Tutorial for „Using Jupyter Notebooks for Re-training Machine Learning Models"

A Jupyter Notebook is provided to generate/retrain classification models for six transporter proteins (BCRP, BSEP, OATP1B1, OATP1B3, MRP3, P-gp). Four different classifiers can be selected without an extensive descriptor selection and hyperparameter search as they have been pre-selected. An in-house data set (provided by the user) can be combined with the available UNIVIE data set(s) to extend the chemical space of the model.

Getting Started

- Download files from the repository

File name	**Description**
standardise.py	Standardizer
models.yml	Virtual Environment
RDKIT_Descriptors.txt	List of selected descriptors
Data folder:	
*NameOfTransporter*_Univie.sdf	Training set provided by the University of Vienna
*NameOfTransporter*_ChEMBL28.sdf	Test set provided by the University of Vienna

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

  `conda env create -f model.yml`

  `conda activate models`
- 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

Download Repository: https://docs.github.com/en/repositories/creating-and-managing-repositories/cloning-a-repository

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/manage-environments.html

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

Jupyter Notebook

There are four sections available in the code:

1. **Data Collection**
2. **Data Set Preparation for ML Task**
3. **Applicability Domain**
4. **Model Generation & Evaluation**

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.

Procedure:

"Step 1: Data Collection"

The "Data Collection"- step includes some code cells to extend the provided data set as well as prepare it prior to model building. If you want to add additional data, you have to change the file name at the variable "Intern_Data". Please, be aware that only the SDF format can be used.

```
# Please add the name of your file
Intern_Data = "BSEP_ChEMBL28.sdf"
```

!!! If you don't want to add additional data, you can skip Step 1!!!

You can modify the code to customize Step 1, but it is not mandatory. It automatically, calculates important parameters such as SMILES and InChIs as well as removes stereoisomers and duplicates. At the end of this step a SDF with the training set will be generated.

"Step 2: Data Set Preparation for ML Task"

At cell "Read SDF":

Add the name of your training set under "Training set*".

```
# Please add the name of your training set
molecules = Chem.ForwardSDMolSupplier("BSEP_Univie.sdf", sanitize=False)
```

Add the name of your test set under “Test set*”.

```
# Please add the name of your test set
test_molecules = Chem.ForwardSDMolSupplier("BSEP_ChEMBL28.sdf", sanitize=False)
```

Further code cells in this section are available for customization such as the standardization step as well as the selection of the descriptors. But it is not mandatory to modify them.

“Step 3: Applicability Domain”

The Applicability Domain is assessed to check if the test set is within the chemical space of the model. If this is the case, the compound is marked as in domain, otherwise it is marked out of domain. A file is generated namely:” Outlier_Compounds.sdf” which includes all compounds which are out of domain.

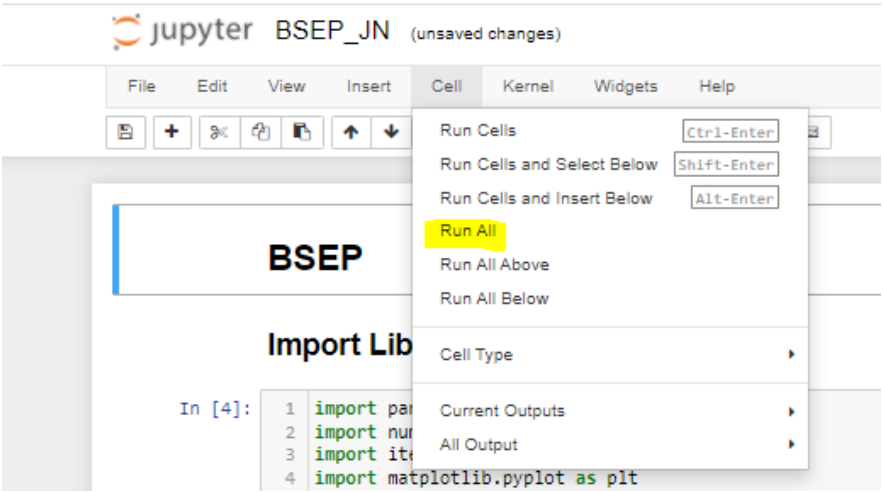
At “Add your test data for comparison” fill in the name of your test set file.

“Step 4: Model Generation & Evaluation”

Add the name of the training data set.

Provides the different machine learning approaches, namely Logistic Regression, Support Vector Machine, Random Forest as well as k-nearest neighbor. No customization is necessary at that step.

You can run the whole Jupyter Notebook now and analyze the results in the current dropdown fields:



What will be saved and what can you analyze:

If you provide additional data and used it to combine it with the UNIVIE data set, a new file will be created, namely "*NameOfTransporter*_Training_Set.sdf". Containing the new compounds as well as the UNIVIE data. The file is stored in the folder *data*.

Generated files including results are stored in the folder *results*.

At section "Applicability Domain", all compounds which are out of domain will be saved in the file "*NameOfTransporter*_Outlier_Compounds.sdf" and in the folder *results*. The structure of the compounds will be also visualized in the notebook.

For each classifier, you will receive statistical performance metrics for the evaluation of the models and the models will be saved as pkl-files in the folder *results*. Wrongly predicted compounds will be saved as an SDF file. These compounds can also be visualized within the Jupyter Notebook. Statistical metrics from the cross validation and the external test set validation can be selected separately. A CSV-file will be generated including the information from the prediction and the applicability domain run.