Scaffold analysis in Python with RDKit and pandas

Dr. Samo Turk

BioMed X Innovation Center, Heidelberg

Python

Python (http://www.python.org/) very popular programming language especially in science.

pandas

Pandas (http://pandas.pydata.org/) is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
No need for R!

RDKit

RDKit (http://www.rdkit.org/) is an open source chemistry toolkit.

IPython

IPython (http://ipython.org/) interactive python shell. Has web-based interactive computational environment IPython Notebook.

This are not slides but interactive tutorial! https://github.com/Team-SKI/snippets

In [1]:
```
import pandas as pd
import rdkit.Chem as Chem
from rdkit.Chem import PandasTools
from rdkit.Chem import Draw
from rdkit.Chem import Descriptors
from rdkit.Chem.Draw import IPythonConsole # Enables RDKit IPython integration
```

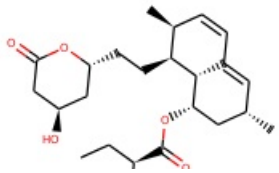
IPython and RDKit

RDKit provides IPython integration

In [2]:
```
mol = Chem.MolFromSmiles('O=C(O[C@@H]1[C@H]3C(=C/[C@H](C)C1)\C=C/[C@@H]([C@@H]3CC[C@H]2OC(=O)C[C@H
](O)C2)C)[C@@H](C)CC')
```

In [3]:
```
mol
```

Out[3]:



In [4]:
```
Descriptors.NumHDonors(mol)
```

Out[4]:
```
1
```

In [5]:
```
Descriptors.MolLogP(mol)
```

Out[5]:
```
4.195500000000004
```

RDKit and pandas

PandasTools.py

Load 'approved drugs' downloaded from www.drugbank.ca:
`% time` is a ipython magic function that tells you how much time did certain operation take to finish. It will be used to give you a feeling about speed of certain functions

In [6]:
```
% time cpds = PandasTools.LoadSDF('approved.sdf', includeFingerprints=False)
```

```
CPU times: user 6.48 s, sys: 16.7 ms, total: 6.5 s
Wall time: 6.5 s
```

In [7]:
```
cpds.columns
```

Out[7]:
```
Index([u'ALOGPS_LOGP', u'ALOGPS_LOGS', u'ALOGPS_SOLUBILITY', u'BRANDS', u'CHEMICAL_FORMULA', u'DRU
GBANK_ID', u'DRUG_GROUPS', u'EXACT_MASS', u'GENERIC_NAME', u'ID', u'INCHI_IDENTIFIER', u'INCHI_KEY
', u'IUPAC_NAME', u'JCHEM_ACCEPTOR_COUNT', u'JCHEM_ACIDIC_PKA', u'JCHEM_BASIC_PKA', u'JCHEM_DONOR_
COUNT', u'JCHEM_LOGP', u'JCHEM_PHYSIOLOGICAL_CHARGE', u'JCHEM_POLARIZABILITY', u'JCHEM_POLAR_SURFA
CE_AREA', u'JCHEM_REFRACTIVITY', u'JCHEM_ROTATABLE_BOND_COUNT', u'MOLECULAR_WEIGHT', u'SALTS', u'S
MILES', u'SYNONYMS', u'ROMol'], dtype=object)
```

In [8]:
```
len(cpds)
```

Out[8]:
```
1485
```

Assign the values of molnames and smiles (makes it easier to use this notebook on other sets with different col names)

In [9]:
```
molnames = 'DRUGBANK_ID'
smiles = 'SMILES'
```

Keep only columns 'DRUGBANK_ID', 'SMILES' and 'ROMol'

In [10]:
```
cpds = cpds[[molnames, smiles, 'ROMol']]
```

In [11]:
```
cpds.columns
```

Out[11]:
```
Index([u'DRUGBANK_ID', u'SMILES', u'ROMol'], dtype=object)
```

In [12]:
```
cpds.head(2)
```

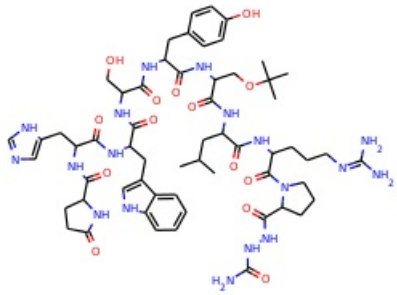
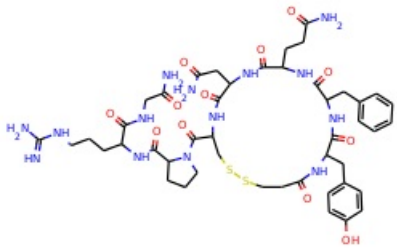
Out[12]:

	DRUGBANK_ID	SMILES
0	DB00014	CC(C)CC(NC(=O)C(COC(C)(C)C)NC(=O)C(Cc1ccc(O)cc1)NC(=O)C(CO)NC(=O)C(Cc1c[nH]c2ccccc12)NC(=O)C(Cc1cnc[nH]1)NC(=O)C1C
1	DB00035	N=C(N)NCCCC(NC(=O)C1CCCN1C(=O)C1CSSCCC(=O)NC(Cc2ccc(O)cc2)C(=O)NC(Cc2ccccc2)C(=O)NC(CC

Look at two columns and only first 2 lines:

In [13]:
```
cpds[[molnames, 'ROMol']].head(2)
```

Out[13]:

	DRUGBANK_ID	ROMol
0	DB00014	
1	DB00035	

Remove lines with NaN (empty) values and duplicates

In [14]:
```
cpds = cpds.dropna()
cpds = cpds.drop_duplicates(molnames)
cpds = cpds.drop_duplicates(smiles)
len(cpds)
```

Out[14]: 1462

Descriptors

Add some descriptors

In [15]:
```
from rdkit.Chem import Descriptors
cpds['logp'] = cpds['ROMol'].map(Descriptors.MolLogP)
cpds['mw'] = cpds['ROMol'].map(Descriptors.MolWt)
```

Remove compounds with logp >= 5 and MW >= 500

In [16]:
```
cpds = cpds[cpds['logp'] <= 5]
cpds = cpds[cpds['mw'] <= 500]
len(cpds)
```

Out[16]: 1143

In [17]:
```
cpds[[molnames, 'logp', 'mw', smiles]].head()
```

Out[17]:

	DRUGBANK_ID	logp	mw	SMILES

6	DB00116	-0.2820	445.436	Nc1nc(=O)c2c([nH]1)NCC(CNc1ccc(C(=O)NC(CCC(=O)O)C(=O)O)cc1)N2
7	DB00117	-0.6359	155.157	NC(Cc1cnc[nH]1)C(=O)O
8	DB00118	-1.9222	399.453	C[S+](CCC(N)C(=O)O)CC1OC(n2cnc3c2ncnc3N)C(O)C1O
9	DB00119	-0.3400	88.062	CC(=O)C(=O)O
10	DB00120	0.6410	165.192	NC(Cc1ccccc1)C(=O)O

Matplotlib and pylab

IPython has matplotlib integration

In [18]:
```
%matplotlib inline
```

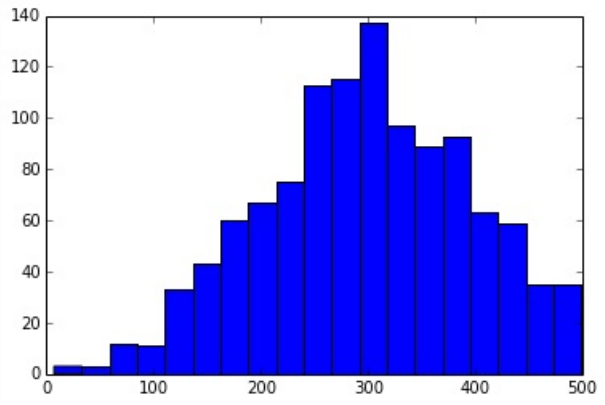
In [19]:
```
import pylab
import numpy as np
import matplotlib as plt
```

Distribution on molecular weights

Bin the data and plot it

In [20]:
```
bins = np.linspace(cpds['mw'].min(), cpds['mw'].max(), 20)
pylab.hist(cpds['mw'], bins)
pylab.show
```

Out[20]: <function matplotlib.pyplot.show>

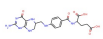


Alternative visualisation of a table

Default takes a lot of space

In [21]:
```
cpds.head(1)
```

Out[21]:

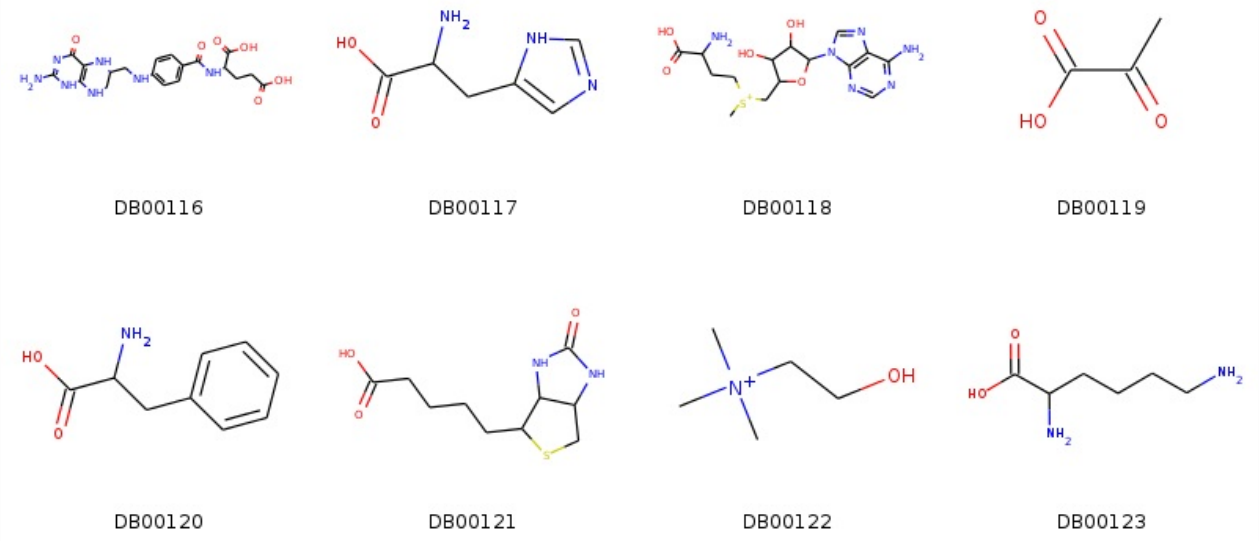
	DRUGBANK_ID	SMILES	ROMol	logp	mw
6	DB00116	Nc1nc(=O)c2c([nH]1)NCC(CNc1ccc(C(=O)NC(CCC(=O)O)C(=O)O)cc1)N2		-0.282	445.436

FrameToGridImage(pandasFrame, legendsCol=, molsPerRow=)

With this function you can visualize a dataframe (or part of it) as a single image

In [22]:
```
PandasTools.FrameToGridImage(cpds.head(8), legendsCol=molnames, molsPerRow=4)
```

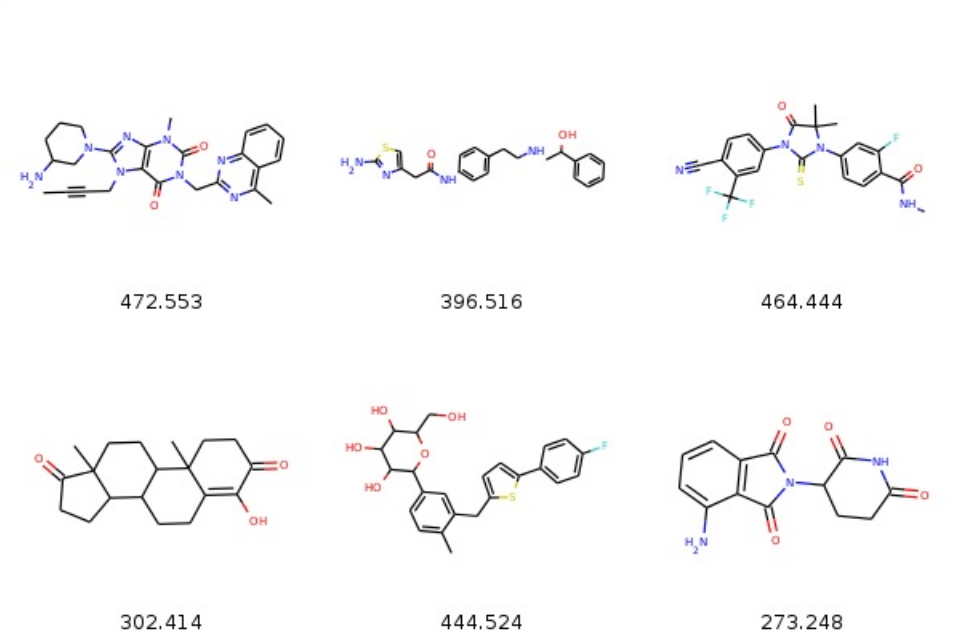
Out[22]:



You can define different colums as legends

In [23]:
```
PandasTools.FrameToGridImage(cpds.tail(6), legendsCol='mw', molsPerRow=3)
```

Out[23]:



Murcko scaffold decomposition Bemis, G. W.; Murcko, M. A. “The Properties of Known Drugs. 1. Molecular Frameworks.” J. Med. Chem. 39:2887-93 (1996).

Decomposition of molecules to scaffolds or generic frameworks

Functionality present in RDKit. Added it to PandasTools

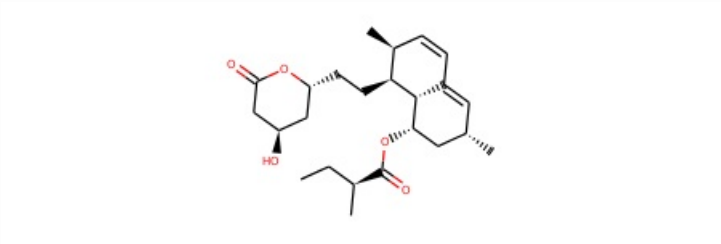
In [24]:
```
from rdkit.Chem.Scaffolds import MurckoScaffold
```

How it works with RDKit:

In [25]:
```
scaffold = MurckoScaffold.GetScaffoldForMol(mol)
generic = MurckoScaffold.MakeScaffoldGeneric(MurckoScaffold.GetScaffoldForMol(mol))
```

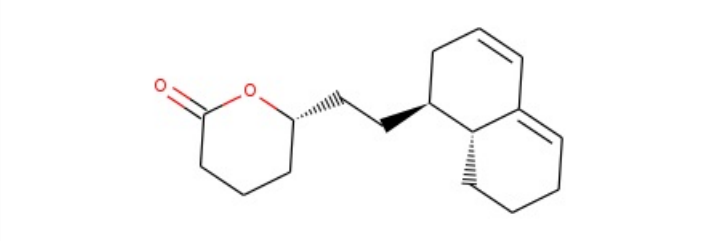
In [26]:
```
mol
```

Out[26]:



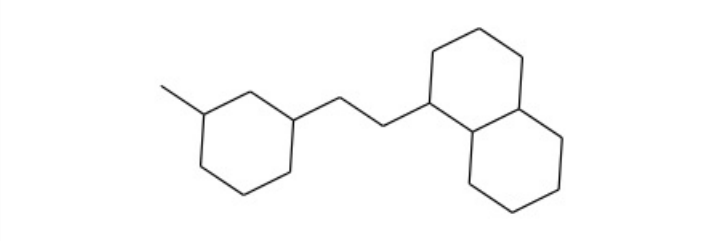
In [27]:
```
scaffold
```

Out[27]:



In [28]:
```
generic
```

Out[28]:



AddMurckoToFrame(pandasFrame, MurcoCol=, Generic=False)

Returns SMILES of scaffolds (or generic frameworks)

In [29]:
```
% time PandasTools.AddMurckoToFrame(cpds)
```

```
CPU times: user 417 ms, sys: 6.67 ms, total: 423 ms
Wall time: 418 ms
```

In [30]:
```
% time PandasTools.AddMurckoToFrame(cpds, MurckoCol='Murcko_GENERIC', Generic=True)
```

```
CPU times: user 987 ms, sys: 3.33 ms, total: 990 ms
Wall time: 979 ms
```

In [31]:
```
cpds.head(1)
```

Out[31]:

	DRUGBANK_ID	SMILES	ROMol	logp	mw	Murcko_
6	DB00116	Nc1nc(=O)c2c([nH]1)NCC(CNc1ccc(C(=O)NC(CCC(=O)O)C(=O)O)cc1)N2		-0.282	445.436	O=c1nc[

Now we can use pandas groupby() functionality and group by scaffolds and create a **new frame** with scaffolds sorted by number of members

In [32]:
```
sortedScaffolds = cpds.groupby(['Murcko_SMILES']).count().sort(smiles, ascending=False)
```

In [33]:
```
sortedScaffolds = sortedScaffolds[[smiles]] # Keep only smiles column
```

```
sortedScaffolds = sortedScaffolds.rename(columns={smiles:'count'}) # rename smiles column to count
sortedScaffolds['Murcko_SMILES'] = sortedScaffolds.index # actual SMILES are only in index column,
 move it
sortedScaffolds.head()
```

Out[33]:

	count	Murcko_SMILES
Murcko_SMILES		
c1ccccc1	112	c1ccccc1
	111	
O=C1C=CC2C(=C1)CCC1C3CCCC3CCC21	17	O=C1C=CC2C(=C1)CCC1C3CCCC3CCC21
O=C1C=C2CCC3C4CCCC4CCC3C2CC1	12	O=C1C=C2CCC3C4CCCC4CCC3C2CC1
O=C1CN=C(c2ccccc2)c2ccccc2N1	12	O=C1CN=C(c2ccccc2)c2ccccc2N1

Add RDKit's ROMol column to scaffolds dataframe so we can visualize it

In [34]:
```
PandasTools.AddMoleculeColumnToFrame(sortedScaffolds, smilesCol='Murcko_SMILES')
```

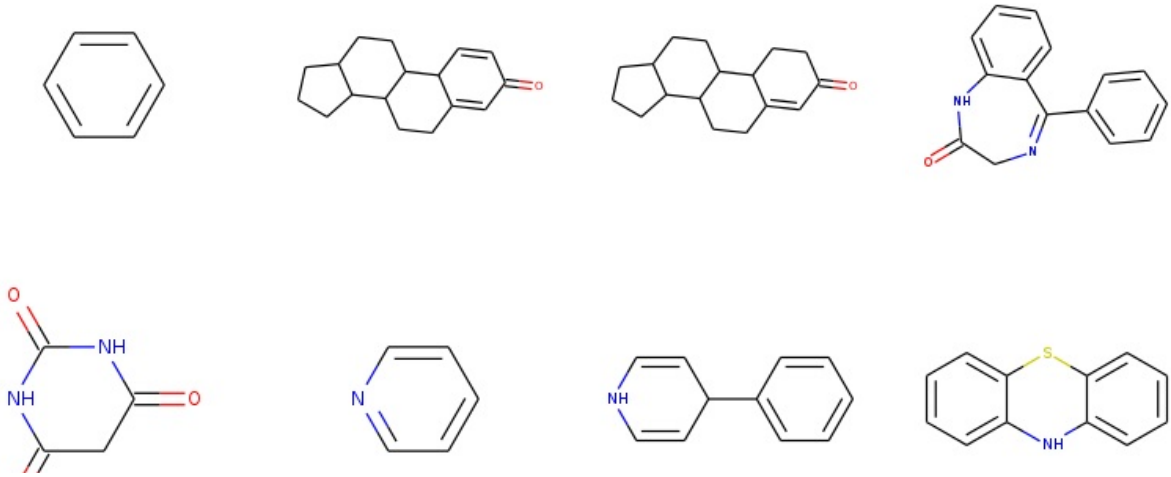
In [35]:
```
sortedScaffolds.head(1)
```

Out[35]:

	count	Murcko_SMILES	ROMol
Murcko_SMILES			
c1ccccc1	112	c1ccccc1	

In [36]:
```
PandasTools.FrameToGridImage(sortedScaffolds.dropna().head(8), molsPerRow=4) #dropna drops compoun
ds without scaffold
```

Out[36]:



We can also retrieve all compounds with certain scaffold from original table
Benzodiazepine scaffold is #4, SMILES: *O=C1CN=C(c2ccccc2)c2ccccc2N1*

In [37]:
```
cpds[cpds['Murcko_SMILES'] == 'O=C1CN=C(c2ccccc2)c2ccccc2N1'].head(1)
```

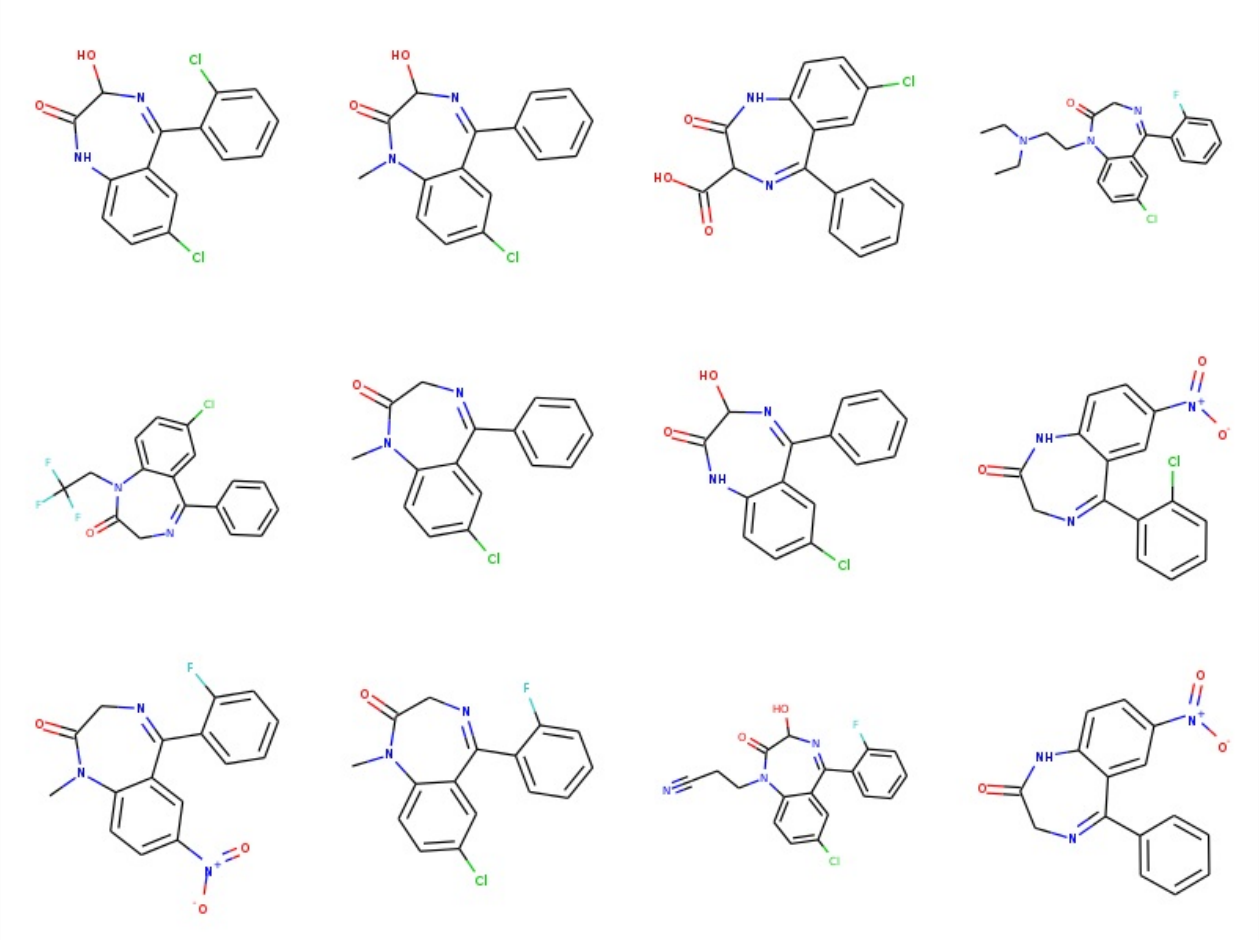
Out[37]:

	DRUGBANK_ID	SMILES	ROMol	logp	mw	Murcko_SMILES	M
74	DB00186	O=C1Nc2ccc(Cl)cc2C(c2ccccc2Cl)=NC1O		3.1013	321.163	O=C1CN=C(c2ccccc2)c2ccccc2N1	C

Get all of them and show them as grid image

In [38]:
```
PandasTools.FrameToGridImage(cpds[cpds['Murcko_SMILES'] == 'O=C1CN=C(c2ccccc2)c2ccccc2N1'], molsPe
rRow=4)
```

Out[38]:



Aligning compounds to scaffolds

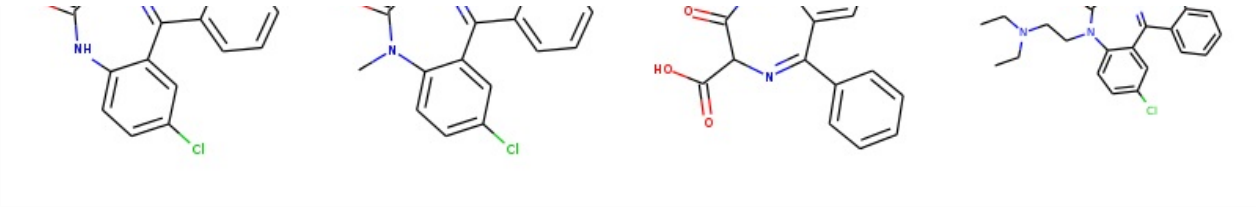
AlignToScaffold(dataframe, molCol=, scaffoldCol=)

In [39]:
```
somemols = cpds.groupby('Murcko_SMILES').get_group('O=C1CN=C(c2ccccc2)c2ccccc2N1')
```

Note how molecules are not alignied

In [40]:
```
PandasTools.FrameToGridImage(somemols.head(4), molsPerRow=4)
```

Out[40]:

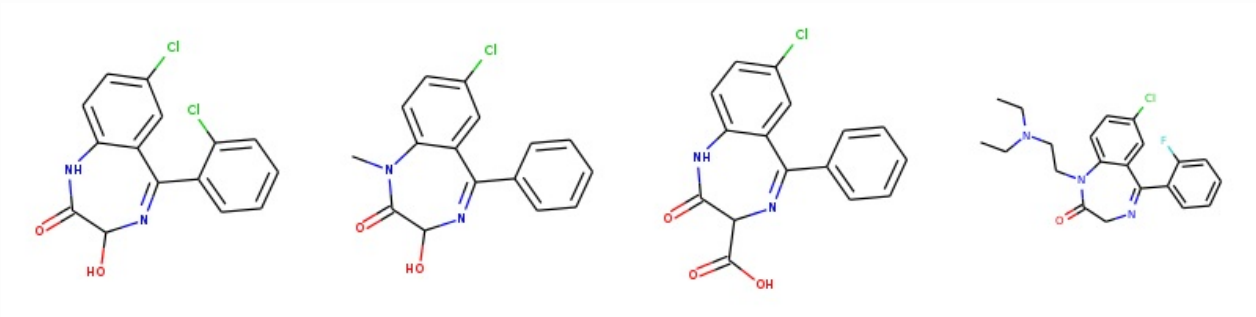


Align them to scaffold

In [41]:
```
PandasTools.AlignToScaffold(somemols, molCol='ROMol', scaffoldCol='Murcko_SMILES')
```

In [42]:
```
PandasTools.FrameToGridImage(somemols.head(4), molsPerRow=4)
```

Out[42]:



Check our GitHub https://github.com/Team-SKI/snippets

Thank you!