## Method overview

Briefly, we performed a Feature-Based Molecular Networking ([FBMN](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/)) workflow on the untargeted metabolomics peak height data using [MZmine3](https://github.com/mzmine/mzmine) as the feature identifier and [GNPS2](http://www.gnps2.org) as the analysis platform for FBMN. An MZmine3 workflow was used to generate a list of features (mzrt values obtained from extracted ion chromatograms containing chromatographic peaks within a narrow m/z range) and filtered to remove isotopes. For each feature, the most intense fragmentation spectrum was uploaded to the web-based mass spectrometry identification tool GNPS2 (Global Natural Products Social Molecular Networking). When a sample mass spectrum identified by MZmine3 matches one deposited within the GNPS2 database, a putative identification is made. [FBMN](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/) (Feature-based Molecular Networking) links all features - including those without MSMS database matches - by structural similarity which can be visualized with a program such as [Cytoscape](https://cytoscape.org/) to help in the identification of unknowns. It should be noted that a spectra match to a database spectra is not a definitive identification of the feature - it could be an isomer with a similar fragmentation, an ion with a close but not exact m/z but similar fragmentation pattern, or an in source degradation product of another larger molecule (degradation product may look similar to database match).

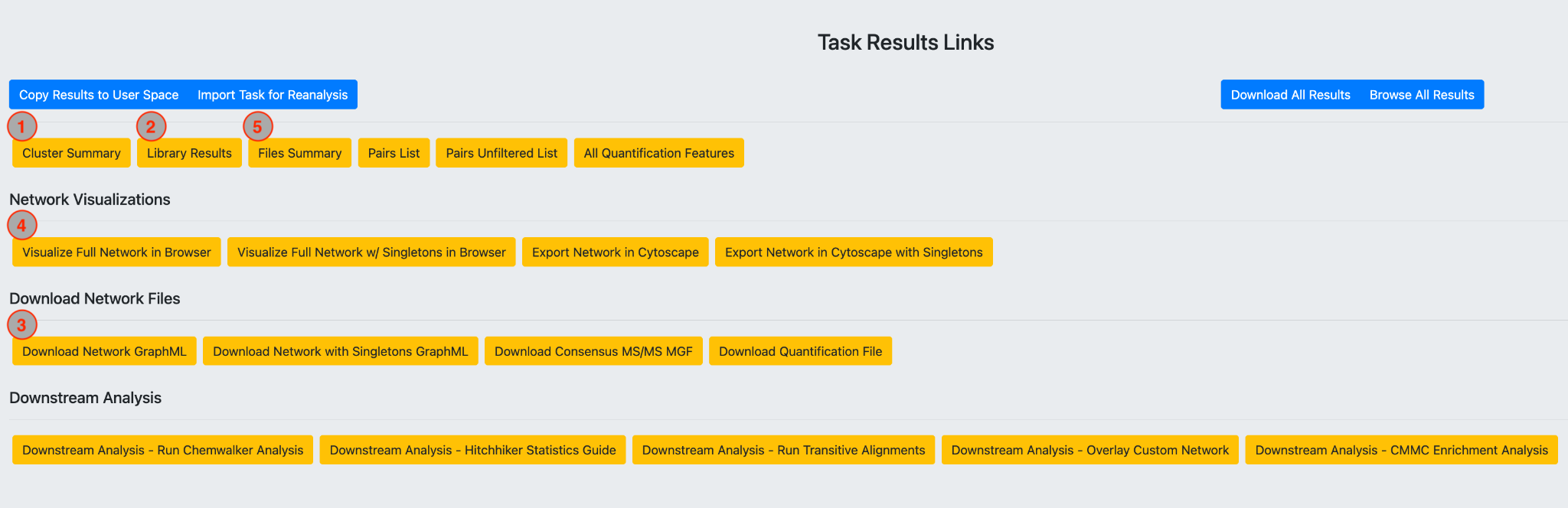
## Connecting to and navigating GNPS2

Your data and analysis results have been uploaded to GNPS2, which allows for interactive exploration and analysis of untargeted metabolomics features generated with our workflow. To get your GNPS2 account, go here: <https://gnps2.org/user/signup>. **Note**: You will need a token to get an account. Please ask [Christa Pennacchio](mailto:cppennacchio@lbl.gov), [Katherine Louie](mailto:kblouie@lbl.gov), [Thomas Harwood](mailto:tharwood@lbl.gov), [Brandon Kieft](mailto:bkieft@lbl.gov), or [Benjamin Bowen](mailto:bpbowen@lbl.gov) for a token.

Once you have your account set up, you can navigate to your FBMN results page using the link provided in the zipped folder file (<project>\_<polarity>\_gnps2-page-link.txt).

From that page, you can interact with a view your data in a number of ways, including:

1. View boxplots of feature abundance per experimental treatment with “Cluster Summary”:
2. View a table of all library hits with “Library Results” (same as the zipped folder file <project>\_<polarity>\_gnps2-fbmn-library-results.tsv):
3. Download your network for viewing in Cytoscape and pie-charts with “Download Network GraphML” (same as the zipped folder file <project>\_<polarity>\_gnps2-fbmn-network.graphml):
4. Or, visualize the network directly on your browser via the GNPS2 server with “Visualize Full Network in Browser”:
5. View Raw Data and EICs with “Files Summary”:



## Zip archive of your results from Google Drive

### For each of your polarities, the following files will be in the corresponding subfolder:

<project>\_<polarity>\_batch-params.xml MZmine parameters

<project>\_<polarity>\_filelist.txt List of all mzML files

<project>\_<polarity>\_gnps2-fbmn-library-results.tsv GNPS2 results table

<project>\_<polarity>\_gnps2-fbmn-network.graphml FBMN (network) file

<project>\_<polarity>\_gnps2-fbmn-task.txt FBMN taskID

<project>\_<polarity>\_gnps2-page-link.txt GNPS2 link to results

<project>\_<polarity>\_metadata.tab Project metadata table

<project>\_<polarity>\_mzmine-sbatch.sbatch MZmine submission

<project>\_<polarity>\_mzmine.sh MZmine runner

<project>\_<polarity>\_peak-height.csv MZmine **peak height** table

<project>\_<polarity>\_peak-height-filtered-\*\*\*.csv ^Background signal removed

<project>\_<polarity>\_peak-height-filtering-summary.csv ^Details on feature filter

<project>\_<polarity>\_quant.csv MZmine **peak area** table

<project>\_<polarity>-mzmine.err MZmine log

<project>\_<polarity>-mzmine.out MZmine done checker

<project>\_<polarity>.mgf MZmine spectral file

### Tips and details on the results files

The GNPS2 outputs (i.e., \*gnps2-fbmn-library-results.tsv and \*\_gnps2-fbmn-network.graphml)are derived from features in the “peak height” (\*peak-height.csv) and “peak area” (\*quant.csv) tables. For example, in your project’s POS mode GNPS2 network, if feature 300 is identified as an interesting metabolite, you can look at feature 300 in the tables and explore their intensity levels across your samples.

Other potentially useful files are the: <project>\_<polarity>\_mzmine-sbatch.sbatch and <project>\_<polarity>\_batch-params.xml files: these files let you reproduce the MZmine run exactly should you choose to re-run it on your own. These are the exact commands that were used for peak detection.

The <project>\_<polarity>.mgf file can be used to search your spectra for candidate hits using third party tools. It is mappable to both your peak tables and the GNPS2 results/network.

When checking out the <project>\_<polarity>\_gnps2-fbmn-library-results.tsv table, note that there can be multiple hits per feature. This is typical when a single MSMS can have multiple plausible identifications. Hits are generally ranked in descending order of quality, but note that the topmost hit for a particular feature is not always the “best” or “correct” one.

You can also use the <project>\_<polarity>\_peak-height.csv table (or the filtered version, see details below) to calculate fold change between treatments or another statistical method you are interested in. Paint this data into the molecular network using the cross reference attributes below.

### Features in the different results files can be cross-referenced by the following:

While the numeric feature IDs themselves are synchronous through the untargeted metabolomics analysis (i.e., a feature’s name is consistent and carried through), the header of the feature column is not always identical between output tables. In case your data munging or merging code or manual effort is failing, here are cross-references:

1. In <project>\_<polarity>\_peak-height.csv, features are in column “row ID”
2. In <project>\_<polarity>\_gnps2-fbmn-library-results.tsv, features are in column “#Scan#”.
3. In the table viewer below the FBMN when you click the “Visualize Full Network in Browser” button in GNPS2, the features are in column “id”.
4. If you open <project>\_<polarity>\_gnps2-fbmn-network.graphml in CytoScape or another tool, in the table viewer below the network the features are in column “shared name” or “name”.

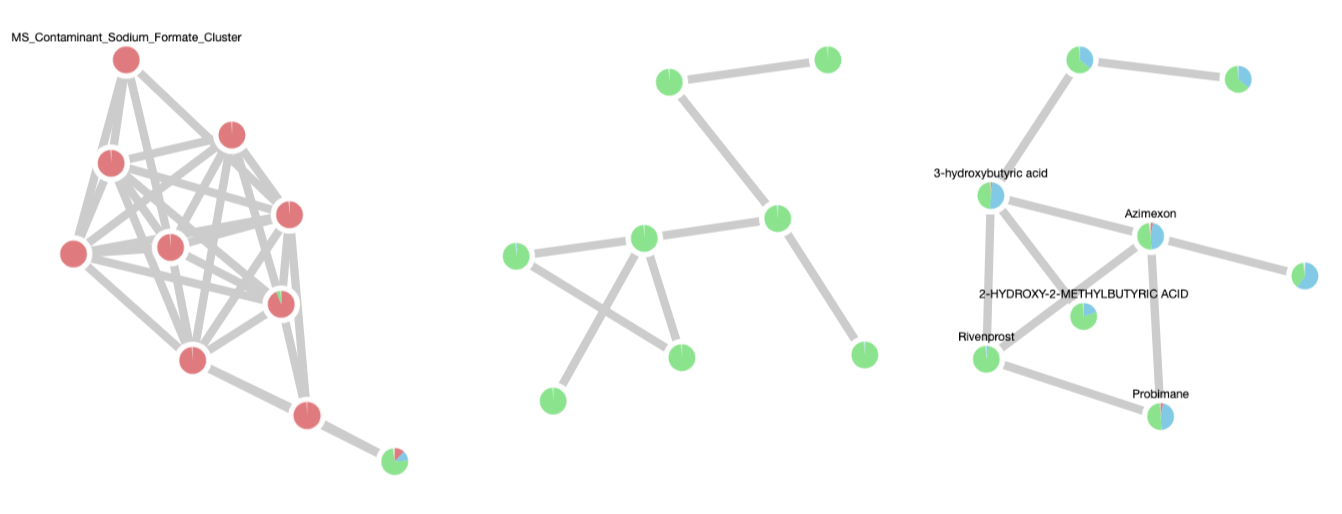
### Filtering features abundant in the extraction control

While the <project>\_<polarity>\_peak-height.csv file contains all features that were identified with the untargeted metabolomics run, some features may not have an intensity above a designated background (e.g., extraction control) and therefore should be removed as potentially spurious hits. This is done automatically in the <project>\_<polarity>\_peak-height-filtered-\*\*\*.csv file. This filtered peak height file includes some convenient and basic filters of the untargeted features (in this order):

1. Features are removed if they fall outside the reliable solvent migration time of the chromatography column (defaulted to 0.5-10 mins for non-polar and 0.8-17.5 mins for polar)
2. Features are removed if they do not have a maximum peak height (across all experimental samples) that is [NUMBER] times larger than (defaulted to 3X) the maximum peak height of all background samples (usually extraction controls).
3. Feature peak heights of 0.0 are replaced with ⅔ of the smallest peak height value in the dataset (to aid in downstream statistical analysis).

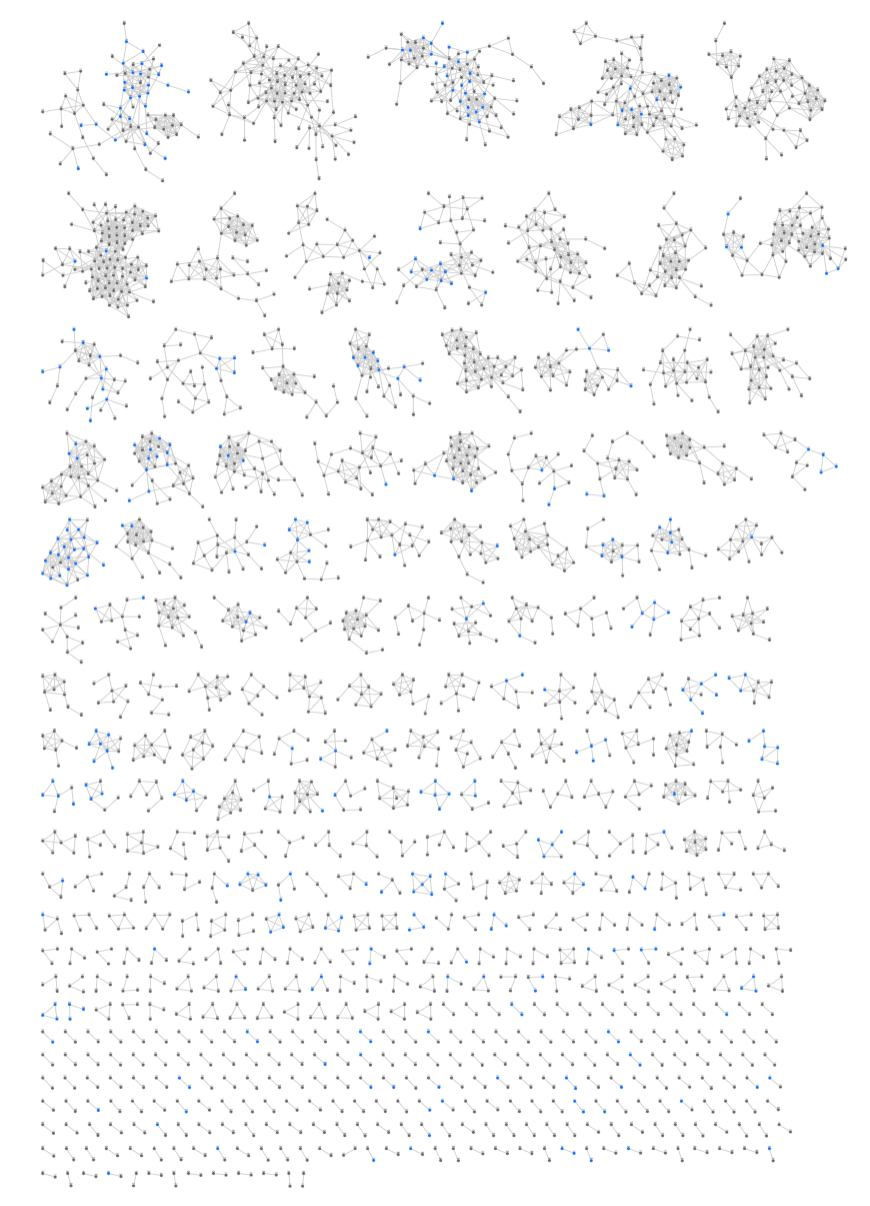
## Interpreting the FBMN (network) from GNPS2

When viewing the FBMN encoded in <project>\_<polarity>\_gnps2-fbmn-network.graphml (e.g., in Cytoscape or on GNPS2’s interactive page), it is a good idea to color code the nodes in your network by categorical observations (Figure 1). The most important observation is to label the signals that are found in your blank and/or control samples (typically labeled with ‘ExCtrl’ or ‘TxCtrl’ in the file names). These signals are often as much as half the detected features and result from contaminants that have leached from solvents, HPLC capillaries, vial septa, and other materials exposed to the samples during the extraction processing and/or LCMS run. One way to do this in GNPS2 is to navigate to “Visualize Full Network in Browser” and use the “Pie Configuration” tool to set all “ExCtrl/ExBlk” (extraction control/blank) samples to one color and all experimental samples to another color(s). Features (points) on the network with mostly or all control/blank signal are likely background.



***Figure 1:*** *An example of how to choose pie chart colors using the GNPS2 interface (left) and an example of how the nodes in the network can be visualized by eye as “experimental” in mostly blue/green or “background” in mostly red (right). Ignore the features that are mostly found in extraction controls.*

For the full network for each polarity, each node has a numeric “ID” that can be traced back to the GNPS2 results table and can potentially be mapped to a known compound name. By default, features with an acceptable match to a known compound will be colored blue, while those without a confident match will be colored gray (Figure 2). To see the identify of the match, you can use the “Network Controls” section to change the node labels to “Library Match” (Figure 3). Note that some clusters in the network will have dense representation in databases and some will have sparse.



***Figure 2.*** *Each node in the GNPS2 network corresponds to a feature. There can be 1000s of features. Typically, these correspond to 100s of metabolites and the degeneracy is due to various physical phenomena associated with ionization of metabolites.*

### 

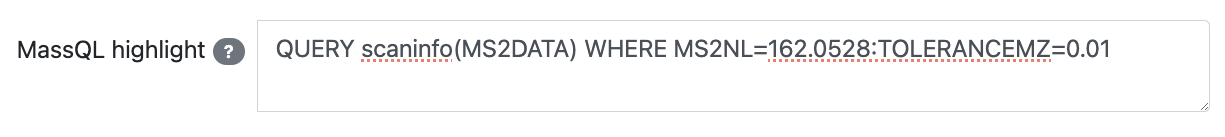
***Figure 3.*** *Visualizing the identity of known metabolites in the feature-based molecular network.*

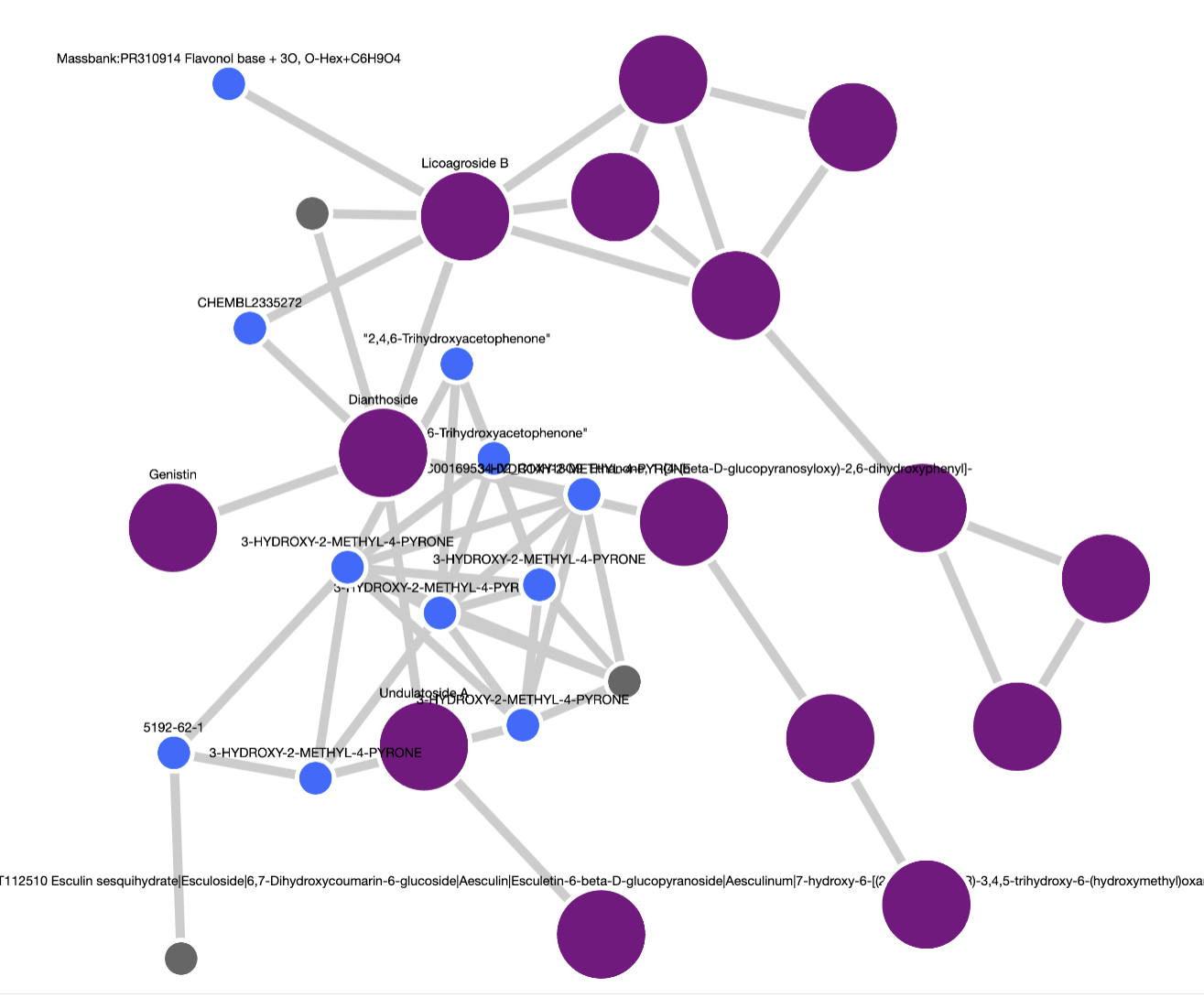
## 

## Methodological notes and tips

### Identification of metabolites using the Mass Spectrometry Query Language

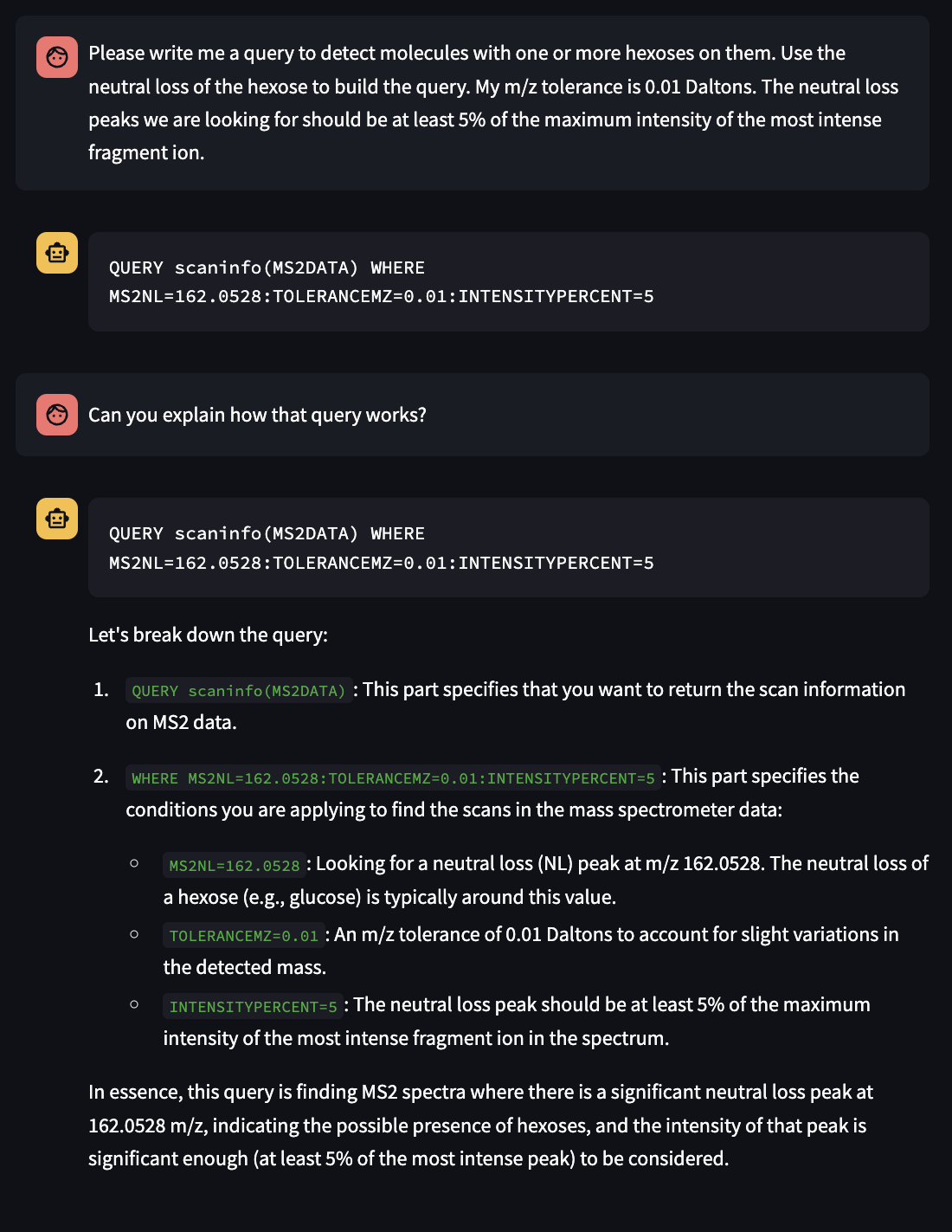
[MassQL Documentation](https://mwang87.github.io/MassQueryLanguage_Documentation/)





[MassQL LLM/AI Chatbot and Explainer](https://massql-analysis.gnps2.org/MassQL_Chatbot)

[MassQL Compendium](https://massql.gnps2.org/compendium/)



### Making pie-charts in Cytoscape

Please refer to “Reproducible molecular networking of untargeted mass spectrometry data using GNPS” in [Nature Protocols 2020: DOI: 10.1038/s41596-020-0317-5](https://www.nature.com/articles/s41596-020-0317-5).

### Why do my features not have an identification?

The vast majority of features are not identified typically. It can be because they are understudied molecules with no reference signals available or because they are artifacts of the ionization process.

### Adducts and Degeneracy

Often there will be multiple peaks at the same retention time. These degenerate peaks can be from adducts and in source degradation. For deadducting, considering using MS-FLO: <https://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b04372>

This tool will look for co-eluting adducts and label them.

* Note: Isotopes have been filtered out of the feature table.

### Ion Suppression

The artifacts of ion suppression and non-linear dependencies using electrospray ionization are well known yet poorly understood.  [Recent work aims](https://pubs.acs.org/doi/abs/10.1021/acs.analchem.0c04113) to correct these artifacts. In practice, internal standards can help to identify fold-change requirements that yield meaningful comparisons.

### Ways to reduce/refine total features

The features identified by MZmine3 have been automatically filtered for you by the methods described above to reduce background signal and remove suspect features (see the file <project>\_<polarity>\_peak-height-filtered-\*\*\*.csv).

Additional filters that may be useful on a per-project basis:

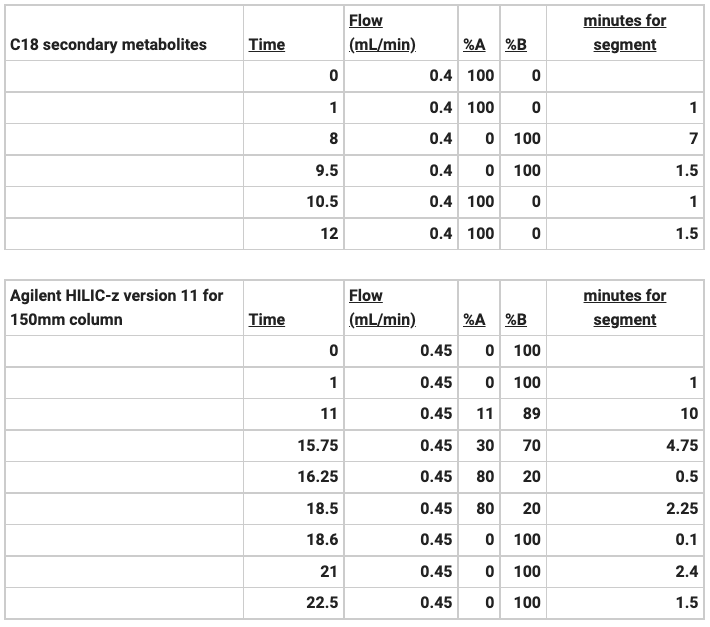
1. Increase the value of the treatment samples against background (e.g., 5X instead of the standard 3X)
2. Require minimum intensity in treatment groups of interest
   1. Calculate the average intensity for each treatment group
   2. Remove features where the average intensity in treatment groups of interest is less than a threshold (typically ~1e6 or ~1e4, depending on which machine was used to run your samples. In general, if the file names in your project zip folder have “IQX” or “IDX” in them, use 1e4; otherwise, 1e6.)
3. Calculate a log-fold-change between treatments for each feature to determine which features were most impacted by experimental design

### Timetables for C18/nonpolar and HILICZ/polar runs

For C18, the gradient starts at 1 min and ends at 8 - isocratic elution from 8-10 minutes. Ignore anything AFTER 10 minutes and anything BEFORE 0.5 minutes (at/before solvent front).

For HILICZ, the gradient starts at 1 min and ends at 16.25 - isocratic elution from 16.25-18.5 minutes. Ignore anything AFTER 17.5 minutes and anything BEFORE 0.8 minutes (at/before solvent front).

See table below for details:



Don’t hesitate to ask any questions. We’ll do our best to get back to you as soon as possible.

Ben