS Peaks to Pathways module is to predict biological activity directly from peak list data, thereby bypassing metabolite identification. Following selecting an organism library, the predicted pathway activity of your data is listed in a table, identifying the top-pathways that are enriched in the uploaded data, ranked by the gamma-adjusted p-values. The table consists of the total number of hits, the raw p-value (Fisher's or Hypergeometric), the EASE score, and the gamma-adjusted p-value (for permutations) per pathway. From the example data in the screenshot below, we can see that Tryptophan metabolism, Ascorbate (Vitamin C) and Aldarate Metabolism, and Aminosugars metabolism are the top three enriched pathways in the example data.

#### Predicted pathway activity profiles based on Mummichog:

Click [View](#) under Match Details to view the compounds within each pathway (with matched compounds highlighted). The detailed pathways and matched compounds tables can be downloaded using the links [at the bottom](#).

[Explore Results in Network](#)

Pathway Name	Total	Hits (all)	Hits (sig.)	Fisher's Pvalue	EASE Score	Gamma Pvalue	Match Details
Tryptophan metabolism	94	64	21	0.0045504	0.0098086	0.0046682	<a href="#">View</a>
Ascorbate (Vitamin C) and Aldarate Metabolism	29	18	9	0.0026117	0.010691	0.0046835	<a href="#">View</a>
Aminosugars metabolism	69	29	12	0.0038443	0.011655	0.0047003	<a href="#">View</a>
Nitrogen metabolism	6	4	4	0.0012414	0.022604	0.0048951	<a href="#">View</a>
N-Glycan biosynthesis	48	14	7	0.0080322	0.032406	0.0050767	<a href="#">View</a>
Pyrimidine metabolism	70	43	14	0.020992	0.045203	0.0053243	<a href="#">View</a>
Vitamin B3 (nicotinate and nicotinamide) metabolism	28	19	8	0.015895	0.049225	0.0054047	<a href="#">View</a>
Sialic acid metabolism	107	28	10	0.025643	0.062907	0.0056878	<a href="#">View</a>
Alanine and Aspartate Metabolism	30	17	7	0.0272	0.08133	0.0060936	<a href="#">View</a>
Glutathione Metabolism	19	10	5	0.025367	0.099714	0.0065288	<a href="#">View</a>
Hexose phosphorylation	20	18	7	0.037539	0.10359	0.0066246	<a href="#">View</a>
Arginine and Proline Metabolism	45	31	10	0.051168	0.10984	0.0067823	<a href="#">View</a>
Glycosphingolipid biosynthesis - ganglioseries	62	11	5	0.039494	0.13447	0.0074426	<a href="#">View</a>
Glutamate metabolism	15	11	5	0.039494	0.13447	0.0074426	<a href="#">View</a>
Methionine and cysteine metabolism	94	46	13	0.075854	0.13715	0.0075184	<a href="#">View</a>
Purine metabolism	80	51	14	0.082458	0.1436	0.007704	<a href="#">View</a>
Parathio degradation	6	4	3	0.022858	0.16222	0.0082678	<a href="#">View</a>
Vitamin B9 (folate) metabolism	33	12	5	0.057555	0.17336	0.0086255	<a href="#">View</a>
Glycosphingolipid biosynthesis - globoseries	16	8	4	0.045914	0.17647	0.0087285	<a href="#">View</a>
Starch and Sucrose Metabolism	33	15	5	0.13495	0.30626	0.014409	<a href="#">View</a>

Download Result Tables: [Pathway Hits](#) [Compound Hits](#)

The pathway results table is available for download from the “[Pathway Hits](#)” link on the bottom right corner of the page. A second table is also available from the “[Compound Hits](#)” link for download that contains all matched metabolites from the user's uploaded list of m/z features. This table lists the query mass, the matched compound, the matched form, and the mass difference between the query and the matched compound.

Click “View” under the Match Details header to view which compounds in the pathway were matched with your data. Compounds highlighted in red represent significant hits, and blue compound names represent non-significant yet present hits.



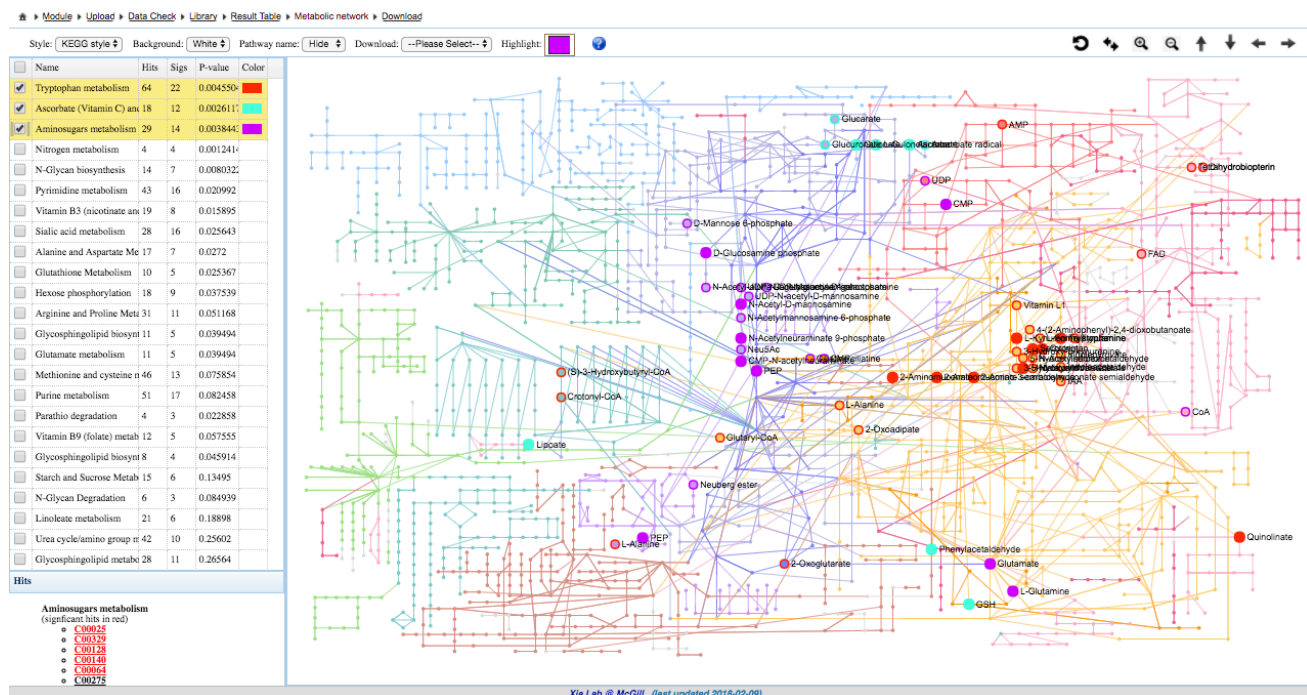
The colored compounds indicate potential matches from the user's input, with red colors indicating significant hits and blue colors indicating non-significant hits.

Pathway	Metabolites
Tryptophan metabolism	<p> <span style="color: red;">C00025</span>; C00024; C00027; <span style="color: blue;">C00026</span>; C00021; <span style="color: red;">C00020</span>; <span style="color: red;">CE2119</span>; C00028;  <span style="color: red;">C03722</span>; <span style="color: red;">C05647</span>; <span style="color: red;">C05645</span>; <span style="color: red;">CE1395</span>; <span style="color: red;">C05643</span>; <span style="color: red;">C05640</span>; <span style="color: red;">C00780</span>; <span style="color: red;">C05648</span>;            C00704; C01342; <span style="color: blue;">C00010</span>; C00014; <span style="color: blue;">C00016</span>; C00019; C15605; C00067;            thbpt4acam; <span style="color: red;">C05651</span>; C03512; <span style="color: red;">C05653</span>; <span style="color: red;">C02693</span>; <span style="color: red;">C05660</span>; <span style="color: red;">C00398</span>; <span style="color: red;">CE5982</span>;  <span style="color: red;">C00643</span>; <span style="color: red;">C02220</span>; C00078; C00978; C00877; <span style="color: red;">C05642</span>; <span style="color: red;">CE5980</span>; C01252;  <span style="color: red;">C05637</span>; <span style="color: red;">C02406</span>; <span style="color: red;">C00108</span>; <span style="color: red;">C00272</span>; <span style="color: red;">CE1916</span>; C01652; <span style="color: red;">C02470</span>; C01144;  <span style="color: red;">C06212</span>; <span style="color: red;">C06213</span>; <span style="color: red;">CE2949</span>; <span style="color: red;">CE2948</span>; <span style="color: red;">CE3140</span>; <span style="color: red;">CE2122</span>; C00479; <span style="color: red;">CE2947</span>;            C01717; <span style="color: red;">CE3092</span>; <span style="color: red;">CE6205</span>; <span style="color: red;">CE5899</span>; C03161; <span style="color: red;">C00268</span>; <span style="color: red;">C00632</span>; <span style="color: red;">CE2152</span>;  <span style="color: red;">CE2153</span>; <span style="color: red;">C02700</span>; <span style="color: red;">CE2095</span>; <span style="color: red;">C00328</span>; <span style="color: red;">CE3087</span>; <span style="color: red;">CE3086</span>; <span style="color: red;">C00322</span>; <span style="color: red;">C04409</span>;            C01352; <span style="color: red;">C00051</span>; C00058; C00030; <span style="color: red;">C00331</span>; C00332; <span style="color: red;">C00936</span>; <span style="color: red;">C03824</span>;  <span style="color: red;">C05636</span>; <span style="color: red;">C05635</span>; <span style="color: red;">C05634</span>; <span style="color: red;">C05639</span>; <span style="color: red;">C05638</span>; <span style="color: red;">C10164</span>; <span style="color: red;">C00637</span>; <span style="color: red;">C03227</span>;  <span style="color: red;">C01598</span>; <span style="color: red;">C00525</span>; <span style="color: red;">C00527</span>; <span style="color: red;">C00041</span>; C03024; <span style="color: blue;">C00954</span> </p>

OK



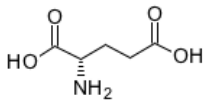
**Step 7:** To further visualize your results, click the “Explore Results in Network” button on the top of the pathway table. This will load the global metabolic network, where you can visually assess the global peak matching patterns of your data and interactively zoom into a particular candidate compound to examine all of its matched isotopic or adduct forms. The metabolic network visualization is based on the KEGG global metabolic network (KEGGscape) and has been manually curated. The global metabolic network page consists of three sections; the **top** section contains a toolbar for user-enabled editing, the **left** section contains pathway analysis results, and the main **central** section contains the metabolic network. Here, the metabolites of significantly enriched pathways are represented as nodes on the network, and their size is based on their associated pathway p-value. For example, click the empty box next to a pathway to view the matched compounds hits on the network (see screenshot below). Empty nodes represent compounds detected in your data but not significant, while solid nodes represent significantly enriched features detected in your data.



In the bottom left corner of the network visualization page is a box containing the KEGG compound identifiers for all matched compounds in a selected pathway. For instance, after highlighting the Aminosugars metabolism pathway, the box lists all of the pathway’s compounds that match with the user’s data, with significant hits highlighted in red. Clicking on a compound name from this box will link you directly to the KEGG page for that compound. For instance, if we click “C00025”, which is

the first red compound in the box, a new tab opens in the browser that takes us to KEGG (screenshot below).

**KEGG** COMPOUND: C00025 Help

<b>Entry</b>	C00025	Compound
<b>Name</b>	L-Glutamate; L-Glutamic acid; L-Glutaminic acid; Glutamate	
<b>Formula</b>	C5H9NO4	
<b>Exact mass</b>	147.0532	
<b>Mol weight</b>	147.1293	
<b>Structure</b>	 C00025	

**All links**

- Ontology (1)
- KEGG BRITE (1)
- Pathway (57)
- KEGG PATHWAY (48)
- KEGG MODULE (9)
- Drug (5)
- KEGG DRUG (1)
- KEGG ENVIRON (3)
- CHEMBL (1)
- Chemical substance (45)
- PubChem (1)
- ChEBI (1)
- 3DMET (1)
- HMDB (1)
- KNApSacK (1)
- MASSBANK (37)

Returning to the MetaboAnalyst global metabolic network view, if we double-click a highlighted node, we will see corresponding metabolite information such as adduct matches and fold-change/t-score values (screenshot below).

**D-Mannose 6-phosphate**

- M+H[1+]: 4.011266683
- M+Na[1+]: 1.805258786

Further, the toolbar at the top of the page provides interactive options for changing the background color of the network (black or white), changing the network view style (KEGG, expression, or plain), and highlighting of user-specified pathways in any color.

Style:  Background:  Pathway name:  Download:  Highlight:

To use the highlight feature, click the coloured box at the top of the toolbar, and then use the palette to select a color, then select “choose”. You can then click the pathways to highlight the matched compounds in the selected color. Mouse scrolling for zooming-in and out of the network is also

enabled. Further, the maps can be downloaded as PNG or SVG files for publication/report purposes. Overall, this metabolic network visualization provides an opportunity for you to visually explore your results, as well as provides a global metabolic context for the significantly enriched pathways in your data.

**Step 8:** On the top of the network visualization, click “Download” to view the tables generated throughout the MS Peaks to Pathways module, as well as to download the Analysis Report, which contains all the details of each step of your analysis as well as all of the results. You can also directly download all the results in a zip file by clicking “Download.zip”.

#### Result Download

Please download the PDF Analysis Report and results (tables and images) below. The "Download.zip" contains all the files in your home directory. The PDF report may not be generated sometimes. You can try to re-generate PDF using an alternative approach using the button below.

[Regenerate](#) [Analysis Report](#)

<a href="#">Download.zip</a>	<a href="#">mummichog_matched_compound_all.csv</a>
<a href="#">Rhistory.R</a>	<a href="#">mummichog_pathway_enrichment.csv</a>

[Logout](#)

-- End of tutorial --