

# Molecular Networking and *in-silico* MS/MS Database : a workflow to dereplicate and visualize results

## I LC-MS/MS analysis of crude extract

- use ramp energy gradient or combined energies (ex: 15,30 and 45 eV) for optimal MS/MS spectrum coverage

## II Conversion of proprietary to .mzXML format

- use ProteoWizard (<http://proteowizard.sourceforge.net/>)

## III Generate Molecular Networks on GNPS

- GNPS server : <http://gnps.ucsd.edu>
- follow instructions at <https://bix-lab.ucsd.edu/display/Public/Molecular+Networking+Documentation> for optimal parameters and MN visualization in Cytoscape
- Cytoscape is available at <http://www.cytoscape.org/>

## IV Fetch clustered data from the MN on GNPS

- hit «Download Clustered Data». You will get a folder containing files as described on **Fig. 1**

- the MN attributes file appears as an .out file in the folder. Let's call it **cytoscape\_attributes.out**
- the clustered spectra appears as a .mgf file in the folder. Let's call it **your\_spectra.mgf**

## V Library search using Tremolo

- Tremolo runs under Linux. If you don't have a Linux-based system you can easily install Ubuntu on your Windows or Mac OS via a virtual machine. See instructions here : <http://www.wikihow.com/Install-Ubuntu-on-VirtualBox>

- download Tremolo <http://proteomics.ucsd.edu/Software/Tremolo/>

- in order to easily perform library search using Tremolo and merge results within the MN attributes file we wrote a python script (treat.py) and a bash script (run.sh). These scripts and the UNPD\_ISDB files should be placed in the folders (respect folders names) as described in **Fig. 2**

- to adjust the library search parameters you need to edit the **run.sh** file.

The important parameters are seen on **Fig. 3**

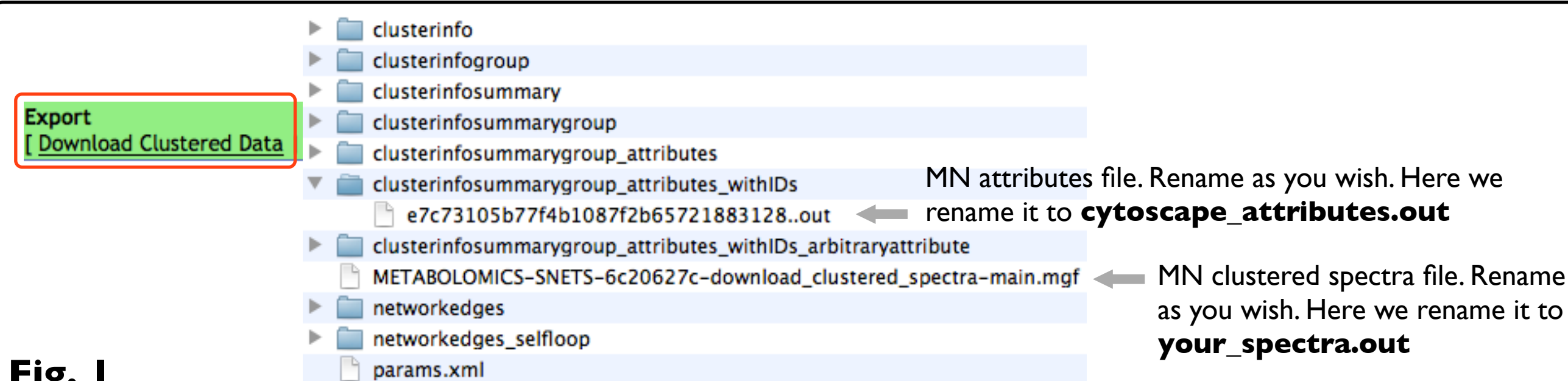
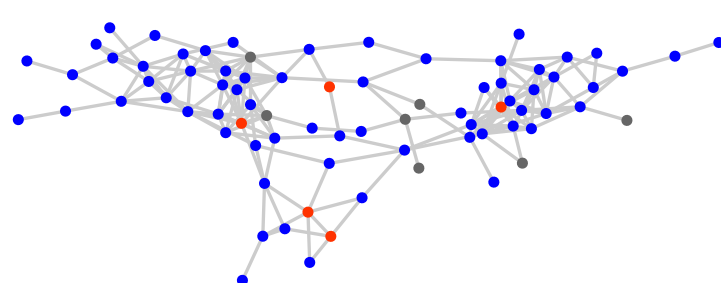
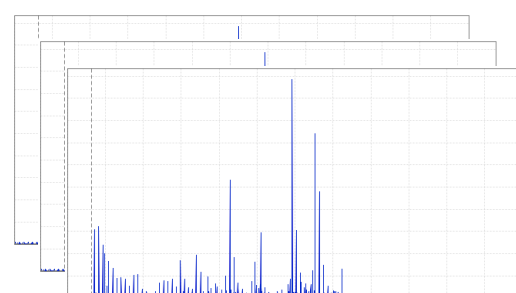
- **TOLERANCE**:  $\pm$  tolerance for parent mass search in Da. Set a small tolerance for dereplication using parent ion mass as prefilter, keeping in mind the resolution of your data. Increase to the wanted range for variable dereplication search (ex: 100 or 200 Da) Caution as this will also increase calculation times !

- **SCORE\_THRESHOLD**: should be kept low when using *in-silico* DB. Typically 0.2 to 0.3.

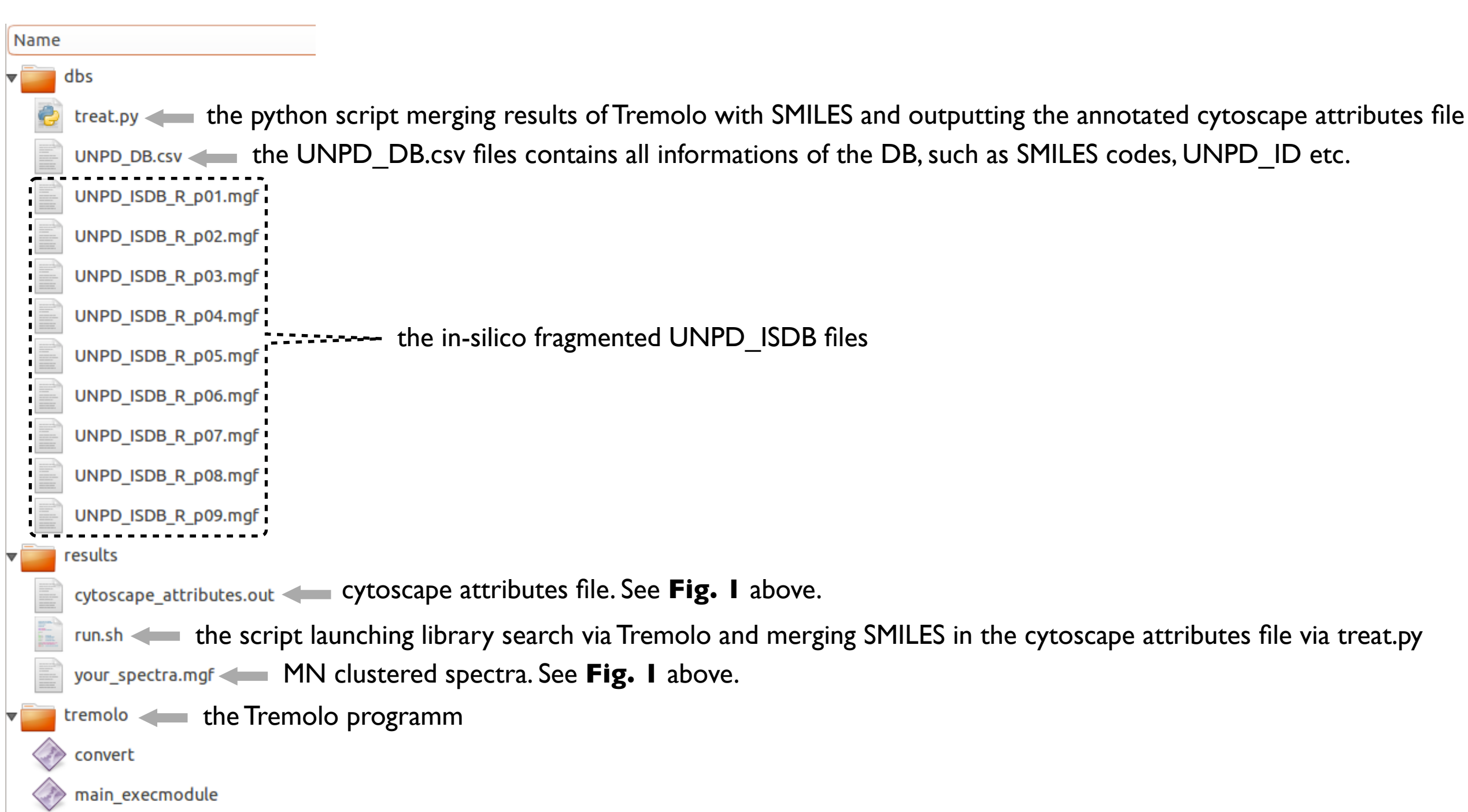
- **TOP\_K\_RESULTS**: Defines the maximal number of results returned

To launch the search open a terminal window and navigate to the results folder. Type:

```
bash run.sh your_spectra.mgf cytoscape_attributes.out results.out
```



**Fig. 1**



**Fig. 2** Organization of the files & folders to run the ISDB library search

```
# Set the tolerance to be used
TOLERANCE=0.005
# Score threshold in Da
SCORE_THRESHOLD=0.2
# Top K results
TOP_K_RESULTS=5
```

**Fig. 3** Library search parameters to edit in run.sh

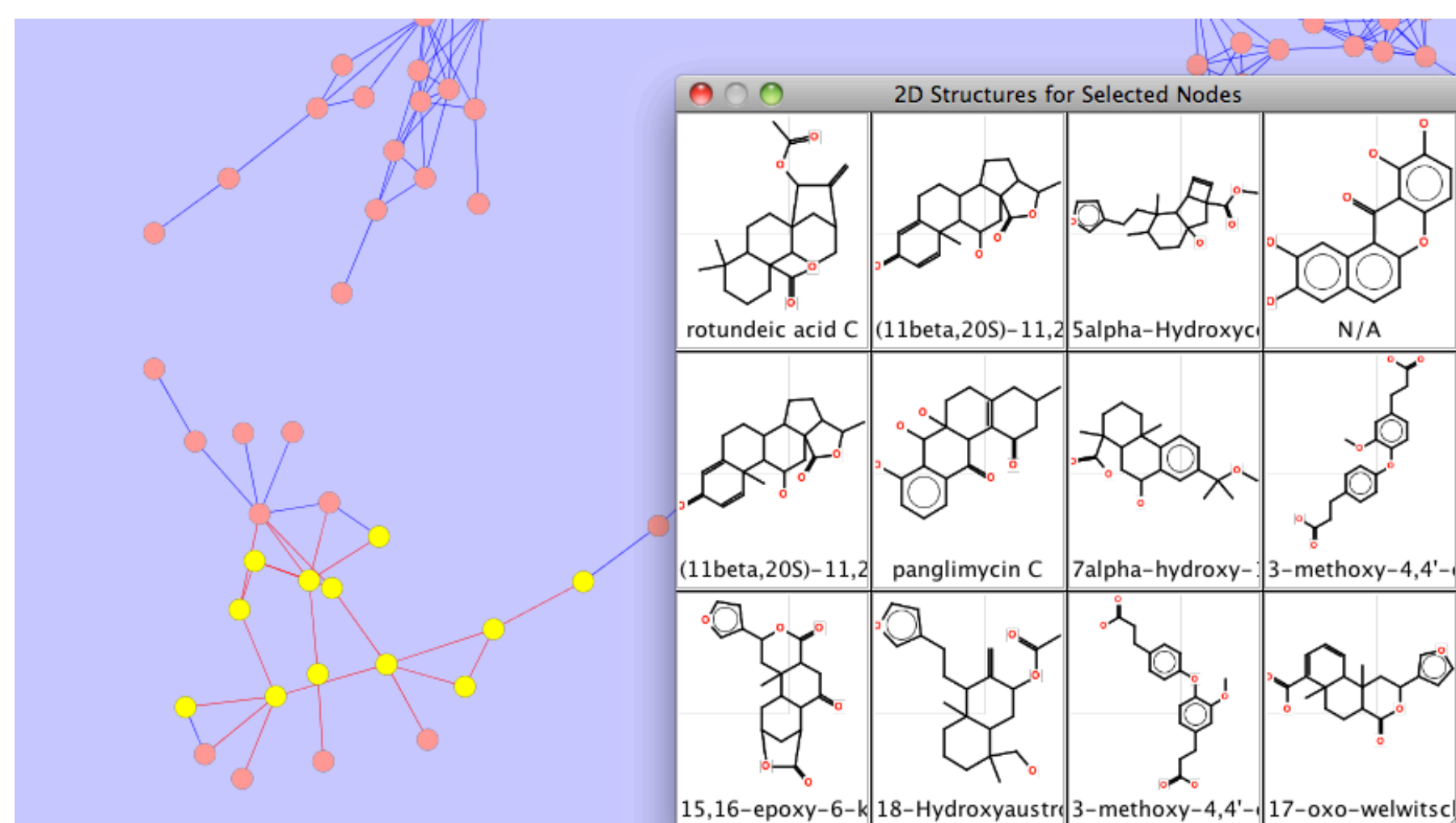
## VIII Visualize dereplication results in Cytoscape

- Install chemViz plugin (<http://apps.cytoscape.org/apps/chemviz>)
- load the results.out file as attribute file with corresponding SMILES and IDs
- select nodes of interest, right click and under *Chemoinformatic tools* select *Show structures window*.

# Start exploring the network !



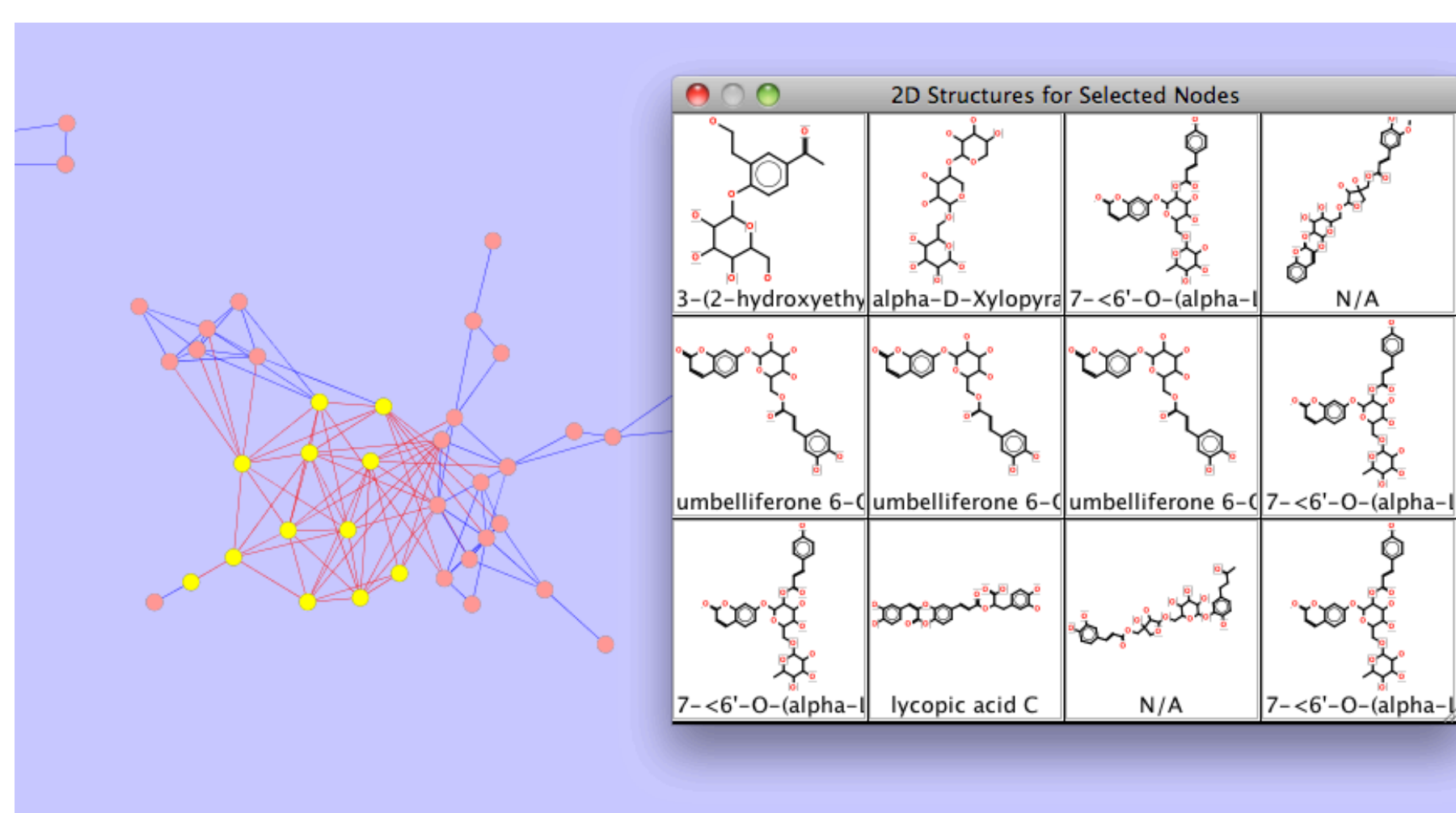
## Examples of results visualization



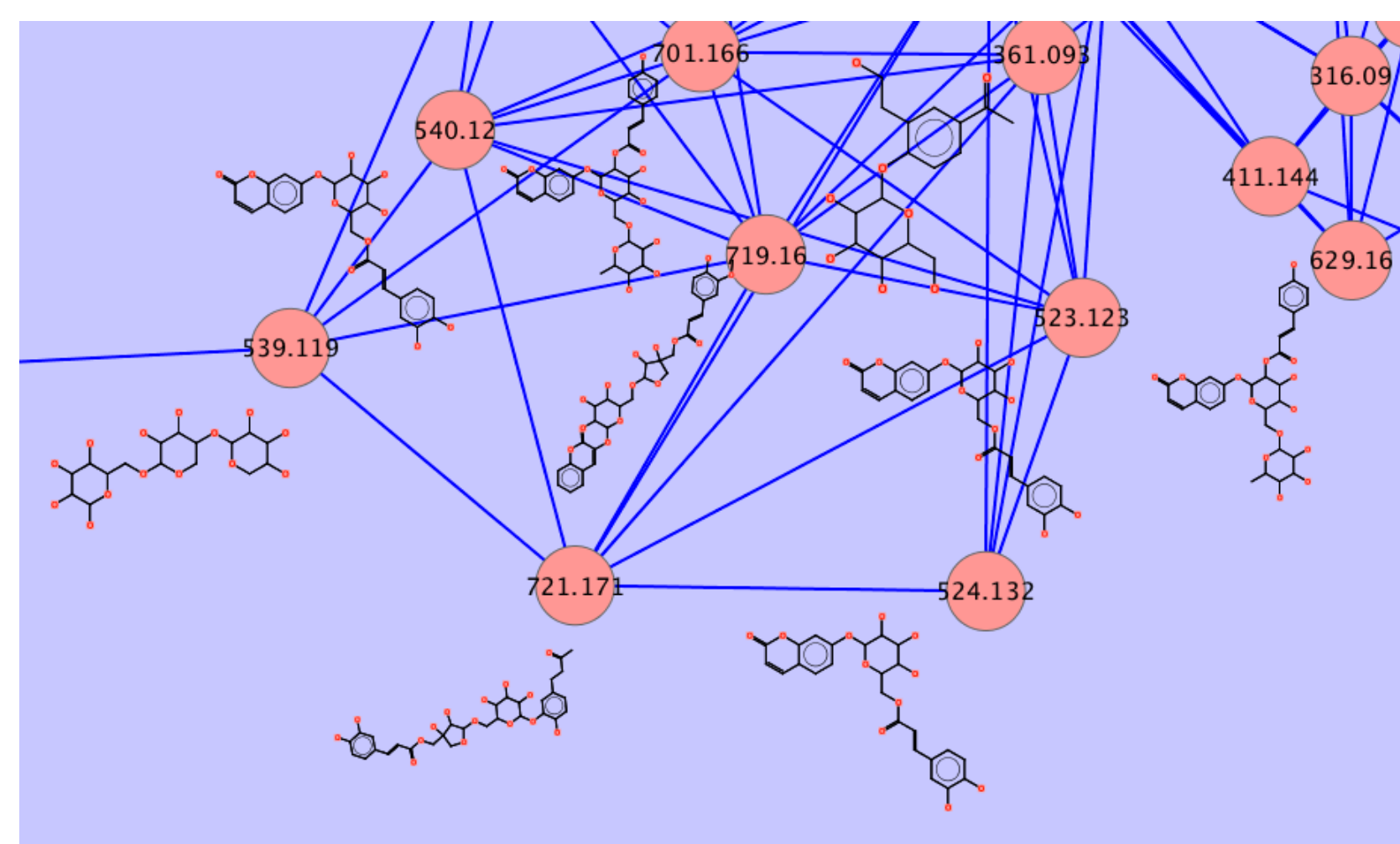
**Fig. A** - variable dereplication against the UNPD-ISDB indicates the possible presence of diterpenoids for this extract of *Salvia* sp.

ID	Attribute	Molecula...	Molecula...	XLogP	HBond Acc...	HBond Donors	2D Image
1555	SMILES	CC=C(C) /C(=O)O C1CC=C 2CC3(O) OC(=O)C (=C3C(O)	362	0,592	6	2	
1769	SMILES	CC(=O)O C1C(=C) C2CCC3 C45COC( O)(C(O)C 4C(O)(C	388	1,451	6	2	
1853	SMILES	CC(=O)O CC(=O)C1 (O)CCC2 C3CCCC4 =CC(=O) CCC4(C)	402	0,451	6	1	
1857	SMILES	CC1OC(C (O)C(O)C 1O)C2c3c ccc(O)c3C (=O)c4c( O)cc(CO)c	402	0,147	6	6	

**Fig. B** - structures can be viewed as a table which can in turn be searched for substructures or exported.



**Fig. C** - here caffeoyl and coumarin glycosylated compounds seem to hide under this cluster.



**Fig. D** - structures can also be directly displayed on top of the nodes

### Notes

- Outliers in the panel of proposed structures should be spotted (see **Fig. A** for ex.) and attention should be focussed on compounds of a common class or bearing common structural functionalities that can lead to similar MS/MS fragmentation data (sugars, aliphatic side chains etc ...)
- Merging orthogonal informations : phylogenetic data, comparison of hits logP with experimental retention times, input of exact mass for the molecular formula determination is the key to assess the relevance of the hits.
- More than an strict dereplication tool, using variable dereplication mode against the ISDB should be seen as an exploratory tool allowing to gain a feeling of the chemistry behind a cluster of metabolites.