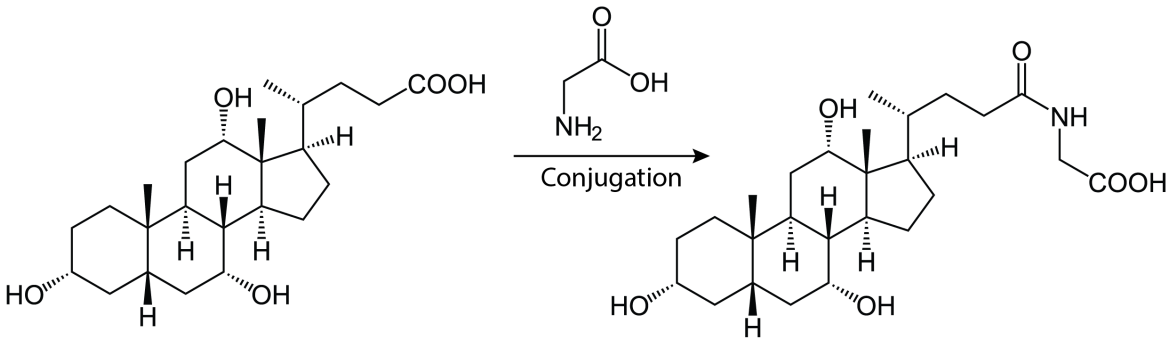
Tutorial: in silico library generation

The following tutorial provides step-by-step instructions for in silico library generation as described in the accompanying white paper.

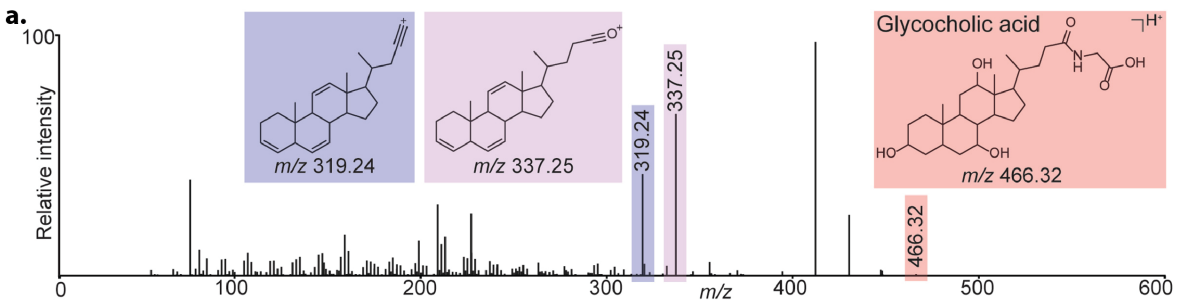
1. Construct a MassQL query

MassQL is a programmatic language for searching MS/MS data. In this example, we will be searching for all spectra that contain the core structure of a trihydroxylated bile acid. All trihydroxylated bile acids share this core structure, which can then be modified at a carboxyl group to create different bile acids.



Trihydroxylated Bile Acid Core **Modified Trihydroxylated Bile Acid**

To search for these molecules, we identified two MS2 peaks from the core structure that are shared by all trihydroxylated bile acids. The following figure shows these two peaks and the massql query designed to search for them.



b.

MassQL query:

QUERY scaninfo(MS2DATA) WHERE
MS2PROD=337.25:TOLERANCEMZ=0.01:INTENSITYPERCENT=5 AND
MS2PROD=319.24:TOLERANCEMZ=0.01:INTENSITYPERCENT=5

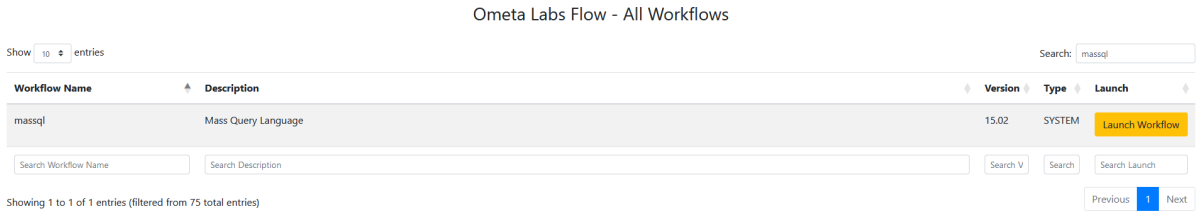
English translation:

Returning scan information on MS2:
Find MS2 peak at m/z 337.25 with a 0.01 m/z tolerance and a minimum intensity of 5% relative to the base peak
Find MS2 peak at m/z 319.24 with a 0.01 m/z tolerance and a minimum intensity of 5% relative to the base peak

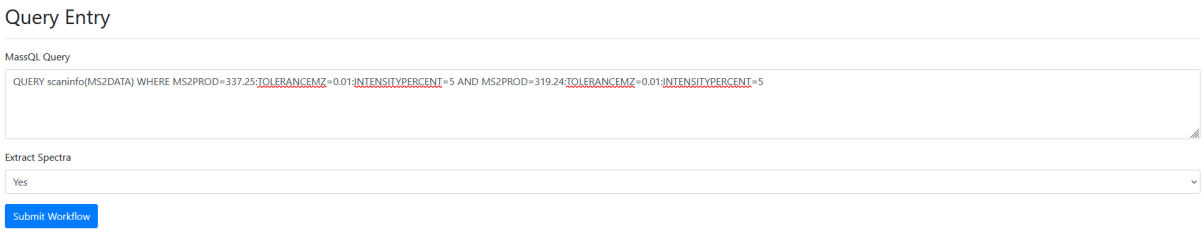
For more information on how to construct a MassQL query, see our MassQL documentation

2. Run MassQL

To run MassQL, navigate to the massql workflow in the workflows tab. This is easily done by typing "massql" into the search bar.



Clicking the Launch Workflow button will bring up the massql workflow input page. Here, you can choose which files to search (in the Input Data Folder) and provide optional network .graphml files and sample metadata. Most importantly, you'll input your MassQL query in the query entry box.



Here is the text version of that query:

```
QUERY scaninfo(MS2DATA) WHERE
MS2PROD=337.25:TOLERANCEMZ=0.01:INTENSITYPERCENT=5 AND
MS2PROD=319.24:TOLERANCEMZ=0.01:INTENSITYPERCENT=5
```

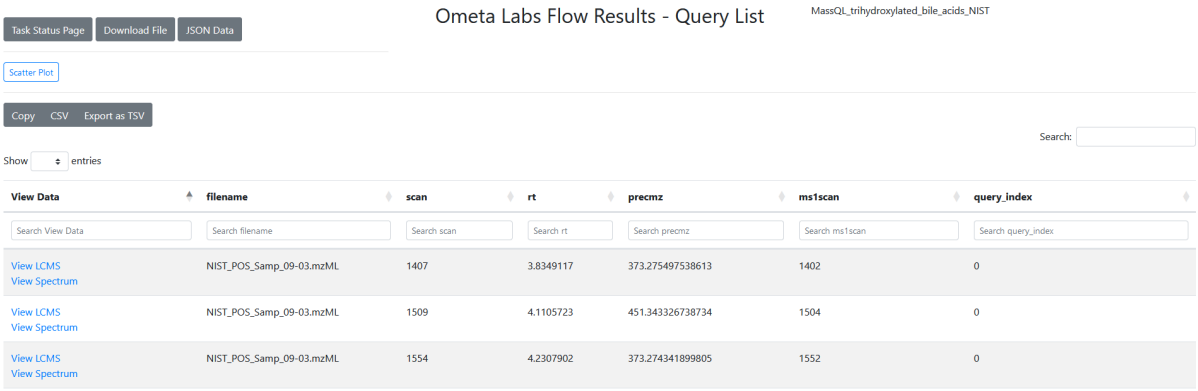
MassQL from molecular networks

You can auto-populate the MassQL workflow with the results of a molecular networking job using the "Downstream Analysis - MassQL Query" button on your network's Task Status page. You can also visualize MassQL queries directly from your network dashboard! This feature will highlight all the nodes in your network that match your MassQL query. For more information on how to use this feature, see the MassQL documentation.

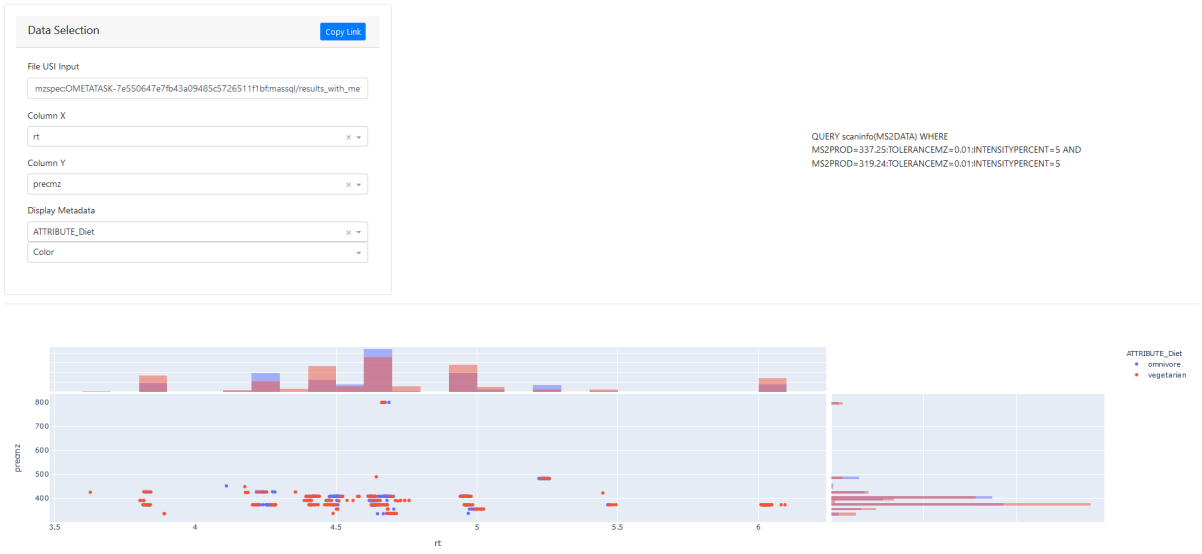
Click "Submit Workflow" to start the analysis.

3. Export MassQL results

MassQL will provide a list of spectra in input files that matched your query. If this is your first time constructing a MassQL query, we recommend you spend some time navigating through these spectra. You can find the list of all matched spectra using the "Query List" button on your results page. This will not only bring up all matches, but allow you to see the view the XIC trace and MS2 spectrum for each match using the "View LCMS" and "View Spectrum" links.



If you provided metadata on the files you searched, you can also visualize the distribution of your matches by clicking on the "Downstream Analysis - Scatter Plot Results with Metadata" button. For example, here is a scatter plot showing the precursor m/z and retention times of all the trihydroxylated bile acid matches in fecal samples from omnivore and vegetarian volunteers.



Once you're confident that your results are accurate, you can export all matched spectra as .mzML or .mgf files by clicking "Browse All Results", navigating to the folder labeled "massql" and downloading the *extracted_[task number].mzML* and *extracted_[task number].mgf* files.

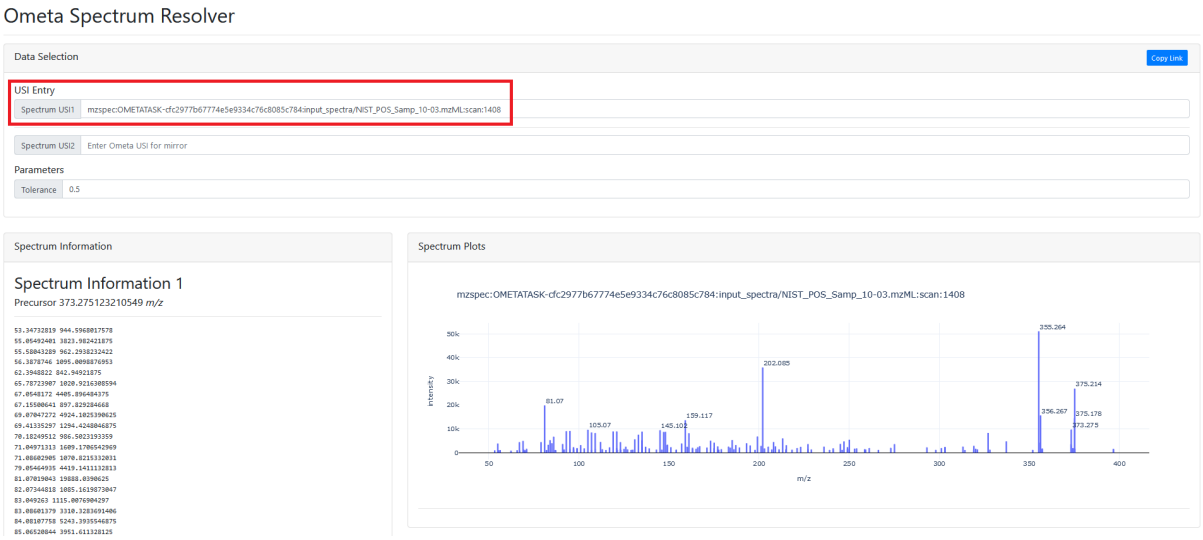
Note: if you don't see these files, go back to your task status page and make sure you set "Extract Spectra" to "Yes" in your MassQL workflow settings.

4. Explore matched spectra using MASST

Before creating a spectral library from your MassQL results, it can be helpful to see where these spectra occur in other data. The search tool MASST can rapidly identify all spectra that match an experimental spectrum on a database level, giving you an idea of where that spectra exists in all your other data. There are two main ways to use MASST in Ometa Flow:

If you want to look at a single spectra:

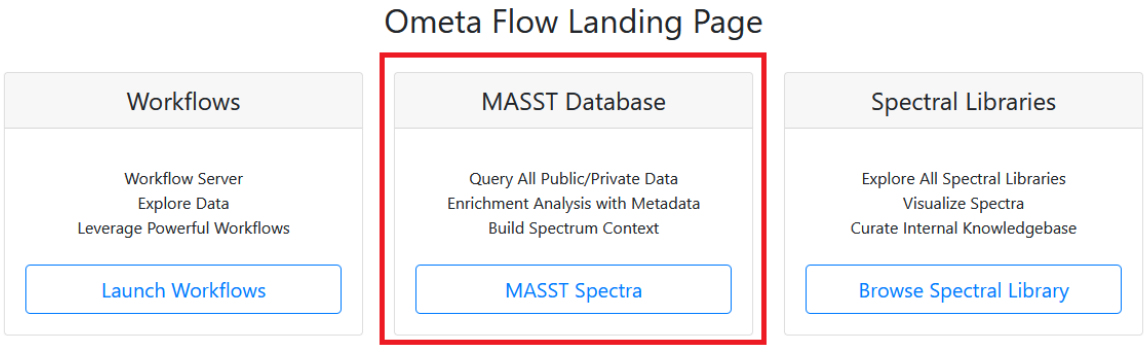
To MASST search a single spectrum, find a spectrum of interest in your query list and click the "View Spectrum" link to pull up that Spectrum Resolver. The Spectrum Resolver provides a spectrum plot, text list of spectrum peaks, and the spectrum USI.



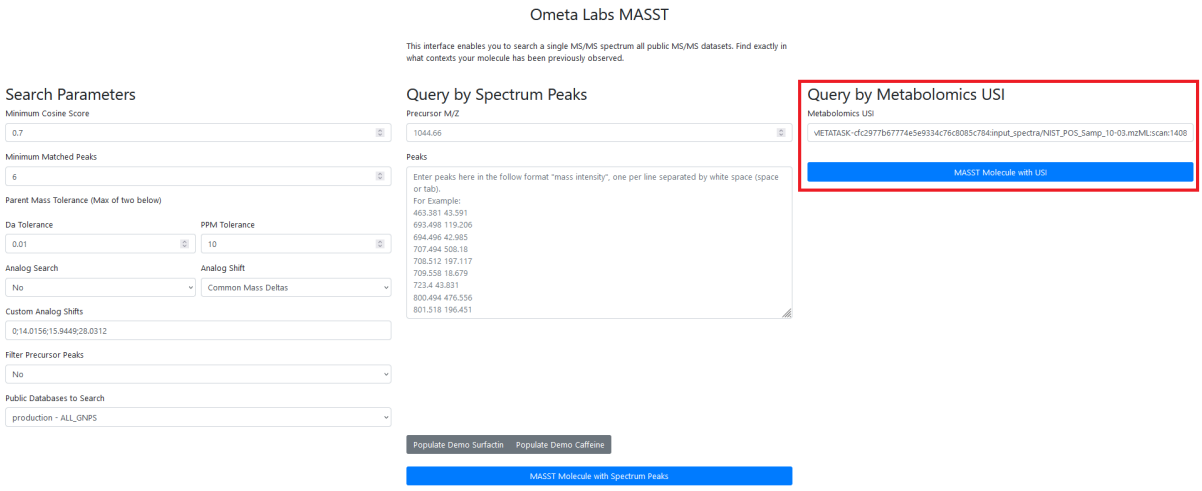
Spectrum USI

USI stands for Universal Spectrum Identifier and is a unique identifier for your spectrum. This unique identifier can be used across the Ometa Flow platform and will always point to this specific spectrum. For more information, check out our USI documentation.

To search this spectrum, copy the spectrum USI (highlighted above) and navigate to the MASST Database box on the Ometa Flow Homepage.



Clicking "MASST Spectra" will bring you to the Ometa Labs MASST page. On the left, you can edit your search parameters and choose which database to search (for more information, check out the MASST documentation). Paste your USI into the Metabolomics USI box on the far right and click "MASST Molecule with USI".



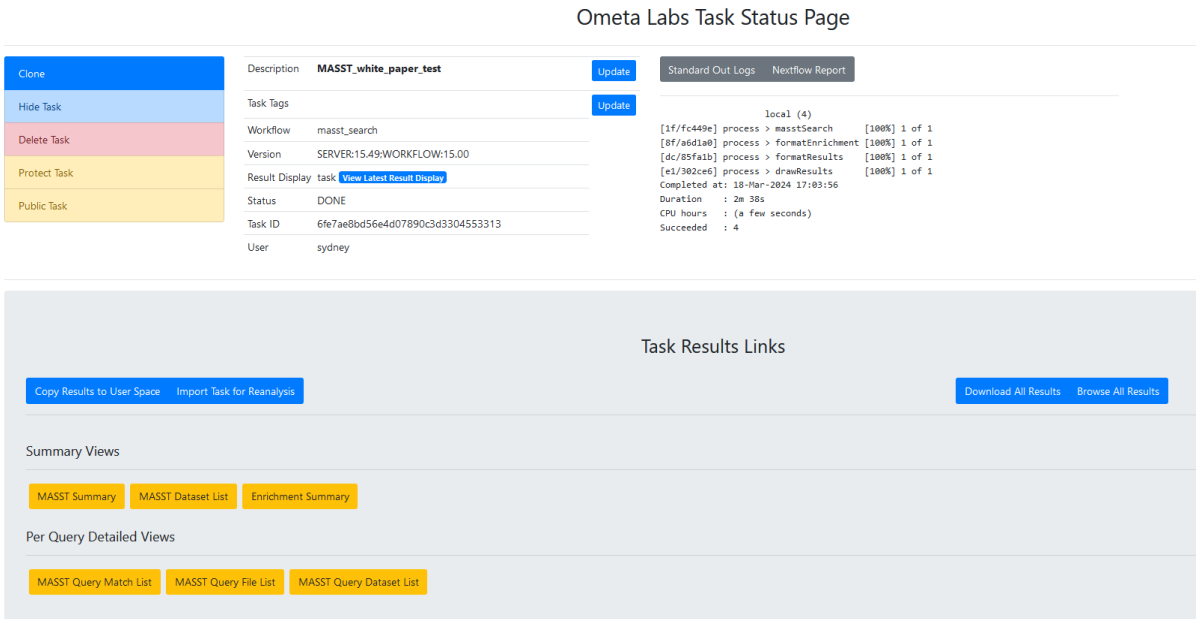
This will bring up the results page with the list of datasets, files, and specific scans that match your spectrum of interest.

> **MASST Results**
>
> Your MASST results will depend on which database you search. The most basic version will provide a list of filenames that contained matched scans. If your database has curated metadata, you may also be able to sort matches by organism, phenotype, sample collection method, etc. For questions on how to get the most out of your MASST searches, contact us.

If you want to look at all your spectra:

To MASST search all the spectra from your MassQL query, you can use the `masst_search` workflow. To do so, find `masst_search` on your Workflows page and click "Launch Workflow". On the Workflow Input page, add a job description and select the *extracted_[task number].mgf* file you downloaded from your MassQL results. Set your search parameters and then click "Submit Workflow".

This workflow will search all your query spectra against the database you chose and find all matching datasets, files, and scans. Summary Views will provide combined results for all your query spectra, while you can find matches to individual query spectra in the Per Query Detailed Views.



Interpreting MASST Results

MASST results can be a great way to quickly validate your MassQL queries. Let's take the spectra from our trihydroxylated bile acid query as an example. Bile acids are regularly found in fecal samples from various animals. If a MASST search of one of my query spectra brings up matches in fecal samples, that's great, but if all the matches are to marine sponges, maybe I need to keep refining my MassQL query.

5. Create an in silico reference library

Once you're comfortable with your MassQL results, you can add your query spectra into an in silico library. We recommend creating a new library for your in silico results so that they are easy to differentiate from other types of library spectra.

Creating new libraries

By default, creating new libraries is only available to admins. You can either ask your admin to create a library for you using the `lc_annotate_library_create` workflow, or you can ask your admin account to give your user access to this workflow.

To upload your spectra as a batch, use the `lc_annotate_spectrum_batch` workflow. For this

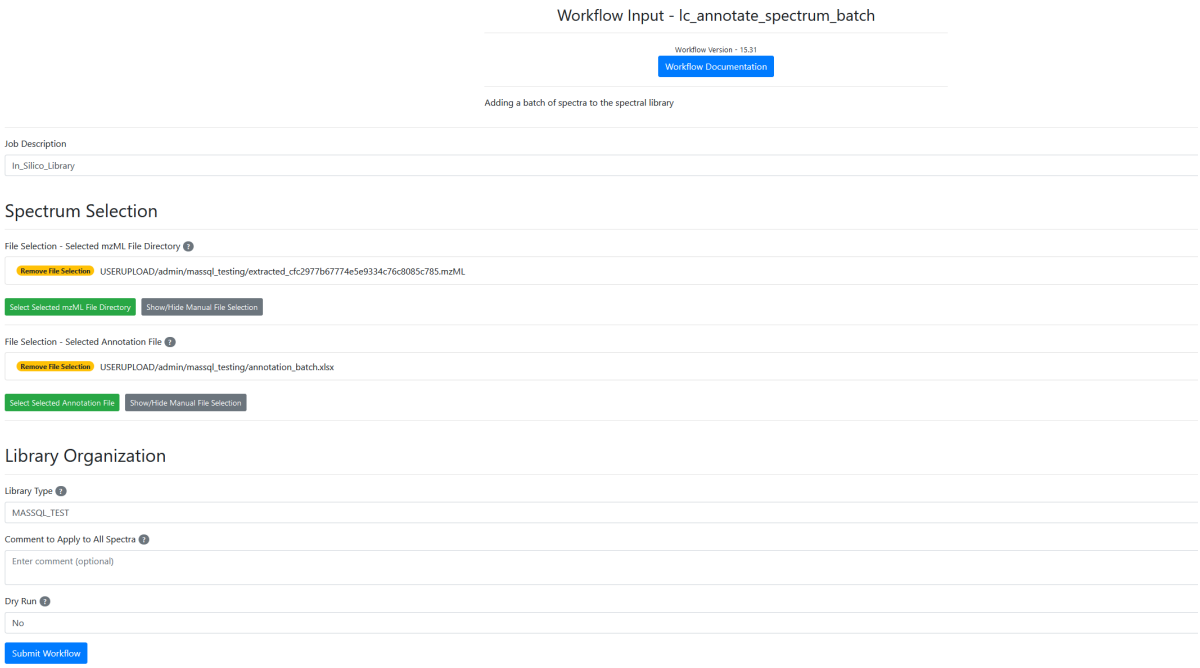
workflow, you'll need to upload your spectra as the *extracted_[task number].mzML* file you downloaded from your MassQL results. You will also need to upload an annotation file with information for each library spectrum. The minimum required information is below:

Parameter	Description
filename	Name of .mzML file containing the library spectra (in this case, *extracted_[task number].mzML*)
spectrum_id	Scan number of library spectrum. The scan number for each spectra can be found in the Extracted List from your MassQL results under the "new_scan" column
compound_name	Compound name. This can be anything, from a chemical name to "Candidate Compound 1"

Filtering spectra

If there are MassQL results you don't want to add to your library, just don't include them in your library annotation file. All scans not included in the annotation file will be ignored.

To see what other information can be included and see an example annotation file, see the library documentation. To check whether your files are compatible without actually adding the spectra, you can keep the "Dry Run" option as "Yes". If you want add your spectra to the library, change this option to "No".



This workflow will add in your library spectra along with any information you've provided on data collection, adducts, structures, etc.

Naming in silico libraries

In most cases, you won't be able to provide an exact compound name or structure with just the information from your MassQL query. However, you can still include useful information in the compound name. For example, since we know the exact mass of our bile acid core from our trihydroxylated bile acid query, we can subtract the mass of the core from the precursor m/z to get a "delta mass" corresponding to the mass of the conjugation. Thus, our compound name could be "Candidate trihydroxylated bile acid (delta mass: 76.64)". If you later learn more information about a specific library compound, you can update this information using the `lc_annotate_update_spectrum` workflow.

Once you've created your in-silico library, you can use this reference library in all other Ometa Flow workflows.

This tutorial is adapted from results presented in the following publication: Mohanty, et. al. The underappreciated diversity of bile acid modifications. Cell 0, (2024).