Documentation – Sandbox - Dataset Generation Workflow

For further questions please don't hesitate and contact: fabian.kaiser@univie.ac.at

Introduction

We created a KNIME workflow (WF) which handles all steps of dataset generation automatically, the workflow only needs minimal input which can be configured by using a graphical user interface. The finished datasets can be downloaded and used for further modeling with the Jupyter Notebook of the Sandbox.

Setting Up KNIME

Knime is a free open-source data science software which allows the easy creation of workflows without the need of any programming skills.

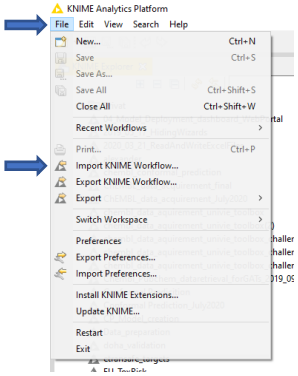
You can download KNIME here:
https://www.knime.com/downloads/download-knime

Choose the correct version for your operating system and follow the installation instructions.

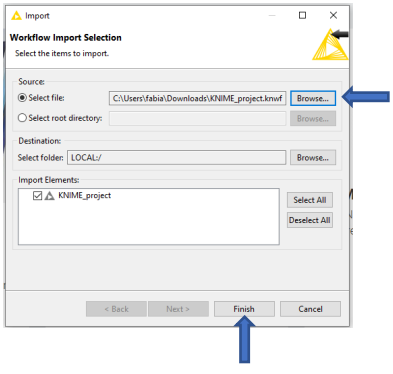
Import the Workflow

You will be provided with the *data_gathering_sandbox.knwf* file which includes all the workflow data.

Import the Workflow by clicking:
File → Import KNIME Workflow…



The "Import" window will open. Select the provided WF file and click "Finish".

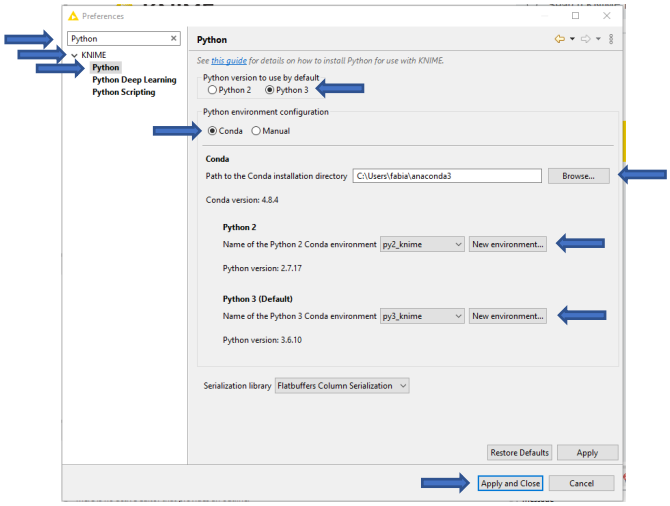


Set-Up Python Environment

The standardization of the compounds requires certain python libraries therefore KNIME needs to access the previously installed Environment.

Set up the correct environment by clicking:

File → Preferences
The preferences window will open.



On the top left side of the “Preferences” window you can search for the term “Python”.
Click on the tab “KNIME” and then the sub-tab “Python”.

Now select the previously created Python Environment by choosing the correct directory. The drop-down menus for “Python 2” and “Python 3” should now be filled with the environment.

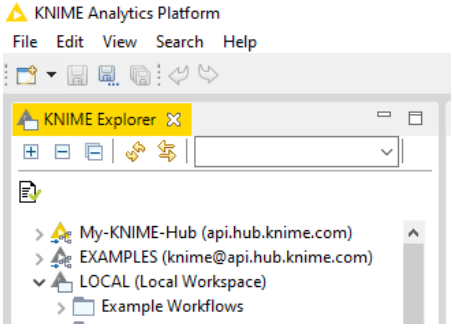
Once finished click “Apply and Close”.

Set-up local ChEMBL Database

The Database needs to be set up before you execute the workflow. A detailed description is given at: https://hub.docker.com/r/pharminfovienna/sandbox

Access the Workflow

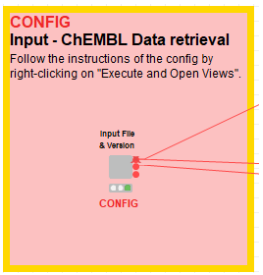
You can access the WF by clicking on it in the KNIME Explorer tab. This can be found on the left side of your screen. After clicking on it the WF will open in one of the tabs in the middle of the screen.



Usage/Configuration of the Workflow

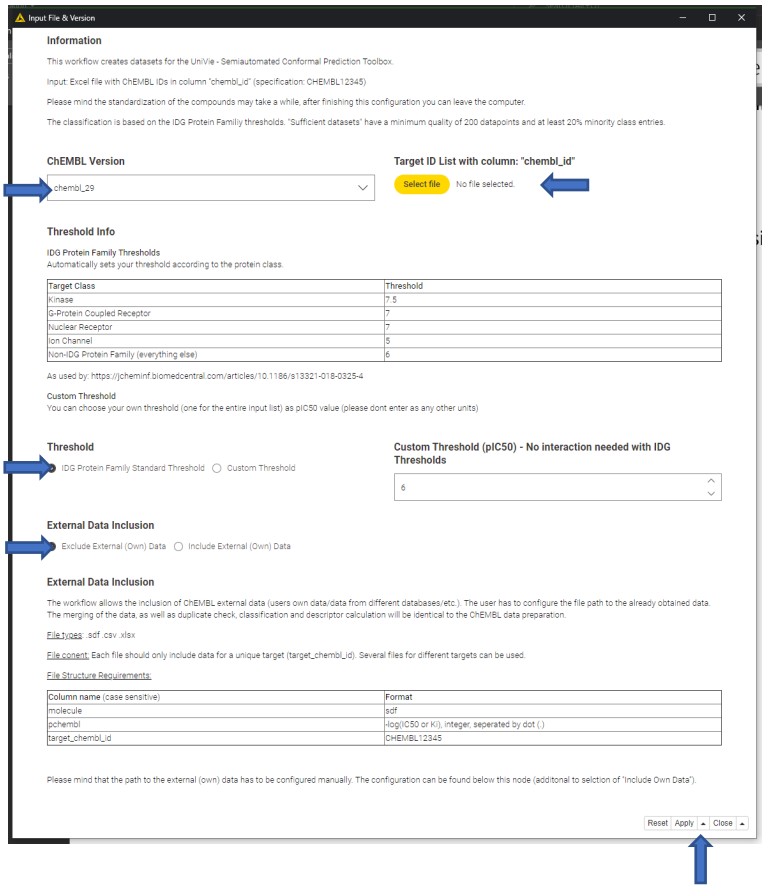
Parts of the workflow are color coded: A red box means that the user has to configure something. Please configure the red boxes in the same order as listed in this documentation.

1. ChEMBL Data retrieval



Right click on the grey “Input File & Version” node → choose “Execute and Open Views”.

A configuration window will open, please follow the instructions in the config-dialogue.



Choose **ChEMBL Version** 29

Target List Input
Click on the yellow “Select File” button and choose your input file.

Input-File Format
.xlsx file (Excel), any file name with at least the column:
chembl_id
(case sensitive, e.g., CHEMBL1234)

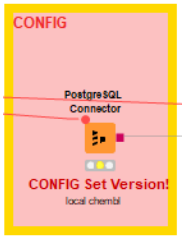
Threshold (manual possible)

External Data Inclusion
The workflow can merge your in-house data with the ChEMBL data to create a combined dataset with the same data preparation procedure.

The data has to follow some requirements which can also be found in the configuration dialogue. (see step 3)

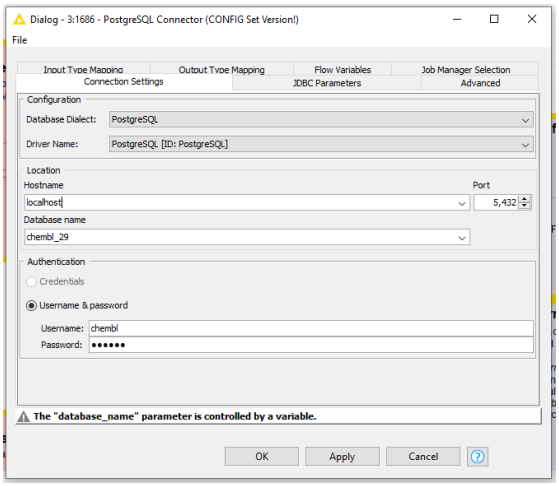
After finishing the configuration click on “Apply” and “Close” the window.

2. Configure SQL Connection



This step can only be performed if the local ChEMBL Database is already installed.

Right-click on the node “PostgreSQL Connector” → click on “Configure” → a dialogue window will open



Select the following options:

Database Dialect: PostgreSQL
Driver Name: PostgreSQL [ID: PostgreSQL]

Hostname: localhost (might differ if you diverted from the provided installation guide)
Port: 5432
Database name: chembl_29

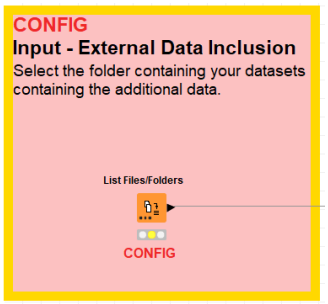
Username: chembl
Password: chembl

After choosing the correct options click on: “Apply” and close the window.

3. Configure External Data Input

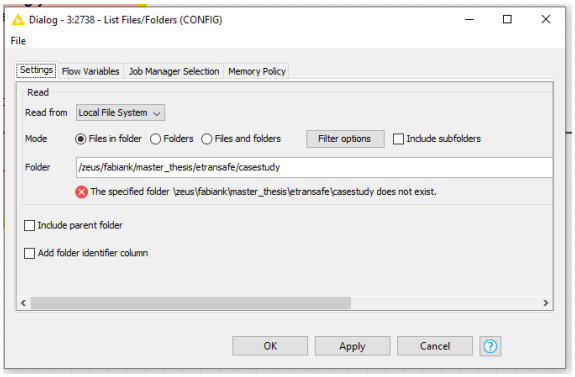
Please perform the configuration step, even if no external data is added, for means of error prevention. If no external data is added just enter the path to your e.g., /home directory.

The workflow allows the inclusion of ChEMBL external data (users own data/data from different databases/etc.). The merging of the data, as well as duplicate check, classification and descriptor calculation (data processing) will be identical to the ChEMBL data preparation.



If the option “Include External (Own) Data” was selected in step one, please follow this step:

Right-click the “List Files/Folder” node → choose the option “Configure”



Select the file path with folder linking to your data then click “Apply”.

If no external data will be used enter any valid file path e.g. /home and then click “Apply”.

Data Requirements

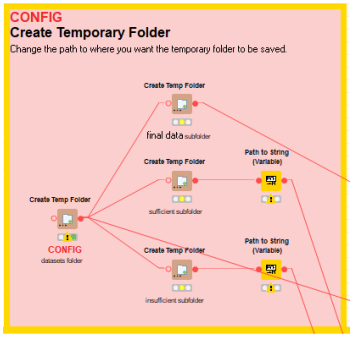
File types: .sdf .csv .xlsx

File content: Each unique target should be one separate file. Several files for different targets can be used.

File Structure Requirements:

Column name (case sensitive)	**Format**
molecule	sdf
pchembl	-log(IC50 or Ki), integer, seperated by dot (.)
target_chembl_id	CHEMBL12345

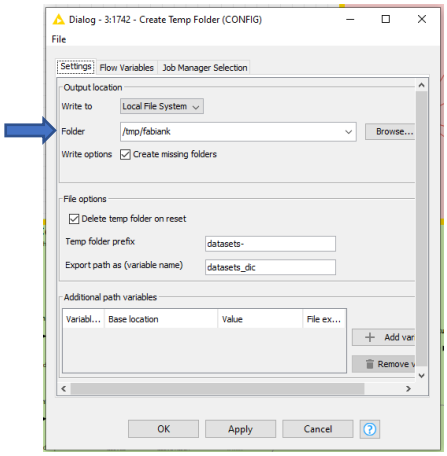
4. Configuration of Temporary Folder Structure



This configuration step is not mandatory. By standard the files will be temporarily saved under /tmp/fabiank (Tested on Linux & Windows).

If wanted the path of the temporary directory can be changed by:

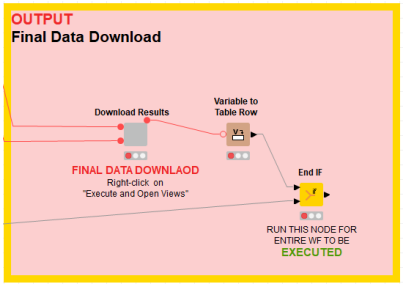
Right-clicking on the “Create Temp Folder” node and choosing the option “Configure”.



In the configuration dialogue the temporary folder path can be adapted. After changing the path, click “Apply” and close the window.

(Please only perform the configuration on the most left “Create Temp Folder” node which is marked with the word “CONFIG”).

Execution of the Workflow

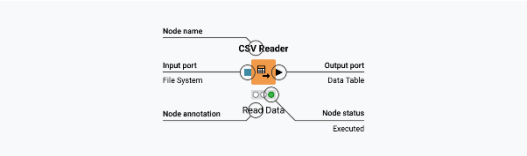


Once every configuration step is performed, the WF can be executed.

To execute the WF:
Right-click the "End IF" node and click "Execute".

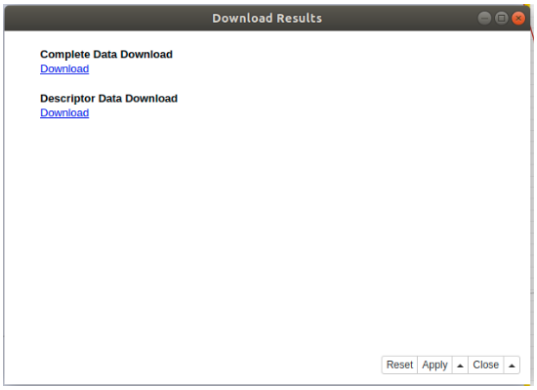
The execution of the workflow may take some time due to the structure standardization process.

The execution is finished once the node status (red dot) under the End IF node changes to green.



Once the node status is green:
Right-click the grey "Download Results" node and choose "Interactive View: Download Results".

The results can be obtained by clicking on the "Download" hyperlink. The files will be located at the Download folder of your computer.



Info: Complete Data includes:
- .csv file (all, training, test), with activity & descriptor columns → Ready for machine learning
- .sdf files with molecule, activity (binary classification & pchembl), threshold, InChI-Key → Ready for Data Science, calculate own descriptors etc. …

Descriptor Data:
only includes the .csv files → Ready for machine learning

<u>The output will have the following format:</u>

- all_data.tar.gz (Compressed File) → Unpack
  - datasets (Folder)
    - final_data (Includes "sufficient" dataset with the correct folder structure/format for usage with the Modeling Sandbox.
      - .csv files (all, training, test)
        with activity & descriptor columns → Ready for machine learning
      - Configuration.csv → Includes the Information needed for the Modeling Sandbox.
    - sufficient (Includes all datasets which apply to minimum dataset quality rules)
    - insufficient (If some datasets do not apply to minimum dataset quality rules this folder will be created and include those datasets).

(Info: At the end of the folder name a unique number code is added to make sure that no existing data is overwritten e.g.: datasets-1d3050f0d030).