qsar

November 5, 2022

1 QSAR Modelling of PTP1B Receptor Ligands

The goal of this tutorial will be to develop a machine learning model that can capture the Quantitative Structre-Activity Relationships (QSARs) from the data set we extracted from Papyrus previously. This time we will need a little more control than the `scaffviz` package would allow so we load the data file directly into `pandas.DataFrame`:

[5]:
```
# mount google drive
from google.colab import drive
drive.mount('/content/drive')

# define work directory to store data
DATA_ROOT = '/content/drive/MyDrive/DrugExDemo/' # or wherever you want the␣
 ↪generated files to live on your GoogleDrive
import os
os.makedirs(DATA_ROOT, exist_ok=True)
os.chdir(DATA_ROOT)

# fetch pretrained model
os.makedirs("./data/drugex/models/pretrained/", exist_ok=True)
! wget -nc -P './data/drugex/models/pretrained/' 'https://zenodo.org/record/
 ↪7096859/files/DrugEx_v2_PT_Papyrus05.5.zip'
! unzip -n './data/drugex/models/pretrained/DrugEx_v2_PT_Papyrus05.5.zip' -d './
 ↪data/drugex/models/pretrained/DrugEx_v2_PT_Papyrus05.5'

# install dependencies
! git clone https://github.com/martin-sicho/drugex-demo
! pip install -r drugex-demo/requirements.txt

# verify where we are working
os.getcwd()
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
File ‘./data/drugex/models/pretrained/DrugEx_v2_PT_Papyrus05.5.zip’ already
there; not retrieving.

Archive:  ./data/drugex/models/pretrained/DrugEx_v2_PT_Papyrus05.5.zip
```

```
fatal: destination path 'drugex-demo' already exists and is not an empty
directory.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting drugex@ git+https://github.com/martin-sicho/DrugEx-CDDG.git@master
  Cloning https://github.com/martin-sicho/DrugEx-CDDG.git (to revision master)
to /tmp/pip-install-rcgf3nro/drugex_8d41dea94f0a41a589b8964e4295179e
  Running command git clone -q https://github.com/martin-sicho/DrugEx-CDDG.git
/tmp/pip-install-rcgf3nro/drugex_8d41dea94f0a41a589b8964e4295179e
Collecting papyrus-scaffold-visualizer@ git+https://github.com/martin-
sicho/papyrus-scaffold-visualizer.git@v0.2.0
  Cloning https://github.com/martin-sicho/papyrus-scaffold-visualizer.git (to
revision v0.2.0) to /tmp/pip-install-rcgf3nro/papyrus-scaffold-
visualizer_e62d1ac9ba804c2e8db50432ba355439
  Running command git clone -q https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git /tmp/pip-install-rcgf3nro/papyrus-scaffold-
visualizer_e62d1ac9ba804c2e8db50432ba355439
  Running command git checkout -q ff4f2e885a3973f90a0d9864dfa00abed493f78d
Collecting molplotly==1.1.4
  Downloading molplotly-1.1.4.tar.gz (15 kB)
Collecting mols2grid==1.0.0
  Downloading mols2grid-1.0.0-py2.py3-none-any.whl (100 kB)
     |                  | 100 kB 4.5 MB/s
Requirement already satisfied: numpy>=1.19 in
/usr/local/lib/python3.7/dist-packages (from drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (1.21.6)
Requirement already satisfied: scikit-learn>=1.0.2 in
/usr/local/lib/python3.7/dist-packages (from drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (1.0.2)
Requirement already satisfied: pandas>=1.2.2 in /usr/local/lib/python3.7/dist-
packages (from drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (1.3.5)
Requirement already satisfied: torch>=1.7.0 in /usr/local/lib/python3.7/dist-
packages (from drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (1.12.1+cu113)
Requirement already satisfied: matplotlib>=2.0 in /usr/local/lib/python3.7/dist-
packages (from drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (3.2.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages
(from drugex@ git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r
drugex-demo/requirements.txt (line 1)) (4.64.1)
Collecting rdkit-pypi
  Downloading
rdkit_pypi-2022.9.1-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(29.5 MB)
     |                  | 29.5 MB 1.5 MB/s
```

```
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-
packages (from drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (1.2.0)
Collecting optuna
  Downloading optuna-3.0.3-py3-none-any.whl (348 kB)
     |                  | 348 kB 32.3 MB/s
Collecting gitpython
  Downloading GitPython-3.1.29-py3-none-any.whl (182 kB)
     |                  | 182 kB 56.7 MB/s
Requirement already satisfied: xgboost in /usr/local/lib/python3.7/dist-
packages (from drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (0.90)
Collecting papyrus_scripts@ git+https://github.com/OlivierBeq/Papyrus-
scripts.git@master
  Cloning https://github.com/OlivierBeq/Papyrus-scripts.git (to revision master)
to /tmp/pip-install-rcgf3nro/papyrus-scripts_eace7a589c4649788a0cec309347e1c8
  Running command git clone -q https://github.com/OlivierBeq/Papyrus-scripts.git
/tmp/pip-install-rcgf3nro/papyrus-scripts_eace7a589c4649788a0cec309347e1c8
Collecting sklearn
  Downloading sklearn-0.0.tar.gz (1.1 kB)
Collecting prodec@ https://github.com/OlivierBeq/ProDEC/tarball/master
  Downloading https://github.com/OlivierBeq/ProDEC/tarball/master
     / 60 kB 1.3 MB/s
  Installing build dependencies … done
  Getting requirements to build wheel … done
    Preparing wheel metadata … done
Collecting upsetplot@ https://github.com/OlivierBeq/UpSetPlot/tarball/master
  Downloading https://github.com/OlivierBeq/UpSetPlot/tarball/master
     \ 429 kB 2.7 MB/s
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-
packages (from papyrus_scripts@ git+https://github.com/OlivierBeq/Papyrus-
scripts.git@master->papyrus-scaffold-visualizer@ git+https://github.com/martin-
sicho/papyrus-scaffold-visualizer.git@v0.2.0->-r drugex-demo/requirements.txt
(line 4)) (2.23.0)
Requirement already satisfied: natsort in /usr/local/lib/python3.7/dist-packages
(from papyrus_scripts@ git+https://github.com/OlivierBeq/Papyrus-
scripts.git@master->papyrus-scaffold-visualizer@ git+https://github.com/martin-
sicho/papyrus-scaffold-visualizer.git@v0.2.0->-r drugex-demo/requirements.txt
(line 4)) (5.5.0)
Collecting mordred
  Downloading mordred-1.2.0.tar.gz (128 kB)
     |                  | 128 kB 67.5 MB/s
Collecting swifter
  Downloading swifter-1.3.4.tar.gz (830 kB)
     |                  | 830 kB 62.3 MB/s
Collecting pystow
  Downloading pystow-0.4.6-py3-none-any.whl (35 kB)
Requirement already satisfied: tabulate in /usr/local/lib/python3.7/dist-
```

```
packages (from papyrus_scripts@ git+https://github.com/OlivierBeq/Papyrus-
scripts.git@master->papyrus-scaffold-visualizer@ git+https://github.com/martin-
sicho/papyrus-scaffold-visualizer.git@v0.2.0->-r drugex-demo/requirements.txt
(line 4)) (0.8.10)
Collecting orjson
  Downloading
orjson-3.8.1-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (272 kB)
     |                 | 272 kB 65.4 MB/s
Requirement already satisfied: psutil in /usr/local/lib/python3.7/dist-
packages (from prodec@
https://github.com/OlivierBeq/ProDEC/tarball/master->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (5.4.8)
Collecting dash>=2.0.0
  Downloading dash-2.6.2-py3-none-any.whl (9.8 MB)
     |                 | 9.8 MB 45.3 MB/s
Collecting werkzeug>=2.0.0
  Downloading Werkzeug-2.2.2-py3-none-any.whl (232 kB)
     |                 | 232 kB 40.6 MB/s
Collecting jupyter-dash>=0.4.2
  Downloading jupyter_dash-0.4.2-py3-none-any.whl (23 kB)
Requirement already satisfied: plotly>=5.0.0 in /usr/local/lib/python3.7/dist-
packages (from molplotly==1.1.4->-r drugex-demo/requirements.txt (line 2))
(5.5.0)
Requirement already satisfied: ipykernel in /usr/local/lib/python3.7/dist-
packages (from molplotly==1.1.4->-r drugex-demo/requirements.txt (line 2))
(5.3.4)
Requirement already satisfied: nbformat in /usr/local/lib/python3.7/dist-
packages (from molplotly==1.1.4->-r drugex-demo/requirements.txt (line 2))
(5.7.0)
Requirement already satisfied: jinja2>=2.11.0 in /usr/local/lib/python3.7/dist-
packages (from mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(2.11.3)
Requirement already satisfied: ipywidgets<8,>=7 in
/usr/local/lib/python3.7/dist-packages (from mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (7.7.1)
Collecting dash-html-components==2.0.0
  Downloading dash_html_components-2.0.0-py3-none-any.whl (4.1 kB)
Collecting dash-table==5.0.0
  Downloading dash_table-5.0.0-py3-none-any.whl (3.9 kB)
Collecting dash-core-components==2.0.0
  Downloading dash_core_components-2.0.0-py3-none-any.whl (3.8 kB)
Collecting flask-compress
  Downloading Flask_Compress-1.13-py3-none-any.whl (7.9 kB)
Requirement already satisfied: Flask>=1.0.4 in /usr/local/lib/python3.7/dist-
packages (from dash>=2.0.0->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (1.1.4)
```

```
Requirement already satisfied: click<8.0,>=5.1 in /usr/local/lib/python3.7/dist-
packages (from Flask>=1.0.4->dash>=2.0.0->molplotly==1.1.4->-r drugex-
demo/requirements.txt (line 2)) (7.1.2)
Requirement already satisfied: itsdangerous<2.0,>=0.24 in
/usr/local/lib/python3.7/dist-packages (from
Flask>=1.0.4->dash>=2.0.0->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (1.1.0)
Collecting Flask>=1.0.4
  Downloading Flask-2.2.2-py3-none-any.whl (101 kB)
     |                  | 101 kB 12.5 MB/s
Collecting click>=8.0
  Downloading click-8.1.3-py3-none-any.whl (96 kB)
     |                  | 96 kB 6.2 MB/s
Collecting jinja2>=2.11.0
  Downloading Jinja2-3.1.2-py3-none-any.whl (133 kB)
     |                  | 133 kB 71.3 MB/s
Collecting itsdangerous>=2.0
  Downloading itsdangerous-2.1.2-py3-none-any.whl (15 kB)
Requirement already satisfied: importlib-metadata>=3.6.0 in
/usr/local/lib/python3.7/dist-packages (from
Flask>=1.0.4->dash>=2.0.0->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (4.13.0)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-
packages (from importlib-
metadata>=3.6.0->Flask>=1.0.4->dash>=2.0.0->molplotly==1.1.4->-r drugex-
demo/requirements.txt (line 2)) (3.9.0)
Requirement already satisfied: typing-extensions>=3.6.4 in
/usr/local/lib/python3.7/dist-packages (from importlib-
metadata>=3.6.0->Flask>=1.0.4->dash>=2.0.0->molplotly==1.1.4->-r drugex-
demo/requirements.txt (line 2)) (4.1.1)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in
/usr/local/lib/python3.7/dist-packages (from
ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(3.0.3)
Requirement already satisfied: traitlets>=4.3.1 in
/usr/local/lib/python3.7/dist-packages (from
ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(5.1.1)
Requirement already satisfied: widgetsnbextension~=3.6.0 in
/usr/local/lib/python3.7/dist-packages (from
ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(3.6.1)
Requirement already satisfied: ipython-genutils~=0.2.0 in
/usr/local/lib/python3.7/dist-packages (from
ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(0.2.0)
Requirement already satisfied: ipython>=4.0.0 in /usr/local/lib/python3.7/dist-
packages (from ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
```

```
demo/requirements.txt (line 3)) (7.9.0)
Requirement already satisfied: jupyter-client in /usr/local/lib/python3.7/dist-
packages (from ipykernel->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (6.1.12)
Requirement already satisfied: tornado>=4.2 in /usr/local/lib/python3.7/dist-
packages (from ipykernel->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (5.1.1)
Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages
(from ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (4.8.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.7/dist-
packages (from ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (0.2.0)
Collecting jedi>=0.10
  Downloading jedi-0.18.1-py2.py3-none-any.whl (1.6 MB)
     |                  | 1.6 MB 44.1 MB/s
Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-
packages (from ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-
packages (from ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (0.7.5)
Requirement already satisfied: prompt-toolkit<2.1.0,>=2.0.0 in
/usr/local/lib/python3.7/dist-packages (from
ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (2.0.10)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.7/dist-packages (from
ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (57.4.0)
Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-
packages (from ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (2.6.1)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in
/usr/local/lib/python3.7/dist-packages (from
jedi>=0.10->ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (0.8.3)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.7/dist-
packages (from jinja2>=2.11.0->mols2grid==1.0.0->-r drugex-demo/requirements.txt
(line 3)) (2.0.1)
Collecting nest-asyncio
  Downloading nest_asyncio-1.5.6-py3-none-any.whl (5.2 kB)
Collecting retrying
  Downloading retrying-1.3.3.tar.gz (10 kB)
Collecting ansi2html
  Downloading ansi2html-1.8.0-py3-none-any.whl (16 kB)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-
packages (from matplotlib>=2.0->drugex@ git+https://github.com/martin-
```

```
sicho/DrugEx-CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (0.11.0)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python3.7/dist-packages (from matplotlib>=2.0->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.7/dist-packages (from matplotlib>=2.0->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (1.4.4)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/usr/local/lib/python3.7/dist-packages (from matplotlib>=2.0->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (3.0.9)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-
packages (from pandas>=1.2.2->drugex@ git+https://github.com/martin-
sicho/DrugEx-CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (2022.5)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.7/dist-
packages (from plotly>=5.0.0->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (8.1.0)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages
(from plotly>=5.0.0->molplotly==1.1.4->-r drugex-demo/requirements.txt (line 2))
(1.15.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages
(from prompt-
toolkit<2.1.0,>=2.0.0->ipython>=4.0.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r
drugex-demo/requirements.txt (line 3)) (0.2.5)
Requirement already satisfied: scipy>=1.1.0 in /usr/local/lib/python3.7/dist-
packages (from scikit-learn>=1.0.2->drugex@ git+https://github.com/martin-
sicho/DrugEx-CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (1.7.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.7/dist-packages (from scikit-learn>=1.0.2->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (3.1.0)
Collecting MarkupSafe>=2.0
  Downloading
MarkupSafe-2.1.1-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (25
kB)
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-
packages (from widgetsnbextension~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r
drugex-demo/requirements.txt (line 3)) (5.5.0)
Requirement already satisfied: pyzmq>=17 in /usr/local/lib/python3.7/dist-
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets<8,>=7->mol
s2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (23.2.1)
Requirement already satisfied: jupyter-core>=4.4.0 in
/usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension
~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt
(line 3)) (4.11.2)
Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-
```

```
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets<8,>=7->mol
s2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (1.8.0)
Requirement already satisfied: terminado>=0.8.1 in
/usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension
~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt
(line 3)) (0.13.3)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-
packages (from notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets<8,>=7->mol
s2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (5.6.1)
Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-
packages (from terminado>=0.8.1->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipy
widgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(0.7.0)
Collecting brotli
  Downloading Brotli-1.0.9-cp37-cp37m-manylinux1_x86_64.whl (357 kB)
     |                  | 357 kB 68.0 MB/s
Collecting gitdb<5,>=4.0.1
  Downloading gitdb-4.0.9-py3-none-any.whl (63 kB)
     |                  | 63 kB 1.7 MB/s
Collecting smmap<6,>=3.0.1
  Downloading smmap-5.0.0-py3-none-any.whl (24 kB)
Requirement already satisfied: networkx==2.* in /usr/local/lib/python3.7/dist-
packages (from mordred->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (2.6.3)
Requirement already satisfied: entrypoints>=0.2.2 in
/usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (0.4)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-
packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets
<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (0.7.1)
Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-
packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets
<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (0.6.0)
Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages
(from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.6.0->ipywidgets<8,>=7->m
ols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3)) (5.0.1)
Requirement already satisfied: mistune<2,>=0.8.1 in
/usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (0.8.4)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgets
nbextension~=3.6.0->ipywidgets<8,>=7->mols2grid==1.0.0->-r drugex-
demo/requirements.txt (line 3)) (1.5.0)
Requirement already satisfied: jsonschema>=2.6 in /usr/local/lib/python3.7/dist-
```

```
packages (from nbformat->molplotly==1.1.4->-r drugex-demo/requirements.txt (line
2)) (4.3.3)
Requirement already satisfied: fastjsonschema in /usr/local/lib/python3.7/dist-
packages (from nbformat->molplotly==1.1.4->-r drugex-demo/requirements.txt (line
2)) (2.16.2)
Requirement already satisfied: attrs>=17.4.0 in /usr/local/lib/python3.7/dist-
packages (from jsonschema>=2.6->nbformat->molplotly==1.1.4->-r drugex-
demo/requirements.txt (line 2)) (22.1.0)
Requirement already satisfied: importlib-resources>=1.4.0 in
/usr/local/lib/python3.7/dist-packages (from
jsonschema>=2.6->nbformat->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (5.10.0)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
/usr/local/lib/python3.7/dist-packages (from
jsonschema>=2.6->nbformat->molplotly==1.1.4->-r drugex-demo/requirements.txt
(line 2)) (0.18.1)
Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-
packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.6.0->ip
ywidgets<8,>=7->mols2grid==1.0.0->-r drugex-demo/requirements.txt (line 3))
(0.5.1)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-
packages (from optuna->drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (21.3)
Collecting cmaes>=0.8.2
  Downloading cmaes-0.8.2-py3-none-any.whl (15 kB)
Collecting alembic>=1.5.0
  Downloading alembic-1.8.1-py3-none-any.whl (209 kB)
     |                | 209 kB 61.1 MB/s
Requirement already satisfied: sqlalchemy>=1.3.0 in
/usr/local/lib/python3.7/dist-packages (from optuna->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (1.4.42)
Requirement already satisfied: PyYAML in /usr/local/lib/python3.7/dist-packages
(from optuna->drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (6.0)
Collecting colorlog
  Downloading colorlog-6.7.0-py2.py3-none-any.whl (11 kB)
Collecting cliff
  Downloading cliff-3.10.1-py3-none-any.whl (81 kB)
     |                | 81 kB 10.8 MB/s
Collecting Mako
  Downloading Mako-1.2.3-py3-none-any.whl (78 kB)
     |                | 78 kB 7.1 MB/s
Requirement already satisfied: greenlet!=0.4.17 in
/usr/local/lib/python3.7/dist-packages (from sqlalchemy>=1.3.0->optuna->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (1.1.3.post0)
Collecting stevedore>=2.0.1
```

```
  Downloading stevedore-3.5.2-py3-none-any.whl (50 kB)
     |                  | 50 kB 7.3 MB/s
Requirement already satisfied: PrettyTable>=0.7.2 in
/usr/local/lib/python3.7/dist-packages (from cliff->optuna->drugex@
git+https://github.com/martin-sicho/DrugEx-CDDG.git@master->-r drugex-
demo/requirements.txt (line 1)) (3.4.1)
Collecting cmd2>=1.0.0
  Downloading cmd2-2.4.2-py3-none-any.whl (147 kB)
     |                  | 147 kB 52.9 MB/s
Collecting pbr!=2.1.0,>=2.0.0
  Downloading pbr-5.11.0-py2.py3-none-any.whl (112 kB)
     |                  | 112 kB 61.3 MB/s
Collecting autopage>=0.4.0
  Downloading autopage-0.5.1-py3-none-any.whl (29 kB)
Collecting pyperclip>=1.6
  Downloading pyperclip-1.8.2.tar.gz (20 kB)
Collecting pickle5
  Downloading
pickle5-0.0.12-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl (256 kB)
     |                  | 256 kB 55.3 MB/s
Requirement already satisfied: Pillow in /usr/local/lib/python3.7/dist-
packages (from rdkit-pypi->drugex@ git+https://github.com/martin-sicho/DrugEx-
CDDG.git@master->-r drugex-demo/requirements.txt (line 1)) (7.1.2)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.7/dist-packages (from requests->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-
packages (from requests->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.7/dist-packages (from requests->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (1.24.3)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.7/dist-packages (from requests->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (2022.9.24)
Collecting psutil
  Downloading psutil-5.9.3-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86_64
.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (291 kB)
     |                  | 291 kB 59.3 MB/s
Requirement already satisfied: dask[dataframe]>=2.10.0 in
```

```
/usr/local/lib/python3.7/dist-packages (from swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (2022.2.0)
Requirement already satisfied: cloudpickle>=0.2.2 in
/usr/local/lib/python3.7/dist-packages (from swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (1.5.0)
Requirement already satisfied: toolz>=0.8.2 in /usr/local/lib/python3.7/dist-
packages (from dask[dataframe]>=2.10.0->swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (0.12.0)
Requirement already satisfied: fsspec>=0.6.0 in /usr/local/lib/python3.7/dist-
packages (from dask[dataframe]>=2.10.0->swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (2022.10.0)
Requirement already satisfied: partd>=0.3.10 in /usr/local/lib/python3.7/dist-
packages (from dask[dataframe]>=2.10.0->swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (1.3.0)
Requirement already satisfied: locket in /usr/local/lib/python3.7/dist-packages
(from partd>=0.3.10->dask[dataframe]>=2.10.0->swifter->papyrus_scripts@
git+https://github.com/OlivierBeq/Papyrus-scripts.git@master->papyrus-scaffold-
visualizer@ git+https://github.com/martin-sicho/papyrus-scaffold-
visualizer.git@v0.2.0->-r drugex-demo/requirements.txt (line 4)) (1.0.0)
Building wheels for collected packages: drugex, papyrus-scaffold-visualizer,
papyrus-scripts, prodec, upsetplot, molplotly, mordred, pyperclip, retrying,
sklearn, swifter
  Building wheel for drugex (setup.py) … done
  Created wheel for drugex: filename=drugex-3.2.0-py3-none-any.whl size=11599995
sha256=a8ad1e12e3d3e53430c9f0ad1cc4b7a4a396966f0b7f644f9d89b3803a9e57b5
  Stored in directory: /tmp/pip-ephem-wheel-cache-
gy91px2u/wheels/a1/0e/fa/7e538fa81bdfab41f4b8ba576d20eaa36bedd87ccd7bfec4f8
  Building wheel for papyrus-scaffold-visualizer (setup.py) … done
  Created wheel for papyrus-scaffold-visualizer:
filename=papyrus_scaffold_visualizer-0.2.0-py3-none-any.whl size=12960
sha256=a109868db43130e0ff35b0b5992afc644c85a8b4d819d6b207fc91cab8bda64e
  Stored in directory: /tmp/pip-ephem-wheel-cache-
gy91px2u/wheels/f6/ba/43/791fe3545a9132773948556fe2eb428b78054589dc482d8f54
  Building wheel for papyrus-scripts (setup.py) … done
  Created wheel for papyrus-scripts:
filename=papyrus_scripts-0.0.2.dev0-py3-none-any.whl size=67671
sha256=4f944bb49282548e57c533e9c6133c8370687a0e39f5664ea1f1a11cb51a9a41
  Stored in directory: /tmp/pip-ephem-wheel-cache-
```

```
gy91px2u/wheels/21/ff/e1/2dfa3afa7830fceb3e04db76e7ad7733f064310f9516a11bea
  Building wheel for prodec (PEP 517) … done
  Created wheel for prodec: filename=prodec-1.0.2.post3-py3-none-any.whl
size=52665
sha256=992adcfaa2eacd33ee6d27ca7ce889cf0f41b058453ded7a5ff7750ae838fca8
  Stored in directory: /tmp/pip-ephem-wheel-cache-
gy91px2u/wheels/26/32/22/62c3caf3cab13b3e470bd9259ae811e2c929ac3f5595f99111
  Building wheel for upsetplot (setup.py) … done
  Created wheel for upsetplot: filename=UpSetPlot-0.7.dev1-py3-none-any.whl
size=21430
sha256=2d8746e3f058fa39beb0ba3bf0d7f4074c8bf800b5093f1ddefef4bd1c4bf527
  Stored in directory: /tmp/pip-ephem-wheel-cache-
gy91px2u/wheels/53/e0/5c/6e251495ed03bf62e04fed2ed92d1fca8c5e586650896f35f2
  Building wheel for molplotly (setup.py) … done
  Created wheel for molplotly: filename=molplotly-1.1.4-py3-none-any.whl
size=13230
sha256=842df7ad8d63cf17ea1fe0cb5563183b47c11a513ba1e5aaaf567130c0de1168
  Stored in directory: /root/.cache/pip/wheels/4c/67/f7/022d8010193733af123e5327
c03775e7d85767ca35b66f79e6
  Building wheel for mordred (setup.py) … done
  Created wheel for mordred: filename=mordred-1.2.0-py3-none-any.whl size=176725
sha256=7536e64f669eeb541a33fe6483956f7d5d093f0e0dd921765188ae1e5a309b18
  Stored in directory: /root/.cache/pip/wheels/02/c0/2e/e7e3d63b431777712ebc128b
c4deb9ac5cb19afc7c1ea341ec
  Building wheel for pyperclip (setup.py) … done
  Created wheel for pyperclip: filename=pyperclip-1.8.2-py3-none-any.whl
size=11137
sha256=abab3a80b36aabdabb579205b0f3defeae59f110e73ef7c5a0065caa93027b48
  Stored in directory: /root/.cache/pip/wheels/9f/18/84/8f69f8b08169c7bae2dde6bd
7daf0c19fca8c8e500ee620a28
  Building wheel for retrying (setup.py) … done
  Created wheel for retrying: filename=retrying-1.3.3-py3-none-any.whl
size=11447
sha256=aaa1cfd8cfa119ce21bad9fdf4119e6f9c186b40789fac38894784505d84ac79
  Stored in directory: /root/.cache/pip/wheels/f9/8d/8d/f6af3f7f9eea3553bc2fe6d5
3e4b287dad18b06a861ac56ddf
  Building wheel for sklearn (setup.py) … done
  Created wheel for sklearn: filename=sklearn-0.0-py2.py3-none-any.whl size=1310
sha256=3fa496f9396d2859cbd4a9104d0dbf7a857f34533bb4c9fa9bd134c56c05337f
  Stored in directory: /root/.cache/pip/wheels/46/ef/c3/157e41f5ee1372d1be90b09f
74f82b10e391eaacca8f22d33e
  Building wheel for swifter (setup.py) … done
  Created wheel for swifter: filename=swifter-1.3.4-py3-none-any.whl size=16322
sha256=5627ace386365883056df80a8c351bcd88d6d083895f593f89d5bc3eae9c1029
  Stored in directory: /root/.cache/pip/wheels/29/a7/0e/3a8f17ac69d759e1e9364711
4bc9bdc95957e5b0cbfd405205
Successfully built drugex papyrus-scaffold-visualizer papyrus-scripts prodec
upsetplot molplotly mordred pyperclip retrying sklearn swifter
```

```
Installing collected packages: MarkupSafe, jedi, jinja2, werkzeug, itsdangerous,
click, Flask, brotli, pyperclip, pbr, flask-compress, dash-table, dash-html-
components, dash-core-components, stevedore, smmap, retrying, psutil, pickle5,
orjson, nest-asyncio, Mako, dash, cmd2, autopage, ansi2html, upsetplot, swifter,
pystow, prodec, mordred, jupyter-dash, gitdb, colorlog, cmaes, cliff, alembic,
sklearn, rdkit-pypi, papyrus-scripts, optuna, molplotly, gitpython, papyrus-
scaffold-visualizer, mols2grid, drugex
  Attempting uninstall: MarkupSafe
    Found existing installation: MarkupSafe 2.0.1
    Uninstalling MarkupSafe-2.0.1:
      Successfully uninstalled MarkupSafe-2.0.1
  Attempting uninstall: jinja2
    Found existing installation: Jinja2 2.11.3
    Uninstalling Jinja2-2.11.3:
      Successfully uninstalled Jinja2-2.11.3
  Attempting uninstall: werkzeug
    Found existing installation: Werkzeug 1.0.1
    Uninstalling Werkzeug-1.0.1:
      Successfully uninstalled Werkzeug-1.0.1
  Attempting uninstall: itsdangerous
    Found existing installation: itsdangerous 1.1.0
    Uninstalling itsdangerous-1.1.0:
      Successfully uninstalled itsdangerous-1.1.0
  Attempting uninstall: click
    Found existing installation: click 7.1.2
    Uninstalling click-7.1.2:
      Successfully uninstalled click-7.1.2
  Attempting uninstall: Flask
    Found existing installation: Flask 1.1.4
    Uninstalling Flask-1.1.4:
      Successfully uninstalled Flask-1.1.4
  Attempting uninstall: psutil
    Found existing installation: psutil 5.4.8
    Uninstalling psutil-5.4.8:
      Successfully uninstalled psutil-5.4.8
Successfully installed Flask-2.2.2 Mako-1.2.3 MarkupSafe-2.1.1 alembic-1.8.1
ansi2html-1.8.0 autopage-0.5.1 brotli-1.0.9 click-8.1.3 cliff-3.10.1 cmaes-0.8.2
cmd2-2.4.2 colorlog-6.7.0 dash-2.6.2 dash-core-components-2.0.0 dash-html-
components-2.0.0 dash-table-5.0.0 drugex-3.2.0 flask-compress-1.13 gitdb-4.0.9
gitpython-3.1.29 itsdangerous-2.1.2 jedi-0.18.1 jinja2-3.1.2 jupyter-dash-0.4.2
molplotly-1.1.4 mols2grid-1.0.0 mordred-1.2.0 nest-asyncio-1.5.6 optuna-3.0.3
orjson-3.8.1 papyrus-scaffold-visualizer-0.2.0 papyrus-scripts-0.0.2.dev0
pbr-5.11.0 pickle5-0.0.12 prodec-1.0.2.post3 psutil-5.9.3 pyperclip-1.8.2
pystow-0.4.6 rdkit-pypi-2022.9.1 retrying-1.3.3 sklearn-0.0 smmap-5.0.0
stevedore-3.5.2 swifter-1.3.4 upsetplot-0.7.dev1 werkzeug-2.2.2
```

[5]: `'/content/drive/MyDrive/DrugExDemo'`

[6]:
```
import pandas as pd

df_all = pd.read_table("data/PTP1B_LIGANDS.tsv")
df_all.columns
```

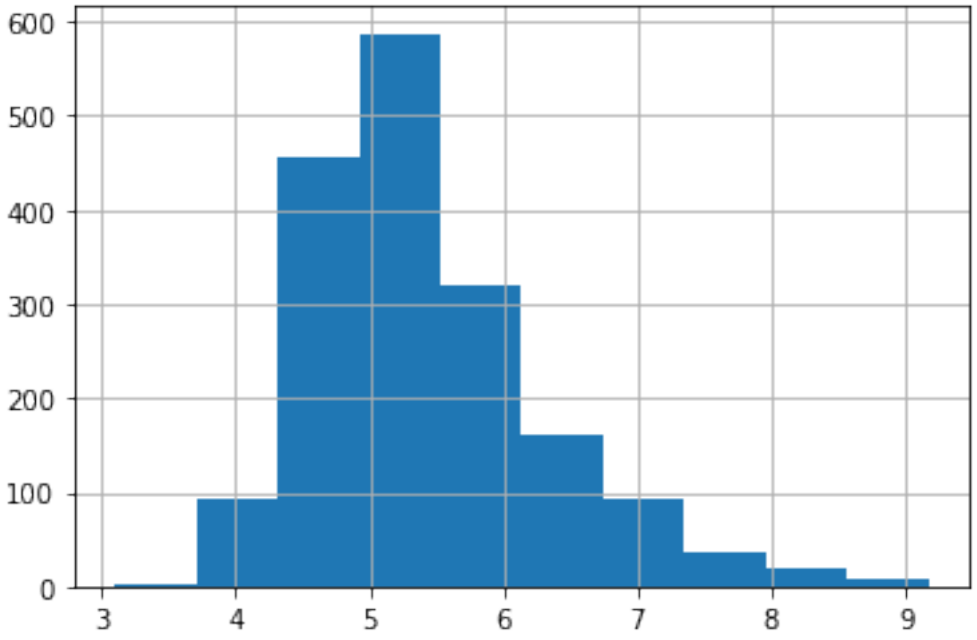
[6]:
```
Index(['Activity_ID', 'Quality', 'source', 'CID', 'SMILES', 'connectivity',
       'InChIKey', 'InChI', 'InChI_AuxInfo', 'target_id',
       …
       'Descriptor_MorganFP_1016', 'Descriptor_MorganFP_1017',
       'Descriptor_MorganFP_1018', 'Descriptor_MorganFP_1019',
       'Descriptor_MorganFP_1020', 'Descriptor_MorganFP_1021',
       'Descriptor_MorganFP_1022', 'Descriptor_MorganFP_1023', 'TSNE_1',
       'TSNE_2'],
      dtype='object', length=1056)
```

1.1 Assigning Activity Classes

As you can see it has all the feature we previously added to it, including the t-SNE embedding. But in this part of the tutorial we are more interested in the pChEMBL activities. Let's create a histogram of these values to get an idea of a distribution:

[8]:
```
df_all.pchembl_value_Median.hist()
```

[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc009cbe850>
```



We have a nice normal distribution of values as would be expected, but we also have a fairly balanced

ration between inactive and active compounds, which is always a good thing when we are trying to build machine learning models. For example, if we take into account the previously introduced 6.5 threshold that distinguishes actives from inactives, we get reasonably balanced subsets:

[9]:
```
activity_class_mask = df_all.pchembl_value_Median >= 6.5
sum(activity_class_mask) / len(activity_class_mask)
```

[9]:
```
0.11455981941309255
```

Therefore, this split of the data leaves us with about 61% of active molecules. Therefore, we do not need to think much about what to do here for now, but note that for a lot of other data sets the ratio of actives could be much lower and then we might want to think about using some data balancing techniques.

We will now create a new column in this data set that will serve as the endpoint we want to model (the assignment to an activity class based on molecular structure):

[10]:
```
df_all['ActivityClass'] = activity_class_mask
df_all.head()
```

[10]:

	Activity_ID	Quality	source	CID	SMILES	connectivity	InChIKey
0	AADVGRCQGZZZNH_on_P18031_WT	High	ChEMBL30	CHEMBL363338;44397840;CHEMBL363338;44397840;CH…	CC(C)CC(CC(=O)C(Cc1ccc(OC(F)(F)C(=O)O)cc1)NC(=…	AADVGRCQGZZZNH	AADVGRCQGZZZNH-UHFFFAOYSA-N
1	AAGZFGSRAZMRCS_on_P18031_WT	High	ChEMBL30	CHEMBL511000;CHEMBL511000;44563281;CHEMBL51100…	CC1(C)CCC2(C(=O)Nc3ccc(C(=O)O)cc3)CCC3(C)C(=CC…	AAGZFGSRAZMRCS	AAGZFGSRAZMRCS-UHFFFAOYSA-N
2	AATACZKOPAGNPL_on_P18031_WT	High	ChEMBL30	CHEMBL4160013	CCOc1ccc(OC(=O)CSc2nnc(-c3cc(O)c(O)cc3)o2)cc1	AATACZKOPAGNPL	AATACZKOPAGNPL-UHFFFAOYSA-N
3	AAWBMDOJDQKGGQ_on_P18031_WT	High	ChEMBL30	CHEMBL3759405	CCCOc1ccc(-c2cccc(-c3noc(Cc4c[nH]c5ccccc45)n3)…	AAWBMDOJDQKGGQ	AAWBMDOJDQKGGQ-UHFFFAOYSA-N
4	ABDQNRCWGSGNBQ_on_P18031_WT	High	ChEMBL30	CHEMBL197929	O=C(O)COc1ccc(S(=O)(=O)N(Cc2ccc(C(F)(F)P(=O)(O…	ABDQNRCWGSGNBQ	ABDQNRCWGSGNBQ-UHFFFAOYSA-N

```
                                                InChI  \
0  InChI=1S/C45H47F2N3O10/c1-27(2)22-30(41(48)53)…
1  InChI=1S/C37H53NO4/c1-32(2)18-20-37(31(42)38-2…
2  InChI=1S/C18H16N2O6S/c1-2-24-12-4-6-13(7-5-12)…
3  InChI=1S/C25H22N4O2/c1-2-14-30-19-12-10-17(11-…
4  InChI=1S/C21H20F2NO8PS2/c22-21(23,33(27,28)29)…

                                    InChI_AuxInfo  target_id  …  \
0  "AuxInfo=1/1/N:1,3,37,36,38,49,56,48,55,35,39,…  P18031_WT  …
1  "AuxInfo=1/1/N:1,3,34,35,28,22,40,12,17,11,18,…  P18031_WT  …
2  "AuxInfo=1/0/N:1,2,24,5,27,6,26,23,18,11,17,4,…  P18031_WT  …
3  "AuxInfo=1/0/N:1,2,25,24,11,26,23,10,12,7,30,6…  P18031_WT  …
4  "AuxInfo=1/1/N:31,30,16,27,17,26,7,35,8,34,32,…  P18031_WT  …

  Descriptor_MorganFP_1017 Descriptor_MorganFP_1018 Descriptor_MorganFP_1019  \
0                      0.0                      1.0                      1.0
1                      0.0                      0.0                      1.0
2                      0.0                      0.0                      0.0
3                      0.0                      0.0                      0.0
4                      0.0                      0.0                      0.0

  Descriptor_MorganFP_1020  Descriptor_MorganFP_1021 Descriptor_MorganFP_1022  \
0                      0.0                       0.0                      0.0
1                      0.0                       0.0                      0.0
2                      0.0                       0.0                      0.0
3                      0.0                       0.0                      0.0
4                      0.0                       0.0                      0.0

  Descriptor_MorganFP_1023     TSNE_1     TSNE_2 ActivityClass
0                      0.0  29.735216  -4.313994         False
1                      0.0  -0.945723  55.179680         False
2                      0.0  -0.630247 -24.217190         False
3                      0.0  17.455896 -43.069510         False
4                      0.0  37.058834   5.535724         False

[5 rows x 1057 columns]
```

1.2 Creating a Test Set

A lot has been written about selecting appropriate test sets for machine learning and indeed that should always be an important first step. The purpose of the test set is to get us an idea on future/prospective performance of our model on unknown data. In QSAR, a popular choice is to make a 'time split' of the data based on publication year:

[11]:
```
cutoff_year = 2015
sum(df_all.Year >= cutoff_year) / len(df_all)
```

[11]: 0.3182844243792325

If we train our model on data before year 2015, we could get an idea about how it could perform on data that was not known yet at that time. This approach also has plenty of caveats, one being that the data can follow a different development pattern over time. For example, it can happen that many more chemically novel molecules were found after 2015 making it much harder for the model when previously many chemically related structurues were explored. Fluctuations like these over time can make it harder to find an appropriate split and can result in some unfair comparison so one should always think twice when applying such a split. Here we base our decision solely on the fact that split on year 2015 lands around 20% of the data set in the test set, which is good enough for us, but does not mean it should be good enough for you in every situation. Let's now save our 'Train' and 'Test' set assignment into the data set:

[12]:
```
df_all["TimeSplit"] = (df_all.Year >= cutoff_year).apply(lambda x : "Test" if x␣
 ↪else "Train")
df_all
```

[12]:

	Activity_ID	Quality	source	CID	SMILES	connectivity
0	AADVGRCQGZZZNH_on_P18031_WT	High	ChEMBL30	CHEMBL363338;44397840;CHEMBL363338;44397840;CH…	CC(C)CC(CC(=O)C(Cc1ccc(OC(F)(F)C(=O)O)cc1)NC(=…	AADVGRCQGZZZNH
1	AAGZFGSRAZMRCS_on_P18031_WT	High	ChEMBL30	CHEMBL511000;CHEMBL511000;44563281;CHEMBL51100…	CC1(C)CCC2(C(=O)Nc3ccc(C(=O)O)cc3)CCC3(C)C(=CC…	AAGZFGSRAZMRCS
2	AATACZKOPAGNPL_on_P18031_WT	High	ChEMBL30	CHEMBL4160013	CCOc1ccc(OC(=O)CSc2nnc(-c3cc(O)c(O)cc3)o2)cc1	AATACZKOPAGNPL
3	AAWBMDOJDQKGGQ_on_P18031_WT	High	ChEMBL30	CHEMBL3759405		
4	ABDQNRCWGSGNBQ_on_P18031_WT	High	ChEMBL30	CHEMBL197929		
…	…	…	…	…		
1767	ZZLZHWBNIIKVOG_on_P18031_WT	High	ChEMBL30	CHEMBL3974642;CHEMBL3903731;CHEMBL3974642;CHEM…		
1768	ZZQGXJVGZKQIGN_on_P18031_WT	High	ChEMBL30	CHEMBL3402418		
1769	ZZTBNXDDDGFULL_on_P18031_WT	High	ChEMBL30	CHEMBL1778901		
1770	ZZTYPLSBNNGEIS_on_P18031_WT	High	ChEMBL30	CHEMBL486986		
1771	ZZVOLWIBVIYSBV_on_P18031_WT	High	ChEMBL30	CHEMBL58354		

```
3     CCCOc1ccc(-c2cccc(-c3noc(Cc4c[nH]c5ccccc45)n3)…  AAWBMDOJDQKGGQ
4     O=C(O)COc1ccc(S(=O)(=O)N(Cc2ccc(C(F)(F)P(=O)(O…  ABDQNRCWGSGNBQ
…                                                   …             …
1767  CC(C)CC(c1c(O)c(C=O)c(O)c(C=O)c1O)C1CCC(C)C2(C…  ZZLZHWBNIIKVOG
1768  O=C(OC1CSSC1)c1ccc(CCCCCCOC(=O)c2c(Br)ccc(Br)c…  ZZQGXJVGZKQIGN
1769         CCCCOc1ccc(C(=O)C=Cc2c(OC)cc(O)c(Br)c2)cc1  ZZTBNXDDDGFULL
1770  CC1CCC2(C(=O)O)CCC3(C)C(=CCC4C5(C)CCC(O)C(C)(C…  ZZTYPLSBNNGEIS
1771  CC(C)(C)OC(=O)COc1ccc(CC2=C(c3ccccc3)c3ccccc3C…  ZZVOLWIBVIYSBV

                          InChIKey  \
0     AADVGRCQGZZZNH-UHFFFAOYSA-N
1     AAGZFGSRAZMRCS-UHFFFAOYSA-N
2     AATACZKOPAGNPL-UHFFFAOYSA-N
3     AAWBMDOJDQKGGQ-UHFFFAOYSA-N
4     ABDQNRCWGSGNBQ-UHFFFAOYSA-N
…                               …
1767  ZZLZHWBNIIKVOG-UHFFFAOYSA-N
1768  ZZQGXJVGZKQIGN-UHFFFAOYSA-N
1769  ZZTBNXDDDGFULL-UHFFFAOYSA-N
1770  ZZTYPLSBNNGEIS-UHFFFAOYSA-N
1771  ZZVOLWIBVIYSBV-UHFFFAOYSA-N

                                                    InChI  \
0     InChI=1S/C45H47F2N3O10/c1-27(2)22-30(41(48)53)…
1     InChI=1S/C37H53NO4/c1-32(2)18-20-37(31(42)38-2…
2     InChI=1S/C18H16N2O6S/c1-2-24-12-4-6-13(7-5-12)…
3     InChI=1S/C25H22N4O2/c1-2-14-30-19-12-10-17(11-…
4     InChI=1S/C21H20F2NO8PS2/c22-21(23,33(27,28)29)…
…                                                   …
1767  InChI=1S/C28H38O5/c1-14(2)11-17(22-24(32)18(12…
1768  InChI=1S/C23H24Br2O4S2/c24-18-10-11-21(25)20(1…
1769  InChI=1S/C20H21BrO4/c1-3-4-11-25-16-8-5-14(6-9…
1770  InChI=1S/C30H48O4/c1-18-10-15-30(24(32)33)17-1…
1771  InChI=1S/C29H26O5/c1-29(2,3)34-25(30)18-33-21-…

                                         InChI_AuxInfo  target_id  …  \
0     "AuxInfo=1/1/N:1,3,37,36,38,49,56,48,55,35,39,…  P18031_WT  …
1     "AuxInfo=1/1/N:1,3,34,35,28,22,40,12,17,11,18,…  P18031_WT  …
2     "AuxInfo=1/0/N:1,2,24,5,27,6,26,23,18,11,17,4,…  P18031_WT  …
3     "AuxInfo=1/0/N:1,2,25,24,11,26,23,10,12,7,30,6…  P18031_WT  …
4     "AuxInfo=1/1/N:31,30,16,27,17,26,7,35,8,34,32,…  P18031_WT  …
…                                                   …         …  …
1767  "AuxInfo=1/0/N:1,3,23,32,33,25,21,20,27,28,4,1…  P18031_WT  …
1768  "AuxInfo=1/0/N:15,16,14,17,13,11,30,10,31,26,2…  P18031_WT  …
1769  "AuxInfo=1/0/N:1,17,2,3,8,24,13,7,25,12,4,23,1…  P18031_WT  …
1770  "AuxInfo=1/1/N:1,24,25,18,12,30,33,14,15,3,27,…  P18031_WT  …
1771  "AuxInfo=1/0/N:1,3,4,20,19,21,25,26,18,22,24,2…  P18031_WT  …
```

```
     Descriptor_MorganFP_1018 Descriptor_MorganFP_1019  \
0                         1.0                      1.0
1                         0.0                      1.0
2                         0.0                      0.0
3                         0.0                      0.0
4                         0.0                      0.0
…                         …                        …
1767                      0.0                      1.0
1768                      0.0                      1.0
1769                      0.0                      0.0
1770                      0.0                      1.0
1771                      0.0                      0.0

     Descriptor_MorganFP_1020 Descriptor_MorganFP_1021  \
0                         0.0                      0.0
1                         0.0                      0.0
2                         0.0                      0.0
3                         0.0                      0.0
4                         0.0                      0.0
…                         …                        …
1767                      0.0                      0.0
1768                      0.0                      0.0
1769                      0.0                      0.0
1770                      0.0                      0.0
1771                      0.0                      0.0

      Descriptor_MorganFP_1022 Descriptor_MorganFP_1023     TSNE_1     TSNE_2  \
0                          0.0                      0.0  29.735216  -4.313994
1                          0.0                      0.0  -0.945723  55.179680
2                          0.0                      0.0  -0.630247 -24.217190
3                          0.0                      0.0  17.455896 -43.069510
4                          0.0                      0.0  37.058834   5.535724
…                          …                        …          …          …
1767                       0.0                      0.0   0.197739  21.740646
1768                       0.0                      0.0 -11.149357  -5.444707
1769                       0.0                      0.0 -35.552700 -15.787386
1770                       0.0                      0.0   1.356204  52.808704
1771                       0.0                      0.0  10.432344 -11.464129

     ActivityClass TimeSplit
0            False     Train
1            False     Train
2            False      Test
3            False      Test
4            False     Train
…            …       …
```

```
1767         False      Test
1768         False      Test
1769         False     Train
1770         False     Train
1771         False     Train

[1772 rows x 1058 columns]
```

Now, we can use the `scaffviz` package again to visualize how different our test set is from the training set. This will be useful to assess how difficult the test set is for the resulting model. We start by saving our annotated data set into a new file and wrapping it as a `DataSetTSV` instance for plotting:

[13]:
```
from scaffviz.data.dataset import DataSetTSV

dataset = DataSetTSV(data=df_all, path='data/PTP1B_LIGANDS_qsar.tsv')
```

Next, we can just drop it into the `Plot.plot` method again:

Note: We should still have the embedding calculated from the previous exercise, but if you do not have that information anymore, it will be recalculated.

[15]:
```
from scaffviz.depiction.plot import Plot
from scaffviz.clustering.manifold import TSNE

plot = Plot(dataset, TSNE())
plot.plot(
    recalculate=False,
    color_by="TimeSplit",
    color_style="groups",
    card_data=["pchembl_value_Median", "all_doc_ids", "source"],
    title_data="Activity_ID",
    viewport_height=800,
    port=9191
)
```

```
<IPython.core.display.Javascript object>
```

In the plot, we can see that some of the test compounds are in similar parts of chemical space as the training compounds, but there are also a few 'lonely' compounds that cluster together. So the split we selected actually seems to have a good balance between compounds that are challenging (more distant from the training data) and easier (closer to the training data).

1.3 Model Training

First, we seperate the testing part of the data set and start preparing our training set for modelling:

[16]:
```
df_train = df_all[df_all.TimeSplit == "Train"]
len(df_train)
```

[16]:
```
1208
```

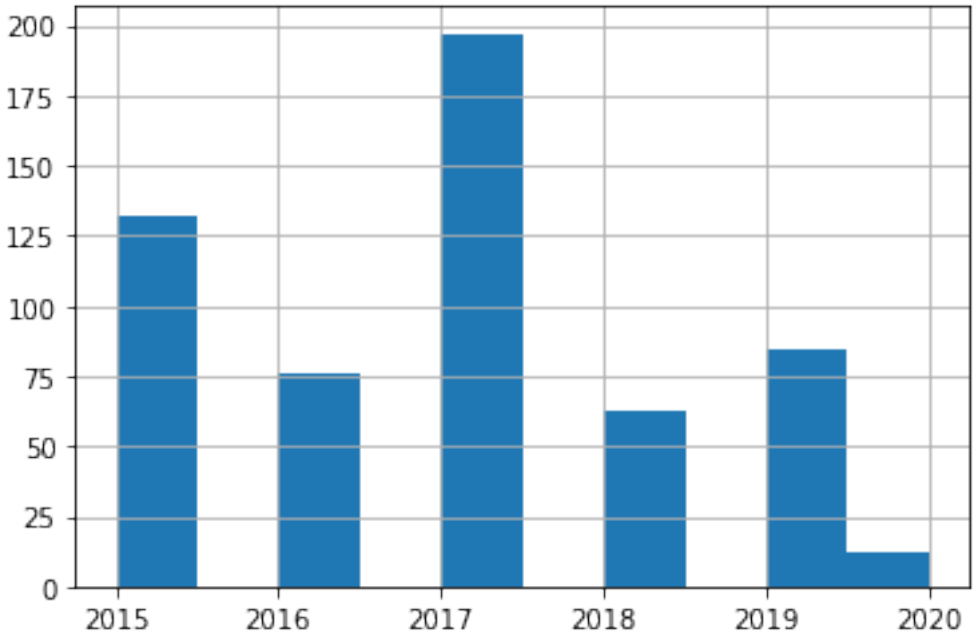
[17]:
```
df_test = df_all[df_all.TimeSplit == "Test"]
len(df_test)
```

[17]:
```
564
```

And finally a quick sanity check that we truly seperated the data correctly by year:

[18]:
```
df_test.Year.hist()
```

[18]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbfebe37e90>
```



Next, we extract the descriptors we will use to derive the QSAR model. This is how we will relate structural information with the endpoint variable (the `ActivityClass` group). The model will need two types of inputs:

1. The data describing molecular structure. In our case, these will be the fingerprints we also used to make the t-SNE plot above.
2. The categorization to `ActivityClass`. This will be used to fit the model and create the mathematical links between the structure (the fingerprints) and the categorization variable (`ActivityClass`).

All descriptor column names start with the `Descriptor_` string so that allows us to extract them easily from the data:

[19]:
```
X = df_train[df_train.columns[df_train.columns.str.startswith("Descriptor_")]]
X
```

[19]:

	Descriptor_MorganFP_0	Descriptor_MorganFP_1	Descriptor_MorganFP_2	\
0	0.0	1.0	0.0	
1	0.0	0.0	0.0	
4	0.0	0.0	0.0	
5	0.0	0.0	0.0	
7	0.0	0.0	0.0	
…	…	…	…	
1764	0.0	1.0	1.0	
1765	0.0	1.0	0.0	
1769	0.0	0.0	0.0	
1770	0.0	0.0	0.0	
1771	0.0	1.0	0.0	

	Descriptor_MorganFP_3	Descriptor_MorganFP_4	Descriptor_MorganFP_5	\
0	1.0	0.0	0.0	
1	0.0	0.0	0.0	
4	0.0	0.0	0.0	
5	0.0	0.0	0.0	
7	0.0	0.0	0.0	
…	…	…	…	
1764	0.0	1.0	0.0	
1765	1.0	0.0	0.0	
1769	0.0	0.0	0.0	
1770	0.0	0.0	0.0	
1771	0.0	0.0	0.0	

	Descriptor_MorganFP_6	Descriptor_MorganFP_7	Descriptor_MorganFP_8	\
0	0.0	0.0	0.0	
1	0.0	0.0	0.0	
4	0.0	0.0	0.0	
5	0.0	0.0	0.0	
7	0.0	0.0	0.0	
…	…	…	…	
1764	0.0	0.0	0.0	
1765	0.0	0.0	0.0	
1769	0.0	0.0	0.0	
1770	0.0	0.0	0.0	
1771	0.0	0.0	0.0	

	Descriptor_MorganFP_9	…	Descriptor_MorganFP_1014	\
0	0.0	…	0.0	
1	0.0	…	0.0	
4	0.0	…	0.0	
5	0.0	…	0.0	

```
7                       0.0  …                       0.0
…                       …  …                       …
1764                    0.0  …                       0.0
1765                    0.0  …                       0.0
1769                    0.0  …                       0.0
1770                    0.0  …                       0.0
1771                    0.0  …                       0.0

      Descriptor_MorganFP_1015  Descriptor_MorganFP_1016  \
0                          0.0                       0.0
1                          0.0                       0.0
4                          0.0                       0.0
5                          0.0                       0.0
7                          0.0                       0.0
…                          …                         …
1764                       0.0                       0.0
1765                       0.0                       0.0
1769                       0.0                       0.0
1770                       0.0                       0.0
1771                       0.0                       0.0

      Descriptor_MorganFP_1017  Descriptor_MorganFP_1018  \
0                          0.0                       1.0
1                          0.0                       0.0
4                          0.0                       0.0
5                          0.0                       0.0
7                          0.0                       0.0
…                          …                         …
1764                       0.0                       1.0
1765                       1.0                       0.0
1769                       0.0                       0.0
1770                       0.0                       0.0
1771                       0.0                       0.0

      Descriptor_MorganFP_1019  Descriptor_MorganFP_1020  \
0                          1.0                       0.0
1                          1.0                       0.0
4                          0.0                       0.0
5                          0.0                       0.0
7                          0.0                       1.0
…                          …                         …
1764                       1.0                       0.0
1765                       1.0                       0.0
1769                       0.0                       0.0
1770                       1.0                       0.0
1771                       0.0                       0.0
```

```
      Descriptor_MorganFP_1021  Descriptor_MorganFP_1022  \
0                          0.0                       0.0
1                          0.0                       0.0
4                          0.0                       0.0
5                          0.0                       0.0
7                          0.0                       0.0
…                          …                         …
1764                       0.0                       0.0
1765                       0.0                       0.0
1769                       0.0                       0.0
1770                       0.0                       0.0
1771                       0.0                       0.0

      Descriptor_MorganFP_1023
0                          0.0
1                          0.0
4                          0.0
5                          0.0
7                          0.0
…                          …
1764                       0.0
1765                       0.0
1769                       0.0
1770                       0.0
1771                       0.0

[1208 rows x 1024 columns]
```

The classification is just the `ActivityClass` column:

[20]:
```
y = df_train.ActivityClass
y
```

[20]:
```
0       False
1       False
4       False
5       False
7       False
        …
1764    False
1765    False
1769    False
1770    False
1771    False
Name: ActivityClass, Length: 1208, dtype: bool
```

Finally, we can train our classifier. That alone is usually a lengthy process of trial and error during which you may want to tune the kind of machine learning algorithms you want to use and their

hyper-parameters (i.e. `n_estimators` for the `ExtraTreesClassifier` algorithm we will be using). However, you may also want to explore different validation strategies as well as data balancing methods as hinted at previously. To make this tutorial as simple as possible, we will just opt for a simple n-fold cross-validation approach.

First, let us define our model:

[21]:
```
from sklearn.ensemble import ExtraTreesClassifier

classifier = ExtraTreesClassifier(n_estimators=250)
```

We will fix the parameters of this models for simplicity and dive right into cross-validation. Under this scheme, the model is trained multiple times and its performance validated n times:

The purpose of this exercise is to give us a good idea of the algorithm's ability to learn the QSAR patterns in the data and to accurately predict the `ActivityClass` for unseen data points. It is a common strategy to compare multiple models or the same model with different hyper-parameters. The `scikit-learn` documentation provides and excellent example that allows us to just plugin our classifier into the workflow:

[22]:
```
# adapted from https://scikit-learn.org/stable/modules/cross_validation.html

import matplotlib.pyplot as plt
import numpy as np

from sklearn.ensemble import ExtraTreesClassifier
from sklearn.metrics import auc
from sklearn.metrics import RocCurveDisplay
from sklearn.model_selection import StratifiedKFold

# Run classifier with cross-validation and plot ROC curves
cv = StratifiedKFold(n_splits=5)

tprs = []
aucs = []
mean_fpr = np.linspace(0, 1, 100)

fig, ax = plt.subplots()
for i, (train, test) in enumerate(cv.split(X, y)):
    classifier.fit(X.iloc[train], y.iloc[train])
    viz = RocCurveDisplay.from_estimator(
        classifier,
        X.iloc[test],
        y.iloc[test],
        name="ROC fold {}".format(i),
        alpha=0.3,
        lw=1,
        ax=ax,
    )
```

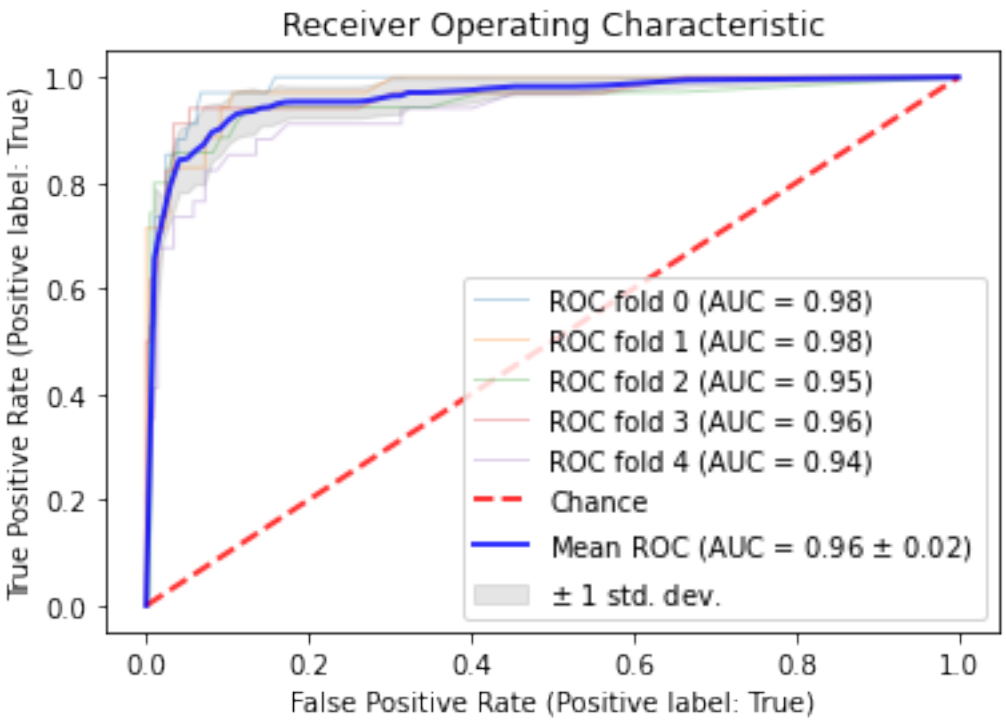
```
    interp_tpr = np.interp(mean_fpr, viz.fpr, viz.tpr)
    interp_tpr[0] = 0.0
    tprs.append(interp_tpr)
    aucs.append(viz.roc_auc)

ax.plot([0, 1], [0, 1], linestyle="--", lw=2, color="r", label="Chance",␣
 ↪alpha=0.8)

mean_tpr = np.mean(tprs, axis=0)
mean_tpr[-1] = 1.0
mean_auc = auc(mean_fpr, mean_tpr)
std_auc = np.std(aucs)
ax.plot(
    mean_fpr,
    mean_tpr,
    color="b",
    label=r"Mean ROC (AUC = %0.2f $\pm$ %0.2f)" % (mean_auc, std_auc),
    lw=2,
    alpha=0.8,
)

std_tpr = np.std(tprs, axis=0)
tprs_upper = np.minimum(mean_tpr + std_tpr, 1)
tprs_lower = np.maximum(mean_tpr - std_tpr, 0)
ax.fill_between(
    mean_fpr,
    tprs_lower,
    tprs_upper,
    color="grey",
    alpha=0.2,
    label=r"$\pm$ 1 std. dev.",
)

ax.set(
    xlim=[-0.05, 1.05],
    ylim=[-0.05, 1.05],
    title="Receiver Operating Characteristic",
)
ax.legend(loc="lower right")
plt.show()
```



Good news! Our model on average performed much better than random (red dashed line) and it was also stable across folds (small variation from the mean ROC in blue). Therefore, we can now be confident that this model could also perform well on our test data. All we have to do is find out at this point.

1.4 Validation on the Test Set

Finally, it is time to use our test set and find out how predictive our model really is. First, we train the model on the entire training set:

[23]:
```
model = classifier.fit(X, y)
```

and then we make the predictions for our test data:

[24]:
```
X_test = df_test[df_test.columns[df_test.columns.str.startswith("Descriptor_")]]

predictions = classifier.predict(X_test)
predictions[0:10]
```

[24]:
```
array([False, False, False, False, False, False, False, False, False,
       False])
```

This gives us directly the labels the model thinks our test data should have, but we can also get a more fine-grained idea by extracting the probabilites as well:

[25]:
```
predictions_proba = classifier.predict_proba(X_test)[:,1]
predictions_proba[0:10]
```

[25]:
```
array([0.012, 0.088, 0.052, 0.02 , 0.072, 0.092, 0.036, 0.004, 0.068,
       0.104])
```

These are numbers between 0 and 1 that indicate how the model is confident about the compound being active, large values mean the model is fairly confident about the compound being an A2A receptor binder.

Finally, we compare these predictions with the true values from the test set:

[26]:
```
y_test = df_test.ActivityClass
y_test
```

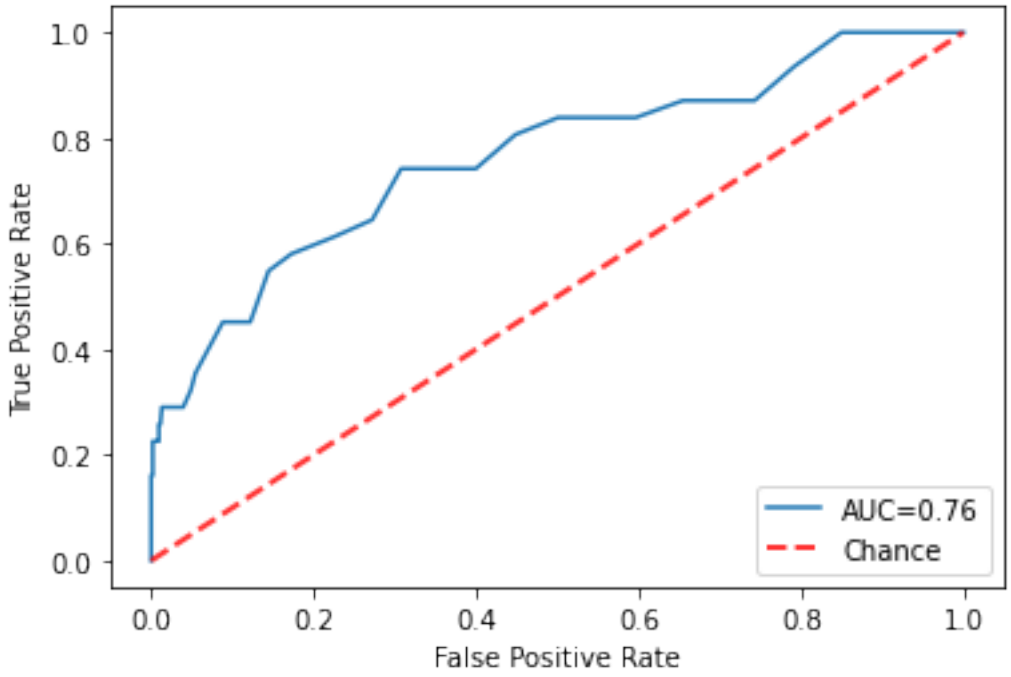
[26]:
```
2       False
3       False
6        True
21      False
26      False
        …
1751    False
1762    False
1766    False
1767    False
1768    False
Name: ActivityClass, Length: 564, dtype: bool
```

The probabilities effectively give us a ranking of the compounds from the most likely actives to the least likely and we can use them to draw an ROC curve like we saw during cross-validation:

[27]:
```
from sklearn.metrics import roc_curve, roc_auc_score

fpr, tpr, _ = roc_curve(y_test,  predictions_proba)
auc = roc_auc_score(y_test, predictions_proba)

plt.plot(fpr,tpr, label=f"AUC={auc:2.2}")
plt.plot([0, 1], [0, 1], linestyle="--", lw=2, color="r", label="Chance",␣
 ↪alpha=0.8)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```



Clearly, our model has difficulties in this data set and even though the AUC is reasonably high, the compounds ranked the highest (left part of the curve) are not always the actives. This is very common to see in virtual screening. The highest ranking compounds are not always the active ones, but we can still see the model gives us significant enrichment when compared with a random model that would draw compounds by chance (red line).

We can calculate many classification metrics this way and evaluate various aspects of our model. For example, we can directly use the predictions to calculate Matthew's correlation coefficient that gives us a balanced score of how the classifier can label the data:

[28]:
```
from sklearn.metrics import matthews_corrcoef

matthews_corrcoef(y_test, predictions)
```

[28]:
```
0.30322287021579997
```

The best score the model could obtain is 1 so even though this is not ideal, it is still not bad because the value of 0 would be a random model here. With this, let's deem our model fit for use in future endeavors and train the final version on the whole data set:

[29]:
```
X = dataset.getDescriptors()
y = dataset.getSubset(('ActivityClass',)).iloc[:,0]
```

[30]:
```
classifier.fit(X, y)
classifier
```

[30]:
```
ExtraTreesClassifier(n_estimators=250)
```

This is our final model so we might want to save it for future use:

[31]:
```
import joblib
import os

data_dir = 'data/qsar/models/'
data_file = os.path.join(data_dir, 'PTP1B_CLS_ET_250.pickle')
os.makedirs(data_dir, exist_ok=True)
joblib.dump(classifier, data_file)
```

[31]:
```
['data/qsar/models/PTP1B_CLS_ET_250.pickle']
```

This means we can easily load it whenever we need it in this tutorial, which will come right in the next section:

[32]:
```
model = joblib.load(data_file)
model
```

[32]:
```
ExtraTreesClassifier(n_estimators=250)
```

[ ]:
```

```