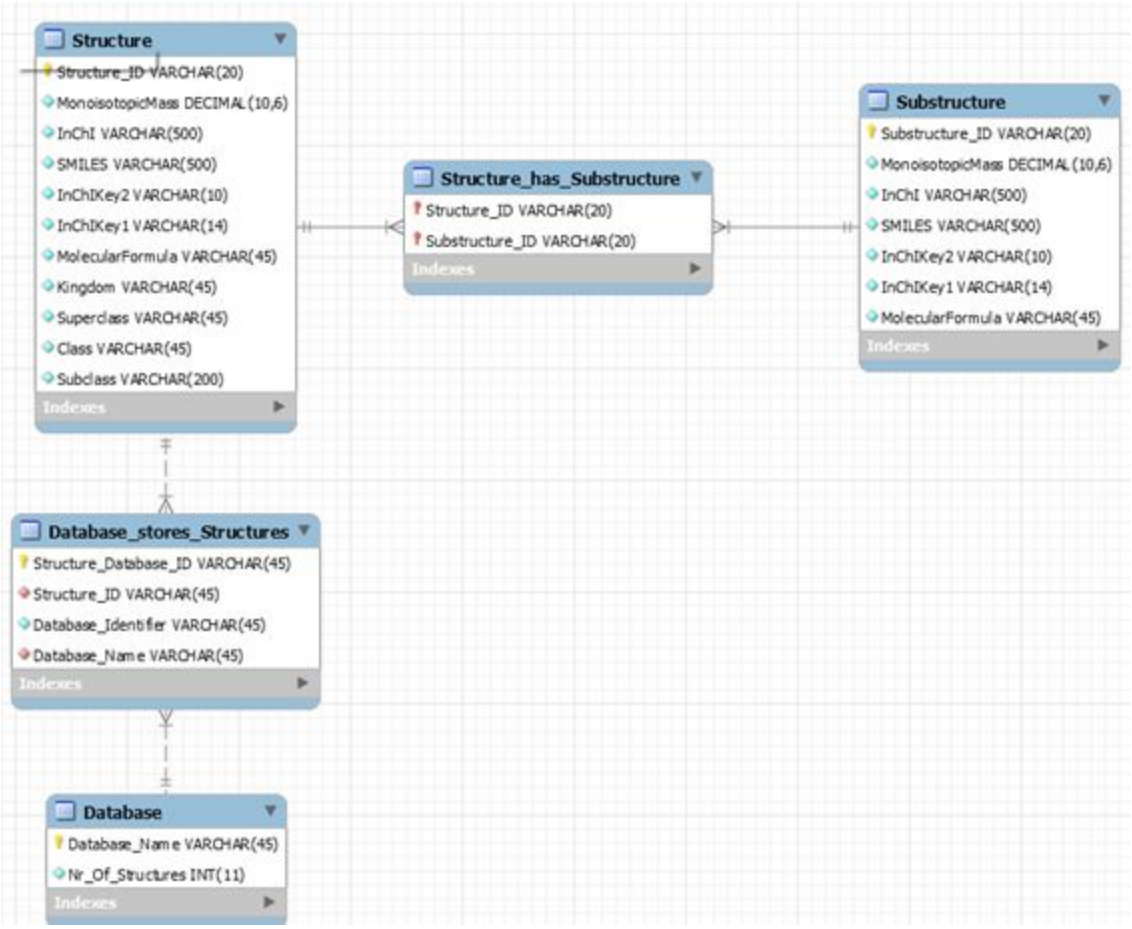
Natural product expansion Guide

Summary
In order to expand the already existing NP database created by Sam Stokman this guide was made.

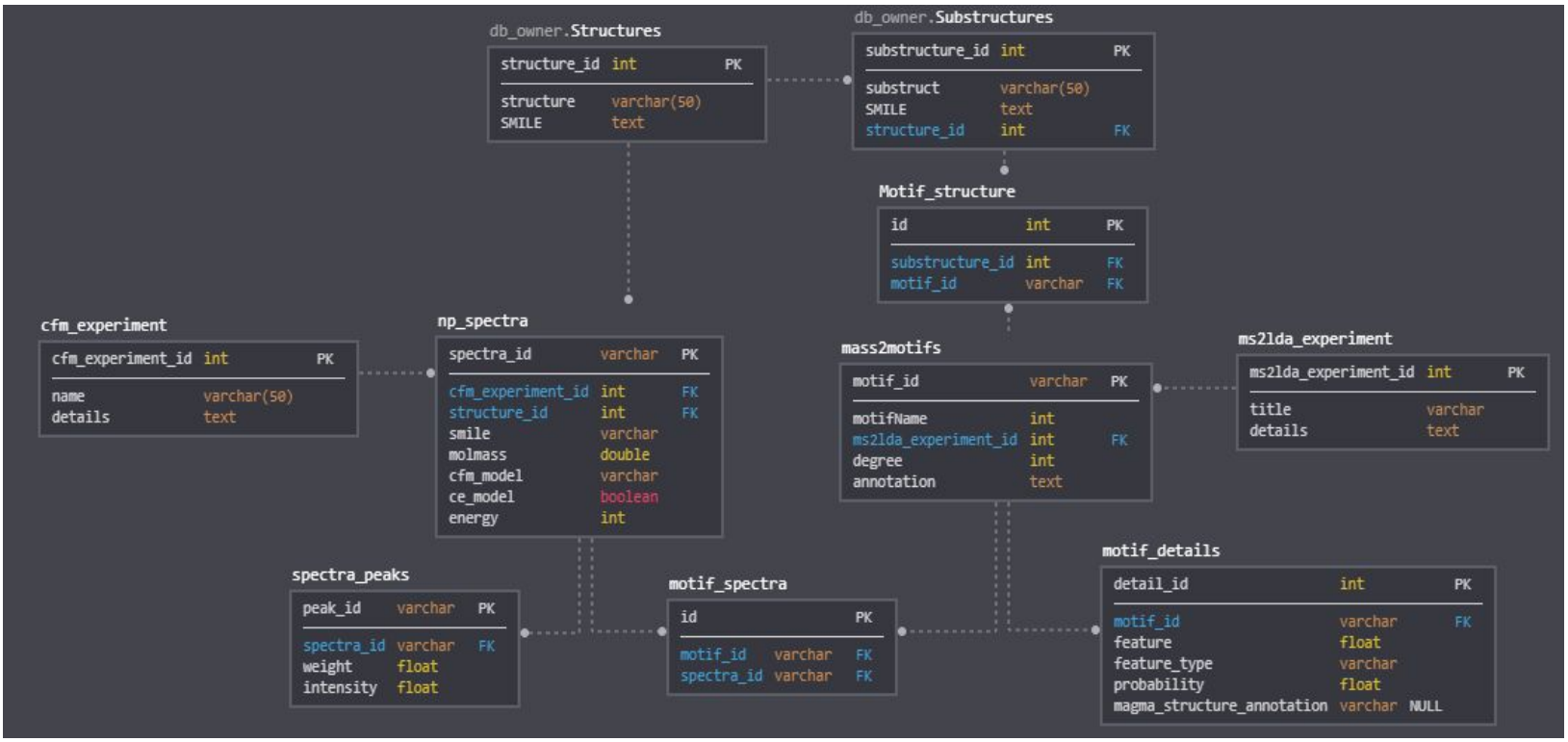
The NP DB contains data of around 320.000 NPs in the data structure below



For the project Natural Products Plug and Play with Chemical Substructures, a lot of spectra data was generated with the data from the NP DB. These spectra were then used to create motifs with the MS2LDA tool (**http://ms2lda.org/**), with all this data a need arose to store the data that is deemed valuable. Thus this database expansion was made along with this guide to help people understand what is added and how to add or remove data from the new tables.

Adding Tables

With this expansion 6 new tables will be added to the NP database, 3 tables for the spectra and 3 tables for the MS2LDA data. Once added to the database are structured as shown in the picture below.



List new tables:

-cfm_experiment
-np_spectra
-spectra_peaks
-ms2lda_experiment
-mass2motifs
-motif_details

To add the tables all that needs to be done is download the script (link) and sql files and run it
This can be done by typing *python /path/to/script/np_db_expansion.py /path/to/sqlFiles/ /path/to/database/Natural_Product_database.sqlite .*
This should have added the tables listed above to the database.To test this enter the database and us the following query: “query to see what's in the table”.
With this done we can move to the next step which is filling the table with info.

WARNING: running this script overwrites existing tables in the database. So make sure that you only do this to get the tables or clear all tables. To remove 1 experiment scroll down to the removing data section.

Adding data

Adding data can be achieved by running 1 of 2 scripts depending on the input data.
In case spectra data needs to be added run the script db_filler_NP_spectra.py.
In order to run this script three inputs are needed

- first you will need a mfg file containing spectra. Preferably created with the cfm_pipeline
- a file containing info about the experiment should contain a title,cfm_model,ce_mode and description. This info should be in csv format:

  title,someTitle
  cfm_model,modelUsed
  ce_mode,true
  description,description of experiment

  A template can be found in the repository. Also keep in mind the ce_mode should be either true or false.
- And the path to the NPDB.

To run it type: *Python /path/to/script/db_filler_NP_spectra.py /path/to/spectra/Spectra_of_choice.mgf path/to/experiment/info/cfm_experiment_info.txt /path/to/database/Natural_Product_database.sqlite*

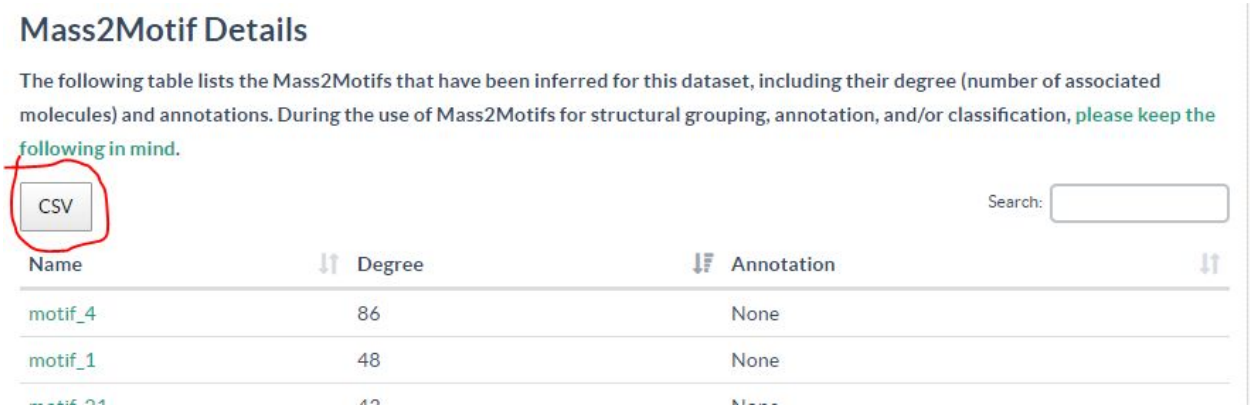
In the case of MS2LDA data run the script called db_filler_mass2motifs.py.
For this script you will need four inputs

- Experiment info which is a csv file containing info on the experiment should contain a Name and the Details of the experiment obtained from MS2LDA, picture below shows how its shown on the site when looking at the summary page. Just create a two line csv file (or use the template from the repo) containing:

  Name,nameOfFile
  Details,someInfoAboutExperiment

  And add the path to this file to the command.

- Next you want the path to a file containing the Motif data. This can be obtained by going to the experiment you want and look at the summary page and press the Csv button below Mass2Motif details



Mass2Motif Details

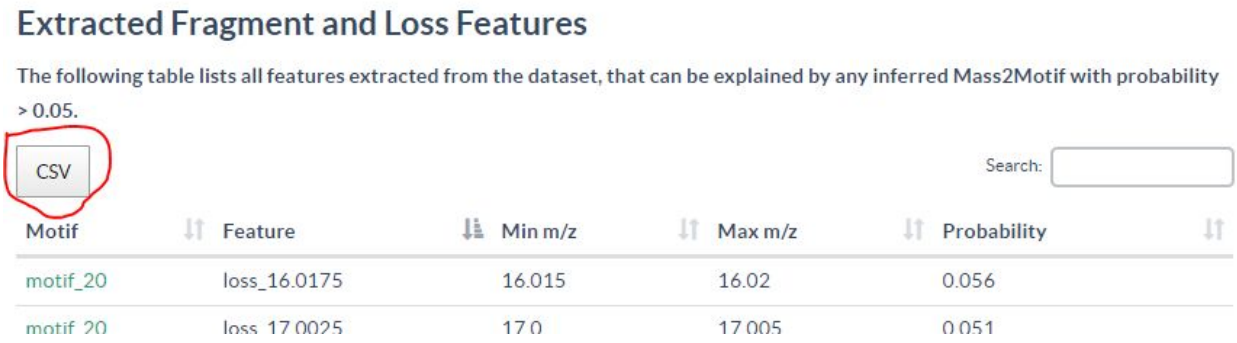
The following table lists the Mass2Motifs that have been inferred for this dataset, including their degree (number of associated molecules) and annotations. During the use of Mass2Motifs for structural grouping, annotation, and/or classification, please keep the following in mind.

CSV

Search:

Name	Degree	Annotation
motif_4	86	None
motif_1	48	None

- Next you want to get the motif details, this can be obtained from the summary page as well this time we look for the Extracted Fragment and Loss Features. Again to obtain this info click the CSV button.



Extracted Fragment and Loss Features

The following table lists all features extracted from the dataset, that can be explained by any inferred Mass2Motif with probability > 0.05.

CSV

Search:

Motif	Feature	Min m/z	Max m/z	Probability
motif_20	loss_16.0175	16.015	16.02	0.056
motif_20	loss_17.0025	17.0	17.005	0.051

- Lastly you will need the path to the NP database again.

To run it type: *Python /path/to/script/db_filler_mass2motifs.py /path/to/experimentInfo/experiment_info.csv /path/to/motifs/motifs.csv /path/to/features/motif_features.csv /path/to/database/Natural_Product_database.sqlite*

Once this is done you can sit back while the data is uploaded to the database. For the mass2motif one this is a matter of seconds. However with larger mgf files the cfm_spectra file can take hours, so keep that in mind and maybe run in in a screen. (https://linuxize.com/post/how-to-use-linux-screen/)

This should cover how to upload data to the database.

Removing Data

In order to remove data from the tables you can run the np_db_experiment_remover.py
This script requires a path to the database followed by either cfm or ms2lda (depending on what you which to remove) and an experiment_id. Based on the input it should remove the entire selected experiment.