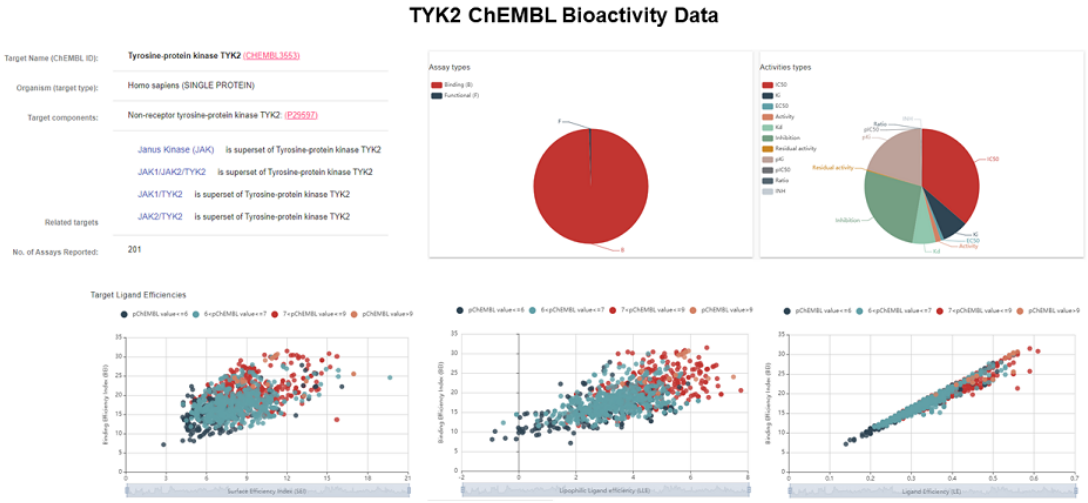
2_TYK2_Chembl_Bioactivity_Data

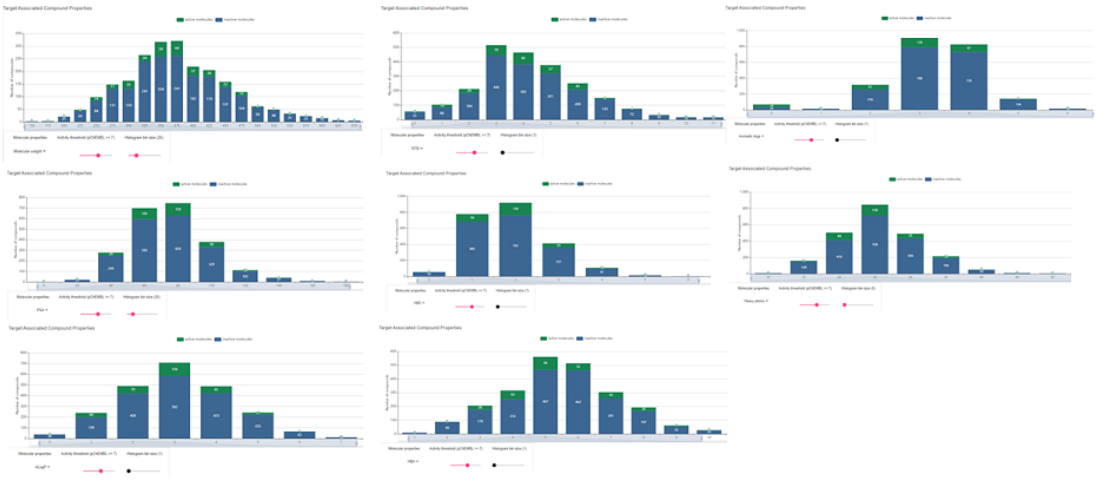
April 8, 2022

[1]:
```
from chembl_webresource_client.settings import Settings
from chembl_webresource_client.new_client import new_client

import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
activities = new_client.activity
target = 'CHEMBL3553'

tyk2_activities = activities.filter(target_chembl_id=target,␣
 ↪pchembl_value__isnull=False)
print(f"{len(tyk2_activities)} molecules for {target}")
```

```
1857 molecules for CHEMBL3553
```





[3]:
```
tyk2_activities_df = pd.DataFrame.from_dict(tyk2_activities)
tyk2_activities_df.head()
```

[3]:

	activity_comment	activity_id	activity_properties	assay_chembl_id	assay_description	assay_type	assay_variant_accession	assay_variant_mutation	bao_endpoint	bao_format	…	target_organism	target_pref_name	target_tax_id	text_value	toid	type	units	uo_units	upper_value	value
0	None	506158	[]	CHEMBL819922	Inhibition of Tyrosine kinase 2 kinase	B	None	None	BAO_0000190	BAO_0000357	…	Homo sapiens	Tyrosine-protein kinase TYK2	9606	None	None	IC50	uM			
1	None	1765466	[]	CHEMBL871539	Inhibition of Tyk2 by HTRF kinase assay	B	None	None	BAO_0000190	BAO_0000357	…	Homo sapiens	Tyrosine-protein kinase TYK2	9606	None	None	IC50	uM			
2	None	1765494	[]	CHEMBL871539	Inhibition of Tyk2 by HTRF kinase assay	B	None	None	BAO_0000190	BAO_0000357	…	Homo sapiens	Tyrosine-protein kinase TYK2	9606	None	None	IC50	uM			
3	None	2112048	[]	CHEMBL934199	Inhibition of Tyk2 by HTRF assay	B	None	None	BAO_0000190	BAO_0000357	…	Homo sapiens	Tyrosine-protein kinase TYK2	9606	None	None	IC50	uM			
4	None	2137279	[]	CHEMBL937052	Inhibition of Tyk2	B	None	None	BAO_0000190	BAO_0000357	…	Homo sapiens	Tyrosine-protein kinase TYK2	9606	None	None	IC50	nM			

```
0  UO_0000065        None  0.001
1  UO_0000065        None   2.73
2  UO_0000065        None   1.83
3  UO_0000065        None    1.2
4  UO_0000065        None    1.0

[5 rows x 45 columns]
```

[4]:
```
tyk2_activities
```

[4]:

```
[{'activity_comment': None, 'activity_id': 506158, 'activity_properties': [],
'assay_chembl_id': 'CHEMBL819922', 'assay_description': 'Inhibition of Tyrosine
kinase 2 kinase', 'assay_type': 'B', 'assay_variant_accession': None,
'assay_variant_mutation': None, 'bao_endpoint': 'BAO_0000190', 'bao_format':
'BAO_0000357', 'bao_label': 'single protein format', 'canonical_smiles':
'CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1', 'data_validity_comment': None,
'data_validity_description': None, 'document_chembl_id': 'CHEMBL1135866',
'document_journal': 'Bioorg. Med. Chem. Lett.', 'document_year': 2002,
'ligand_efficiency': {'bei': '29.09', 'le': '0.53', 'lle': '5.01', 'sei':
'14.62'}, 'molecule_chembl_id': 'CHEMBL21156', 'molecule_pref_name': None,
'parent_molecule_chembl_id': 'CHEMBL21156', 'pchembl_value': '9.00',
'potential_duplicate': True, 'qudt_units':
'http://www.openphacts.org/units/Nanomolar', 'record_id': 25954, 'relation':
'=', 'src_id': 1, 'standard_flag': True, 'standard_relation': '=',
'standard_text_value': None, 'standard_type': 'IC50', 'standard_units': 'nM',
'standard_upper_value': None, 'standard_value': '1.0', 'target_chembl_id':
'CHEMBL3553', 'target_organism': 'Homo sapiens', 'target_pref_name': 'Tyrosine-
protein kinase TYK2', 'target_tax_id': '9606', 'text_value': None, 'toid': None,
'type': 'IC50', 'units': 'uM', 'uo_units': 'UO_0000065', 'upper_value': None,
'value': '0.001'}, {'activity_comment': None, 'activity_id': 1765466,
'activity_properties': [], 'assay_chembl_id': 'CHEMBL871539',
'assay_description': 'Inhibition of Tyk2 by HTRF kinase assay', 'assay_type':
'B', 'assay_variant_accession': None, 'assay_variant_mutation': None,
'bao_endpoint': 'BAO_0000190', 'bao_format': 'BAO_0000357', 'bao_label': 'single
protein format', 'canonical_smiles':
'CNc1ncc2cc(-c3cc(C(=O)Nc4cc(C(F)(F)F)ccc4OC4CCN(C)CC4)ccc3C)ccc2n1',
'data_validity_comment': None, 'data_validity_description': None,
'document_chembl_id': 'CHEMBL1149344', 'document_journal': 'J. Med. Chem.',
'document_year': 2006, 'ligand_efficiency': {'bei': '10.12', 'le': '0.19',
'lle': '-0.83', 'sei': '7.01'}, 'molecule_chembl_id': 'CHEMBL386661',
'molecule_pref_name': None, 'parent_molecule_chembl_id': 'CHEMBL386661',
'pchembl_value': '5.56', 'potential_duplicate': False, 'qudt_units':
'http://www.openphacts.org/units/Nanomolar', 'record_id': 565637, 'relation':
'=', 'src_id': 1, 'standard_flag': True, 'standard_relation': '=',
'standard_text_value': None, 'standard_type': 'IC50', 'standard_units': 'nM',
'standard_upper_value': None, 'standard_value': '2730.0', 'target_chembl_id':
'CHEMBL3553', 'target_organism': 'Homo sapiens', 'target_pref_name': 'Tyrosine-
```

```
protein kinase TYK2', 'target_tax_id': '9606', 'text_value': None, 'toid': None,
'type': 'IC50', 'units': 'uM', 'uo_units': 'UO_0000065', 'upper_value': None,
'value': '2.73'}, {'activity_comment': None, 'activity_id': 1765494,
'activity_properties': [], 'assay_chembl_id': 'CHEMBL871539',
'assay_description': 'Inhibition of Tyk2 by HTRF kinase assay', 'assay_type':
'B', 'assay_variant_accession': None, 'assay_variant_mutation': None,
'bao_endpoint': 'BAO_0000190', 'bao_format': 'BAO_0000357', 'bao_label': 'single
protein format', 'canonical_smiles':
'Cc1ccc(C(=O)Nc2cccc(C(F)(F)F)c2)cc1-c1ccc2nc(NCCN3CCOCC3)ncc2c1',
'data_validity_comment': None, 'data_validity_description': None,
'document_chembl_id': 'CHEMBL1149344', 'document_journal': 'J. Med. Chem.',
'document_year': 2006, 'ligand_efficiency': {'bei': '10.71', 'le': '0.20',
'lle': '0.12', 'sei': '7.23'}, 'molecule_chembl_id': 'CHEMBL215943',
'molecule_pref_name': None, 'parent_molecule_chembl_id': 'CHEMBL215943',
'pchembl_value': '5.74', 'potential_duplicate': False, 'qudt_units':
'http://www.openphacts.org/units/Nanomolar', 'record_id': 565656, 'relation':
'=', 'src_id': 1, 'standard_flag': True, 'standard_relation': '=',
'standard_text_value': None, 'standard_type': 'IC50', 'standard_units': 'nM',
'standard_upper_value': None, 'standard_value': '1830.0', 'target_chembl_id':
'CHEMBL3553', 'target_organism': 'Homo sapiens', 'target_pref_name': 'Tyrosine-
protein kinase TYK2', 'target_tax_id': '9606', 'text_value': None, 'toid': None,
'type': 'IC50', 'units': 'uM', 'uo_units': 'UO_0000065', 'upper_value': None,
'value': '1.83'}, {'activity_comment': None, 'activity_id': 2112048,
'activity_properties': [], 'assay_chembl_id': 'CHEMBL934199',
'assay_description': 'Inhibition of Tyk2 by HTRF assay', 'assay_type': 'B',
'assay_variant_accession': None, 'assay_variant_mutation': None, 'bao_endpoint':
'BAO_0000190', 'bao_format': 'BAO_0000357', 'bao_label': 'single protein
format', 'canonical_smiles':
'Cc1cccc(C)c1-n1c(=O)c2cnc(Nc3ccc(N4CCN(C)CC4)cc3)nc2n2c3ccccc3nc12',
'data_validity_comment': None, 'data_validity_description': None,
'document_chembl_id': 'CHEMBL1143254', 'document_journal': 'J. Med. Chem.',
'document_year': 2008, 'ligand_efficiency': {'bei': '11.16', 'le': '0.20',
'lle': '1.23', 'sei': '7.08'}, 'molecule_chembl_id': 'CHEMBL410295',
'molecule_pref_name': None, 'parent_molecule_chembl_id': 'CHEMBL410295',
'pchembl_value': '5.92', 'potential_duplicate': False, 'qudt_units':
'http://www.openphacts.org/units/Nanomolar', 'record_id': 701224, 'relation':
'=', 'src_id': 1, 'standard_flag': True, 'standard_relation': '=',
'standard_text_value': None, 'standard_type': 'IC50', 'standard_units': 'nM',
'standard_upper_value': None, 'standard_value': '1200.0', 'target_chembl_id':
'CHEMBL3553', 'target_organism': 'Homo sapiens', 'target_pref_name': 'Tyrosine-
protein kinase TYK2', 'target_tax_id': '9606', 'text_value': None, 'toid': None,
'type': 'IC50', 'units': 'uM', 'uo_units': 'UO_0000065', 'upper_value': None,
'value': '1.2'}, '…(remaining elements truncated)…']
```

[5]:
```
tyk2_activities_df_reduced = tyk2_activities_df[['molecule_chembl_id',␣
↪'molecule_pref_name', 'canonical_smiles',
```

```
                                                    'pchembl_value',␣
↪'standard_type', 'standard_relation', 'standard_value',
                                                    'standard_units',␣
↪'potential_duplicate',
                                                    'target_pref_name',␣
↪'target_organism',
                                                    'assay_type',␣
↪'assay_description']]

tyk2_activities_df_reduced.head(10)
```

[5]:

	molecule_chembl_id	molecule_pref_name	canonical_smiles	pchembl_value	standard_type	standard_relation	standard_value	standard_units	potential_duplicate	target_pref_name	target_organism	assay_type	assay_description
0	CHEMBL21156	None	CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1	9.00	IC50	=	1.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
1	CHEMBL386661	None	CNc1ncc2cc(-c3cc(C(=O)Nc4cc(C(F)(F)F)ccc4OC4CC…	5.56	IC50	=	2730.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
2	CHEMBL215943	None	Cc1ccc(C(=O)Nc2cccc(C(F)(F)F)c2)cc1-c1ccc2nc(N…	5.74	IC50	=	1830.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
3	CHEMBL410295	None	Cc1cccc(C)c1-n1c(=O)c2cnc(Nc3ccc(N4CCN(C)CC4)c…	5.92	IC50	=	1200.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
4	CHEMBL21156	None	CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1	9.00	IC50	=	1.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
5	CHEMBL509032	TAE-684	COc1cc(N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc(Cl)c(N…	5.64	IC50	=	2309.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
6	CHEMBL495727	AT-9283	O=C(Nc1c[nH]nc1-c1nc2ccc(CN3CCOCC3)cc2[nH]1)NC…	9.00	IC50	=	1.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
7	CHEMBL21156	None	CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1	9.00	IC50	=	1.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
8	CHEMBL514409	HESPERADIN	CCS(=O)(=O)Nc1ccc2c(c1)/C(=C(/Nc1ccc(CN3CCCCC3…	7.12	Kd	=	75.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain
9	CHEMBL221959	TOFACITINIB	C[C@@H]1CCN(C(=O)CC#N)C[C@@H]1N(C)c1ncnc2[nH]c…	6.60	Kd	=	250.0	nM	0	Tyrosine-protein kinase TYK2	Homo sapiens	B	Inhibition of human TYK2 JH2 domain

```
   potential_duplicate              target_pref_name target_organism  \
0                 True  Tyrosine-protein kinase TYK2    Homo sapiens
1                False  Tyrosine-protein kinase TYK2    Homo sapiens
2                False  Tyrosine-protein kinase TYK2    Homo sapiens
3                False  Tyrosine-protein kinase TYK2    Homo sapiens
4                 True  Tyrosine-protein kinase TYK2    Homo sapiens
5                False  Tyrosine-protein kinase TYK2    Homo sapiens
6                False  Tyrosine-protein kinase TYK2    Homo sapiens
7                False  Tyrosine-protein kinase TYK2    Homo sapiens
8                False  Tyrosine-protein kinase TYK2    Homo sapiens
9                False  Tyrosine-protein kinase TYK2    Homo sapiens

  assay_type                                   assay_description
0          B             Inhibition of Tyrosine kinase 2 kinase
1          B            Inhibition of Tyk2 by HTRF kinase assay
2          B            Inhibition of Tyk2 by HTRF kinase assay
3          B                   Inhibition of Tyk2 by HTRF assay
4          B                                 Inhibition of Tyk2
5          B  Inhibition of Tel-fused TYK2 kinase-mediated m…
6          B                                 Inhibition of Tyk2
7          B                                 Inhibition of TYK2
8          B                     Binding affinity to human TYK2
9          B                           Binding affinity to TYK2
```

[6]:
```
tyk2_activities_df_reduced.shape
```

[6]:
```
(1857, 13)
```

[7]:
```
tyk2_activities_df_reduced[tyk2_activities_df_reduced['molecule_chembl_id'] ==␣
↪'CHEMBL21156']
```

[7]:
```
    molecule_chembl_id molecule_pref_name  \
0          CHEMBL21156               None
4          CHEMBL21156               None
7          CHEMBL21156               None
527        CHEMBL21156               None
528        CHEMBL21156               None

                                canonical_smiles pchembl_value  \
0    CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1          9.00
4    CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1          9.00
7    CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1          9.00
527  CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1          8.30
528  CC(C)(C)c1nc2c3ccc(F)cc3c3c(=O)[nH]ccc3c2[nH]1          5.77

    standard_type standard_relation standard_value standard_units  \
0            IC50                 =            1.0             nM
```

```
4            IC50                  =            1.0             nM
7            IC50                  =            1.0             nM
527          IC50                  =            5.0             nM
528          IC50                  =         1700.0             nM

     potential_duplicate             target_pref_name target_organism  \
0                   True  Tyrosine-protein kinase TYK2    Homo sapiens
4                   True  Tyrosine-protein kinase TYK2    Homo sapiens
7                  False  Tyrosine-protein kinase TYK2    Homo sapiens
527                False  Tyrosine-protein kinase TYK2    Homo sapiens
528                False  Tyrosine-protein kinase TYK2    Homo sapiens

    assay_type                                   assay_description
0            B             Inhibition of Tyrosine kinase 2 kinase
4            B                                  Inhibition of Tyk2
7            B                                  Inhibition of TYK2
527          B  Inhibition of Tyk2 (unknown origin) using 25 u…
528          B  Inhibition of Tyk2 (unknown origin) using 1 mM…
```

DEUCRAVACITINIB/LB7/ CHEMBL4435170/BMS-986165 potent, selective, oral, once daily allosteric inhibitor targeting the pseudokinase domain of TYK2, in phase III clinical trials for treatment of autoimmune diseases (active psoriatic arthritis)

[8]:
```
tyk2_activities_df_reduced[tyk2_activities_df_reduced['molecule_pref_name'] ==␣
↪'DEUCRAVACITINIB']
```

[8]:
```
     molecule_chembl_id molecule_pref_name  \
1393      CHEMBL4435170    DEUCRAVACITINIB
1429      CHEMBL4435170    DEUCRAVACITINIB
1457      CHEMBL4435170    DEUCRAVACITINIB
1458      CHEMBL4435170    DEUCRAVACITINIB
1462      CHEMBL4435170    DEUCRAVACITINIB
1463      CHEMBL4435170    DEUCRAVACITINIB
1599      CHEMBL4435170    DEUCRAVACITINIB
1600      CHEMBL4435170    DEUCRAVACITINIB
1778      CHEMBL4435170    DEUCRAVACITINIB
1799      CHEMBL4435170    DEUCRAVACITINIB
1827      CHEMBL4435170    DEUCRAVACITINIB

                                     canonical_smiles pchembl_value  \
1393  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70
1429  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          7.89
1457  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          8.70
1458  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          8.05
1462  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          9.70
1463  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70
1599  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70
```

```
1600  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          7.89
1778  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          9.70
1799  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          8.28
1827  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…          7.75

     standard_type standard_relation standard_value standard_units  \
1393            Ki                 =           0.02             nM
1429          IC50                 =           13.0             nM
1457          IC50                 =            2.0             nM
1458          IC50                 =            9.0             nM
1462          IC50                 =            0.2             nM
1463            Ki                 =           0.02             nM
1599            Ki                 =           0.02             nM
1600          EC50                 =           13.0             nM
1778          IC50                 =            0.2             nM
1799          IC50                 =            5.3             nM
1827          IC50                 =           18.0             nM

      potential_duplicate              target_pref_name target_organism  \
1393                 True  Tyrosine-protein kinase TYK2    Homo sapiens
1429                False  Tyrosine-protein kinase TYK2    Homo sapiens
1457                False  Tyrosine-protein kinase TYK2    Homo sapiens
1458                False  Tyrosine-protein kinase TYK2    Homo sapiens
1462                 True  Tyrosine-protein kinase TYK2    Homo sapiens
1463                False  Tyrosine-protein kinase TYK2    Homo sapiens
1599                 True  Tyrosine-protein kinase TYK2    Homo sapiens
1600                False  Tyrosine-protein kinase TYK2    Homo sapiens
1778                False  Tyrosine-protein kinase TYK2    Homo sapiens
1799                False  Tyrosine-protein kinase TYK2    Homo sapiens
1827                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description
1393          B  Inhibition of TYK2 in human Jurkat cells asses…
1429          B  Inhibition of TYK2 in human whole blood assess…
1457          B  Inhibition of TYK2 in human PBMC assessed as d…
1458          B  Inhibition of TYK2 in human PBMC assessed as d…
1462          B  Allosteric inhibition of fluorescein labeled p…
1463          B  Inhibition of fluorescein labeled probe bindin…
1599          B  Binding affinity to TYK2 pseudokinase domain (…
1600          B  Inhibition of TYK2 in human whole blood assess…
1778          B  Inhibition of TYK2 JH2 domain (unknown origin)…
1799          B  Inhibition of TYK2 in IFN-alpha stimulated hum…
1827          B  Inhibition of TYK2 in human whole blood assess…
```

[9]:
```
tyk2_activities_df_reduced[tyk2_activities_df_reduced['molecule_chembl_id'] ==␣
↪'CHEMBL4435170']
```

[9]:

	molecule_chembl_id	molecule_pref_name	canonical_smiles	pchembl_value
1393	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70
1429	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89
1457	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.70
1458	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.05
1462	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	9.70
1463	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70
1599	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70
1600	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89
1778	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	9.70
1799	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.28
1827	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.75

	standard_type	standard_relation	standard_value	standard_units
1393	Ki	=	0.02	nM
1429	IC50	=	13.0	nM
1457	IC50	=	2.0	nM
1458	IC50	=	9.0	nM
1462	IC50	=	0.2	nM
1463	Ki	=	0.02	nM
1599	Ki	=	0.02	nM
1600	EC50	=	13.0	nM
1778	IC50	=	0.2	nM
1799	IC50	=	5.3	nM
1827	IC50	=	18.0	nM

	potential_duplicate	target_pref_name	target_organism
1393	True	Tyrosine-protein kinase TYK2	Homo sapiens
1429	False	Tyrosine-protein kinase TYK2	Homo sapiens
1457	False	Tyrosine-protein kinase TYK2	Homo sapiens
1458	False	Tyrosine-protein kinase TYK2	Homo sapiens
1462	True	Tyrosine-protein kinase TYK2	Homo sapiens
1463	False	Tyrosine-protein kinase TYK2	Homo sapiens
1599	True	Tyrosine-protein kinase TYK2	Homo sapiens

```
1600                False  Tyrosine-protein kinase TYK2    Homo sapiens
1778                False  Tyrosine-protein kinase TYK2    Homo sapiens
1799                False  Tyrosine-protein kinase TYK2    Homo sapiens
1827                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                  assay_description
1393          B  Inhibition of TYK2 in human Jurkat cells asses…
1429          B  Inhibition of TYK2 in human whole blood assess…
1457          B  Inhibition of TYK2 in human PBMC assessed as d…
1458          B  Inhibition of TYK2 in human PBMC assessed as d…
1462          B  Allosteric inhibition of fluorescein labeled p…
1463          B  Inhibition of fluorescein labeled probe bindin…
1599          B  Binding affinity to TYK2 pseudokinase domain (…
1600          B  Inhibition of TYK2 in human whole blood assess…
1778          B  Inhibition of TYK2 JH2 domain (unknown origin)…
1799          B  Inhibition of TYK2 in IFN-alpha stimulated hum…
1827          B  Inhibition of TYK2 in human whole blood assess…
```

[10]:
```
tyk2_activ_df_sorted = tyk2_activities_df_reduced.
↪sort_values(by=['molecule_chembl_id','pchembl_value'],
                                                        ascending=True)
tyk2_activ_df_sorted
```

[10]:
```
     molecule_chembl_id molecule_pref_name  \
118            CHEMBL10          SB-203580
30        CHEMBL1076700               None
1110      CHEMBL1078178        MOMELOTINIB
27        CHEMBL1080159               None
19        CHEMBL1081290               None
…                   …                  …
117        CHEMBL608154               None
13         CHEMBL608533        MIDOSTAURIN
50         CHEMBL608533        MIDOSTAURIN
95         CHEMBL608533        MIDOSTAURIN
77            CHEMBL941           IMATINIB

                                      canonical_smiles pchembl_value  \
118   C[S+]([O-])c1ccc(-c2nc(-c3ccc(F)cc3)c(-c3ccncc…          5.70
30    Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          7.01
1110  N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
27    Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
19          COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47
…                                                  …          …
117       COc1c(Cl)cc2c([nH]c3cnccc32)c1NC(=O)c1cccnc1C          6.30
13    CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
50    CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
95    CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
```

```
77    Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc…          5.06

     standard_type standard_relation standard_value standard_units  \
118             Kd                =         2000.0             nM
30           IC50                 =           97.0             nM
1110            Kd                =          401.0             nM
27           IC50                 =          200.0             nM
19           IC50                 =         3400.0             nM
…             …                   …             …            …
117             Kd                =          500.0             nM
13              Kd                =          250.0             nM
50              Kd                =          250.0             nM
95              Kd                =          250.0             nM
77              Kd                =         8700.0             nM

      potential_duplicate              target_pref_name target_organism  \
118                 False  Tyrosine-protein kinase TYK2    Homo sapiens
30                  False  Tyrosine-protein kinase TYK2    Homo sapiens
1110                False  Tyrosine-protein kinase TYK2    Homo sapiens
27                  False  Tyrosine-protein kinase TYK2    Homo sapiens
19                  False  Tyrosine-protein kinase TYK2    Homo sapiens
…                   …                              …            …
117                 False  Tyrosine-protein kinase TYK2    Homo sapiens
13                   True  Tyrosine-protein kinase TYK2    Homo sapiens
50                   True  Tyrosine-protein kinase TYK2    Homo sapiens
95                  False  Tyrosine-protein kinase TYK2    Homo sapiens
77                  False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description
118           B  Binding constant for TYK2(JH2domain-pseudokina…
30            B  Inhibition of GST-tagged TYK2 assessed as inhi…
1110          B  Kinobeads (epsilon), multiple immobilized ATP-…
27            B  Inhibition of GST-tagged TYK2 assessed as inhi…
19            B  Inhibition of GST-tagged TYK2 assessed as inhi…
…         …                                                   …
117           B  Binding constant for TYK2(JH2domain-pseudokina…
13            B  Binding constant for TYK2(Kin.Dom.2/JH1 - cata…
50            B      Binding affinity to TYK2(JH1domain-catalytic)
95            B  Binding constant for TYK2(JH1domain-catalytic)…
77            B  Binding constant for TYK2(JH2domain-pseudokina…

[1857 rows x 13 columns]
```

[11]:
```
tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
↪'CHEMBL4435170']
```

[11]:

	molecule_chembl_id	molecule_pref_name	canonical_smiles	pchembl_value	standard_type	standard_relation	standard_value	standard_units	potential_duplicate	target_pref_name	target_organism
1393	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70	Ki	=	0.02	nM	True	Tyrosine-protein kinase TYK2	Homo sapiens
1463	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70	Ki	=	0.02	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1599	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70	Ki	=	0.02	nM	True	Tyrosine-protein kinase TYK2	Homo sapiens
1827	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.75	IC50	=	18.0	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1429	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89	IC50	=	13.0	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1600	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89	EC50	=	13.0	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1458	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.05	IC50	=	9.0	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1799	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.28	IC50	=	5.3	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1457	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.70	IC50	=	2.0	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1462	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	9.70	IC50	=	0.2	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens
1778	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	9.70	IC50	=	0.2	nM	False	Tyrosine-protein kinase TYK2	Homo sapiens

```
1799                False  Tyrosine-protein kinase TYK2    Homo sapiens
1457                False  Tyrosine-protein kinase TYK2    Homo sapiens
1462                 True  Tyrosine-protein kinase TYK2    Homo sapiens
1778                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description
1393          B  Inhibition of TYK2 in human Jurkat cells asses…
1463          B  Inhibition of fluorescein labeled probe bindin…
1599          B  Binding affinity to TYK2 pseudokinase domain (…
1827          B  Inhibition of TYK2 in human whole blood assess…
1429          B  Inhibition of TYK2 in human whole blood assess…
1600          B  Inhibition of TYK2 in human whole blood assess…
1458          B  Inhibition of TYK2 in human PBMC assessed as d…
1799          B  Inhibition of TYK2 in IFN-alpha stimulated hum…
1457          B  Inhibition of TYK2 in human PBMC assessed as d…
1462          B  Allosteric inhibition of fluorescein labeled p…
1778          B  Inhibition of TYK2 JH2 domain (unknown origin)…
```

[12]:
```
#remove potential duplicated activity entries
tyk2_activ_df_sorted.
 ↪drop(tyk2_activ_df_sorted[tyk2_activ_df_sorted['potential_duplicate'] ==␣
 ↪True].index, inplace = True)
tyk2_activ_df_sorted.shape
```

[12]:
```
(1820, 13)
```

[13]:
```
tyk2_activ_df_sorted
```

[13]:
```
     molecule_chembl_id molecule_pref_name  \
118            CHEMBL10          SB-203580
30        CHEMBL1076700               None
1110      CHEMBL1078178        MOMELOTINIB
27        CHEMBL1080159               None
19        CHEMBL1081290               None
…                   …                  …
116        CHEMBL603469       LESTAURTINIB
115        CHEMBL603469       LESTAURTINIB
117        CHEMBL608154               None
95         CHEMBL608533        MIDOSTAURIN
77            CHEMBL941           IMATINIB

                                     canonical_smiles pchembl_value  \
118   C[S+]([O-])c1ccc(-c2nc(-c3ccc(F)cc3)c(-c3ccncc…          5.70
30    Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          7.01
1110  N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
27    Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
19          COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47
```

```
…                                                    …          …
116   C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6…          6.04
115   C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6…          7.82
117       COc1c(Cl)cc2c([nH]c3cnccc32)c1NC(=O)c1cccnc1C          6.30
95    CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
77    Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc…          5.06

     standard_type standard_relation standard_value standard_units  \
118             Kd                 =         2000.0             nM
30            IC50                 =           97.0             nM
1110            Kd                 =          401.0             nM
27            IC50                 =          200.0             nM
19            IC50                 =         3400.0             nM
…              …                   …             …             …
116             Kd                 =          910.0             nM
115             Kd                 =           15.0             nM
117             Kd                 =          500.0             nM
95              Kd                 =          250.0             nM
77              Kd                 =         8700.0             nM

      potential_duplicate              target_pref_name target_organism  \
118                 False  Tyrosine-protein kinase TYK2    Homo sapiens
30                  False  Tyrosine-protein kinase TYK2    Homo sapiens
1110                False  Tyrosine-protein kinase TYK2    Homo sapiens
27                  False  Tyrosine-protein kinase TYK2    Homo sapiens
19                  False  Tyrosine-protein kinase TYK2    Homo sapiens
…                   …                             …             …
116                 False  Tyrosine-protein kinase TYK2    Homo sapiens
115                 False  Tyrosine-protein kinase TYK2    Homo sapiens
117                 False  Tyrosine-protein kinase TYK2    Homo sapiens
95                  False  Tyrosine-protein kinase TYK2    Homo sapiens
77                  False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description
118           B  Binding constant for TYK2(JH2domain-pseudokina…
30            B  Inhibition of GST-tagged TYK2 assessed as inhi…
1110          B  Kinobeads (epsilon), multiple immobilized ATP-…
27            B  Inhibition of GST-tagged TYK2 assessed as inhi…
19            B  Inhibition of GST-tagged TYK2 assessed as inhi…
…           …                                                 …
116           B  Binding constant for TYK2(JH2domain-pseudokina…
115           B  Binding constant for TYK2(JH1domain-catalytic)…
117           B  Binding constant for TYK2(JH2domain-pseudokina…
95            B  Binding constant for TYK2(JH1domain-catalytic)…
77            B  Binding constant for TYK2(JH2domain-pseudokina…

[1820 rows x 13 columns]
```

[14]:
```
#Remove duplicates of remaining molecules after aggregating the pChembl values

#duplicates_df = tyk2_activ_df_sorted.duplicated('molecule_chembl_id')
#duplicates_df
tyk2_activ_df_sorted['chembl_id_duplicate'] = tyk2_activ_df_sorted.
 ↪duplicated('molecule_chembl_id')
tyk2_activ_df_sorted[['molecule_chembl_id','chembl_id_duplicate']]
```

[14]:

	molecule_chembl_id	chembl_id_duplicate
118	CHEMBL10	False
30	CHEMBL1076700	False
1110	CHEMBL1078178	False
27	CHEMBL1080159	False
19	CHEMBL1081290	False
…	…	…
116	CHEMBL603469	False
115	CHEMBL603469	True
117	CHEMBL608154	False
95	CHEMBL608533	False
77	CHEMBL941	False

[1820 rows x 2 columns]

[15]:
```
tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
 ↪'CHEMBL4435170']
```

[15]:

	molecule_chembl_id	molecule_pref_name	canonical_smiles	pchembl_value	standard_type	standard_relation	standard_value	standard_units
1463	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	10.70				
1827	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.75				
1429	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89				
1600	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	7.89				
1458	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.05				
1799	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.28				
1457	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	8.70				
1778	CHEMBL4435170	DEUCRAVACITINIB	[2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…	9.70				

```
1463           Ki                =           0.02            nM
1827         IC50                =           18.0            nM
1429         IC50                =           13.0            nM
1600         EC50                =           13.0            nM
1458         IC50                =            9.0            nM
1799         IC50                =            5.3            nM
1457         IC50                =            2.0            nM
1778         IC50                =            0.2            nM

      potential_duplicate             target_pref_name target_organism  \
1463                False  Tyrosine-protein kinase TYK2    Homo sapiens
1827                False  Tyrosine-protein kinase TYK2    Homo sapiens
1429                False  Tyrosine-protein kinase TYK2    Homo sapiens
1600                False  Tyrosine-protein kinase TYK2    Homo sapiens
1458                False  Tyrosine-protein kinase TYK2    Homo sapiens
1799                False  Tyrosine-protein kinase TYK2    Homo sapiens
1457                False  Tyrosine-protein kinase TYK2    Homo sapiens
1778                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description  \
1463          B  Inhibition of fluorescein labeled probe bindin…
1827          B  Inhibition of TYK2 in human whole blood assess…
1429          B  Inhibition of TYK2 in human whole blood assess…
1600          B  Inhibition of TYK2 in human whole blood assess…
1458          B  Inhibition of TYK2 in human PBMC assessed as d…
1799          B  Inhibition of TYK2 in IFN-alpha stimulated hum…
1457          B  Inhibition of TYK2 in human PBMC assessed as d…
1778          B  Inhibition of TYK2 JH2 domain (unknown origin)…

      chembl_id_duplicate
1463                False
1827                 True
1429                 True
1600                 True
1458                 True
1799                 True
1457                 True
1778                 True
```

[16]:
```
duplicates =␣
 ↪set(tyk2_activ_df_sorted[tyk2_activ_df_sorted['chembl_id_duplicate'] ==␣
 ↪True]['molecule_chembl_id'])
duplicates
```

[16]:
```
{'CHEMBL1241674',
 'CHEMBL1287853',
 'CHEMBL1421',
```

```
'CHEMBL1650951',
'CHEMBL1721885',
'CHEMBL1789941',
'CHEMBL1908395',
'CHEMBL1908397',
'CHEMBL1983111',
'CHEMBL2035187',
'CHEMBL2105759',
'CHEMBL21156',
'CHEMBL221959',
'CHEMBL2385096',
'CHEMBL2386629',
'CHEMBL2386633',
'CHEMBL2386635',
'CHEMBL2386636',
'CHEMBL2387110',
'CHEMBL2387112',
'CHEMBL2387118',
'CHEMBL2387119',
'CHEMBL2387124',
'CHEMBL2387126',
'CHEMBL2387127',
'CHEMBL2387221',
'CHEMBL2387222',
'CHEMBL2387223',
'CHEMBL2387224',
'CHEMBL2387225',
'CHEMBL288441',
'CHEMBL3301607',
'CHEMBL3622821',
'CHEMBL3655081',
'CHEMBL3763184',
'CHEMBL3763213',
'CHEMBL3763252',
'CHEMBL3763697',
'CHEMBL3763991',
'CHEMBL3764030',
'CHEMBL3764167',
'CHEMBL3764277',
'CHEMBL3764383',
'CHEMBL3764637',
'CHEMBL3765517',
'CHEMBL3765822',
'CHEMBL388978',
'CHEMBL3906967',
'CHEMBL4062680',
'CHEMBL4062758',
```

```
'CHEMBL4068357',
'CHEMBL4069942',
'CHEMBL4070262',
'CHEMBL4071399',
'CHEMBL4075453',
'CHEMBL4076947',
'CHEMBL4080904',
'CHEMBL4084436',
'CHEMBL4092116',
'CHEMBL4092191',
'CHEMBL4093872',
'CHEMBL4097446',
'CHEMBL4099854',
'CHEMBL4105148',
'CHEMBL4128181',
'CHEMBL4238926',
'CHEMBL4278757',
'CHEMBL4279157',
'CHEMBL4279883',
'CHEMBL4281304',
'CHEMBL4282627',
'CHEMBL4283351',
'CHEMBL4283724',
'CHEMBL4285530',
'CHEMBL4285755',
'CHEMBL4285841',
'CHEMBL4286006',
'CHEMBL428690',
'CHEMBL4287081',
'CHEMBL4287153',
'CHEMBL4287761',
'CHEMBL4289149',
'CHEMBL4289426',
'CHEMBL4289875',
'CHEMBL4290130',
'CHEMBL4291200',
'CHEMBL4291611',
'CHEMBL4293510',
'CHEMBL4293619',
'CHEMBL4293907',
'CHEMBL4295039',
'CHEMBL4434711',
'CHEMBL4435047',
'CHEMBL4435170',
'CHEMBL4437714',
'CHEMBL4438107',
'CHEMBL4438202',
```

'CHEMBL4438296',
'CHEMBL4439957',
'CHEMBL4440718',
'CHEMBL4440767',
'CHEMBL4442827',
'CHEMBL4443010',
'CHEMBL4444169',
'CHEMBL4444178',
'CHEMBL4453441',
'CHEMBL4453827',
'CHEMBL4454109',
'CHEMBL4459585',
'CHEMBL4460194',
'CHEMBL4460368',
'CHEMBL4466139',
'CHEMBL4469812',
'CHEMBL4474801',
'CHEMBL4476830',
'CHEMBL4517542',
'CHEMBL4519857',
'CHEMBL4526283',
'CHEMBL4530719',
'CHEMBL4532948',
'CHEMBL4537678',
'CHEMBL4543066',
'CHEMBL4544320',
'CHEMBL4547009',
'CHEMBL4561123',
'CHEMBL4561663',
'CHEMBL4571920',
'CHEMBL4579439',
'CHEMBL4583185',
'CHEMBL4585272',
'CHEMBL4637587',
'CHEMBL4641006',
'CHEMBL4643392',
'CHEMBL4643899',
'CHEMBL4645110',
'CHEMBL4645258',
'CHEMBL475251',
'CHEMBL4761365',
'CHEMBL4776752',
'CHEMBL4777175',
'CHEMBL4777528',
'CHEMBL4777648',
'CHEMBL4780057',
'CHEMBL4780233',

```
 'CHEMBL4780315',
 'CHEMBL4782449',
 'CHEMBL4784391',
 'CHEMBL4784838',
 'CHEMBL4785512',
 'CHEMBL4785957',
 'CHEMBL4788122',
 'CHEMBL4789015',
 'CHEMBL4789639',
 'CHEMBL4792231',
 'CHEMBL4792694',
 'CHEMBL4792780',
 'CHEMBL4793760',
 'CHEMBL4795231',
 'CHEMBL4796821',
 'CHEMBL4797596',
 'CHEMBL4799019',
 'CHEMBL4799031',
 'CHEMBL4799559',
 'CHEMBL4799830',
 'CHEMBL4800399',
 'CHEMBL482967',
 'CHEMBL495727',
 'CHEMBL502835',
 'CHEMBL509032',
 'CHEMBL522892',
 'CHEMBL535',
 'CHEMBL572878',
 'CHEMBL574738',
 'CHEMBL590109',
 'CHEMBL601719',
 'CHEMBL603469'}
```

[17]:
```
#calculate mean, max, min pchembl value of duplicates

agg_activ_dp = [(m,
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
 ↪m]['pchembl_value'].astype(float).mean(), 2),
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
 ↪m]['pchembl_value'].astype(float).max(), 2),
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
 ↪m]['pchembl_value'].astype(float).min(), 2)
 )
 for m in duplicates]
agg_activ_dp
```

```
[17]: [('CHEMBL4800399', 7.72, 8.07, 7.38),
       ('CHEMBL4476830', 7.22, 8.82, 5.62),
       ('CHEMBL4530719', 8.12, 9.49, 7.38),
       ('CHEMBL4537678', 7.59, 8.89, 6.28),
       ('CHEMBL509032', 6.29, 7.11, 5.64),
       ('CHEMBL4643392', 7.5, 8.31, 6.68),
       ('CHEMBL4645110', 7.14, 7.92, 6.36),
       ('CHEMBL4797596', 8.05, 9.15, 7.2),
       ('CHEMBL4799559', 8.18, 9.52, 7.16),
       ('CHEMBL1421', 6.19, 6.96, 5.75),
       ('CHEMBL4443010', 7.83, 10.04, 6.61),
       ('CHEMBL4637587', 6.39, 6.89, 5.89),
       ('CHEMBL4434711', 7.73, 10.15, 6.09),
       ('CHEMBL4789639', 8.72, 10.7, 7.24),
       ('CHEMBL1908395', 7.28, 8.24, 6.33),
       ('CHEMBL2387110', 7.46, 8.52, 6.41),
       ('CHEMBL4788122', 7.58, 8.85, 6.74),
       ('CHEMBL4547009', 7.48, 9.48, 6.09),
       ('CHEMBL4526283', 7.9, 9.04, 6.75),
       ('CHEMBL2386633', 6.58, 7.92, 5.25),
       ('CHEMBL4453827', 6.73, 8.1, 5.36),
       ('CHEMBL4517542', 7.67, 9.13, 6.21),
       ('CHEMBL4777528', 7.92, 8.21, 7.62),
       ('CHEMBL2387225', 7.8, 9.0, 6.61),
       ('CHEMBL4285530', 5.59, 6.3, 5.14),
       ('CHEMBL4793760', 8.2, 9.6, 7.23),
       ('CHEMBL4789015', 7.79, 8.19, 7.39),
       ('CHEMBL4099854', 7.3, 9.15, 5.44),
       ('CHEMBL4583185', 7.97, 9.24, 6.7),
       ('CHEMBL4585272', 7.61, 9.54, 6.46),
       ('CHEMBL4062680', 8.18, 9.22, 7.14),
       ('CHEMBL4444178', 7.23, 9.1, 4.3),
       ('CHEMBL4780057', 7.39, 8.35, 6.65),
       ('CHEMBL2387224', 7.37, 8.32, 6.42),
       ('CHEMBL4293907', 7.79, 9.89, 6.96),
       ('CHEMBL3622821', 5.33, 5.33, 5.33),
       ('CHEMBL4279157', 7.48, 8.15, 7.0),
       ('CHEMBL502835', 6.86, 7.43, 6.3),
       ('CHEMBL4281304', 6.35, 7.18, 5.92),
       ('CHEMBL3764167', 6.84, 7.48, 6.19),
       ('CHEMBL4080904', 7.88, 8.92, 6.85),
       ('CHEMBL1287853', 6.81, 7.68, 6.35),
       ('CHEMBL4289149', 5.44, 6.32, 4.9),
       ('CHEMBL4460194', 6.88, 8.28, 5.48),
       ('CHEMBL3763184', 6.1, 7.04, 5.16),
       ('CHEMBL4460368', 8.51, 10.82, 7.05),
       ('CHEMBL2387112', 7.44, 8.51, 6.36),
```

```
('CHEMBL572878', 6.06, 6.7, 5.18),
('CHEMBL4643899', 7.43, 8.28, 6.58),
('CHEMBL2035187', 7.36, 7.42, 7.3),
('CHEMBL3763252', 8.04, 8.82, 7.27),
('CHEMBL2387127', 7.65, 8.6, 6.71),
('CHEMBL4128181', 5.64, 6.34, 4.93),
('CHEMBL4777648', 7.85, 8.89, 6.95),
('CHEMBL4287153', 6.53, 7.57, 6.0),
('CHEMBL2387126', 7.59, 8.47, 6.7),
('CHEMBL495727', 7.78, 9.0, 6.55),
('CHEMBL4444169', 8.06, 9.31, 6.8),
('CHEMBL4466139', 7.64, 9.47, 6.42),
('CHEMBL3301607', 6.88, 7.18, 5.58),
('CHEMBL4440718', 8.77, 10.15, 6.89),
('CHEMBL2387119', 7.52, 8.48, 6.56),
('CHEMBL3765517', 7.78, 8.92, 6.64),
('CHEMBL4287081', 6.74, 7.77, 6.22),
('CHEMBL4291200', 6.99, 8.1, 6.35),
('CHEMBL4282627', 5.85, 6.96, 5.21),
('CHEMBL4289875', 6.41, 7.62, 5.77),
('CHEMBL2386629', 7.49, 8.47, 6.51),
('CHEMBL2387221', 7.84, 8.85, 6.84),
('CHEMBL4780233', 7.91, 9.57, 6.63),
('CHEMBL4796821', 8.07, 9.32, 6.97),
('CHEMBL4439957', 8.37, 10.46, 7.07),
('CHEMBL4453441', 7.77, 9.7, 6.57),
('CHEMBL21156', 7.69, 9.0, 5.77),
('CHEMBL4544320', 7.81, 9.02, 6.6),
('CHEMBL4777175', 8.24, 9.36, 7.28),
('CHEMBL4792231', 7.72, 8.05, 7.39),
('CHEMBL4532948', 8.25, 10.07, 7.09),
('CHEMBL2387124', 7.77, 8.74, 6.8),
('CHEMBL603469', 6.93, 7.82, 6.04),
('CHEMBL2387222', 7.74, 8.8, 6.68),
('CHEMBL4084436', 8.05, 9.52, 6.58),
('CHEMBL3765822', 7.7, 8.68, 6.72),
('CHEMBL4795231', 7.69, 8.12, 7.27),
('CHEMBL4062758', 8.09, 9.4, 6.77),
('CHEMBL4440767', 6.3, 6.89, 5.7),
('CHEMBL4438296', 7.74, 9.59, 6.42),
('CHEMBL4286006', 6.01, 7.3, 5.31),
('CHEMBL3655081', 6.35, 6.8, 5.9),
('CHEMBL4579439', 7.75, 9.66, 6.43),
('CHEMBL3763991', 8.18, 9.04, 7.32),
('CHEMBL2105759', 7.46, 8.06, 7.21),
('CHEMBL4238926', 8.31, 8.32, 8.3),
('CHEMBL4295039', 7.35, 8.3, 6.82),
```

```
('CHEMBL4438107', 7.61, 9.89, 6.12),
('CHEMBL288441', 5.27, 5.35, 5.19),
('CHEMBL4278757', 6.87, 7.82, 6.32),
('CHEMBL2387118', 7.35, 8.36, 6.34),
('CHEMBL4070262', 7.72, 9.0, 6.44),
('CHEMBL4105148', 6.66, 8.29, 5.03),
('CHEMBL4285841', 7.97, 8.22, 7.7),
('CHEMBL4543066', 7.72, 9.54, 6.54),
('CHEMBL4442827', 8.44, 10.74, 7.2),
('CHEMBL4761365', 7.05, 8.07, 6.03),
('CHEMBL4785512', 7.62, 7.85, 7.38),
('CHEMBL4780315', 7.75, 9.27, 6.47),
('CHEMBL4076947', 8.16, 9.15, 7.18),
('CHEMBL4097446', 8.07, 8.92, 7.23),
('CHEMBL4784838', 8.06, 9.29, 7.14),
('CHEMBL4438202', 7.66, 8.8, 6.51),
('CHEMBL1908397', 6.54, 6.8, 6.28),
('CHEMBL574738', 5.72, 5.75, 5.68),
('CHEMBL2386636', 7.39, 8.77, 6.01),
('CHEMBL221959', 7.3, 8.44, 5.92),
('CHEMBL4289426', 7.39, 8.7, 6.52),
('CHEMBL590109', 5.35, 5.7, 5.0),
('CHEMBL4561123', 7.82, 9.82, 6.63),
('CHEMBL4776752', 7.5, 8.85, 6.34),
('CHEMBL388978', 8.28, 9.69, 5.51),
('CHEMBL2386635', 8.2, 9.4, 7.0),
('CHEMBL4092116', 7.78, 9.22, 6.35),
('CHEMBL4459585', 7.8, 7.82, 7.77),
('CHEMBL4437714', 7.46, 9.42, 6.33),
('CHEMBL4799031', 8.12, 9.32, 7.28),
('CHEMBL2385096', 6.68, 7.7, 5.66),
('CHEMBL4068357', 6.34, 6.89, 5.79),
('CHEMBL4283724', 5.88, 6.66, 5.47),
('CHEMBL4291611', 7.19, 8.4, 6.48),
('CHEMBL4283351', 6.86, 7.89, 6.29),
('CHEMBL475251', 7.26, 7.8, 6.72),
('CHEMBL4435170', 8.62, 10.7, 7.75),
('CHEMBL2387223', 6.86, 8.8, 5.92),
('CHEMBL535', 6.12, 6.44, 5.8),
('CHEMBL3764277', 7.98, 9.01, 6.96),
('CHEMBL4293619', 6.43, 7.5, 5.82),
('CHEMBL4641006', 7.5, 8.26, 6.75),
('CHEMBL4561663', 8.89, 10.22, 7.82),
('CHEMBL601719', 6.09, 6.68, 5.7),
('CHEMBL4799019', 8.18, 9.6, 7.04),
('CHEMBL4285755', 6.56, 7.66, 5.85),
('CHEMBL428690', 6.56, 7.46, 5.46),
```

```
('CHEMBL4474801', 7.71, 9.85, 6.42),
('CHEMBL4571920', 8.47, 10.62, 6.87),
('CHEMBL1983111', 6.82, 7.9, 5.73),
('CHEMBL4435047', 7.88, 9.92, 6.5),
('CHEMBL1721885', 5.7, 6.12, 5.29),
('CHEMBL4071399', 8.33, 9.3, 7.36),
('CHEMBL522892', 6.39, 6.82, 5.66),
('CHEMBL3764030', 7.54, 8.26, 6.82),
('CHEMBL4784391', 7.67, 8.72, 6.79),
('CHEMBL4799830', 7.52, 8.8, 6.37),
('CHEMBL1789941', 8.34, 9.4, 5.85),
('CHEMBL4519857', 6.88, 7.64, 6.11),
('CHEMBL4454109', 7.55, 9.32, 6.47),
('CHEMBL4785957', 7.9, 8.92, 6.89),
('CHEMBL4279883', 7.33, 8.42, 6.43),
('CHEMBL4287761', 7.63, 8.22, 7.16),
('CHEMBL4792780', 8.2, 9.28, 7.33),
('CHEMBL1650951', 5.51, 5.77, 5.09),
('CHEMBL4469812', 7.12, 8.74, 5.51),
('CHEMBL4293510', 6.83, 7.6, 6.39),
('CHEMBL4792694', 7.66, 7.96, 7.36),
('CHEMBL4093872', 7.96, 9.22, 6.7),
('CHEMBL3906967', 7.42, 7.43, 7.42),
('CHEMBL4092191', 8.3, 9.15, 7.46),
('CHEMBL3763697', 6.55, 7.14, 5.96),
('CHEMBL4782449', 7.97, 9.3, 7.06),
('CHEMBL482967', 6.54, 7.1, 5.99),
('CHEMBL4290130', 6.43, 8.1, 5.38),
('CHEMBL4069942', 8.04, 9.22, 6.85),
('CHEMBL3764383', 7.18, 8.01, 6.36),
('CHEMBL4075453', 7.79, 9.3, 6.28),
('CHEMBL3763213', 7.98, 9.26, 6.7),
('CHEMBL4645258', 7.16, 7.64, 6.68),
('CHEMBL3764637', 7.36, 8.2, 6.51),
('CHEMBL1241674', 6.01, 6.55, 5.47)]
```

[18]:
```
#calculate mean, max, min pchembl value of duplicates

agg_activ_dp = [(m,
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
↪m]['pchembl_value'].astype(float).mean(), 2),
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
↪m]['pchembl_value'].astype(float).max(), 2),
  np.round(tyk2_activ_df_sorted[tyk2_activ_df_sorted['molecule_chembl_id'] ==␣
↪m]['pchembl_value'].astype(float).min(), 2))
                for m in duplicates]
```

```
agg_activ_dp_df = pd.DataFrame(agg_activ_dp,
                               columns=['molecule_chembl_id',␣
 ↪'mean_pchembl_value', 'max_pchembl_value', 'min_pchembl_value'])

agg_activ_dp_df
```

[18]:
```
    molecule_chembl_id  mean_pchembl_value  max_pchembl_value  \
0        CHEMBL4800399                7.72               8.07
1        CHEMBL4476830                7.22               8.82
2        CHEMBL4530719                8.12               9.49
3        CHEMBL4537678                7.59               8.89
4         CHEMBL509032                6.29               7.11
..                 …                   …                  …
171      CHEMBL4075453                7.79               9.30
172      CHEMBL3763213                7.98               9.26
173      CHEMBL4645258                7.16               7.64
174      CHEMBL3764637                7.36               8.20
175      CHEMBL1241674                6.01               6.55

     min_pchembl_value
0                 7.38
1                 5.62
2                 7.38
3                 6.28
4                 5.64
..                 …
171               6.28
172               6.70
173               6.68
174               6.51
175               5.47

[176 rows x 4 columns]
```

[19]:
```
#remove duplicates and keep first match

new_activ_df = tyk2_activ_df_sorted.drop_duplicates('molecule_chembl_id',␣
 ↪keep='first')
new_activ_df = new_activ_df.reset_index(drop=True)
new_activ_df
```

[19]:
```
     molecule_chembl_id molecule_pref_name  \
0              CHEMBL10          SB-203580
1         CHEMBL1076700               None
2         CHEMBL1078178        MOMELOTINIB
3         CHEMBL1080159               None
4         CHEMBL1081290               None
```

```
…                …                …
1497       CHEMBL601719         CRIZOTINIB
1498       CHEMBL603469       LESTAURTINIB
1499       CHEMBL608154               None
1500       CHEMBL608533        MIDOSTAURIN
1501          CHEMBL941           IMATINIB

                                      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
2     N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
3     Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
4           COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47
…                                                  …           …
1497  C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)…          5.70
1498  C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6…          6.04
1499      COc1c(Cl)cc2c([nH]c3cnccc32)c1NC(=O)c1cccnc1C          6.30
1500  CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
1501  Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc…          5.06

     standard_type standard_relation standard_value standard_units  \
0               Kd                =         2000.0             nM
1             IC50                =           97.0             nM
2               Kd                =          401.0             nM
3             IC50                =          200.0             nM
4             IC50                =         3400.0             nM
…             …                   …             …             …
1497            Kd                =         2000.0             nM
1498            Kd                =          910.0             nM
1499            Kd                =          500.0             nM
1500            Kd                =          250.0             nM
1501            Kd                =         8700.0             nM

      potential_duplicate             target_pref_name target_organism  \
0                   False  Tyrosine-protein kinase TYK2    Homo sapiens
1                   False  Tyrosine-protein kinase TYK2    Homo sapiens
2                   False  Tyrosine-protein kinase TYK2    Homo sapiens
3                   False  Tyrosine-protein kinase TYK2    Homo sapiens
4                   False  Tyrosine-protein kinase TYK2    Homo sapiens
…                   …                              …              …
1497                False  Tyrosine-protein kinase TYK2    Homo sapiens
1498                False  Tyrosine-protein kinase TYK2    Homo sapiens
1499                False  Tyrosine-protein kinase TYK2    Homo sapiens
1500                False  Tyrosine-protein kinase TYK2    Homo sapiens
1501                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…
2             B  Kinobeads (epsilon), multiple immobilized ATP-…
3             B  Inhibition of GST-tagged TYK2 assessed as inhi…
4             B  Inhibition of GST-tagged TYK2 assessed as inhi…
…           …                                                 …
1497          B  Binding constant for TYK2(JH2domain-pseudokina…
1498          B  Binding constant for TYK2(JH2domain-pseudokina…
1499          B  Binding constant for TYK2(JH2domain-pseudokina…
1500          B  Binding constant for TYK2(JH1domain-catalytic)…
1501          B  Binding constant for TYK2(JH2domain-pseudokina…

      chembl_id_duplicate
0                   False
1                   False
2                   False
3                   False
4                   False
…                   …
1497                False
1498                False
1499                False
1500                False
1501                False

[1502 rows x 14 columns]
```

[20]:
```
new_activ_df.shape
```

[20]:
```
(1502, 14)
```

[21]:
```
#make a copy of the dataframe just in case
import copy
new_activ_df_dc = copy.deepcopy(new_activ_df)
```

[22]:
```
#merge the nonduplicated dataframe with aggregated values of duplicates␣
↪dataframe
merged_activ_df = new_activ_df.merge(agg_activ_dp_df, how='left',␣
↪on='molecule_chembl_id')
merged_activ_df
```

[22]:
```
     molecule_chembl_id molecule_pref_name  \
0              CHEMBL10          SB-203580
1         CHEMBL1076700               None
2         CHEMBL1078178        MOMELOTINIB
3         CHEMBL1080159               None
4         CHEMBL1081290               None
```

```
…                …                …
1497       CHEMBL601719         CRIZOTINIB
1498       CHEMBL603469       LESTAURTINIB
1499       CHEMBL608154               None
1500       CHEMBL608533        MIDOSTAURIN
1501          CHEMBL941           IMATINIB

                                      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
2     N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
3     Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
4           COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47
…                                                   …           …
1497  C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)…          5.70
1498  C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6…          6.04
1499      COc1c(Cl)cc2c([nH]c3cnccc32)c1NC(=O)c1cccnc1C          6.30
1500  CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
1501  Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc…          5.06

     standard_type standard_relation standard_value standard_units  \
0               Kd                =         2000.0             nM
1             IC50                =           97.0             nM
2               Kd                =          401.0             nM
3             IC50                =          200.0             nM
4             IC50                =         3400.0             nM
…             …                   …             …             …
1497            Kd                =         2000.0             nM
1498            Kd                =          910.0             nM
1499            Kd                =          500.0             nM
1500            Kd                =          250.0             nM
1501            Kd                =         8700.0             nM

      potential_duplicate             target_pref_name target_organism  \
0                   False  Tyrosine-protein kinase TYK2    Homo sapiens
1                   False  Tyrosine-protein kinase TYK2    Homo sapiens
2                   False  Tyrosine-protein kinase TYK2    Homo sapiens
3                   False  Tyrosine-protein kinase TYK2    Homo sapiens
4                   False  Tyrosine-protein kinase TYK2    Homo sapiens
…                   …                              …              …
1497                False  Tyrosine-protein kinase TYK2    Homo sapiens
1498                False  Tyrosine-protein kinase TYK2    Homo sapiens
1499                False  Tyrosine-protein kinase TYK2    Homo sapiens
1500                False  Tyrosine-protein kinase TYK2    Homo sapiens
1501                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…
2             B  Kinobeads (epsilon), multiple immobilized ATP-…
3             B  Inhibition of GST-tagged TYK2 assessed as inhi…
4             B  Inhibition of GST-tagged TYK2 assessed as inhi…
…           …                                                 …
1497          B  Binding constant for TYK2(JH2domain-pseudokina…
1498          B  Binding constant for TYK2(JH2domain-pseudokina…
1499          B  Binding constant for TYK2(JH2domain-pseudokina…
1500          B  Binding constant for TYK2(JH1domain-catalytic)…
1501          B  Binding constant for TYK2(JH2domain-pseudokina…

      chembl_id_duplicate  mean_pchembl_value  max_pchembl_value  \
0                   False                 NaN                NaN
1                   False                 NaN                NaN
2                   False                 NaN                NaN
3                   False                 NaN                NaN
4                   False                 NaN                NaN
…                   …                     …                  …
1497                False                6.09               6.68
1498                False                6.93               7.82
1499                False                 NaN                NaN
1500                False                 NaN                NaN
1501                False                 NaN                NaN

      min_pchembl_value
0                   NaN
1                   NaN
2                   NaN
3                   NaN
4                   NaN
…                   …
1497               5.70
1498               6.04
1499                NaN
1500                NaN
1501                NaN

[1502 rows x 17 columns]
```

[23]:
```
#verify merge
merged_activ_df[merged_activ_df['molecule_chembl_id'] == 'CHEMBL4435170']
```

[23]:
```
     molecule_chembl_id molecule_pref_name  \
1171      CHEMBL4435170    DEUCRAVACITINIB

                                    canonical_smiles pchembl_value  \
```

```
1171  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70

     standard_type standard_relation standard_value standard_units  \
1171            Ki                 =           0.02             nM

      potential_duplicate              target_pref_name target_organism  \
1171                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                    assay_description  \
1171          B  Inhibition of fluorescein labeled probe bindin…

      chembl_id_duplicate  mean_pchembl_value  max_pchembl_value  \
1171                False                8.62               10.7

      min_pchembl_value
1171               7.75
```

[24]:
```
merged_activ_df_dc = copy.deepcopy(merged_activ_df)
```

[25]:
```
#fill NaNs in the aggregated columns with original pchembl values for the␣
 ↪original non-duplicated chembl ids

idxs = merged_activ_df[pd.isna(merged_activ_df['mean_pchembl_value'])].index
fill_NaN = merged_activ_df['pchembl_value'].iloc[idxs]
merged_activ_df['mean_pchembl_value'].iloc[idxs] = fill_NaN
merged_activ_df['max_pchembl_value'].iloc[idxs] = fill_NaN
merged_activ_df['min_pchembl_value'].iloc[idxs] = fill_NaN
merged_activ_df
```

```
c:\users\lac_2019\anaconda3\envs\deepchem\lib\site-
packages\pandas\core\indexing.py:1637: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
```

[25]:
```
     molecule_chembl_id molecule_pref_name  \
0              CHEMBL10          SB-203580
1         CHEMBL1076700               None
2         CHEMBL1078178        MOMELOTINIB
3         CHEMBL1080159               None
4         CHEMBL1081290               None
…                   …                  …
1497       CHEMBL601719         CRIZOTINIB
1498       CHEMBL603469       LESTAURTINIB
1499       CHEMBL608154               None
```

```
1500       CHEMBL608533        MIDOSTAURIN
1501          CHEMBL941           IMATINIB

                                      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
2     N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
3     Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
4           COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47
…                                                  …           …
1497  C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)…          5.70
1498  C[C@]12O[C@H](C[C@]1(O)CO)n1c3ccccc3c3c4c(c5c6…          6.04
1499      COc1c(Cl)cc2c([nH]c3cnccc32)c1NC(=O)c1cccnc1C          6.30
1500  CO[C@@H]1[C@H](N(C)C(=O)c2ccccc2)C[C@H]2O[C@]1…          6.60
1501  Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc…          5.06

     standard_type standard_relation standard_value standard_units  \
0               Kd                =         2000.0             nM
1             IC50                =           97.0             nM
2               Kd                =          401.0             nM
3             IC50                =          200.0             nM
4             IC50                =         3400.0             nM
…              …                  …             …             …
1497            Kd                =         2000.0             nM
1498            Kd                =          910.0             nM
1499            Kd                =          500.0             nM
1500            Kd                =          250.0             nM
1501            Kd                =         8700.0             nM

      potential_duplicate             target_pref_name target_organism  \
0                   False  Tyrosine-protein kinase TYK2    Homo sapiens
1                   False  Tyrosine-protein kinase TYK2    Homo sapiens
2                   False  Tyrosine-protein kinase TYK2    Homo sapiens
3                   False  Tyrosine-protein kinase TYK2    Homo sapiens
4                   False  Tyrosine-protein kinase TYK2    Homo sapiens
…                   …                               …              …
1497                False  Tyrosine-protein kinase TYK2    Homo sapiens
1498                False  Tyrosine-protein kinase TYK2    Homo sapiens
1499                False  Tyrosine-protein kinase TYK2    Homo sapiens
1500                False  Tyrosine-protein kinase TYK2    Homo sapiens
1501                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…
2             B  Kinobeads (epsilon), multiple immobilized ATP-…
3             B  Inhibition of GST-tagged TYK2 assessed as inhi…
```

```
4             B  Inhibition of GST-tagged TYK2 assessed as inhi…
…           …                                                 …
1497          B  Binding constant for TYK2(JH2domain-pseudokina…
1498          B  Binding constant for TYK2(JH2domain-pseudokina…
1499          B  Binding constant for TYK2(JH2domain-pseudokina…
1500          B  Binding constant for TYK2(JH1domain-catalytic)…
1501          B  Binding constant for TYK2(JH2domain-pseudokina…

      chembl_id_duplicate mean_pchembl_value max_pchembl_value  \
0                   False               5.70              5.70
1                   False               7.01              7.01
2                   False               6.40              6.40
3                   False               6.70              6.70
4                   False               5.47              5.47
…                   …                  …                 …
1497                False               6.09              6.68
1498                False               6.93              7.82
1499                False               6.30              6.30
1500                False               6.60              6.60
1501                False               5.06              5.06

     min_pchembl_value
0                 5.70
1                 7.01
2                 6.40
3                 6.70
4                 5.47
…                 …
1497               5.7
1498              6.04
1499              6.30
1500              6.60
1501              5.06

[1502 rows x 17 columns]
```

[26]:
```
idxs = merged_activ_df[pd.isna(merged_activ_df['mean_pchembl_value'])].index
idxs
```

[26]:
```
Int64Index([], dtype='int64')
```

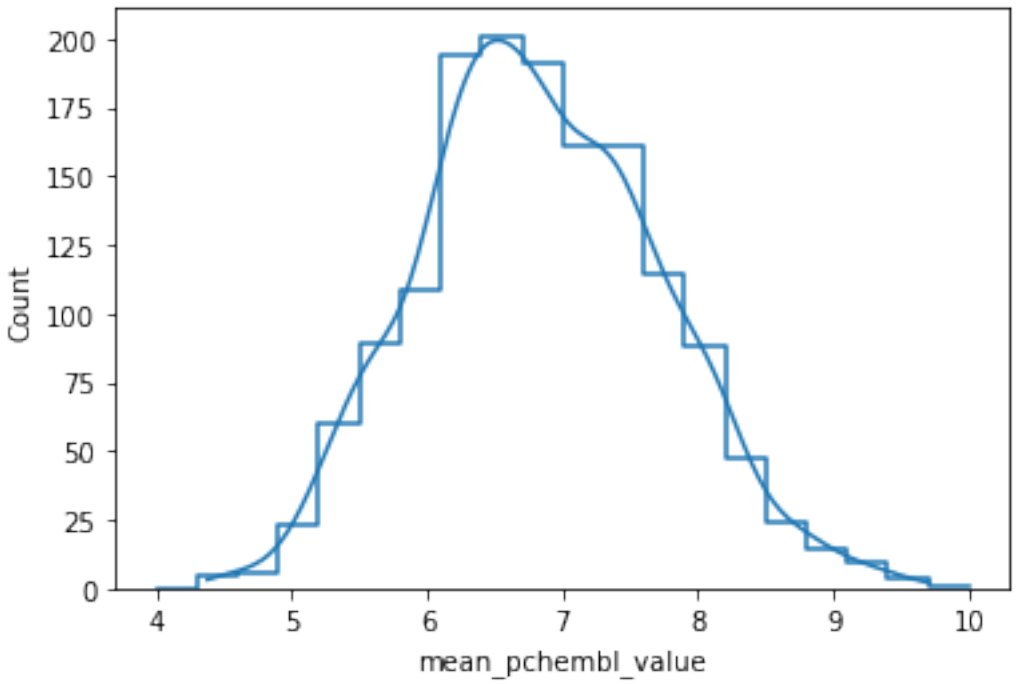
[27]:
```
#verify correct fill
merged_activ_df[merged_activ_df['molecule_chembl_id'] == 'CHEMBL4435170']
```

[27]:
```
     molecule_chembl_id molecule_pref_name  \
1171      CHEMBL4435170    DEUCRAVACITINIB
```

```
                                          canonical_smiles pchembl_value  \
1171  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70

     standard_type standard_relation standard_value standard_units  \
1171            Ki                 =           0.02             nM

      potential_duplicate             target_pref_name target_organism  \
1171                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                   assay_description  \
1171          B  Inhibition of fluorescein labeled probe bindin…

      chembl_id_duplicate mean_pchembl_value max_pchembl_value  \
1171                False               8.62              10.7

     min_pchembl_value
1171              7.75
```

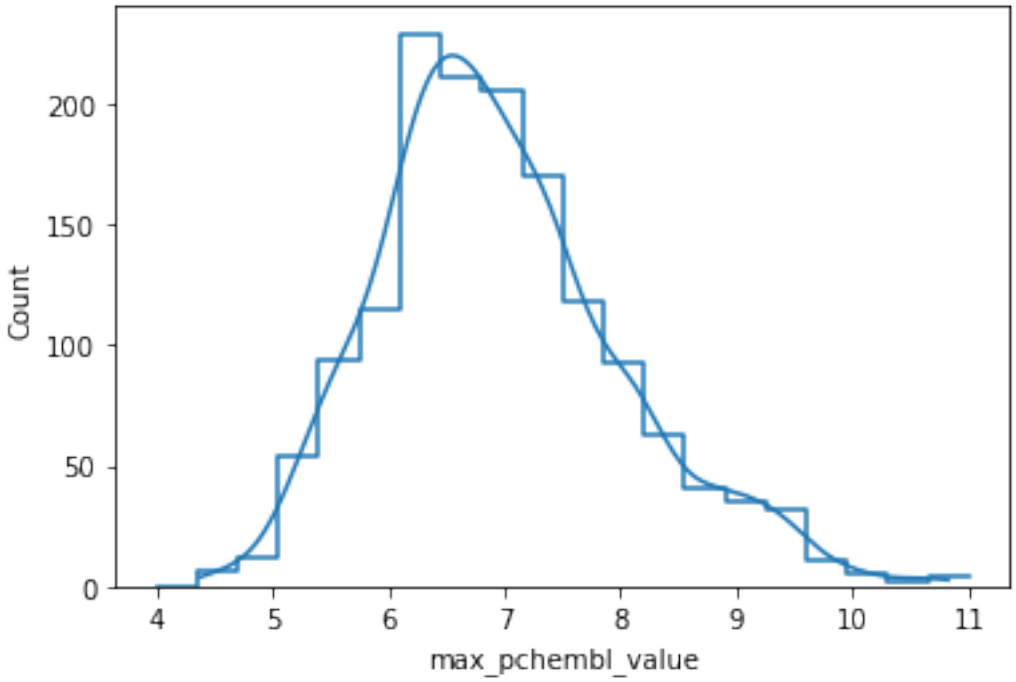
[28]:
```
sns.histplot(data=merged_activ_df, x=merged_activ_df['mean_pchembl_value'].
↪astype(float),
             binrange=[4,10], bins=20, element='step', fill=False, kde=True)
```

[28]:
```
<AxesSubplot:xlabel='mean_pchembl_value', ylabel='Count'>
```



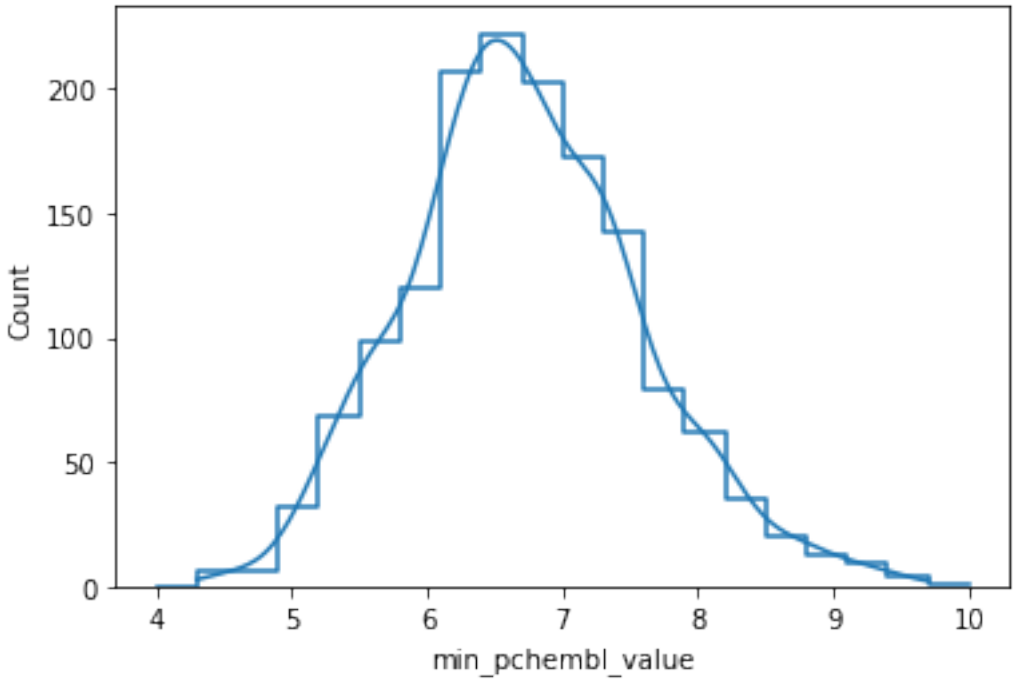
[29]:
```
sns.histplot(data=merged_activ_df, x=merged_activ_df['max_pchembl_value'].
↪astype(float),
             binrange=[4,11], bins=20, element='step', fill=False, kde=True)
```

[29]: <AxesSubplot:xlabel='max_pchembl_value', ylabel='Count'>



[30]:
```
sns.histplot(data=merged_activ_df, x=merged_activ_df['min_pchembl_value'].
↪astype(float),
             binrange=[4,10], bins=20, element='step', fill=False, kde=True)
```

[30]: <AxesSubplot:xlabel='min_pchembl_value', ylabel='Count'>



[31]:
```
from rdkit import Chem
from rdkit.Chem.Scaffolds import MurckoScaffold

mol_list = [Chem.MolFromSmiles(m) for m in merged_activ_df['canonical_smiles']]
scaffolds = []
for m in mol_list:
    try:
        core = MurckoScaffold.GetScaffoldForMol(m)
    except:
        continue
    scaffolds.append(core)

core_smiles = [Chem.MolToSmiles(core) for core in scaffolds]
merged_activ_df['core_smiles'] = core_smiles
merged_activ_df.head()
```

[31]:

	molecule_chembl_id	molecule_pref_name	canonical_smiles	pchembl_value
0	CHEMBL10	SB-203580	C[S+]([O-])c1ccc(-c2nc(-c3ccc(F)cc3)c(-c3ccncc…	5.70
1	CHEMBL1076700	None		
2	CHEMBL1078178	MOMELOTINIB		
3	CHEMBL1080159	None		
4	CHEMBL1081290	None		

```
1  Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          7.01
2  N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2…          6.40
3  Cc1cc(Nc2nc3cccc(-c4cc(F)c(CN5CCOCC5)c(F)c4)c3…          6.70
4        COc1cc(Nc2nc3cccc(-c4ccccc4)c3o2)cc(OC)c1OC          5.47

  standard_type standard_relation standard_value standard_units  \
0            Kd                =         2000.0             nM
1          IC50                =           97.0             nM
2            Kd                =          401.0             nM
3          IC50                =          200.0             nM
4          IC50                =         3400.0             nM

   potential_duplicate              target_pref_name target_organism  \
0                False  Tyrosine-protein kinase TYK2    Homo sapiens
1                False  Tyrosine-protein kinase TYK2    Homo sapiens
2                False  Tyrosine-protein kinase TYK2    Homo sapiens
3                False  Tyrosine-protein kinase TYK2    Homo sapiens
4                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…
2          B  Kinobeads (epsilon), multiple immobilized ATP-…
3          B  Inhibition of GST-tagged TYK2 assessed as inhi…
4          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
2                False               6.40             6.40             6.40
3                False               6.70             6.70             6.70
4                False               5.47             5.47             5.47

                                    core_smiles
0     c1ccc(-c2nc(-c3ccccc3)c(-c3ccncc3)[nH]2)cc1
1  c1ccc(Nc2nc3cccc(-c4ccc(CN5CCOCC5)cc4)c3o2)cc1
2        c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2)cc1
3  c1ccc(Nc2nc3cccc(-c4ccc(CN5CCOCC5)cc4)c3o2)cc1
4             c1ccc(Nc2nc3cccc(-c4ccccc4)c3o2)cc1
```

[32]:
```
merged_activ_df[merged_activ_df['molecule_chembl_id'] == 'CHEMBL4435170']
```

[32]:
```
      molecule_chembl_id molecule_pref_name  \
1171      CHEMBL4435170    DEUCRAVACITINIB

                                      canonical_smiles pchembl_value  \
1171  [2H]C([2H])([2H])NC(=O)c1nnc(NC(=O)C2CC2)cc1Nc…         10.70
```

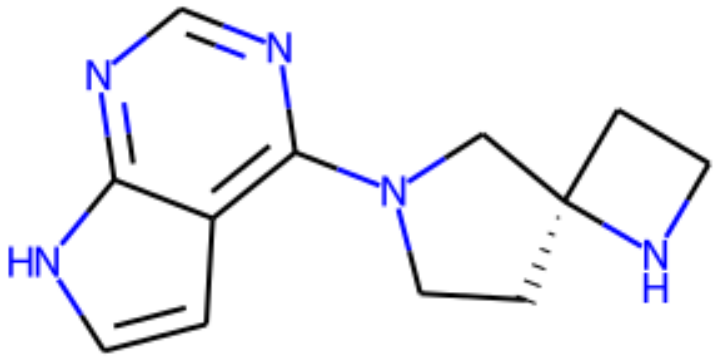
```
     standard_type standard_relation standard_value standard_units  \
1171            Ki                =           0.02             nM

      potential_duplicate             target_pref_name target_organism  \
1171                False  Tyrosine-protein kinase TYK2    Homo sapiens

     assay_type                                    assay_description  \
1171          B  Inhibition of fluorescein labeled probe bindin…

      chembl_id_duplicate mean_pchembl_value max_pchembl_value  \
1171                False               8.62              10.7

     min_pchembl_value                                      core_smiles
1171              7.75  O=C(Nc1cc(Nc2cccc(-c3nc[nH]n3)c2)cnn1)C1CC1
```

[33]:
```
merged_activ_df.iloc[1167]['core_smiles']
```

[33]:
```
'c1nc(N2CC[C@@]3(CCN3)C2)c2cc[nH]c2n1'
```

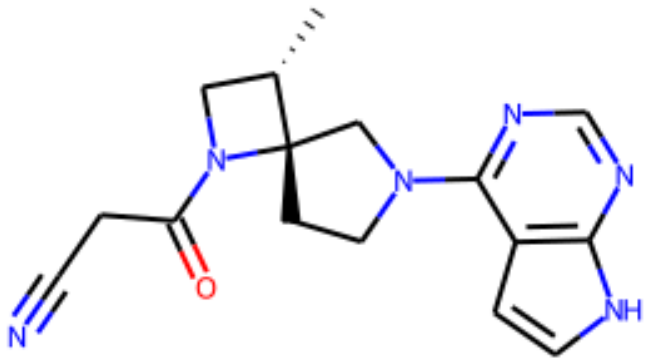
[34]:
```
Chem.MolFromSmiles(merged_activ_df.iloc[1167]['core_smiles'])
```

[34]:



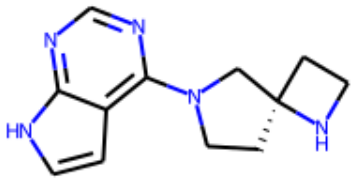
[35]:
```
Chem.MolFromSmiles(merged_activ_df.iloc[1167]['canonical_smiles'])
```

[35]:



[36]:
```
from rdkit.Chem.Draw import MolsToGridImage
from rdkit.Chem import PandasTools

deucravacitinib_core = Chem.
 ↪MolFromSmarts('O=C(Nc1cc(Nc2cccc(-c3nc[nH]n3)c2)cnn1)C1CC1')
deucravacitinib_core
#match_list = [m.GetSubstructMatch(deucravacitinib_core) for m in mol_list]
#match_list[1167]

#MolsToGridImage(mol_list, highlightAtomLists=match_list, molsPerRow=4,␣
 ↪maxMols=8)
deu = Chem.MolFromSmiles(merged_activ_df.iloc[1167]['core_smiles'])
match = [deu.GetSubstructMatch(deucravacitinib_core)]
MolsToGridImage([deu], highlightAtomLists=match)
```

[36]:



[37]:
```
#need to remove the H from smarts to work
core = Chem.MolFromSmarts('O=C(Nc1cc(Nc2cccc(-c3nc[nH]n3)c2)cnn1)C1CC1')
core
match_list = [m.GetSubstructMatch(core) for m in mol_list]
```

```
#[m for m in match_list if len(m) >0]
res = list(filter(lambda i: len(i) !=0, match_list))
res
```

[37]:
```
[]
```

[38]:
```
for atom in core.GetAtoms():
    if atom.GetIsAromatic() and atom.GetAtomicNum()!=6 and atom.
 ↪GetNumExplicitHs():
        atom.SetNoImplicit(True)
        atom.SetNumExplicitHs(0)
print(Chem.MolToSmarts(core))

match_list = [m.GetSubstructMatch(core) for m in mol_list]

#[m for m in match_list if len(m) >0]
match_list_filtered = list(filter(lambda i: len(i) !=0, match_list))
match_list_filtered
```

```
O=C(-,:Nc1cc(-,:Nc2cccc(-c3nc[n&H1]n3)c2)cnn1)C1CC1
```

[38]:
```
[]
```

[39]:
```
PandasTools.AddMoleculeColumnToFrame(merged_activ_df, 'canonical_smiles', 'Mol')
```

[40]:
```
PandasTools.AddMoleculeColumnToFrame(merged_activ_df, 'core_smiles', 'Scaffold')
```

[41]:
```
merged_activ_df[['Mol', 'Scaffold']].head()
```

[41]:

	Mol	Scaffold
0		
1		
2		
3		
4		

[42]:
```
merged_activ_df.to_pickle('../data/chembl_bioactivity_data.pkl')
```

[43]:
```
chembl_bioactivity_df = pd.read_pickle('../data/chembl_bioactivity_data.pkl')
chembl_bioactivity_df.head(2)
```

[43]:
```
  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…
```

[ ]: