Download PaDEL-Descriptor

In [1]:
```
#https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi to fetch sm
iles from pubchem id, also to convert pubchem to chemblids
! wget https://github.com/dataprofessor/bioinformatics/raw/master/p
adel.zip
! wget https://github.com/dataprofessor/bioinformatics/raw/master/p
adel.sh
```

```
'wget' is not recognized as an internal or external command,
operable program or batch file.
'wget' is not recognized as an internal or external command,
operable program or batch file.
```

In [2]:
```
!unzip padel.zip
```

```
'unzip' is not recognized as an internal or external command,
operable program or batch file.
```

Load bioactivity data

The curated ChEMBL bioactivity data that has been pre-processed from Parts 1 and 2 of this Project that essentially contain the pIC50 value will be used for similarity analysis

In [1]:
```
import pandas as pd
df1=pd.read_csv('fdb_classification_bioactivity_pIC50_onlyactive_in
active.csv')
df1.info()
df1.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   chembl_id           297 non-null    object
 1   molecule_pref_name  244 non-null    object
 2   target_chembl_id    297 non-null    object
 3   target_pref_name    297 non-null    object
```

```
 4   target_organism     243 non-null    object
 5   class               297 non-null    object
 6   pIC50               297 non-null    float64
 7   pubchem_id          297 non-null    int64
 8   molecular_weight    297 non-null    float64
 9   hbd_count           297 non-null    int64
 10  hba_count           297 non-null    int64
 11  xlogp               296 non-null    float64
 12  natural             297 non-null    int64
dtypes: float64(3), int64(4), object(6)
memory usage: 30.3+ KB
```

Out[1]:

	chembl_id	molecule_pref_name	target_chembl_id	target_pref_name	t
0	CHEMBL165	RESVERATROL	CHEMBL399	HeLa	
1	CHEMBL19224	PAPAVERINE	CHEMBL613633	Ileum	
2	CHEMBL50	QUERCETIN	CHEMBL2362975	No relevant target	
3	CHEMBL107	COLCHICINE	CHEMBL3879801	NON-PROTEIN TARGET	
4	CHEMBL441687	GLYCYRRHIZIN	CHEMBL3746	11-beta-hydroxysteroid dehydrogenase 2	

In [2]:

```
df_smi=pd.read_csv('fdb_297_onlyactive_inactive_smiles.csv')
df_smi.info()
df_smi.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   molecule_pref_name  297 non-null    object
 1   pubchem_id          297 non-null    int64
 2   smiles              297 non-null    object
dtypes: int64(1), object(2)
memory usage: 7.1+ KB
```

Out[2]:

	molecule_pref_name	pubchem_id	smiles
0	1-Aminopropan-2-ol	4	CC(CN)O
1	PROTOCATECHUIC ACID	72	C1=CC(=C(C=C1C(=O)O)O)O
2	PHENYL PROPIONIC ACID	107	C1=CC=C(C=C1)CCC(=O)O
3	GAMMA-AMINOBUTYRIC ACID	119	C(CC(=O)O)CN

	molecule_pref_name	pubchem_id	smiles
4	PARAHYDROXYBENZYL ALCOHOL	125	C1=CC(=CC=C1CO)O

In [3]:
```
df_chem_pub=df_smi.merge(df1, on=['pubchem_id'], how="inner")
df_chem_pub.head()
```

Out[3]:

	molecule_pref_name_x	pubchem_id	smiles	che
0	1-Aminopropan-2-ol	4	CC(CN)O	CHEMBL
1	PROTOCATECHUIC ACID	72	C1=CC(=C(C=C1C(=O)O)O)O	CHEMBI
2	PHENYL PROPIONIC ACID	107	C1=CC=C(C=C1)CCC(=O)O	CHEM
3	GAMMA-AMINOBUTYRIC ACID	119	C(CC(=O)O)CN	CHE
4	PARAHYDROXYBENZYL ALCOHOL	125	C1=CC(=CC=C1CO)O	CHEMBL

In [4]:
```
df_chem_pub.to_csv('pubchem_chembl_smiles_fbd_activity.csv', index=
False)
df_chem_pub.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 297 entries, 0 to 296
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   molecule_pref_name_x  297 non-null    object
 1   pubchem_id            297 non-null    int64
 2   smiles                297 non-null    object
 3   chembl_id             297 non-null    object
 4   molecule_pref_name_y  244 non-null    object
 5   target_chembl_id      297 non-null    object
 6   target_pref_name      297 non-null    object
 7   target_organism       243 non-null    object
 8   class                 297 non-null    object
 9   pIC50                 297 non-null    float64
 10  molecular_weight      297 non-null    float64
 11  hbd_count             297 non-null    int64
 12  hba_count             297 non-null    int64
 13  xlogp                 296 non-null    float64
 14  natural               297 non-null    int64
```

```
dtypes: float64(3), int64(4), object(8)
memory usage: 37.1+ KB
```

In [5]:
```
selection = ['smiles','chembl_id']
df2_selection = df_chem_pub[selection]
df2_selection.to_csv('molecule.smi', sep='\t', index=False, header=
False)
```

In [6]:
```
! cat molecule.smi | head -5
```

```
C[C@@]12CC[C@H]3[C@]([C@@]14[C@H](O4)C(=O)O[C@H]2C5=COC=C5)(C(=
O)C[C@@H]6[C@@]37COC(=O)C[C@@H]7OC6(C)C)C       CHEMBL517449

C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O        CHEMBL5
0

COC1=CC2=C(C=CN=C2C=C1)[C@H]([C@@H]3C[C@@H]4CCN3C[C@@H]4C=C)O
CHEMBL170

CC1=C2C[C@@H](CC[C@]2(CCC1)C)C(C)(C)O   CHEMBL477900

CC(=O)N[C@H]1CCC2=CC(=C(C(=C2C3=CC=C(C(=O)C=C13)OC)OC)OC)OC
CHEMBL107
```

In [7]:
```
! cat molecule.smi | wc -l
```

```
382
```

Calculate pubchem fingerprint descriptors using Padel Descriptor Software

Calculate PaDEL descriptors

In [11]:
```
! cat padel.sh
```

```
java -Xms1G -Xmx1G -Djava.awt.headless=true -jar ./PaDEL-Descri
ptor/PaDEL-Descriptor.jar -removesalt -standardizenitro -finger
prints -descriptortypes ./PaDEL-Descriptor/PubchemFingerprinte
r.xml -dir ./ -file descriptors_output.csv
```

In [12]:

```
! bash padel.sh
```

```
Processing CHEMBL517449 in molecule.smi (1/382).
Processing CHEMBL477900 in molecule.smi (4/382).
Processing CHEMBL50 in molecule.smi (2/382).
Processing CHEMBL170 in molecule.smi (3/382).
Processing CHEMBL130 in molecule.smi (7/382).
Processing CHEMBL107 in molecule.smi (5/382).
Processing CHEMBL165 in molecule.smi (8/382).
Processing CHEMBL150 in molecule.smi (6/382).
Processing CHEMBL297453 in molecule.smi (9/382).
Processing CHEMBL146 in molecule.smi (10/382).
Processing CHEMBL108861 in molecule.smi (11/382).
Processing CHEMBL47386 in molecule.smi (35/382).
Processing CHEMBL19224 in molecule.smi (12/382).
Processing CHEMBL44 in molecule.smi (13/382).
Processing CHEMBL532 in molecule.smi (14/382).
Processing CHEMBL545 in molecule.smi (15/382).
Processing CHEMBL48310 in molecule.smi (16/382).
Processing CHEMBL45967 in molecule.smi (17/382).
Processing CHEMBL441687 in molecule.smi (18/382).
Processing CHEMBL222021 in molecule.smi (19/382).
Processing CHEMBL1157 in molecule.smi (20/382).
Processing CHEMBL517016 in molecule.smi (21/382).
Processing CHEMBL196 in molecule.smi (22/382).
Processing CHEMBL311498 in molecule.smi (23/382).
Processing CHEMBL31574 in molecule.smi (24/382).
Processing CHEMBL583912 in molecule.smi (25/382).
Processing CHEMBL250450 in molecule.smi (26/382).
Processing CHEMBL151 in molecule.smi (27/382).
Processing CHEMBL7983 in molecule.smi (28/382).
Processing CHEMBL3 in molecule.smi (29/382).
Processing CHEMBL117 in molecule.smi (30/382).
Processing CHEMBL1672002 in molecule.smi (31/382).
Processing CHEMBL416 in molecule.smi (32/382).
Processing CHEMBL576 in molecule.smi (33/382).
Processing CHEMBL508894 in molecule.smi (34/382).
Processing CHEMBL16645 in molecule.smi (36/382).
Processing CHEMBL28 in molecule.smi (37/382). Average speed: 3.
92 s/mol.
Processing CHEMBL7303 in molecule.smi (38/382). Average speed:
3.99 s/mol.
Processing CHEMBL286727 in molecule.smi (39/382). Average spee
d: 2.18 s/mol.
Processing CHEMBL125743 in molecule.smi (40/382). Average spee
d: 1.11 s/mol.
Processing CHEMBL506678 in molecule.smi (41/382). Average spee
d: 0.89 s/mol.
```

```
Processing CHEMBL36327 in molecule.smi (42/382). Average speed:
0.89 s/mol.
Processing CHEMBL221542 in molecule.smi (43/382). Average spee
d: 0.75 s/mol.
Processing CHEMBL465829 in molecule.smi (44/382). Average spee
d: 0.68 s/mol.
Processing CHEMBL45005 in molecule.smi (45/382). Average speed:
0.61 s/mol.
Processing CHEMBL294431 in molecule.smi (46/382). Average spee
d: 0.52 s/mol.
Processing CHEMBL441343 in molecule.smi (47/382). Average spee
d: 0.50 s/mol.
Processing CHEMBL210635 in molecule.smi (48/382). Average spee
d: 0.50 s/mol.
Processing CHEMBL186141 in molecule.smi (50/382). Average spee
d: 0.39 s/mol.
Processing CHEMBL93353 in molecule.smi (49/382). Average speed:
0.42 s/mol.
Processing CHEMBL129795 in molecule.smi (51/382). Average spee
d: 0.37 s/mol.
Processing CHEMBL120568 in molecule.smi (52/382). Average spee
d: 0.37 s/mol.
Processing CHEMBL245067 in molecule.smi (53/382). Average spee
d: 0.34 s/mol.
Processing CHEMBL486625 in molecule.smi (54/382). Average spee
d: 0.34 s/mol.
Processing CHEMBL239243 in molecule.smi (55/382). Average spee
d: 0.33 s/mol.
Processing CHEMBL8145 in molecule.smi (56/382). Average speed:
0.31 s/mol.
Processing CHEMBL275638 in molecule.smi (57/382). Average spee
d: 0.29 s/mol.
Processing CHEMBL82293 in molecule.smi (58/382). Average speed:
0.28 s/mol.
Processing CHEMBL413552 in molecule.smi (60/382). Average spee
d: 0.27 s/mol.
Processing CHEMBL15245 in molecule.smi (59/382). Average speed:
0.27 s/mol.
Processing CHEMBL752 in molecule.smi (61/382). Average speed:
0.27 s/mol.
Processing CHEMBL170190 in molecule.smi (62/382). Average spee
d: 0.27 s/mol.
Processing CHEMBL13883 in molecule.smi (63/382). Average speed:
0.25 s/mol.
Processing CHEMBL96 in molecule.smi (64/382). Average speed: 0.
24 s/mol.
Processing CHEMBL9352 in molecule.smi (65/382). Average speed:
0.24 s/mol.
```

```
Processing CHEMBL267476 in molecule.smi (67/382). Average spee
d: 0.23 s/mol.
Processing CHEMBL364713 in molecule.smi (66/382). Average spee
d: 0.23 s/mol.
Processing CHEMBL25308 in molecule.smi (68/382). Average speed:
0.21 s/mol.
Processing CHEMBL575060 in molecule.smi (69/382). Average spee
d: 0.22 s/mol.
Processing CHEMBL1232207 in molecule.smi (71/382). Average spee
d: 0.21 s/mol.
Processing CHEMBL70518 in molecule.smi (70/382). Average speed:
0.21 s/mol.
Processing CHEMBL537 in molecule.smi (73/382). Average speed:
0.20 s/mol.
Processing CHEMBL226507 in molecule.smi (72/382). Average spee
d: 0.20 s/mol.
Processing CHEMBL14152 in molecule.smi (75/382). Average speed:
0.19 s/mol.
Processing CHEMBL225303 in molecule.smi (74/382). Average spee
d: 0.19 s/mol.
Processing CHEMBL14193 in molecule.smi (76/382). Average speed:
0.19 s/mol.
Processing CHEMBL53566 in molecule.smi (77/382). Average speed:
0.18 s/mol.
Processing CHEMBL8659 in molecule.smi (78/382). Average speed:
0.18 s/mol.
Processing CHEMBL274323 in molecule.smi (79/382). Average spee
d: 0.18 s/mol.
Processing CHEMBL930 in molecule.smi (80/382). Average speed:
0.18 s/mol.
Processing CHEMBL66879 in molecule.smi (81/382). Average speed:
0.17 s/mol.
Processing CHEMBL559945 in molecule.smi (82/382). Average spee
d: 0.17 s/mol.
Processing CHEMBL28626 in molecule.smi (83/382). Average speed:
0.17 s/mol.
Processing CHEMBL55415 in molecule.smi (84/382). Average speed:
0.17 s/mol.
Processing CHEMBL14474 in molecule.smi (85/382). Average speed:
0.17 s/mol.
Processing CHEMBL388558 in molecule.smi (86/382). Average spee
d: 0.16 s/mol.
Processing CHEMBL1485 in molecule.smi (87/382). Average speed:
0.16 s/mol.
Processing CHEMBL17962 in molecule.smi (89/382). Average speed:
0.16 s/mol.
Processing CHEMBL54976 in molecule.smi (88/382). Average speed:
0.16 s/mol.
```

```
Processing CHEMBL379064 in molecule.smi (90/382). Average spee
d: 0.16 s/mol.
Processing CHEMBL12014 in molecule.smi (91/382). Average speed:
0.15 s/mol.
Processing CHEMBL1591973 in molecule.smi (92/382). Average spee
d: 0.15 s/mol.
Processing CHEMBL3186027 in molecule.smi (93/382). Average spee
d: 0.15 s/mol.
Processing CHEMBL66926 in molecule.smi (94/382). Average speed:
0.15 s/mol.
Processing CHEMBL294199 in molecule.smi (95/382). Average spee
d: 0.15 s/mol.
Processing CHEMBL485259 in molecule.smi (96/382). Average spee
d: 0.15 s/mol.
Processing CHEMBL45068 in molecule.smi (97/382). Average speed:
0.14 s/mol.
Processing CHEMBL76447 in molecule.smi (99/382). Average speed:
0.14 s/mol.
Processing CHEMBL29757 in molecule.smi (98/382). Average speed:
0.14 s/mol.
Processing CHEMBL463088 in molecule.smi (100/382). Average spee
d: 0.14 s/mol.
Processing CHEMBL242273 in molecule.smi (101/382). Average spee
d: 0.14 s/mol.
Processing CHEMBL253896 in molecule.smi (102/382). Average spee
d: 0.14 s/mol.
Processing CHEMBL541 in molecule.smi (103/382). Average speed:
0.14 s/mol.
Processing CHEMBL49732 in molecule.smi (104/382). Average spee
d: 0.14 s/mol.
Processing CHEMBL113 in molecule.smi (105/382). Average speed:
0.13 s/mol.
Processing CHEMBL42710 in molecule.smi (107/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL333306 in molecule.smi (106/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL3356397 in molecule.smi (108/382). Average spe
ed: 0.13 s/mol.
Processing CHEMBL27246 in molecule.smi (109/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL73930 in molecule.smi (110/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL395827 in molecule.smi (111/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL1453648 in molecule.smi (112/382). Average spe
ed: 0.13 s/mol.
Processing CHEMBL111077 in molecule.smi (113/382). Average spee
d: 0.13 s/mol.
```

```
Processing CHEMBL151649 in molecule.smi (114/382). Average spee
d: 0.13 s/mol.
Processing CHEMBL105912 in molecule.smi (115/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL301523 in molecule.smi (117/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL25894 in molecule.smi (116/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL541939 in molecule.smi (118/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL510714 in molecule.smi (119/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL573781 in molecule.smi (120/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL308187 in molecule.smi (121/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL328441 in molecule.smi (122/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL274467 in molecule.smi (123/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL201083 in molecule.smi (124/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL21932 in molecule.smi (125/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL450288 in molecule.smi (126/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL185885 in molecule.smi (127/382). Average spee
d: 0.12 s/mol.
Processing CHEMBL66 in molecule.smi (128/382). Average speed:
0.12 s/mol.
Processing CHEMBL1950582 in molecule.smi (129/382). Average spe
ed: 0.12 s/mol.
Processing CHEMBL232202 in molecule.smi (130/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL451532 in molecule.smi (131/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL25719 in molecule.smi (133/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL486422 in molecule.smi (132/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL108862 in molecule.smi (134/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL8320 in molecule.smi (135/382). Average speed:
0.11 s/mol.
Processing CHEMBL328910 in molecule.smi (136/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL1369384 in molecule.smi (137/382). Average spe
ed: 0.11 s/mol.
```

```
Processing CHEMBL23194 in molecule.smi (138/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL3883497 in molecule.smi (139/382). Average spe
ed: 0.11 s/mol.
Processing CHEMBL14060 in molecule.smi (140/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL52267 in molecule.smi (141/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL424 in molecule.smi (142/382). Average speed:
0.11 s/mol.
Processing CHEMBL15844 in molecule.smi (143/382). Average spee
d: 0.11 s/mol.
Processing CHEMBL2268549 in molecule.smi (144/382). Average spe
ed: 0.11 s/mol.
Processing CHEMBL37537 in molecule.smi (145/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL851 in molecule.smi (146/382). Average speed:
0.10 s/mol.
Processing CHEMBL55285 in molecule.smi (148/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL109341 in molecule.smi (147/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL43185 in molecule.smi (149/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL32571 in molecule.smi (150/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL1644111 in molecule.smi (151/382). Average spe
ed: 0.10 s/mol.
Processing CHEMBL452683 in molecule.smi (152/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL297569 in molecule.smi (153/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL105424 in molecule.smi (154/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL506247 in molecule.smi (155/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL390773 in molecule.smi (156/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL1093743 in molecule.smi (157/382). Average spe
ed: 0.10 s/mol.
Processing CHEMBL6466 in molecule.smi (158/382). Average speed:
0.10 s/mol.
Processing CHEMBL1182 in molecule.smi (159/382). Average speed:
0.10 s/mol.
Processing CHEMBL1661 in molecule.smi (160/382). Average speed:
0.10 s/mol.
Processing CHEMBL13766 in molecule.smi (161/382). Average spee
d: 0.10 s/mol.
```

```
Processing CHEMBL82411 in molecule.smi (162/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL202132 in molecule.smi (163/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL90039 in molecule.smi (164/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL108925 in molecule.smi (165/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL15134 in molecule.smi (166/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL347285 in molecule.smi (167/382). Average spee
d: 0.10 s/mol.
Processing CHEMBL248594 in molecule.smi (168/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL358850 in molecule.smi (169/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL417016 in molecule.smi (170/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL189362 in molecule.smi (171/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL460657 in molecule.smi (172/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL161598 in molecule.smi (173/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL18850 in molecule.smi (174/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL1162144 in molecule.smi (176/382). Average spe
ed: 0.09 s/mol.
Processing CHEMBL18360 in molecule.smi (175/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL18620 in molecule.smi (178/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL462997 in molecule.smi (177/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL303697 in molecule.smi (179/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL14227 in molecule.smi (180/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL324794 in molecule.smi (181/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL54922 in molecule.smi (182/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL503160 in molecule.smi (183/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL25028 in molecule.smi (184/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL1354 in molecule.smi (185/382). Average speed:
0.09 s/mol.
```

```
Processing CHEMBL108545 in molecule.smi (186/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL1229937 in molecule.smi (187/382). Average spe
ed: 0.09 s/mol.
Processing CHEMBL300520 in molecule.smi (188/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL491174 in molecule.smi (189/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL2071440 in molecule.smi (190/382). Average spe
ed: 0.09 s/mol.
Processing CHEMBL469654 in molecule.smi (192/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL281202 in molecule.smi (191/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL1239 in molecule.smi (193/382). Average speed:
0.09 s/mol.
Processing CHEMBL396295 in molecule.smi (194/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL504 in molecule.smi (195/382). Average speed:
0.09 s/mol.
Processing CHEMBL18104 in molecule.smi (197/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL128000 in molecule.smi (196/382). Average spee
d: 0.09 s/mol.
Processing CHEMBL346919 in molecule.smi (198/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL610 in molecule.smi (199/382). Average speed:
0.08 s/mol.
Processing CHEMBL30707 in molecule.smi (200/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL1232797 in molecule.smi (201/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL4210821 in molecule.smi (202/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL3126829 in molecule.smi (203/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL233248 in molecule.smi (204/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL365740 in molecule.smi (206/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL46403 in molecule.smi (205/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL383808 in molecule.smi (207/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL1276010 in molecule.smi (208/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL214321 in molecule.smi (209/382). Average spee
d: 0.08 s/mol.
```

```
Processing CHEMBL24147 in molecule.smi (210/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL218693 in molecule.smi (211/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL293492 in molecule.smi (212/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL161577 in molecule.smi (214/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL863 in molecule.smi (213/382). Average speed:
0.08 s/mol.
Processing CHEMBL558557 in molecule.smi (215/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL47244 in molecule.smi (216/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL561014 in molecule.smi (217/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL195895 in molecule.smi (218/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL1234268 in molecule.smi (219/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL401912 in molecule.smi (220/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL453509 in molecule.smi (222/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL539 in molecule.smi (221/382). Average speed:
0.08 s/mol.
Processing CHEMBL370688 in molecule.smi (223/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL89306 in molecule.smi (224/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL462861 in molecule.smi (225/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL1547 in molecule.smi (226/382). Average speed:
0.08 s/mol.
Processing CHEMBL225153 in molecule.smi (227/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL470671 in molecule.smi (228/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL29411 in molecule.smi (229/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL249592 in molecule.smi (230/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL445206 in molecule.smi (231/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL73639 in molecule.smi (232/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL83159 in molecule.smi (233/382). Average spee
d: 0.08 s/mol.
```

```
Processing CHEMBL925 in molecule.smi (234/382). Average speed:
0.08 s/mol.
Processing CHEMBL1224557 in molecule.smi (235/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL247484 in molecule.smi (236/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL14021 in molecule.smi (237/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL46931 in molecule.smi (238/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL773 in molecule.smi (240/382). Average speed:
0.08 s/mol.
Processing CHEMBL1974890 in molecule.smi (239/382). Average spe
ed: 0.08 s/mol.
Processing CHEMBL689 in molecule.smi (241/382). Average speed:
0.08 s/mol.
Processing CHEMBL227934 in molecule.smi (242/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL25306 in molecule.smi (243/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL16217 in molecule.smi (245/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL452630 in molecule.smi (244/382). Average spee
d: 0.08 s/mol.
Processing CHEMBL574688 in molecule.smi (246/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL88244 in molecule.smi (247/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL445740 in molecule.smi (248/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL401911 in molecule.smi (249/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL195593 in molecule.smi (250/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL472877 in molecule.smi (251/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL320358 in molecule.smi (252/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL29966 in molecule.smi (253/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL3186705 in molecule.smi (255/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL3187012 in molecule.smi (254/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL285123 in molecule.smi (256/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL11608 in molecule.smi (257/382). Average spee
d: 0.07 s/mol.
```

```
Processing CHEMBL14092 in molecule.smi (258/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL108475 in molecule.smi (259/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1911053 in molecule.smi (260/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL460647 in molecule.smi (261/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL108299 in molecule.smi (263/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL263094 in molecule.smi (262/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL47127 in molecule.smi (264/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL15972 in molecule.smi (265/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL66693 in molecule.smi (266/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL276218 in molecule.smi (267/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL153339 in molecule.smi (268/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1814589 in molecule.smi (269/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL448502 in molecule.smi (270/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL205268 in molecule.smi (271/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL173521 in molecule.smi (272/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL366603 in molecule.smi (273/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL449062 in molecule.smi (274/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL256087 in molecule.smi (276/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL399036 in molecule.smi (275/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL481044 in molecule.smi (278/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL71595 in molecule.smi (277/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL108766 in molecule.smi (279/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL325372 in molecule.smi (280/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL253582 in molecule.smi (281/382). Average spee
d: 0.07 s/mol.
```

```
Processing CHEMBL298312 in molecule.smi (283/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL350966 in molecule.smi (282/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL25424 in molecule.smi (284/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL304461 in molecule.smi (285/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL16293 in molecule.smi (286/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL18407 in molecule.smi (287/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL95973 in molecule.smi (288/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1232258 in molecule.smi (289/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL1096927 in molecule.smi (290/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL486193 in molecule.smi (291/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1814588 in molecule.smi (293/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL272485 in molecule.smi (292/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL22585 in molecule.smi (294/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL330546 in molecule.smi (295/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL492828 in molecule.smi (297/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1200559 in molecule.smi (296/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL430341 in molecule.smi (298/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL430091 in molecule.smi (300/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL30018 in molecule.smi (299/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL242383 in molecule.smi (301/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL110309 in molecule.smi (302/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL333179 in molecule.smi (304/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL237994 in molecule.smi (303/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL107498 in molecule.smi (305/382). Average spee
d: 0.07 s/mol.
```

```
Processing CHEMBL1222250 in molecule.smi (306/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL1261 in molecule.smi (307/382). Average speed:
0.07 s/mol.
Processing CHEMBL1233860 in molecule.smi (308/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL440161 in molecule.smi (309/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1614854 in molecule.smi (310/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL2229207 in molecule.smi (311/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL191935 in molecule.smi (312/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL280331 in molecule.smi (313/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL442915 in molecule.smi (315/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL318196 in molecule.smi (314/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL170458 in molecule.smi (316/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL56395 in molecule.smi (317/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL2059292 in molecule.smi (319/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL2251610 in molecule.smi (318/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL446299 in molecule.smi (320/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL18602 in molecule.smi (321/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL154155 in molecule.smi (322/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL2268550 in molecule.smi (323/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL1076637 in molecule.smi (324/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL292303 in molecule.smi (325/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL31561 in molecule.smi (326/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL450072 in molecule.smi (327/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL14184 in molecule.smi (328/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL2105350 in molecule.smi (330/382). Average spe
ed: 0.07 s/mol.
```

```
Processing CHEMBL15605 in molecule.smi (329/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL504760 in molecule.smi (331/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL266158 in molecule.smi (332/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL1187 in molecule.smi (333/382). Average speed:
0.07 s/mol.
Processing CHEMBL146755 in molecule.smi (334/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL3617994 in molecule.smi (335/382). Average spe
ed: 0.07 s/mol.
Processing CHEMBL1114 in molecule.smi (336/382). Average speed:
0.07 s/mol.
Processing CHEMBL473366 in molecule.smi (337/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL276849 in molecule.smi (338/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL3183500 in molecule.smi (339/382). Average spe
ed: 0.06 s/mol.
Processing CHEMBL108030 in molecule.smi (341/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL264141 in molecule.smi (340/382). Average spee
d: 0.07 s/mol.
Processing CHEMBL195827 in molecule.smi (342/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL379630 in molecule.smi (344/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL104875 in molecule.smi (345/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL361197 in molecule.smi (347/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL190927 in molecule.smi (346/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL298717 in molecule.smi (348/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL365316 in molecule.smi (343/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL721 in molecule.smi (349/382). Average speed:
0.06 s/mol.
Processing CHEMBL108778 in molecule.smi (350/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL1162495 in molecule.smi (351/382). Average spe
ed: 0.06 s/mol.
Processing CHEMBL9113 in molecule.smi (352/382). Average speed:
0.06 s/mol.
Processing CHEMBL251280 in molecule.smi (353/382). Average spee
d: 0.06 s/mol.
```

```
Processing CHEMBL486795 in molecule.smi (354/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL1044 in molecule.smi (355/382). Average speed:
0.06 s/mol.
Processing CHEMBL31422 in molecule.smi (356/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL291962 in molecule.smi (357/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL470874 in molecule.smi (358/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL123040 in molecule.smi (359/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL460025 in molecule.smi (360/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL2368547 in molecule.smi (362/382). Average spe
ed: 0.06 s/mol.
Processing CHEMBL1233058 in molecule.smi (361/382). Average spe
ed: 0.06 s/mol.
Processing CHEMBL324846 in molecule.smi (363/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL273782 in molecule.smi (364/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL3360549 in molecule.smi (365/382). Average spe
ed: 0.06 s/mol.
Processing CHEMBL195215 in molecule.smi (366/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL8085 in molecule.smi (367/382). Average speed:
0.06 s/mol.
Processing CHEMBL510309 in molecule.smi (368/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL18549 in molecule.smi (369/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL18893 in molecule.smi (370/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL487213 in molecule.smi (371/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL277871 in molecule.smi (372/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL14253 in molecule.smi (373/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL118504 in molecule.smi (374/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL198877 in molecule.smi (376/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL573448 in molecule.smi (375/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL470670 in molecule.smi (377/382). Average spee
d: 0.06 s/mol.
```

```
Processing CHEMBL107874 in molecule.smi (378/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL326602 in molecule.smi (379/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL271663 in molecule.smi (381/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL108436 in molecule.smi (380/382). Average spee
d: 0.06 s/mol.
Processing CHEMBL192458 in molecule.smi (382/382). Average spee
d: 0.06 s/mol.
Descriptor calculation completed in 23.678 secs . Average spee
d: 0.06 s/mol.
```

In [6]:

```
df3_X = pd.read_csv('descriptors_output.csv')
df3_X.rename(columns = {'Name':'chembl_id'}, inplace = True)
```

In [7]:

```
df3_X
```

Out[7]:

	chembl_id	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3
0	CHEMBL545	1	0	0	0
1	CHEMBL576	1	0	0	0
2	CHEMBL1157	1	1	0	0
3	CHEMBL3	1	1	0	0
4	CHEMBL196	1	1	0	0
...	...	...	...	...	...
377	CHEMBL3617994	1	1	1	0
378	CHEMBL430341	1	1	1	0
379	CHEMBL110309	1	1	1	1
380	CHEMBL2368547	1	1	1	1
381	CHEMBL506247	1	1	1	1

382 rows × 882 columns

In [8]:

```
selection_2 = ['chembl_id','class']
df3_selection = df_chem_pub[selection_2]
df4_pca=df3_X.merge(df3_selection,on=['chembl_id'],how="inner")
df5_tc=df4_pca.drop(['chembl_id','class'], axis =1)
df4_pca.info()
df4_pca.head()
#df5_tc.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 297 entries, 0 to 296
Columns: 883 entries, chembl_id to class
dtypes: int64(881), object(2)
memory usage: 2.0+ MB
```

Out[8]:

	chembl_id	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	Pu
0	CHEMBL576	1	0	0	0	
1	CHEMBL1157	1	1	0	0	
2	CHEMBL3	1	1	0	0	
3	CHEMBL196	1	1	0	0	
4	CHEMBL48310	1	1	0	0	

5 rows × 883 columns

Tanimoto Coefficient

In [9]:

```
df4_pca['pubchem_fp'] = df5_tc[df5_tc.columns[:]].apply(lambda x:
''.join(x.dropna().astype(str)),axis=1)
df4_pca.head()
```

Out[9]:

	chembl_id	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	Pu
0	CHEMBL576	1	0	0	0	
1	CHEMBL1157	1	1	0	0	
2	CHEMBL3	1	1	0	0	
3	CHEMBL196	1	1	0	0	
4	CHEMBL48310	1	1	0	0	

5 rows × 884 columns

Making Pairs of Query and Target for Tanimoto Pairwise Similarity Calculation

In [10]:

```
#df_pca= df_pca.set_index("chembl_id")
import itertools
cc1 = list(itertools.combinations(df4_pca['chembl_id'],2))
df_cc=pd.DataFrame(data=cc1,columns=['Query','Target'])
df_cc.head()
```

Out[10]:

	Query	Target
0	CHEMBL576	CHEMBL1157
1	CHEMBL576	CHEMBL3
2	CHEMBL576	CHEMBL196
3	CHEMBL576	CHEMBL48310
4	CHEMBL576	CHEMBL7303

In [11]:

```
#df4_pca['pubchem_fp'].[df_pca]
df4_pca.loc[df4_pca['chembl_id'] == 'CHEMBL576', 'pubchem_fp'].iloc
[0]
```

Out[11]:

```
'1000000001100000001110000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000011010000000000000000000000100000
0000000000000000001000000000001000000010000000000000000000000000
0001000000000000000000000000001000000000000010000000000000000000
1001000000001000000000000000000000000000000000000000000000000000
0000000000000000000000000000000010000000000000000000000000000000
0000000000000100000000000000000000000000000000000100000000000000
0000000010000000000000000000000000000000000000000010000110000
0000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000
0'
```

Install Rdkit package

In [18]:

```
! conda install -c rdkit rdkit -y
```

```
Collecting package metadata (current_repodata.json): ...workin
g... done
Solving environment: ...working... done

# All requested packages already installed.
```

Generating Bitvectors from Pubchem Fingerprints

In [12]:

```
from rdkit import DataStructs
```

```
for i in range(len(df_cc)):
    ref = df_cc.Query[i]
    bit=df4_pca.loc[df4_pca['chembl_id'] == ref, 'pubchem_fp'].iloc
[0]
    bitvect = DataStructs.CreateFromBitString(bit)
```

In [13]:

```
df_chem_pub.head()
```

Out[13]:

	molecule_pref_name_x	pubchem_id	smiles	che
0	1-Aminopropan-2-ol	4	CC(CN)O	CHEMBL3
1	PROTOCATECHUIC ACID	72	C1=CC(=C(C=C1C(=O)O)O)O	CHEMBI
2	PHENYL PROPIONIC ACID	107	C1=CC=C(C=C1)CCC(=O)O	CHEM
3	GAMMA-AMINOBUTYRIC ACID	119	C(CC(=O)O)CN	CHE
4	PARAHYDROXYBENZYL ALCOHOL	125	C1=CC(=CC=C1CO)O	CHEMBL2

Displaying chemical strs. of 297 active and inactive flavor molecules

In [14]:

```
import requests
import time
from rdkit import Chem
from rdkit.Chem import Draw
mols = [ Chem.MolFromSmiles(x) for x in df_chem_pub['smiles'] ]
Chem.Draw.MolsToGridImage(mols, molsPerRow=4, subImgSize=(200,200),
legends=[str(x) for x in df_chem_pub['pubchem_id']] )
```

```
C:\Users\91809\Anaconda3\lib\site-packages\rdkit\Chem\Draw\IPyt
honConsole.py:188: UserWarning: Truncating the list of molecule
s to be displayed to 50. Change the maxMols value to display mo
re.
  % (maxMols))
```

Out[14]:

4

72

107

119

125

135

176

196

239

243

247

323

326

335

338

342

379

454

460

473

612

660

670

679

750

784

802

931

971

985

996

999

1004

1032

1049

1060

1103

1110

1130

1140

1254

1491

2230

2236

2346

2757

2758

2879

2969

3026

In [17]:
```
df_cc.head()
```

Out[17]:

	Query	Target
0	CHEMBL576	CHEMBL1157
1	CHEMBL576	CHEMBL3
2	CHEMBL576	CHEMBL196
3	CHEMBL576	CHEMBL48310
4	CHEMBL576	CHEMBL7303

In [18]:
```
#type(fp1)
scores = []
from rdkit import DataStructs
for i in range(len(df_cc)):
    ref = df_cc.Query[i]
    comp=df_cc.Target[i]
    bit_ref=df4_pca.loc[df4_pca['chembl_id'] == ref, 'pubchem_fp'].
iloc[0]
    bit_comp=df4_pca.loc[df4_pca['chembl_id'] == comp, 'pubchem_fp'
].iloc[0]
    bitvect_ref = DataStructs.CreateFromBitString(bit_ref)
    bitvect_comp = DataStructs.CreateFromBitString(bit_comp)
    fps_ref=bitvect_ref
    fps_comp=bitvect_comp
    #print(ref,comp)
   # print("Tanimoto    :", round(DataStructs.TanimotoSimilarity(fp
s_ref, fps_comp), 4))
    scores.append(DataStructs.TanimotoSimilarity(fps_ref, fps_comp
))
```

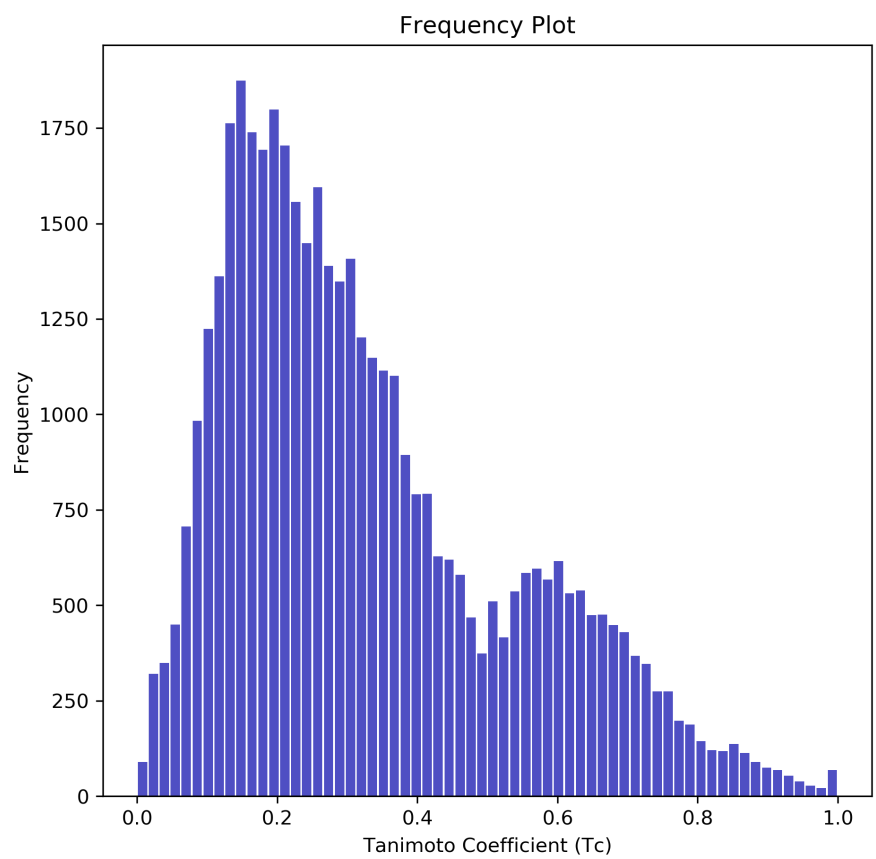
In [34]:
```
len(scores), len(df_cc)
```

Out[34]:
```
(43956, 43956)
```

In [ ]:
```
plt.figure(figsize=(7,7))

sns.countplot(x='natural', data=df_fla,palette=['gold','cornflowerb
lue'])

plt.xlabel('Type of Flavor molecule', fontsize=14, fontweight='bol
d')
plt.ylabel('Frequency', fontsize=14, fontweight='bold')

plt.savefig('plot_synthtic_nativecount_class.png')
```

In [46]:

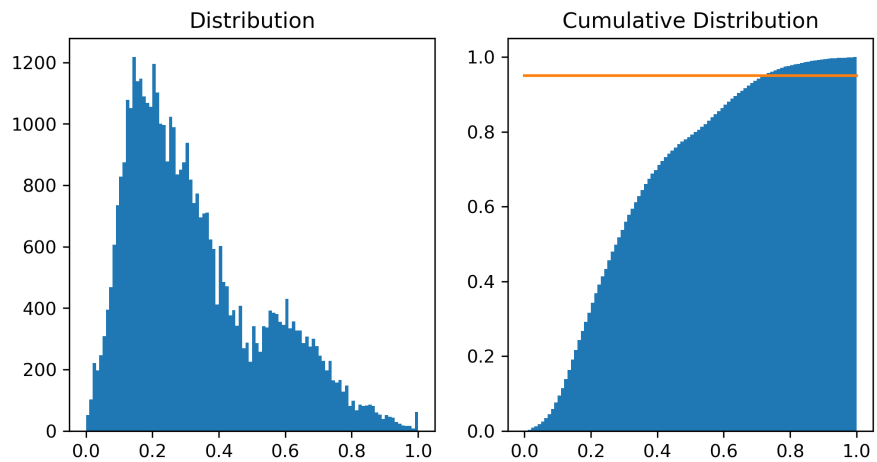
```
import matplotlib.pyplot as plt
fig = plt.figure(figsize=(7,7), dpi=300)
# An "interface" to matplotlib.axes.Axes.hist() method
plt.hist(x=scores, bins='auto', color='#0504aa', alpha=0.7, rwidth=
0.85)
#plt.grid(axis='y', alpha=0.75)

plt.xlabel('Tanimoto Coefficient (Tc)')
plt.ylabel('Frequency')
plt.title('Frequency Plot')
#plt.figure(figsize=(7,7))
plt.savefig('plot_tc_nativecount_class.png')
#plt.text(0.33, r'$\mu=15')
#maxfreq = n.max()
```



In [47]:

```
import matplotlib.pyplot as plt
%matplotlib inline
mybins = [ x * 0.01 for x in range(101)]

fig = plt.figure(figsize=(8,4), dpi=300)
```

```
plt.subplot(1, 2, 1)
plt.title("Distribution")
plt.hist(scores, bins=mybins)

plt.subplot(1, 2, 2)
plt.title("Cumulative Distribution")
plt.hist(scores, bins=mybins, density=True, cumulative=1)
plt.plot([0,1],[0.95,0.95]);
```



In [36]:

```
for i in range(21) :

    thresh = i / 20
    num_similar_pairs = len([x for x in scores if x >= thresh])
    prob = num_similar_pairs / len(scores) * 100
    print("%.3f %8d (%8.4f %%)" % (thresh, num_similar_pairs, round
(prob,4)))
```

```
0.000    43956 (100.0000 %)
0.050    43135 ( 98.1322 %)
0.100    40620 ( 92.4106 %)
0.150    35569 ( 80.9196 %)
0.200    30069 ( 68.4070 %)
0.250    24896 ( 56.6385 %)
0.300    20322 ( 46.2326 %)
0.350    16385 ( 37.2759 %)
0.400    13308 ( 30.2757 %)
0.450    10980 ( 24.9795 %)
0.500     9445 ( 21.4874 %)
0.550     7880 ( 17.9270 %)
0.600     6021 ( 13.6978 %)
0.650     4245 (  9.6574 %)
0.700     2819 (  6.4132 %)
0.750     1734 (  3.9449 %)
```

```
0.800     1051 (  2.3910 %)
0.850      636 (  1.4469 %)
0.900      315 (  0.7166 %)
0.950      121 (  0.2753 %)
1.000       56 (  0.1274 %)
```

In [37]:
```
print("Average:", sum(scores)/len(scores))
```

```
Average: 0.33088463035201027
```

Using molecular fingeprints. we can compute the similarity scores between molecules. However, how should these scores be interpreted? For example, the Tanimoto score between CID 60823 and CID 446155 is computed to be 0.662, but does it mean that the two compounds are similar? How similar is similar? The following analysis would help answer these questions.

From the distribution of the similarity scores among 297 natural compounds, we observe the following:

- If you randomly select two compounds from this dataset, the similarity score between them (computed using the Tanimoto equation and PubChem fingerprints) is ~0.33 on average.
- About 18 % of randomly selected compound pairs have a similarity score greater than 0.55.
- About 9% of randomly selected compound pairs have a similarity score greater than 0.65.

If two compounds have a Tanimoto score of 0.33, it is close to the average Tanimoto score between randomly selected compounds from our dataset and there is a 50% chance that you will get a score of 0.33 or greater just by selecting two compounds from PubChem. Therefore, it is reasonable to consider the two compounds are not similar.

The Tanimoto index may have a value ranging from 0 (for no similarity) to 1 (for identical molecules) and the midpoint of this value range is 0.5. Because of this, a Tanimoto score of **0.55** may not sound great enough to consider two compounds to be similar. However, according to the score distribution curve generated here, only **~17%** of randomly selected compound pairs will have a score greater than this.

In the previous section, we computed the similarity scores between some cholesterol-lowering drugs, and CID 60823 and CID 446155 had a

Tanimoto score of **0.662**. Based on the score distribution curve generated in the second section, we can say that the probablilty of two randomly selected compounds from PubChem having a Tanimoto score greater than 0.662 is **less than 9%**.

The following code cell demonstrates how to find an appropriate similarity score threshold above which a given percentage of the compound pairs will be considered to be similar to each other.

In [40]:
```
scores.sort() # Sort the scores in an increasing order.
```

In [41]:
```
# to find a threshold for top 5% compound pairs (i.e., 95% percenti
le)
print("# total compound pairs:   ", len(scores))
print("# 95% of compound pairs:  ", len(scores) * 0.95)
print("# score at 95% percentile:", scores[ round(len(scores) * 0.9
5) ] )
```

```
# total compound pairs:    43956
# 95% of compound pairs:   42637.32
# score at 95% percentile: 0.7762237762237763
```

PCA and tSNE of active and inactive flavor molecules

In [39]:
```
from sklearn.decomposition import PCA
import time
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from matplotlib.pyplot import figure
import seaborn as sns
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report,confusion_matrix
from sklearn.ensemble import RandomForestClassifier
from sklearn.decomposition import PCA
```

In [25]:
```
pca = PCA(n_components=2)
df4_pca.head()
df5=df4_pca.drop(['chembl_id','class','pubchem_fp'], axis=1)
crds = pca.fit_transform(df5)
```

In [36]:
```
# X = df.drop(['POPULATION'], axis = 1)
# Y = df['POPULATION']
# X = pd.get_dummies(X, prefix_sep='_')
# Y = LabelEncoder().fit_transform(Y)
X = StandardScaler().fit_transform(df5)
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
PCA_df = pd.DataFrame(data = X_pca, columns = ['PC1', 'PC2'])
PCA_df = pd.concat([PCA_df, df4_pca['class']], axis = 1)
PCA_df['class'] = LabelEncoder().fit_transform(PCA_df['class'])
PCA_df.head()
```

Out[36]:

	PC1	PC2	class
0	5.489498	-7.237064	0
1	4.904657	-6.804352	0
2	13.635824	19.299477	0
3	2.144082	-4.545019	0
4	2.082286	-4.406270	0

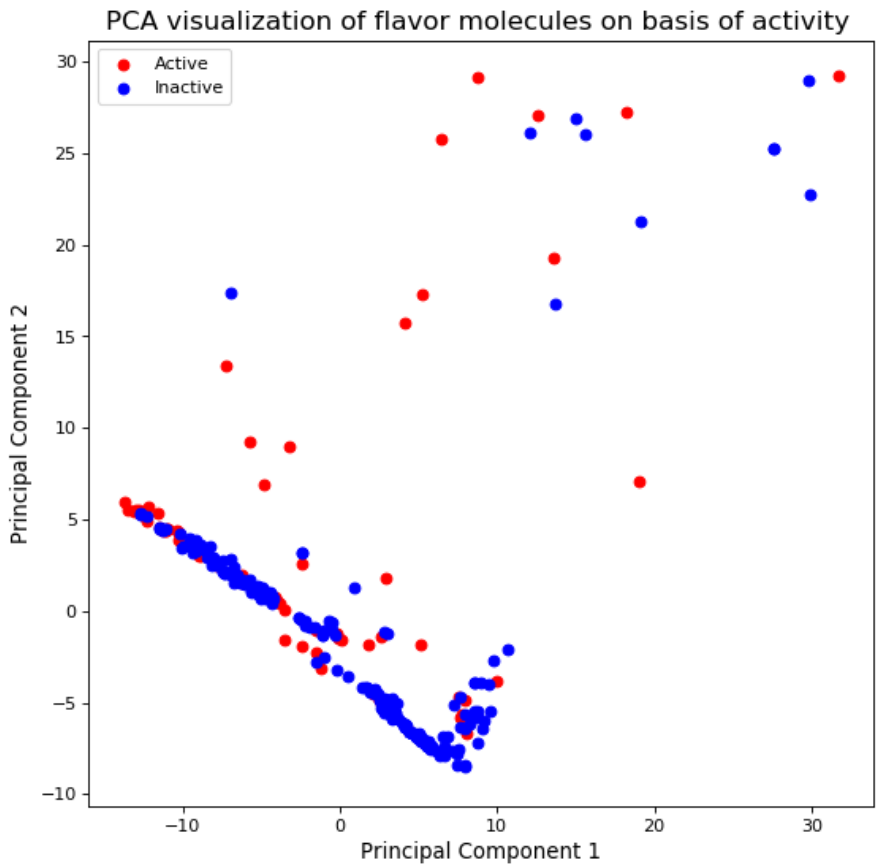
In [21]:
```
set(PCA_df['class'])
```

Out[21]:
```
{0, 1}
```

In [41]:
```
import pandas as pd
import plotly.express as px
from sklearn.decomposition import PCA
from sklearn.datasets import load_boston

# boston = load_boston()
#df = pd.DataFrame(boston.data, columns=boston.feature_names)
n_components = 4

pca = PCA(n_components=4)
X_pca = pca.fit_transform(X)
```

In [42]:
```
figure(num=None, figsize=(8, 8), dpi=80, facecolor='w', edgecolor=
'k')
```

```
classes = [0,1,2]
colors = ['r', 'b',"g"]
for clas, color in zip(classes, colors):
    plt.scatter(PCA_df.loc[PCA_df['class'] == clas, 'PC1'], PCA_df.
loc[PCA_df['class'] == clas, 'PC2'], c = color)

plt.xlabel('Principal Component 1', fontsize = 12)
plt.ylabel('Principal Component 2', fontsize = 12)
plt.title('PCA visualization of flavor molecules on basis of activi
ty ', fontsize = 15)
plt.legend(['Active', 'Inactive'])
#plt.grid()
plt.savefig('pca_fdb.png')
```



In [22]:

```
from sklearn.manifold import TSNE
time_start = time.time()
tsne = TSNE(n_components=2, verbose=1, perplexity=30, n_iter=1500)
X_tsne = tsne.fit_transform(df5)
print('t-SNE done! Time elapsed: {} seconds'.format(time.time()-tim
e_start))
```

```
[t-SNE] Computing 91 nearest neighbors...
```

```
[t-SNE] Indexed 297 samples in 0.027s...
[t-SNE] Computed neighbors for 297 samples in 0.107s...
[t-SNE] Computed conditional probabilities for sample 297 / 297
[t-SNE] Mean sigma: 2.682502
[t-SNE] KL divergence after 250 iterations with early exaggerat
ion: 56.885036
[t-SNE] KL divergence after 1500 iterations: 0.288904
t-SNE done! Time elapsed: 1.3642065525054932 seconds
```

In [27]:

```
set(df4_pca['class'])
```

Out[27]:

```
{'active', 'inactive'}
```

In [23]:

```
color_dict = dict({'active':'red', 'inactive':'yellow','intermediat
e': 'green'})
ax=sns.scatterplot(
    x=X_tsne[:,0], y=X_tsne[:,1],
    data=df4_pca,
    hue="class",
    hue_order=["active","inactive"],
    palette=color_dict,
    legend="brief",
    alpha=0.8
)
ax.set(xlabel = "tSNE1",
      ylabel = "tSNE2",
      title = "tSNE visualization of flavor molecules on basis of a
ctivity")
#plt.figure(figsize=(16,9))
plt.gcf().set_size_inches(10, 8)
plt.savefig("tsne_superpop_4.png")
```