Tutorial for „Using Jupyter Notebook for Retrieving ChEMBL data”

A Jupyter Notebook is provided to retrieve bioactive molecules from the ChEMBL database by using the ChEMBL ID.

Getting Started

File name	Description
UNIVIE_ChEMBL_Query_Jupyter_Notebook .ipynb	Jupyter Notebook
chembl_query_env.yml	Virtual Environment
ChEMBL_Test_File.csv	Example of a list of selected ChEMBL Ids for retrieving data

- Install Anaconda on your device
- Install/activate virtual environment (chembl_query_env.yml)

  conda env create -f chembl_query_env.yml

  conda activate chembl_query_env
- Start the Jupyter Notebook (If you don’t have any experience with using Jupyter Notebook, please look at “Installation Guides/Tutorials” before starting.)

Installation Guides/Tutorials

Anaconda: https://docs.conda.io/projects/conda/en/latest/user-guide/install/index.html

Virtual Environment: https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manageenvironments.html

How to use Jupyter Notebooks: https://jupyter-notebook.readthedocs.io/en/stable/

Jupyter Notebook

Due to the some issues with the new version release of RDKit and Pandas it might happen that the molecules are not displayed within the Jupyter Notebook. To overcome this problem, the Pandastools.py script within the “env” file has to

be replaced with a fixed Pandastools.py version. The fixed script can be found within the folder containing the JN, environment and example csv-files.

How to do it?

1. Activate your conda environment:

conda activate

```
aljosas@A773-PC-N006:~$ conda activate
(base) aljosas@A773-PC-N006:~$
```

The '(base)' indicates that anaconda is activated.

2. Type then:

conda activate chembl_query_env

```
(base) aljosas@A773-PC-N006:~$ conda activate chembl_query_env
(chembl_query_env) aljosas@A773-PC-N006:~$
```

4. Type then:

conda info

```
(chembl_query_env) aljosas@A773-PC-N006:~$ conda info

     active environment : chembl_query_env
    active env location : /home/aljosas/anaconda3/envs/chembl_query_env
            shell level : 2
       user config file : /home/aljosas/.condarc
 populated config files :
          conda version : 4.12.0
    conda-build version : 3.21.4
         python version : 3.8.8.final.0
       virtual packages : __linux=5.15.0=0
                          __glibc=2.31=0
                          __unix=0=0
```

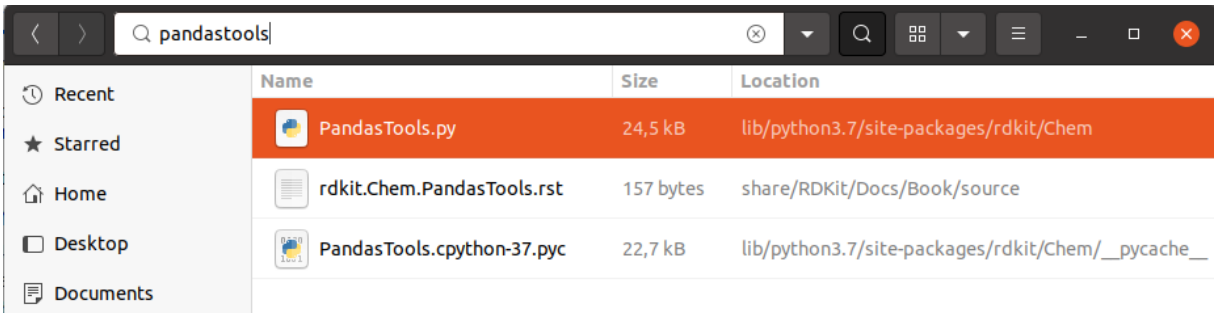
This allows you to see your active env location (see depiction)

5. Replace PandasTools.py:

Within the chembl_query_env file continue with **lib/python3.7/site-packages/rdkit/Chem** where you will find the PandasTools.py script. Replace it with the new one.

“If you already have the chembl_query_env file open, you can just search for the PandasTools.py in the search filed

Follow the path “chembl_query_env/lib/python3.7/site-packages/rdkit/Chem” to the folder and replace the Pandastools.py with the fixed version.



6. Last Step:

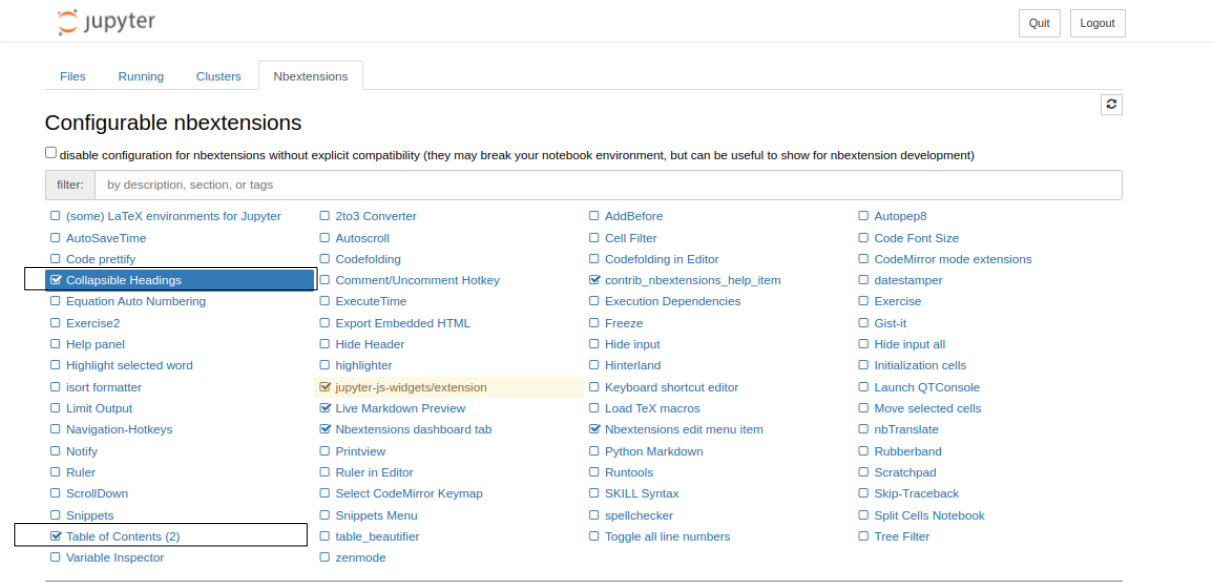
Start your jupyter notebook (see the depiction)

Jupyter Notebook extensions for a better clarity

Please ensure before you open the Jupyter Notebook within “Files” that the “nbextensions” named Collapsible Headings and Table of Contents are tagged.

In the home menu of Jupyter Notebook click on the nbextensions as seen on the depiction below:

This will lead you to the the nbextensions menu where you will be able to tag the collapsible headings and table of content.



Once you set the tags, return back to “Files” and open the the Jupyter Notebook.

There are four sections available in the code:

1. Connect to Postgress

2. Query Data

3. Data Filtering

4. Disconnect to Posgress

Please follow the guide below as described, before running the notebook. A star next to the

title/header indicates that an action from your side is required.

This JN enables the user to retrieve data from ChEMBL. The data can be requested via a SQL query by adding the target ID using a CSV-file, ChEMBL release version and activity threshold in the cell code indicated with a star at the end of the heading.

It can be executed by pushing the "Run ALL" button, once the requested information is added to the corresponding text fields indicated with a star at the heading:"<b> Please enter the name of your csv-file containing your ChEMBL IDs, the desired ChEMBL release, activity threshold and click on 'Run All Below"

**Please run the first three cell codes to activate the libraries, display the RDKit Version/Pandas Version and the search field taskbars. Once the information is added to the JN in the search fields click on the button "Run All Below". *

The individual steps are described by the headings found above each cell code and the resulted data frame is saved as a SDF-file and CSV-file.

If more than 40 targets are within your csv-file, please open your jupyter notebook using the following command to set a new data rate limit:

jupyter notebook --NotebookApp.iopub_data_rate_limit=1.0e10

This will prevent your JN from crushing down.

For further information don't hesitate to contact: aljosa.smajic@univie.ac.at