# 3_TYK2_Affinity_Prediction

April 8, 2022

### 0.0.1 Machine learning for TYK2-inhibitor affinity prediction with Scikit-Learn/DeepChem/RDKit

[1]:
```
import pandas as pd
import numpy as np
```

**Load ChEMBL bioactivity data**

[2]:
```
from rdkit import Chem
from rdkit.Chem.Draw import MolsToGridImage
from rdkit.Chem import PandasTools

chembl_bioactivity_df = pd.read_pickle('../data/chembl_bioactivity_data.pkl')
chembl_bioactivity_df.head(2)
```

[2]:
```
  molecule_chembl_id molecule_pref_name  \
0          CHEMBL10          SB-203580
1      CHEMBL1076700               None

                                  canonical_smiles pchembl_value  \
0  C[S+]([O-])c1ccc(-c2nc(-c3ccc(F)cc3)c(-c3ccncc…          5.70
1  Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          7.01

  standard_type standard_relation standard_value standard_units  \
0            Kd                =         2000.0             nM
1          IC50                =           97.0             nM

   potential_duplicate             target_pref_name target_organism  \
0                False  Tyrosine-protein kinase TYK2    Homo sapiens
1                False  Tyrosine-protein kinase TYK2    Homo sapiens

  assay_type                                   assay_description  \
0          B  Binding constant for TYK2(JH2domain-pseudokina…
1          B  Inhibition of GST-tagged TYK2 assessed as inhi…

   chembl_id_duplicate mean_pchembl_value max_pchembl_value min_pchembl_value  \
0                False               5.70              5.70              5.70
1                False               7.01              7.01              7.01
```

```
                                        core_smiles  \
0     c1ccc(-c2nc(-c3ccccc3)c(-c3ccncc3)[nH]2)cc1
1  c1ccc(Nc2nc3cccc(-c4ccc(CN5CCOCC5)cc4)c3o2)cc1

                                                 Mol  \
0  <img data-content="rdkit/molecule" src="data:i…
1  <img data-content="rdkit/molecule" src="data:i…

                                               Scaffold
0  <img data-content="rdkit/molecule" src="data:i…
1  <img data-content="rdkit/molecule" src="data:i…
```

[3]:
```
chembl_bioactivity_df.shape
```

[3]:
```
(1502, 20)
```

[4]:
```
from rdkit.Chem.SaltRemover import SaltRemover

def unsalt(smiles):
    remover = SaltRemover()
    #print(remover.salts)
    mol = Chem.MolFromSmiles(smiles)
    mol, deleted = remover.StripMolWithDeleted(mol)
    #print([Chem.MolToSmarts(s) for s in deleted])
    return Chem.MolToSmiles(mol, True)

chembl_bioactivity_ml_df = chembl_bioactivity_df[['molecule_chembl_id',␣
↪'canonical_smiles', 'mean_pchembl_value']].copy()

#remove salts
smiles = list(map(lambda i: unsalt(i),␣
↪list(chembl_bioactivity_ml_df['canonical_smiles'])))

chembl_bioactivity_ml_df['smiles']  = smiles
#chembl_bioactivity_ml_df.head()

mols = [Chem.MolFromSmiles(smi) for smi in chembl_bioactivity_ml_df['smiles']] ␣
↪#sanitize=True default
```

### Featurize the ChEMBL dataset

### 1. Use Molecular Descriptors

[5]:
```
import deepchem as dc
```

```
#if use_fragment = True, a total of 208 descriptors are returned to include␣
 ↪fragment binary descriptors like 'fr_'
md_featurizer = dc.feat.RDKitDescriptors(use_fragment = False)

features_md = md_featurizer.featurize(mols)
#features_md is a N x 123 array containing the 123 molecular␣
 ↪descriptors(physiochemical properties) for the 1502 molecules
print(features_md.shape)
features_md[:5]
```

```
(1502, 123)
```

[5]:
```
array([[ 1.33377069e+01, -1.02754413e+00,  1.33377069e+01,
         2.88772971e-01,  5.25196404e-01,  3.77444000e+02,
         3.61316000e+02,  3.77099811e+02,  1.34000000e+02,
         0.00000000e+00,  1.51945536e-01, -6.11646138e-01,
         6.11646138e-01,  1.51945536e-01,  8.51851852e-01,
         1.48148148e+00,  2.11111111e+00,  3.22277374e+01,
         1.00631031e+01,  2.15653172e+00, -2.03749157e+00,
         2.32663597e+00, -1.95909026e+00,  7.90297649e+00,
         6.00587178e-01,  1.83223649e+00,  1.04524911e+03,
         1.88027541e+01,  1.45170915e+01,  1.53335881e+01,
         1.31141935e+01,  8.31838406e+00,  9.87646226e+00,
         6.03030029e+00,  7.40995310e+00,  4.25014686e+00,
         5.18616699e+00,  2.85853311e+00,  3.37789251e+00,
        -2.96000000e+00,  3.14577780e+00,  1.74536410e+01,
         7.34321150e+00,  3.62549170e+00,  1.58876011e+02,
         9.53672839e+00,  1.78973945e+01,  4.89548348e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         4.98397852e+00,  9.37439357e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  7.18395252e+01,
         2.90840416e+01,  1.13878560e+01,  8.94316492e+00,
         1.11758545e+01,  0.00000000e+00,  1.49519356e+01,
         0.00000000e+00,  4.89548348e+00,  6.25576918e+00,
         7.88745787e+01,  0.00000000e+00,  3.39026150e+01,
         0.00000000e+00,  4.39041505e+00,  0.00000000e+00,
         0.00000000e+00,  2.57604546e+01,  1.11758545e+01,
         5.81722084e+00,  0.00000000e+00,  7.79528413e+01,
         0.00000000e+00,  3.39026150e+01,  0.00000000e+00,
         6.46300000e+01,  1.11758545e+01,  8.94316492e+00,
         0.00000000e+00,  5.81722084e+00,  5.82440450e+00,
         3.29736939e+01,  1.21327341e+01,  3.07821905e+01,
         3.63982024e+01,  9.96795704e+00,  4.98397852e+00,
         2.49377605e+01, -1.02754413e+00,  1.29504050e+01,
         0.00000000e+00,  4.20899648e+00,  4.00646120e-01,
         1.75005982e+01,  5.08469334e+00,  0.00000000e+00,
         0.00000000e+00,  4.76190476e-02,  2.70000000e+01,
```

```
  1.00000000e+00,  4.00000000e+00,  0.00000000e+00,
  0.00000000e+00,  0.00000000e+00,  2.00000000e+00,
  2.00000000e+00,  4.00000000e+00,  3.00000000e+00,
  1.00000000e+00,  6.00000000e+00,  4.00000000e+00,
  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
  4.00000000e+00,  4.68220000e+00,  1.04981100e+02],
[ 1.51034474e+01, -5.98462745e-01,  1.51034474e+01,
  4.51967446e-02,  2.52238129e-01,  5.63649000e+02,
  5.28369000e+02,  5.63270796e+02,  2.16000000e+02,
  0.00000000e+00,  2.99786715e-01, -4.22642085e-01,
  4.22642085e-01,  2.99786715e-01,  1.07317073e+00,
  1.82926829e+00,  2.53658537e+00,  1.91455400e+01,
  1.00132730e+01,  2.19718816e+00, -2.26499120e+00,
  2.33483101e+00, -2.35693517e+00,  5.95924950e+00,
  3.32130724e-02,  1.31876590e+00,  1.50871448e+03,
  2.89321462e+01,  2.35979710e+01,  2.35979710e+01,
  1.97788973e+01,  1.37680972e+01,  1.37680972e+01,
  1.05472377e+01,  1.05472377e+01,  7.14241148e+00,
  7.14241148e+00,  4.93437787e+00,  4.93437787e+00,
 -3.86000000e+00,  3.35680838e+00,  2.86608866e+01,
  1.27593639e+01,  6.89871761e+00,  2.37840237e+02,
  2.46875008e+01,  1.71511424e+01,  5.58302014e+00,
  0.00000000e+00,  1.19218384e+01,  0.00000000e+00,
  9.69444691e+00,  8.78083010e+00,  4.98397852e+00,
  0.00000000e+00,  1.21327341e+01,  8.15097646e+01,
  4.85567664e+01,  1.32137639e+01,  2.27293812e+01,
  2.87089456e+01,  0.00000000e+00,  2.01005866e+01,
  0.00000000e+00,  1.98893152e+01,  5.88049221e+01,
  7.68557327e+01,  0.00000000e+00,  1.11269030e+01,
  1.06335772e+01,  2.04828751e+01,  0.00000000e+00,
  0.00000000e+00,  7.41791112e+01,  1.12816194e+01,
  1.85581789e+01,  2.79057133e+01,  5.29480875e+01,
  0.00000000e+00,  2.22266238e+01,  0.00000000e+00,
  8.28700000e+01,  1.16344417e+01,  4.79453718e+00,
  8.78083010e+00,  2.40300463e+01,  6.63254948e+01,
  1.85290295e+01,  1.21327341e+01,  3.03318353e+01,
  3.19853580e+01,  2.05174655e+01,  9.15401389e+00,
  4.15557409e+01,  0.00000000e+00,  2.11749309e+01,
  6.10005208e+00,  4.02353642e+00, -1.31290876e+00,
  1.36282810e+01,  8.70041934e-01,  5.96057223e+00,
  3.99975318e+00,  3.54838710e-01,  4.10000000e+01,
  2.00000000e+00,  8.00000000e+00,  0.00000000e+00,
  1.00000000e+00,  1.00000000e+00,  3.00000000e+00,
  1.00000000e+00,  4.00000000e+00,  7.00000000e+00,
  2.00000000e+00,  1.00000000e+01,  1.00000000e+01,
  0.00000000e+00,  1.00000000e+00,  1.00000000e+00,
  5.00000000e+00,  5.33872000e+00,  1.55422900e+02],
```

```
[ 1.19266022e+01, -2.79009900e-01,  1.19266022e+01,
  1.88983118e-02,  5.97823060e-01,  4.14469000e+02,
  3.92293000e+02,  4.14180424e+02,  1.56000000e+02,
  0.00000000e+00,  2.51600507e-01, -3.77777820e-01,
  3.77777820e-01,  2.51600507e-01,  1.06451613e+00,
  1.80645161e+00,  2.51612903e+00,  1.64775009e+01,
  1.00876096e+01,  2.11079440e+00, -2.28455662e+00,
  2.22497632e+00, -2.35608077e+00,  5.94285074e+00,
  9.58227265e-02,  1.43352478e+00,  1.07047088e+03,
  2.14680446e+01,  1.69143876e+01,  1.69143876e+01,
  1.52246780e+01,  1.00120125e+01,  1.00120125e+01,
  6.95492423e+00,  6.95492423e+00,  4.92611334e+00,
  4.92611334e+00,  3.30016318e+00,  3.30016318e+00,
 -3.96000000e+00,  2.83701353e+00,  2.03183665e+01,
  9.69381941e+00,  5.15322643e+00,  1.80060856e+02,
  2.02703499e+01,  6.54475641e+00,  0.00000000e+00,
  5.94833928e+00,  5.90717973e+00,  0.00000000e+00,
  4.79453718e+00,  9.96795704e+00,  5.26189155e+00,
  0.00000000e+00,  1.21327341e+01,  4.24645695e+01,
  4.17880319e+01,  2.49769132e+01,  9.53140014e+00,
  2.32302916e+01,  5.26189155e+00,  1.52847456e+01,
  0.00000000e+00,  0.00000000e+00,  4.30647315e+01,
  6.63575987e+01,  0.00000000e+00,  1.73266008e+01,
  1.55334869e+01,  1.73231118e+01,  0.00000000e+00,
  0.00000000e+00,  4.87231699e+01,  4.73686295e+00,
  1.13311129e+01,  1.03579887e+01,  6.07941472e+01,
  0.00000000e+00,  1.12573795e+01,  0.00000000e+00,
  1.03170000e+02,  0.00000000e+00,  4.79453718e+00,
  0.00000000e+00,  1.24519361e+01,  1.15117908e+01,
  4.32480425e+01,  5.68738627e+00,  1.83295777e+01,
  3.64010567e+01,  3.76341781e+01,  9.99875451e+00,
  5.40334673e+00,  0.00000000e+00,  2.31168247e+01,
  1.43292960e+01,  4.17057192e+00,  2.14010718e-01,
  1.89455419e+01,  1.69474668e+00,  3.29232807e+00,
  0.00000000e+00,  2.17391304e-01,  3.10000000e+01,
  2.00000000e+00,  8.00000000e+00,  0.00000000e+00,
  1.00000000e+00,  1.00000000e+00,  2.00000000e+00,
  1.00000000e+00,  3.00000000e+00,  7.00000000e+00,
  2.00000000e+00,  8.00000000e+00,  6.00000000e+00,
  0.00000000e+00,  1.00000000e+00,  1.00000000e+00,
  4.00000000e+00,  2.97718000e+00,  1.18170900e+02],
[ 1.50379804e+01, -5.97967642e-01,  1.50379804e+01,
  4.56626816e-02,  3.79810132e-01,  5.06553000e+02,
  4.78329000e+02,  5.06212947e+02,  1.92000000e+02,
  0.00000000e+00,  2.99786715e-01, -4.22642085e-01,
  4.22642085e-01,  2.99786715e-01,  1.08108108e+00,
  1.81081081e+00,  2.48648649e+00,  1.91455375e+01,
```

```
  1.00132730e+01,  2.19750465e+00, -2.26499127e+00,
  2.33496758e+00, -2.35693515e+00,  5.95660885e+00,
  3.32130387e-02,  1.41900038e+00,  1.43833721e+03,
  2.61037190e+01,  2.09766507e+01,  2.09766507e+01,
  1.77957341e+01,  1.20719228e+01,  1.20719228e+01,
  9.33408734e+00,  9.33408734e+00,  6.46671380e+00,
  6.46671380e+00,  4.46804471e+00,  4.46804471e+00,
 -3.82000000e+00,  3.28606205e+00,  2.48559113e+01,
  1.02748353e+01,  5.18584901e+00,  2.13195549e+02,
  1.93707122e+01,  1.71511424e+01,  5.58302014e+00,
  0.00000000e+00,  1.19218384e+01,  0.00000000e+00,
  9.69444691e+00,  8.78083010e+00,  4.98397852e+00,
  0.00000000e+00,  1.21327341e+01,  5.44488426e+01,
  5.61073539e+01,  1.32137639e+01,  2.27293812e+01,
  2.87089456e+01,  0.00000000e+00,  1.47837980e+01,
  0.00000000e+00,  1.34684936e+01,  4.57154093e+01,
  7.68557327e+01,  0.00000000e+00,  1.11269030e+01,
  5.31678860e+00,  2.04828751e+01,  0.00000000e+00,
  0.00000000e+00,  6.10895984e+01,  1.12816194e+01,
  1.85581789e+01,  2.14848917e+01,  5.29480875e+01,
  0.00000000e+00,  2.22266238e+01,  0.00000000e+00,
  7.08400000e+01,  1.16344417e+01,  4.79453718e+00,
  8.78083010e+00,  2.40300463e+01,  5.97807383e+01,
  5.56345149e+00,  1.70326439e+01,  4.44271793e+01,
  1.78900140e+01,  1.03007671e+01,  9.15401389e+00,
  4.13745806e+01,  0.00000000e+00,  2.03005656e+01,
  3.11903302e+00,  4.04020477e+00, -1.27372978e+00,
  1.35714627e+01,  0.00000000e+00,  4.45269334e+00,
  3.41518973e+00,  2.85714286e-01,  3.70000000e+01,
  1.00000000e+00,  7.00000000e+00,  0.00000000e+00,
  1.00000000e+00,  1.00000000e+00,  3.00000000e+00,
  1.00000000e+00,  4.00000000e+00,  6.00000000e+00,
  1.00000000e+00,  9.00000000e+00,  6.00000000e+00,
  0.00000000e+00,  1.00000000e+00,  1.00000000e+00,
  5.00000000e+00,  5.35902000e+00,  1.37936200e+02],
[ 6.02601332e+00,  3.82087113e-01,  6.02601332e+00,
  3.82087113e-01,  4.94825964e-01,  3.76412000e+02,
  3.56252000e+02,  3.76142307e+02,  1.42000000e+02,
  0.00000000e+00,  2.99790840e-01, -4.92633529e-01,
  4.92633529e-01,  2.99790840e-01,  8.21428571e-01,
  1.53571429e+00,  2.17857143e+00,  1.65301435e+01,
  1.00517105e+01,  2.15601000e+00, -2.13518744e+00,
  2.40654304e+00, -1.97471675e+00,  5.90877001e+00,
  3.24463003e-01,  1.73404418e+00,  1.08118303e+03,
  1.95098609e+01,  1.58537094e+01,  1.58537094e+01,
  1.37282079e+01,  8.75198027e+00,  8.75198027e+00,
  6.07788161e+00,  6.07788161e+00,  4.41577518e+00,
```

```
 4.41577518e+00,  3.12426186e+00,  3.12426186e+00,
-3.67000000e+00,  3.15463014e+00,  1.77293300e+01,
 7.52045803e+00,  3.20177590e+00,  1.62226558e+02,
 2.39445284e+01,  5.51670072e+00,  1.70820438e+01,
 5.74951183e+00,  6.01465871e+00,  0.00000000e+00,
 0.00000000e+00,  0.00000000e+00,  4.98397852e+00,
 0.00000000e+00,  4.24645695e+01,  1.16298186e+01,
 2.33835719e+01,  2.13293926e+01,  1.86277398e+01,
 2.28017658e+01,  0.00000000e+00,  4.98397852e+00,
 0.00000000e+00,  0.00000000e+00,  2.66461812e+01,
 6.06636707e+01,  0.00000000e+00,  2.83754385e+01,
 1.95273775e+01,  1.17020450e+01,  1.72485355e+01,
 0.00000000e+00,  2.63133711e+01,  0.00000000e+00,
 0.00000000e+00,  0.00000000e+00,  6.50808216e+01,
 0.00000000e+00,  2.22266238e+01,  0.00000000e+00,
 6.57500000e+01,  0.00000000e+00,  0.00000000e+00,
 0.00000000e+00,  0.00000000e+00,  2.89505805e+01,
 2.22266238e+01,  0.00000000e+00,  3.34621268e+01,
 4.85309365e+01,  1.03007671e+01,  1.86277398e+01,
 2.21792698e+01,  0.00000000e+00,  4.55708278e+00,
 3.17944712e+00,  4.27257556e+00,  1.62069485e+00,
 1.99700100e+01,  0.00000000e+00,  0.00000000e+00,
 4.72091994e+00,  1.36363636e-01,  2.80000000e+01,
 1.00000000e+00,  6.00000000e+00,  0.00000000e+00,
 0.00000000e+00,  0.00000000e+00,  3.00000000e+00,
 1.00000000e+00,  4.00000000e+00,  6.00000000e+00,
 1.00000000e+00,  6.00000000e+00,  6.00000000e+00,
 0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
 4.00000000e+00,  5.26420000e+00,  1.08825700e+02]])
```

[6]:
```
#from rdkit.ML.Descriptors import MoleculeDescriptors

#calculated rdkit descriptors
descriptors = []
descList = []
from rdkit.Chem import Descriptors
for descriptor, function in Descriptors.descList:
    if descriptor.startswith('fr_'):
        continue
    descriptors.append(descriptor)
    descList.append((descriptor, function))
print(descriptors)
print(len(descriptors))
```

```
['MaxEStateIndex', 'MinEStateIndex', 'MaxAbsEStateIndex', 'MinAbsEStateIndex',
'qed', 'MolWt', 'HeavyAtomMolWt', 'ExactMolWt', 'NumValenceElectrons',
'NumRadicalElectrons', 'MaxPartialCharge', 'MinPartialCharge',
```

```
'MaxAbsPartialCharge', 'MinAbsPartialCharge', 'FpDensityMorgan1',
'FpDensityMorgan2', 'FpDensityMorgan3', 'BCUT2D_MWHI', 'BCUT2D_MWLOW',
'BCUT2D_CHGHI', 'BCUT2D_CHGLO', 'BCUT2D_LOGPHI', 'BCUT2D_LOGPLOW',
'BCUT2D_MRHI', 'BCUT2D_MRLOW', 'BalabanJ', 'BertzCT', 'Chi0', 'Chi0n', 'Chi0v',
'Chi1', 'Chi1n', 'Chi1v', 'Chi2n', 'Chi2v', 'Chi3n', 'Chi3v', 'Chi4n', 'Chi4v',
'HallKierAlpha', 'Ipc', 'Kappa1', 'Kappa2', 'Kappa3', 'LabuteASA', 'PEOE_VSA1',
'PEOE_VSA10', 'PEOE_VSA11', 'PEOE_VSA12', 'PEOE_VSA13', 'PEOE_VSA14',
'PEOE_VSA2', 'PEOE_VSA3', 'PEOE_VSA4', 'PEOE_VSA5', 'PEOE_VSA6', 'PEOE_VSA7',
'PEOE_VSA8', 'PEOE_VSA9', 'SMR_VSA1', 'SMR_VSA10', 'SMR_VSA2', 'SMR_VSA3',
'SMR_VSA4', 'SMR_VSA5', 'SMR_VSA6', 'SMR_VSA7', 'SMR_VSA8', 'SMR_VSA9',
'SlogP_VSA1', 'SlogP_VSA10', 'SlogP_VSA11', 'SlogP_VSA12', 'SlogP_VSA2',
'SlogP_VSA3', 'SlogP_VSA4', 'SlogP_VSA5', 'SlogP_VSA6', 'SlogP_VSA7',
'SlogP_VSA8', 'SlogP_VSA9', 'TPSA', 'EState_VSA1', 'EState_VSA10',
'EState_VSA11', 'EState_VSA2', 'EState_VSA3', 'EState_VSA4', 'EState_VSA5',
'EState_VSA6', 'EState_VSA7', 'EState_VSA8', 'EState_VSA9', 'VSA_EState1',
'VSA_EState10', 'VSA_EState2', 'VSA_EState3', 'VSA_EState4', 'VSA_EState5',
'VSA_EState6', 'VSA_EState7', 'VSA_EState8', 'VSA_EState9', 'FractionCSP3',
'HeavyAtomCount', 'NHOHCount', 'NOCount', 'NumAliphaticCarbocycles',
'NumAliphaticHeterocycles', 'NumAliphaticRings', 'NumAromaticCarbocycles',
'NumAromaticHeterocycles', 'NumAromaticRings', 'NumHAcceptors', 'NumHDonors',
'NumHeteroatoms', 'NumRotatableBonds', 'NumSaturatedCarbocycles',
'NumSaturatedHeterocycles', 'NumSaturatedRings', 'RingCount', 'MolLogP',
'MolMR']
123
```

### 2. Use Fingerprints

[7]:
```
fp_featurizer = dc.feat.CircularFingerprint(size=2048)

features_fp = fp_featurizer.featurize(mols)
#features_fp is a N x 2048 array containing the fingerprints for the 1502
 ↪molecules
print(features_fp.shape)
features_fp[:5]
```

```
(1502, 2048)
```

[7]:
```
array([[0., 0., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 0.],
       [0., 0., 0., …, 0., 0., 0.]])
```

### 3. Use Graph Convolutions

[8]:
```
gc_featurizer = dc.feat.ConvMolFeaturizer()
features_graphs = gc_featurizer.featurize(mols)
features_graphs
```

[8]:
```
array([<deepchem.feat.mol_graphs.ConvMol object at 0x7f854b6b5e50>,
       <deepchem.feat.mol_graphs.ConvMol object at 0x7f85c01bdd90>,
       <deepchem.feat.mol_graphs.ConvMol object at 0x7f8622e84890>, …,
       <deepchem.feat.mol_graphs.ConvMol object at 0x7f85410c3750>,
       <deepchem.feat.mol_graphs.ConvMol object at 0x7f85410c3650>,
       <deepchem.feat.mol_graphs.ConvMol object at 0x7f85410c3510>],
      dtype=object)
```

### Dataset preparation

[9]:
```
features = features_fp
labels = chembl_bioactivity_ml_df['mean_pchembl_value']
ids = chembl_bioactivity_ml_df['molecule_chembl_id']

dataset = dc.data.NumpyDataset(X=features, y=labels, ids=ids)

train_dataset, test_dataset = dc.splits.RandomSplitter().
 ↪train_test_split(dataset, seed=42)
```

[10]:
```
train_dataset.get_shape()
```

[10]:
```
((1201, 2048), (1201,), (1201,), (1201,))
```

[11]:
```
test_dataset.get_shape()
```

[11]:
```
((301, 2048), (301,), (301,), (301,))
```

### RandomForestRegressor model

[12]:
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import r2_score
```

[13]:
```
seed = 42
rf_model = RandomForestRegressor()
rf_model.random_state = seed

param_grid = {'oob_score': [True], 'n_estimators':[50, 100, 150, 200, 250]}

grid_search = GridSearchCV(rf_model, param_grid, cv=10, verbose = 1, refit =␣
 ↪True, return_train_score=True, n_jobs = -2)

grid_search.fit(train_dataset.X, train_dataset.y)
```

```
Fitting 10 folds for each of 5 candidates, totalling 50 fits
```

[13]:
```
GridSearchCV(cv=10, estimator=RandomForestRegressor(random_state=42), n_jobs=-2,
             param_grid={'n_estimators': [50, 100, 150, 200, 250],
```

```
                                'oob_score': [True]},
                   return_train_score=True, verbose=1)
```

[14]:
```
grid_search.best_params_
```

[14]:
```
{'n_estimators': 250, 'oob_score': True}
```

[15]:
```
grid_search.best_estimator_
```

[15]:
```
RandomForestRegressor(n_estimators=250, oob_score=True, random_state=42)
```

[16]:
```
grid_search.cv_results_
```

[16]:
```
{'mean_fit_time': array([12.75109618, 25.94488497, 38.75033133, 49.92226384,
57.27423413]),
 'std_fit_time': array([ 1.73536555,  3.20697271,  5.44109877,  7.77227467,
12.86719182]),
 'mean_score_time': array([0.01463439, 0.02689342, 0.03929541, 0.04694223,
0.04528439]),
 'std_score_time': array([0.0022877 , 0.00619353, 0.00864456, 0.01158299,
0.017581  ]),
 'param_n_estimators': masked_array(data=[50, 100, 150, 200, 250],
              mask=[False, False, False, False, False],
        fill_value='?',
             dtype=object),
 'param_oob_score': masked_array(data=[True, True, True, True, True],
              mask=[False, False, False, False, False],
        fill_value='?',
             dtype=object),
 'params': [{'n_estimators': 50, 'oob_score': True},
  {'n_estimators': 100, 'oob_score': True},
  {'n_estimators': 150, 'oob_score': True},
  {'n_estimators': 200, 'oob_score': True},
  {'n_estimators': 250, 'oob_score': True}],
 'split0_test_score': array([0.58432508, 0.59453371, 0.58769327, 0.58698311,
0.58899255]),
 'split1_test_score': array([0.38778558, 0.38354172, 0.38681527, 0.39068304,
0.39093015]),
 'split2_test_score': array([0.511371  , 0.49364357, 0.49402749, 0.49352047,
0.49084331]),
 'split3_test_score': array([0.47914429, 0.47549716, 0.48155062, 0.48529283,
0.48874437]),
 'split4_test_score': array([0.50258585, 0.5195911 , 0.5168272 , 0.51829402,
0.52366279]),
 'split5_test_score': array([0.50238014, 0.5102122 , 0.51291508, 0.50924557,
0.51111861]),
 'split6_test_score': array([0.61784076, 0.61147565, 0.60588248, 0.60720124,
```

```
0.60631674]),
 'split7_test_score': array([0.61595827, 0.60716963, 0.60321666, 0.60552791,
0.60382072]),
 'split8_test_score': array([0.42540469, 0.43449031, 0.44526489, 0.45274896,
0.44869127]),
 'split9_test_score': array([0.45566308, 0.46662133, 0.47157246, 0.47300297,
0.47216653]),
 'mean_test_score': array([0.50824587, 0.50967764, 0.51057654, 0.51225001,
0.5125287 ]),
 'std_test_score': array([0.07379868, 0.07211746, 0.06773264, 0.06654608,
0.06680538]),
 'rank_test_score': array([5, 4, 3, 2, 1], dtype=int32),
 'split0_train_score': array([0.91670663, 0.91771217, 0.91922965, 0.91972257,
0.92078595]),
 'split1_train_score': array([0.92103205, 0.92251799, 0.92345077, 0.92329727,
0.92318562]),
 'split2_train_score': array([0.91567015, 0.91858278, 0.91919125, 0.91962854,
0.92038839]),
 'split3_train_score': array([0.91463003, 0.91892353, 0.91994069, 0.91990314,
0.92025276]),
 'split4_train_score': array([0.91577841, 0.91895397, 0.91973924, 0.91987737,
0.92046801]),
 'split5_train_score': array([0.91084113, 0.91540844, 0.91661181, 0.91700453,
0.91782235]),
 'split6_train_score': array([0.9134142 , 0.91651309, 0.91803527, 0.91895055,
0.91950787]),
 'split7_train_score': array([0.91019995, 0.9131686 , 0.91422691, 0.91485597,
0.91500476]),
 'split8_train_score': array([0.91240486, 0.91650666, 0.91729473, 0.91782821,
0.91848015]),
 'split9_train_score': array([0.91749334, 0.91878353, 0.91965971, 0.92015228,
0.92047063]),
 'mean_train_score': array([0.91481708, 0.91770708, 0.918738  , 0.91912204,
0.91963665]),
 'std_train_score': array([0.00309772, 0.00238439, 0.00231092, 0.00211372,
0.00205785])}
```

[17]:
```
grid_search.best_score_
```

[17]:
```
0.5125287043201121
```

[18]:
```
y_pred_train = grid_search.predict(train_dataset.X)
y_pred_test = grid_search.predict(test_dataset.X)
```

[19]:
```
R2_cv_train = r2_score(train_dataset.y, y_pred_train)
R2_cv_test = r2_score(test_dataset.y, y_pred_test)
```

```
print("RF Train set R2 %f" % R2_cv_train)
print("RF Test set R2 %f" % R2_cv_test)
```

```
RF Train set R2 0.919201
RF Test set R2 0.540480
```

[21]:
```
import deepchem as dc
print("DeepChem: ", dc.__version__)

#deepchem is enabled by/running on TensorFlow GPU platform
import tensorflow as tf
print("TensorFlow: ", tf.__version__)
print("GPUs available: ", tf.config.list_physical_devices('GPU'))

import sklearn
print("Scikit-Learn: ", sklearn.__version__)

import rdkit
print("RDKit: ", rdkit.__version__)

from platform import python_version
print("Python: ", python_version())
print("Numpy: ", np.__version__)
print("Pandas: ", pd.__version__)
```

```
DeepChem:  2.5.0
TensorFlow:  2.4.1
GPUs available:  [PhysicalDevice(name='/physical_device:GPU:0',
device_type='GPU')]
Scikit-Learn:  0.24.2
RDKit:  2021.03.1
Python:  3.7.10
Numpy:  1.19.5
Pandas:  1.2.4
```

[22]:
```
! conda list
```

```
# packages in environment at /home/cv/anaconda3/envs/deepchem:
#
# Name                    Version                   Build  Channel
_libgcc_mutex             0.1                 conda_forge    conda-forge
_openmp_mutex             4.5                       1_gnu    conda-forge
absl-py                   0.12.0                    pypi_0    pypi
argon2-cffi               20.1.0                    pypi_0    pypi
astunparse                1.6.3                     pypi_0    pypi
async-generator           1.10                      pypi_0    pypi
attrs                     21.1.0                    pypi_0    pypi
backcall                  0.2.0                     pypi_0    pypi
```

| | | | |
|---|---|---|---|
| bleach | 3.3.0 | pypi_0 | pypi |
| boost | 1.74.0 | py37h6dcda5c_3 | conda-forge |
| boost-cpp | 1.74.0 | hc6e9bd1_2 | conda-forge |
| bzip2 | 1.0.8 | h7f98852_4 | conda-forge |
| ca-certificates | 2020.12.5 | ha878542_0 | conda-forge |
| cachetools | 4.2.2 | pypi_0 | pypi |
| cairo | 1.16.0 | h6cf1ce9_1008 | conda-forge |
| certifi | 2020.12.5 | py37h89c1867_1 | conda-forge |
| cffi | 1.14.5 | pypi_0 | pypi |
| chardet | 4.0.0 | pypi_0 | pypi |
| chembl-webresource-client | 0.10.4 | pypi_0 | pypi |
| cudatoolkit | 11.0.221 | h6bb024c_0 | |
| cudnn | 8.1.0.77 | h90431f1_0 | conda-forge |
| cycler | 0.10.0 | py_2 | conda-forge |
| decorator | 5.0.7 | pypi_0 | pypi |
| deepchem | 2.5.0 | pyhd8ed1ab_0 | conda-forge |
| defusedxml | 0.7.1 | pypi_0 | pypi |
| docopt | 0.6.2 | pypi_0 | pypi |
| easydict | 1.9 | pypi_0 | pypi |
| entrypoints | 0.3 | pypi_0 | pypi |
| flatbuffers | 1.12 | pypi_0 | pypi |
| fontconfig | 2.13.1 | hba837de_1005 | conda-forge |
| freetype | 2.10.4 | h0708190_1 | conda-forge |
| gast | 0.3.3 | pypi_0 | pypi |
| gettext | 0.19.8.1 | h0b5b191_1005 | conda-forge |
| google-auth | 1.30.0 | pypi_0 | pypi |
| google-auth-oauthlib | 0.4.4 | pypi_0 | pypi |
| google-pasta | 0.2.0 | pypi_0 | pypi |
| greenlet | 1.1.0 | py37hcd2ae1e_0 | conda-forge |
| grpcio | 1.32.0 | pypi_0 | pypi |
| h5py | 2.10.0 | pypi_0 | pypi |
| icu | 68.1 | h58526e2_0 | conda-forge |
| idna | 2.10 | pypi_0 | pypi |
| importlib-metadata | 4.0.1 | py37h89c1867_0 | conda-forge |
| ipykernel | 5.5.4 | pypi_0 | pypi |
| ipython | 7.23.1 | pypi_0 | pypi |
| ipython-genutils | 0.2.0 | pypi_0 | pypi |
| ipywidgets | 7.6.3 | pypi_0 | pypi |
| itsdangerous | 2.0.1 | pypi_0 | pypi |
| jedi | 0.18.0 | pypi_0 | pypi |
| jinja2 | 2.11.3 | pypi_0 | pypi |
| joblib | 1.0.1 | pyhd8ed1ab_0 | conda-forge |
| jpeg | 9d | h36c2ea0_0 | conda-forge |
| jsonschema | 3.2.0 | pypi_0 | pypi |
| jupyter | 1.0.0 | pypi_0 | pypi |
| jupyter-client | 6.1.12 | pypi_0 | pypi |
| jupyter-console | 6.4.0 | pypi_0 | pypi |
| jupyter-core | 4.7.1 | pypi_0 | pypi |

```
jupyterlab-pygments       0.1.2                    pypi_0    pypi
jupyterlab-widgets        1.0.0                    pypi_0    pypi
keras-preprocessing       1.1.2                    pypi_0    pypi
kiwisolver                1.3.1           py37h2527ec5_1    conda-forge
lcms2                     2.12                hddcbb42_0    conda-forge
ld_impl_linux-64          2.33.1              h53a641e_7
libblas                   3.9.0               9_openblas    conda-forge
libcblas                  3.9.0               9_openblas    conda-forge
libffi                    3.3                 he6710b0_2
libgcc-ng                 9.3.0              h2828fa1_19    conda-forge
libgfortran-ng            9.3.0              hff62375_19    conda-forge
libgfortran5              9.3.0              hff62375_19    conda-forge
libglib                   2.68.1              h3e27bee_0    conda-forge
libgomp                   9.3.0              h2828fa1_19    conda-forge
libiconv                  1.16                h516909a_0    conda-forge
liblapack                 3.9.0               9_openblas    conda-forge
libopenblas               0.3.15        pthreads_h8fe5266_0    conda-forge
libpng                    1.6.37              h21135ba_2    conda-forge
libstdcxx-ng              9.3.0              h6de172a_19    conda-forge
libtiff                   4.2.0               hdc55705_1    conda-forge
libuuid                   2.32.1            h7f98852_1000    conda-forge
libwebp-base              1.2.0               h7f98852_2    conda-forge
libxcb                    1.13              h7f98852_1003    conda-forge
libxml2                   2.9.10              h72842e0_4    conda-forge
lz4-c                     1.9.3               h9c3ff4c_0    conda-forge
markdown                  3.3.4                    pypi_0    pypi
markupsafe                1.1.1                    pypi_0    pypi
matplotlib-base           3.4.1           py37hdd32ed1_0    conda-forge
matplotlib-inline         0.1.2                    pypi_0    pypi
mdtraj                    1.9.5                    pypi_0    pypi
mistune                   0.8.4                    pypi_0    pypi
nbclient                  0.5.3                    pypi_0    pypi
nbconvert                 6.0.7                    pypi_0    pypi
nbformat                  5.1.3                    pypi_0    pypi
ncurses                   6.2                 he6710b0_1
nest-asyncio              1.5.1                    pypi_0    pypi
nglview                   3.0.1                    pypi_0    pypi
notebook                  6.3.0                    pypi_0    pypi
numpy                     1.19.5                   pypi_0    pypi
oauthlib                  3.1.0                    pypi_0    pypi
olefile                   0.46              pyh9f0ad1d_1    conda-forge
openjpeg                  2.4.0               hf7af979_0    conda-forge
openssl                   1.1.1k              h7f98852_0    conda-forge
opt-einsum                3.3.0                    pypi_0    pypi
packaging                 20.9                     pypi_0    pypi
pandas                    1.2.4           py37h219a48f_0    conda-forge
pandocfilters             1.4.3                    pypi_0    pypi
parso                     0.8.2                    pypi_0    pypi
```

```
pcre                      8.44                he1b5a44_0    conda-forge
pexpect                   4.8.0                   pypi_0    pypi
pickleshare               0.7.5                   pypi_0    pypi
pillow                    8.1.2             py37h4600e1f_1    conda-forge
pip                       21.0.1            py37h06a4308_0
pixman                    0.40.0              h36c2ea0_0    conda-forge
prometheus-client         0.10.1                  pypi_0    pypi
prompt-toolkit            3.0.18                  pypi_0    pypi
protobuf                  3.15.8                  pypi_0    pypi
pthread-stubs             0.4               h36c2ea0_1001    conda-forge
ptyprocess                0.7.0                   pypi_0    pypi
pyasn1                    0.4.8                   pypi_0    pypi
pyasn1-modules            0.2.8                   pypi_0    pypi
pycairo                   1.20.0            py37hfff247e_1    conda-forge
pycparser                 2.20                    pypi_0    pypi
pygments                  2.9.0                   pypi_0    pypi
pyparsing                 2.4.7               pyh9f0ad1d_0    conda-forge
pyrsistent                0.17.3                  pypi_0    pypi
python                    3.7.10                hdb3f193_0
python-dateutil           2.8.1                     py_0    conda-forge
python_abi                3.7                   1_cp37m    conda-forge
pytz                      2021.1              pyhd8ed1ab_0    conda-forge
pyzmq                     22.0.3                  pypi_0    pypi
qtconsole                 5.1.0                   pypi_0    pypi
qtpy                      1.9.0                   pypi_0    pypi
rd-filters                0.1                     pypi_0    pypi
rdkit                     2021.03.1         py37haf5a968_0    conda-forge
readline                  8.1                 h27cfd23_0
reportlab                 3.5.67            py37h69800bb_0    conda-forge
requests                  2.25.1                  pypi_0    pypi
requests-cache            0.6.4                   pypi_0    pypi
requests-oauthlib         1.3.0                   pypi_0    pypi
rsa                       4.7.2                   pypi_0    pypi
scikit-learn              0.24.2            py37h18a542f_0    conda-forge
scipy                     1.6.3             py37h29e03ee_0    conda-forge
seaborn                   0.11.1                  pypi_0    pypi
send2trash                1.5.0                   pypi_0    pypi
setuptools                52.0.0            py37h06a4308_0
six                       1.15.0            py37h06a4308_0
sqlalchemy                1.4.13            py37h5e8e339_0    conda-forge
sqlite                    3.35.4              hdfb4753_0
tensorboard               2.5.0                   pypi_0    pypi
tensorboard-data-server   0.6.1                   pypi_0    pypi
tensorboard-plugin-wit    1.8.0                   pypi_0    pypi
tensorflow-estimator      2.4.0                   pypi_0    pypi
tensorflow-gpu            2.4.1                   pypi_0    pypi
termcolor                 1.1.0                   pypi_0    pypi
terminado                 0.9.4                   pypi_0    pypi
```

```
testpath                  0.4.4                    pypi_0    pypi
threadpoolctl             2.1.0              pyh5ca1d4c_0    conda-forge
tk                        8.6.10               hbc83047_0
tornado                   6.1               py37h5e8e339_1    conda-forge
traitlets                 5.0.5                    pypi_0    pypi
typing_extensions         3.7.4.3                    py_0    conda-forge
url-normalize             1.4.3                    pypi_0    pypi
urllib3                   1.26.4                   pypi_0    pypi
wcwidth                   0.2.5                    pypi_0    pypi
webencodings              0.5.1                    pypi_0    pypi
werkzeug                  1.0.1                    pypi_0    pypi
wheel                     0.36.2             pyhd3eb1b0_0
widgetsnbextension        3.5.1                    pypi_0    pypi
wrapt                     1.12.1                   pypi_0    pypi
xorg-kbproto              1.0.7              h7f98852_1002    conda-forge
xorg-libice               1.0.10               h7f98852_0    conda-forge
xorg-libsm                1.2.3              hd9c2040_1000    conda-forge
xorg-libx11               1.7.0                h7f98852_0    conda-forge
xorg-libxau               1.0.9                h7f98852_0    conda-forge
xorg-libxdmcp             1.1.3                h7f98852_0    conda-forge
xorg-libxext              1.3.4                h7f98852_1    conda-forge
xorg-libxrender           0.9.10             h7f98852_1003    conda-forge
xorg-renderproto          0.11.1             h7f98852_1002    conda-forge
xorg-xextproto            7.3.0              h7f98852_1002    conda-forge
xorg-xproto               7.0.31             h7f98852_1007    conda-forge
xz                        5.2.5                h7b6447c_0
zipp                      3.4.1              pyhd8ed1ab_0    conda-forge
zlib                      1.2.11               h7b6447c_3
zstd                      1.4.9                ha95c52a_0    conda-forge
```

[ ]: