

Mol2Vec

April 20, 2021

Reference taken from <https://github.com/samoturk/mol2vec/blob/master/examples/02%20-%20Supervised%20machine%20learning%20with%20Mol2vec.ipynb>

```
[1]: !pip install git+https://github.com/samoturk/mol2vec
```

```
Collecting git+https://github.com/samoturk/mol2vec
  Cloning https://github.com/samoturk/mol2vec to
/private/var/folders/3p/mwxtj1wj179dmn25cxrq38380000gn/T/pip-req-build-ed1r2b7f
  Running command git clone -q https://github.com/samoturk/mol2vec
/private/var/folders/3p/mwxtj1wj179dmn25cxrq38380000gn/T/pip-req-build-ed1r2b7f
Requirement already satisfied (use --upgrade to upgrade): mol2vec==0.1 from
git+https://github.com/samoturk/mol2vec in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages
Requirement already satisfied: numpy in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (1.18.5)
Requirement already satisfied: gensim in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (4.0.1)
Requirement already satisfied: tqdm in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (4.47.0)
Requirement already satisfied: joblib in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (0.16.0)
Requirement already satisfied: pandas in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (1.0.5)
Requirement already satisfied: matplotlib in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (3.2.2)
Requirement already satisfied: IPython in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (7.16.1)
Requirement already satisfied: seaborn in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
mol2vec==0.1) (0.10.1)
```

Requirement already satisfied: scipy>=0.18.1 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
gensim->mol2vec==0.1) (1.5.0)

Requirement already satisfied: smart-open>=1.8.1 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
gensim->mol2vec==0.1) (5.0.0)

Requirement already satisfied: pytz>=2017.2 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
pandas->mol2vec==0.1) (2020.1)

Requirement already satisfied: python-dateutil>=2.6.1 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
pandas->mol2vec==0.1) (2.8.1)

Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->mol2vec==0.1) (1.2.0)

Requirement already satisfied: cycycler>=0.10 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->mol2vec==0.1) (0.10.0)

Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->mol2vec==0.1) (2.4.7)

Requirement already satisfied: pygments in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (2.6.1)

Requirement already satisfied: pexpect; sys_platform != "win32" in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (4.8.0)

Requirement already satisfied: appnope; sys_platform == "darwin" in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (0.1.0)

Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (3.0.5)

Requirement already satisfied: pickleshare in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (0.7.5)

Requirement already satisfied: jedi>=0.10 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (0.17.1)

Requirement already satisfied: backcall in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (0.2.0)

Requirement already satisfied: setuptools>=18.5 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (49.2.0.post20200714)

Requirement already satisfied: decorator in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (4.4.2)

```

Requirement already satisfied: traitlets>=4.2 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
IPython->mol2vec==0.1) (4.3.3)
Requirement already satisfied: six>=1.5 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from python-
dateutil>=2.6.1->pandas->mol2vec==0.1) (1.15.0)
Requirement already satisfied: ptyprocess>=0.5 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from pexpect;
sys_platform != "win32"->IPython->mol2vec==0.1) (0.6.0)
Requirement already satisfied: wcwidth in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from prompt-
toolkit!=3.0.0,!<3.0.1,<3.1.0,>=2.0.0->IPython->mol2vec==0.1) (0.2.5)
Requirement already satisfied: parso<0.8.0,>=0.7.0 in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
jedi>=0.10->IPython->mol2vec==0.1) (0.7.0)
Requirement already satisfied: ipython-genutils in
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-packages (from
traitlets>=4.2->IPython->mol2vec==0.1) (0.2.0)
Building wheels for collected packages: mol2vec
  Building wheel for mol2vec (setup.py) ... done
  Created wheel for mol2vec: filename=mol2vec-0.1-py3-none-any.whl
size=14026
sha256=902e226948cc2e52eb4e133b709825aebcfa0609628630978e12722667bd5af1
  Stored in directory:
/private/var/folders/3p/mwxtj1wj179dmn25cxrq38380000gn/T/pip-ephem-wheel-cache-3
6bl0mc2/wheels/28/66/2a/543e91280db70c489a72de9c9fa26dca4caa7113721fb03b86
Successfully built mol2vec

```

```

[2]: import numpy as np
import pandas as pd
import sys, os

sys.path.append(os.path.dirname(os.path.dirname(os.getcwd()))+'/mol2vec')

from rdkit import Chem
from rdkit.Chem import PandasTools
from rdkit.Chem.Draw import IPythonConsole

from mol2vec.features import mol2alt_sentence, mol2sentence, MolSentence,
↳ DfVec, sentences2vec
from gensim.models import word2vec

```

```

RDKit WARNING: [02:43:58] Enabling RDKit 2019.09.3 jupyter extensions
/Users/shrutidudwadkar/opt/anaconda3/lib/python3.8/site-
packages/gensim/similarities/__init__.py:15: UserWarning: The
gensim.similarities.levenshtein submodule is disabled, because the optional
Levenshtein package <https://pypi.org/project/python-Levenshtein/> is
unavailable. Install Levenshtein (e.g. `pip install python-Levenshtein`) to

```

```
suppress this warning.
warnings.warn(msg)
```

Importing the BindingDB_Kd dataset

```
[3]: from tdc.multi_pred import DTI
bindingDB_data = DTI(name = 'BindingDB_Kd')
```

```
Found local copy...
Loading...
Done!
```

```
[4]: bindingDB_cold_split = bindingDB_data.get_split(method = 'cold_split',
↳column_name = 'Drug', seed = 42, frac = [0.7, 0.1, 0.2])
```

```
[5]: train_set = bindingDB_cold_split['train']
```

```
[6]: train_set
```

```
[6]:
```

	Drug_ID	Drug	\
0	4316.0	COc1ccc(CNS(=O)(=O)c2ccc(S(N)(=O)=O)s2)cc1	
1	4293.0	NS(=O)(=O)c1ccc(S(=O)(=O)NCc2cccs2)s1	
2	1612.0	COc1ccc(N2CC(O)c3cc(S(N)(=O)=O)sc3S2(=O)=O)cc1	
3	4369102.0	CCN[C@H]1CN(CCCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21	
4	68844.0	CCN[C@H]1CN(CCCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21	
...	
34696	76311094.0	O=C(NCCCCS(=O)(=O)c1cccc1)c1ccc2nccn2c1	
34697	57406853.0	CCOc1cc(C(C)(C)C)ccc1C1=N[C@@](C)(c2ccc(Cl)cc2...	
34698	58573469.0	CC(C)[C@@H](CS(=O)(=O)C(C)C)N1C(=O)[C@@](C)(CC...	
34699	113557.0	CCCCCCCOC10C(CO)C(O)C(O)C10	
34700	113557.0	CCCCCCCOC10C(CO)C(O)C(O)C10	

	Target_ID	Target	Y
0	P00918	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...	0.490
1	P00918	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...	0.830
2	P00918	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...	0.160
3	P00918	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...	0.320
4	P00918	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...	0.130
...
34696	P43490	MNPAAEAEFNILLATDSYKVTHYKQYPPNTSKVYSYFECREKKTEN...	860.000
34697	NaN	MCNTNMSVPTDGAVTTSQIPASEQETLVRPKPLLLKLLKSVGAQKD...	2.900
34698	NaN	MCNTNMSVPTDGAVTTSQIPASEQETLVRPKPLLLKLLKSVGAQKD...	0.045
34699	P08191	MKRVITLFAVLLMGWSVNAWSFACKTANGTAIPIGGGSANVYVNLA...	17.000
34700	P08191	MKRVITLFAVLLMGWSVNAWSFACKTANGTAIPIGGGSANVYVNLA...	19.000

```
[34701 rows x 5 columns]
```

```
[7]: dataSource = train_set.drop(['Target_ID', 'Drug_ID', 'Y'], axis=1)
dataSource['Mol'] = ''
```

Load the pre-trained Mol2vec model

```
[8]: model = word2vec.Word2Vec.load('model_300dim.pkl')
```

```
[9]: from rdkit import Chem
import tensorflow as tf
import torch

for ind, drug in enumerate(dataSource['Drug']):
    mol = Chem.MolFromSmiles(drug)
    dataSource['Mol'][ind] = mol
```

```
[10]: dataSource['Sentence'] = dataSource.apply(lambda x:
    ↪ MolSentence(mol2alt_sentence(x['Mol'], 1)), axis=1)
```

Method taken from mol2vec and updated as per the changes in the gensim. Tweaked the 'sentences2vec' method provided in Mol2Vec library to make it compatible with gensim 4.0

```
[11]: def sentences2vec(sentences, model, unseen=None):

    #Change made in this line as vocab is removed in gensim 4.0
    #keys = set(model.wv.vocab.keys())
    keys = set(model.wv.key_to_index.keys())

    vec = []
    if unseen:
        unseen_vec = model.wv.word_vec(unseen)

    for sentence in sentences:
        if unseen:
            vec.append(sum([model.wv.word_vec(y) if y in set(sentence) & keys
                            else unseen_vec for y in sentence]))
        else:
            vec.append(sum([model.wv.word_vec(y) for y in sentence
                            if y in set(sentence) & keys]))
    return np.array(vec)
```

```
[12]: dataSource['mol2vec'] = ''
dataSource['mol2vec'] = [DfVec(x) for x in
    ↪ sentences2vec(dataSource['Sentence'], model, unseen='UNK')]
```

```
[13]: X = np.array([x.vec for x in dataSource['mol2vec']])
```

```
[14]: from torch.utils.data import DataLoader

train_loader = DataLoader(dataset=X, shuffle=True, batch_size=256)

[15]: from sklearn.cluster import AgglomerativeClustering
clustering = AgglomerativeClustering(linkage='average', n_clusters=50)

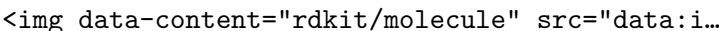
clusters=[]
for i, batch in enumerate(train_loader):
    clustering.fit(batch)
    clusters.extend(clustering.labels_)
```

Grouping the dataframe based on the clusters formed

```
[16]: dataSource['Clusters'] = clusters
by_clusters = dataSource.groupby(['Clusters'])
```

```
[17]: dataSource
```

```
[17]:
```

	Drug \	Target \	Mol \
0	<chem>COc1ccc(CNS(=O)(=O)c2ccc(S(N)(=O)=O)s2)cc1</chem>		
1	<chem>NS(=O)(=O)c1ccc(S(=O)(=O)NCc2cccs2)s1</chem>		
2	<chem>COc1ccc(N2CC(O)c3cc(S(N)(=O)=O)sc3S2(=O)=O)cc1</chem>		
3	<chem>CCN[C@H]1CN(CCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21</chem>		
4	<chem>CCN[C@H]1CN(CCCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21</chem>		
...			
34696	<chem>O=C(NCCCCS(=O)(=O)c1cccc1)c1ccc2nccn2c1</chem>		
34697	<chem>CCOc1cc(C(C)(C)C)ccc1C1=N[C@@](C)(c2ccc(Cl)cc2...</chem>		
34698	<chem>CC(C)[C@@H](CS(=O)(=O)C(C)C)N1C(=O)[C@@](C)(CC...</chem>		
34699	<chem>CCCCCCCOC1OC(CO)C(O)C(O)C1O</chem>		
34700	<chem>CCCCCCCOC1OC(CO)C(O)C(O)C1O</chem>		
...			
34696		<chem>MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...</chem>	
34697		<chem>MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...</chem>	
34698		<chem>MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...</chem>	
34699		<chem>MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...</chem>	
34700		<chem>MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...</chem>	
...			
34696		<chem>MNPAAEAEFNILLATDSYKVTYKQYPPNTSKVYSYFECREKKTEN...</chem>	
34697		<chem>MCNTNMSVPTDGAVTTSQIPASEQETLVRPKPLLLKLLKSVGAQKD...</chem>	
34698		<chem>MCNTNMSVPTDGAVTTSQIPASEQETLVRPKPLLLKLLKSVGAQKD...</chem>	
34699		<chem>MKRVITLFAVLLMGWSVNAWSFACKTANGTAIPIGGGSANVYVNLA...</chem>	
34700		<chem>MKRVITLFAVLLMGWSVNAWSFACKTANGTAIPIGGGSANVYVNLA...</chem>	
0			

```

1      <img data-content="rdkit/molecule" src="data:i...
2      <img data-content="rdkit/molecule" src="data:i...
3      <img data-content="rdkit/molecule" src="data:i...
4      <img data-content="rdkit/molecule" src="data:i...
...
34696 <img data-content="rdkit/molecule" src="data:i...
34697 <img data-content="rdkit/molecule" src="data:i...
34698 <img data-content="rdkit/molecule" src="data:i...
34699 <img data-content="rdkit/molecule" src="data:i...
34700 <img data-content="rdkit/molecule" src="data:i...

```

```

Sentence \
0      (2246728737, 3975275337, 864674487, 2076190208...
1      (847957139, 3275683399, 999334238, 3532053313,...
2      (2246728737, 3975275337, 864674487, 2076190208...
3      (2246728737, 3542456614, 2245384272, 773607102...
4      (2246728737, 3542456614, 2245384272, 773607102...
...
34696 (864942730, 1510328189, 2246699815, 1054767590...
34697 (2246728737, 3542456614, 2245384272, 399408866...
34698 (2246728737, 3537119515, 2245273601, 300433380...
34699 (2246728737, 3542456614, 2245384272, 117312591...
34700 (2246728737, 3542456614, 2245384272, 117312591...

```

```

mol2vec Clusters
0      (300,) dimensional vector      22
1      (300,) dimensional vector      5
2      (300,) dimensional vector      4
3      (300,) dimensional vector      5
4      (300,) dimensional vector      3
...
34696 (300,) dimensional vector      6
34697 (300,) dimensional vector     17
34698 (300,) dimensional vector     15
34699 (300,) dimensional vector     30
34700 (300,) dimensional vector      0

```

[34701 rows x 6 columns]

```

[18]: for cluster, data in by_clusters:
      print(f"First 2 entries for cluster: ", cluster)
      print("-----")
      print(data.head(2))

```

First 2 entries for cluster: 0

Drug \

10 NS(=O)(=O)c1ccc(C(=O)NCc2cc(F)ccc2F)cc1
 14 NS(=O)(=O)c1ccc(C(=O)NCc2c(F)c(F)c(F)c(F)c2F)cc1

Target \

10 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...
 14 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...

Mol \

10 <img data-content="rdkit/molecule" src="data:i...
 14 <img data-content="rdkit/molecule" src="data:i...

Sentence \

10 (847957139, 3275683399, 999334238, 3532053313,...
 14 (847957139, 3275683399, 999334238, 3532053313,...

mol2vec Clusters

10 (300,) dimensional vector 0
 14 (300,) dimensional vector 0

First 2 entries for cluster: 1

Drug \

159 O=C1c2cc(F)ccc2-c2cc(F)c([N+](=O)[O-])cc21
 187 CC(C)NCC(O)c1ccc2ccccc2c1

Target \

159 MSGETTRLTEPQLRELAARGAAELDGATATDMLRWTDETFGDIGGA...
 187 MGQPGNGSAFLLAPNGSHAPDHDVTQERDEVWVGMGIVMSLIVLA...

Mol \

159 <img data-content="rdkit/molecule" src="data:i...
 187 <img data-content="rdkit/molecule" src="data:i...

Sentence \

159 (864942730, 10565946, 3217380708, 1630260676, ...
 187 (2246728737, 3537119515, 2245273601, 263530394...

mol2vec Clusters

159 (300,) dimensional vector 1
 187 (300,) dimensional vector 1

First 2 entries for cluster: 2

Drug \

11 NS(=O)(=O)c1ccc(C(=O)NCc2c(F)cccc2F)cc1
 43 Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc...

Target \

11 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...
 43 HHSTVADGLITTLHYAPKRNKPTVYGVSPNYDKWEMERTDITMKH...

Mol \

11 <img data-content="rdkit/molecule" src="data:i...
 43 <img data-content="rdkit/molecule" src="data:i...

Sentence \

11 (847957139, 3275683399, 999334238, 3532053313,...
 43 (2246728737, 422715066, 3217380708, 3124581743...

mol2vec Clusters

11 (300,) dimensional vector 2
 43 (300,) dimensional vector 2
 First 2 entries for cluster: 3

Drug \

4 CCN[C@H]1CN(CCCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21
 74 COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...

Target \

4 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...
 74 HHSTVADGLITTLHYAPKRKNKPTVYGVSPNYDKWEMERTDITMKH...

Mol \

4 <img data-content="rdkit/molecule" src="data:i...
 74 <img data-content="rdkit/molecule" src="data:i...

Sentence \

4 (2246728737, 3542456614, 2245384272, 773607102...
 74 (2246728737, 3975275337, 864674487, 2076190208...

mol2vec Clusters

4 (300,) dimensional vector 3
 74 (300,) dimensional vector 3
 First 2 entries for cluster: 4

Drug \

2 COc1ccc(N2CC(O)c3cc(S(N)(=O)=O)sc3S2(=O)=O)cc1
 6 CN[C@H]1CN(c2cccc(OC)c2)S(=O)(=O)c2sc(S(N)(=O)=O...

Target \

2 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...
 6 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKP...

Mol \

2 <img data-content="rdkit/molecule" src="data:i...
 6 <img data-content="rdkit/molecule" src="data:i...

Sentence \

2 (2246728737, 3975275337, 864674487, 2076190208...
6 (2246728737, 3824063894, 847961216, 1732628033...

mol2vec Clusters
2 (300,) dimensional vector 4
6 (300,) dimensional vector 4
First 2 entries for cluster: 5

Drug \

1 NS(=O)(=O)c1ccc(S(=O)(=O)NCc2cccs2)s1
3 CCN[C@H]1CN(CCOC)S(=O)(=O)c2sc(S(N)(=O)=O)cc21

Target \

1 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...
3 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...

Mol \

1 <img data-content="rdkit/molecule" src="data:i...
3 <img data-content="rdkit/molecule" src="data:i...

Sentence \

1 (847957139, 3275683399, 999334238, 3532053313,...
3 (2246728737, 3542456614, 2245384272, 773607102...

mol2vec Clusters
1 (300,) dimensional vector 5
3 (300,) dimensional vector 5
First 2 entries for cluster: 6

Drug Target \

15 NS(=O)(=O)c1ccccc1 MASPDWGYDDKNGPEQWSKLYPIANGNNQSPVDIKTSETKHDTSLK...
17 NS(=O)(=O)c1ccccc1 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...

Mol \

15 <img data-content="rdkit/molecule" src="data:i...
17 <img data-content="rdkit/molecule" src="data:i...

Sentence \

15 (847957139, 3275683399, 999334238, 3532053313,...
17 (847957139, 3275683399, 999334238, 3532053313,...

mol2vec Clusters
15 (300,) dimensional vector 6
17 (300,) dimensional vector 6
First 2 entries for cluster: 7

Drug \

24 CN[C@@H](C)C(=O)N[C@H](C(=O)N1C[C@@H](O)c2ccccc...

```

48  Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc...

Target \
24  MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...
48  HHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKH...

Mol \
24  <img data-content="rdkit/molecule" src="data:i...
48  <img data-content="rdkit/molecule" src="data:i...

Sentence \
24  (2246728737, 3824063894, 847961216, 1362425715...
48  (2246728737, 422715066, 3217380708, 3124581743...

mol2vec Clusters
24  (300,) dimensional vector      7
48  (300,) dimensional vector      7
First 2 entries for cluster:  8
-----

Drug \
71  CCN(CC)CCNC(=O)c1c(C) [nH] c(/C=C2\C(=O)Nc3ccc(F...
87  COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...

Target \
71  HHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKH...
87  MRGARGAWDFLCVLLLLLRVQTGSSQPSVSPGEPSPPSIHPGKSDL...

Mol \
71  <img data-content="rdkit/molecule" src="data:i...
87  <img data-content="rdkit/molecule" src="data:i...

Sentence \
71  (2246728737, 3542456614, 2245384272, 592593828...
87  (2246728737, 3975275337, 864674487, 2076190208...

mol2vec Clusters
71  (300,) dimensional vector      8
87  (300,) dimensional vector      8
First 2 entries for cluster:  9
-----

Drug \
7   NS(=O)(=O)c1ccc(C(=O)NCc2ccccc2)cc1
31  Cc1sc(NC(=O)[C@H](C)N)nc1-c1ccccc1

Target \
7   SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...
31  MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...

```

Mol \

7 <img data-content="rdkit/molecule" src="data:i...
31 <img data-content="rdkit/molecule" src="data:i...

Sentence \

7 (847957139, 3275683399, 999334238, 3532053313,...
31 (2246728737, 422715066, 3217380708, 2718013753...

mol2vec Clusters

7 (300,) dimensional vector 9
31 (300,) dimensional vector 9
First 2 entries for cluster: 10

Drug \

8 NS(=O)(=O)c1ccc(C(=O)NCc2ccc(F)cc2)cc1
9 NS(=O)(=O)c1ccc(C(=O)NCc2ccc(F)cc2F)cc1

Target \

8 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKPL...
9 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKPL...

Mol \

8 <img data-content="rdkit/molecule" src="data:i...
9 <img data-content="rdkit/molecule" src="data:i...

Sentence \

8 (847957139, 3275683399, 999334238, 3532053313,...
9 (847957139, 3275683399, 999334238, 3532053313,...

mol2vec Clusters

8 (300,) dimensional vector 10
9 (300,) dimensional vector 10
First 2 entries for cluster: 11

Drug \

219 CC12CC(=O)C3c4ccc(O)cc4CCC3C1CCC2=O
264 NC(=O)CCCCNC(=O)OC[C@H]1O[C@H](n2cnc3c(N)ncn...

Target \

219 MSAVALPRVSGGHDEHGHLEEFRTDPIGLMQRVRDECGDVGTFQLA...
264 MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDR...

Mol \

219 <img data-content="rdkit/molecule" src="data:i...
264 <img data-content="rdkit/molecule" src="data:i...

Sentence \

219 (2246728737, 1861965050, 2976816164, 101278089...

264 (847957139, 2591432844, 2246699815, 3328145258...

```
mol2vec Clusters
219 (300,) dimensional vector 11
264 (300,) dimensional vector 11
First 2 entries for cluster: 12
-----
```

```
Drug \
242 CCN1CCN(Cc2ccc(-c3cc4c(N[C@H](C)c5ccccc5)ncnc4...
439 CC(C)(C)c1cnc(CSc2cnc(NC(=O)C3CCNCC3)s2)o1
```

```
Target \
242 MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFED...
439 MRHSKRTHCPDWDSRESWGHEsyRGSHKRKRSHSSTQENRHCKPH...
```

```
Mol \
242 <img data-content="rdkit/molecule" src="data:i...
439 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
242 (2246728737, 3542456614, 2245384272, 225184566...
439 (2246728737, 3537123720, 2245277810, 994237755...
```

```
mol2vec Clusters
242 (300,) dimensional vector 12
439 (300,) dimensional vector 12
First 2 entries for cluster: 13
-----
```

```
Drug \
247 COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc10CCCN1CCOCC1
500 O=C(NCC(c1cccc1)n1ccnc1)c1ccc(-c2ccc(Cl)cc2)cc1
```

```
Target \
247 MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFED...
500 MFIEAIVLALTALILYSVYSVKSFNTTRPTDPPVYPVTVPFLGHIV...
```

```
Mol \
247 <img data-content="rdkit/molecule" src="data:i...
500 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
247 (2246728737, 3975275337, 864674487, 2076190208...
500 (864942730, 1510328189, 2246699815, 1054767590...
```

```
mol2vec Clusters
247 (300,) dimensional vector 13
500 (300,) dimensional vector 13
First 2 entries for cluster: 14
```

```

-----
Drug \
18 NS(=O)(=O)c1ccc(C(=O)NCCOCCOCCN(CC(=O)[O-])CC(...
41 C[C@H](N)C(=O)Nc1nc2c(s1)CCc1ccccc1-2

Target \
18 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...
41 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...

Mol \
18 <img data-content="rdkit/molecule" src="data:i...
41 <img data-content="rdkit/molecule" src="data:i...

Sentence \
18 (847957139, 3275683399, 999334238, 3532053313,...
41 (2246728737, 3537119515, 2245273601, 265540621...

mol2vec Clusters
18 (300,) dimensional vector 14
41 (300,) dimensional vector 14
First 2 entries for cluster: 15
-----

Drug \
113 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...
443 CN(C)C[C@@H]1CCn2cc(c3ccccc32)C2=C(C(=O)NC2=O)...

Target \
113 MADNFSLHDALSGSGNPNPQGWPGAWGNQPAGAGGYPGASYPGAYP...
443 MRHSKRTHCPDWDSRESWGHESYRGSHKRKRSHSSTQENRHCKPH...

Mol \
113 <img data-content="rdkit/molecule" src="data:i...
443 <img data-content="rdkit/molecule" src="data:i...

Sentence \
113 (2246728737, 3975275337, 864674487, 1978844073...
443 (2246728737, 3818546315, 848128881, 3121777292...

mol2vec Clusters
113 (300,) dimensional vector 15
443 (300,) dimensional vector 15
First 2 entries for cluster: 16
-----

Drug \
172 C=CCc1ccccc1OCC(O)CNC(C)C
388 CN1CCN(c2ccc3nc(-c4c(N)c5c(F)cccc5[nH]c4=O)[nH...

Target \

```

172 MGAGVLVLGASEPGNLSSAAPLPDGAATAARLLVPASPPASLLPPA...
388 MTSSLQRPWRVPWLPWTILLVSTAAASQNQERLCAFKDPYQDLGI...

Mol \

172 <img data-content="rdkit/molecule" src="data:i...
388 <img data-content="rdkit/molecule" src="data:i...

Sentence \

172 (2246997334, 3696402029, 2246703798, 535847852...
388 (2246728737, 3657471097, 2092489639, 202181243...

mol2vec Clusters

172 (300,) dimensional vector 16
388 (300,) dimensional vector 16

First 2 entries for cluster: 17

Drug \

22 CN[C@@H](C)C(=O)N[C@H](C(=O)N1CCC[C@H]1C(=O)NC...
132 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

Target \

22 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...
132 MMLSLNNLQNIINPVIPFVGTIPDQLDPGTLIVIRGHVPSDADRF...

Mol \

22 <img data-content="rdkit/molecule" src="data:i...
132 <img data-content="rdkit/molecule" src="data:i...

Sentence \

22 (2246728737, 3824063894, 847961216, 1362425715...
132 (2246728737, 3975275337, 864674487, 1978844073...

mol2vec Clusters

22 (300,) dimensional vector 17
132 (300,) dimensional vector 17

First 2 entries for cluster: 18

Drug \

229 CC(COc1cccc(-c2nc(-c3ccncc3)c(-c3ccc4c(c3)CCC4...
231 O=NC1CCc2cc(-c3[nH]c(C4CCNCC4)nc3-c3ccncc3)ccc21

Target \

229 MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEE...
231 MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEE...

Mol \

229 <img data-content="rdkit/molecule" src="data:i...
231 <img data-content="rdkit/molecule" src="data:i...

Sentence \

229 (2246728737, 3537119515, 2245273601, 486030630...

231 (864942730, 2380084179, 847336149, 3977339131,...

mol2vec Clusters

229 (300,) dimensional vector 18

231 (300,) dimensional vector 18

First 2 entries for cluster: 19

Drug \

34 C[C@H](N)C(=O)Nc1nc(-c2ccc(Cc3ccccc3)cc2)c(-c2...

58 Cc1ccc(-n2nc(C(C)(C)C)cc2NC(=O)Nc2ccc(OCCN3CCO...

Target \

34 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...

58 HHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKH...

Mol \

34 <img data-content="rdkit/molecule" src="data:i...

58 <img data-content="rdkit/molecule" src="data:i...

Sentence \

34 (2246728737, 3537119515, 2245273601, 265540621...

58 (2246728737, 422715066, 3217380708, 3207567135...

mol2vec Clusters

34 (300,) dimensional vector 19

58 (300,) dimensional vector 19

First 2 entries for cluster: 20

Drug \

75 COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...

210 CC(C)(C)NCC(O)COc1cccc2c1C[C@H](O)[C@H](O)C2

Target \

75 HHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKH...

210 MAPWPHENSSLAPWPDLPNTANTSGLPVPWEAALAGALLAL...

Mol \

75 <img data-content="rdkit/molecule" src="data:i...

210 <img data-content="rdkit/molecule" src="data:i...

Sentence \

75 (2246728737, 3975275337, 864674487, 2076190208...

210 (2246728737, 3537123720, 2245277810, 144568923...

mol2vec Clusters


```

75 (300,) dimensional vector      20
210 (300,) dimensional vector     20
First 2 entries for cluster: 21
-----

Drug \
20 CN[C@@H](C)C(=O)N[C@H](C(=O)N1CCC[C@H]1C(=O)N[...
28 Cc1cc(C)c2ccc(NC(=O)[C@H](C)N)nc2n1

Target \
20 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...
28 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...

Mol \
20 <img data-content="rdkit/molecule" src="data:i...
28 <img data-content="rdkit/molecule" src="data:i...

Sentence \
20 (2246728737, 3824063894, 847961216, 1362425715...
28 (2246728737, 422715066, 3217380708, 4036277955...

mol2vec Clusters
20 (300,) dimensional vector      21
28 (300,) dimensional vector     21
First 2 entries for cluster: 22
-----

Drug \
0 COc1ccc(CNS(=O)(=O)c2ccc(S(N)(=O)=O)s2)cc1
142 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

Target \
0 MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...
142 MAFSGSQAPYLSPAVPFSGTIQGGLQDGLQITVNGTVLSSSGTRFA...

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

Sentence \
0 (2246728737, 3975275337, 864674487, 2076190208...
142 (2246728737, 3975275337, 864674487, 1978844073...

mol2vec Clusters
0 (300,) dimensional vector      22
142 (300,) dimensional vector    22
First 2 entries for cluster: 23
-----

Drug \
200 CC(C)NCC(O)COc1ccc(COCCOC(C)C)cc1

```

474 O=C1Nc2ccccc2/C1=C1/SC(=S)N(C(Cc2ccc(O)cc2)C(=...

Target \

200 MAPWPHENSSLAPWPDLPNTANTSGLPVGPWEAALAGALLAL...

474 MQPMTARFDLFVVGSGFFGLTIAERVATQLDKRVLVLERRPHIGGN...

Mol \

200 <img data-content="rdkit/molecule" src="data:i...

474 <img data-content="rdkit/molecule" src="data:i...

Sentence \

200 (2246728737, 3537119515, 2245273601, 263530394...

474 (864942730, 10565946, 3217380708, 606428793, 2...

mol2vec Clusters

200 (300,) dimensional vector 23

474 (300,) dimensional vector 23

First 2 entries for cluster: 24

Drug \

114 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

325 COCCNC(=O)[C@@H]1CCCN1C(=O)[C@H](N)C1CCCC1

Target \

114 MADNFSLHDALSGSGNPNPQGWPAGWGNQPAGAGGYPGASYPGAYP...

325 MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLQRRVRRANT...

Mol \

114 <img data-content="rdkit/molecule" src="data:i...

325 <img data-content="rdkit/molecule" src="data:i...

Sentence \

114 (2246728737, 3975275337, 864674487, 1978844073...

325 (2246728737, 3975275337, 864674487, 2222621677...

mol2vec Clusters

114 (300,) dimensional vector 24

325 (300,) dimensional vector 24

First 2 entries for cluster: 25

Drug \

90 CC1(C)SSC(C)(C)[C@@H](NC(=O)[C@@H](N)Cc2ccc(O)...

403 CNC(=O)c1cc(Oc2ccc(NC(=O)Nc3ccc(Cl)c(C(F)(F)F)...

Target \

90 MEPAPSAGAELQPPLFANASDAYPSACPSAGANASGPPGARSASSL...

403 MDYDFKAKLAAERERVELDFEYEGCKVGRGTYGHVYKARRKDGKDE...

Mol \

90 <img data-content="rdkit/molecule" src="data:i...
403 <img data-content="rdkit/molecule" src="data:i...

Sentence \

90 (2246728737, 1861965050, 2976816164, 387115942...
403 (2246728737, 3824063894, 847961216, 828793177,...

mol2vec Clusters

90 (300,) dimensional vector 25
403 (300,) dimensional vector 25
First 2 entries for cluster: 26

Drug \

149 CCCCCCCCCC(=O)[O-]
342 CCOC1CC2NCC(C#N)C(NC3CCC(F)C(C1)C3)C2CC1NC(=O)...

Target \

149 MGLEKTVKEKLSFEGVGIHTGEYSKLIHPEKEGTGIRFFKNGVYL...
342 MGAIGLLWLLPLLLSTAAVGSGMGTGQRAGSPAAGPPLQPREPLSY...

Mol \

149 <img data-content="rdkit/molecule" src="data:i...
342 <img data-content="rdkit/molecule" src="data:i...

Sentence \

149 (2246728737, 3542456614, 2245384272, 117312591...
342 (2246728737, 3542456614, 2245384272, 399408866...

mol2vec Clusters

149 (300,) dimensional vector 26
342 (300,) dimensional vector 26
First 2 entries for cluster: 27

Drug \

35 C[C@H](N)C(=O)Nc1nc(-c2cccc(Cc3ccccc3)c2)c(Cc2...
298 COc1ccc2c(c1)c(CC(=O)NCc1ccncc1)c(C)n2C(=O)c1c...

Target \

35 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...
298 MFIEAIVLALTALILYSVYSVKSFNTRPTDPPVYPVTVPFLGHIV...

Mol \

35 <img data-content="rdkit/molecule" src="data:i...
298 <img data-content="rdkit/molecule" src="data:i...

Sentence \

35 (2246728737, 3537119515, 2245273601, 265540621...

298 (2246728737, 3975275337, 864674487, 2076190208...

```
mol2vec Clusters
35 (300,) dimensional vector 27
298 (300,) dimensional vector 27
First 2 entries for cluster: 28
-----
```

```
Drug \
98 O=C(CCCN1CCC2(CC1)C(=O)NCN2c1cccc1)c1ccc(F)cc1
421 CCN1CCN(Cc2ccc(NC(=O)Nc3ccc(Oc4cc(NC)ncn4)cc3)...
```

```
Target \
98 MDPLNLSWYDDDLERQNWSRPFNGSDGKADRPHYNYATLLTLLIA...
421 MAKQYDSVECPFCDEVSKYEKLAKIGQGTGFEVFKARHRKTGQKVA...
```

```
Mol \
98 <img data-content="rdkit/molecule" src="data:i...
421 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
98 (864942730, 1510328189, 2246699815, 1909986973...
421 (2246728737, 3542456614, 2245384272, 225184566...
```

```
mol2vec Clusters
98 (300,) dimensional vector 28
421 (300,) dimensional vector 28
First 2 entries for cluster: 29
-----
```

```
Drug \
141 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...
336 CNC(=O)c1cc(Oc2ccc(NC(=O)Nc3ccc(Cl)c(C(F)(F)F)...
```

```
Target \
141 MAFSGSQAPYLSPAVPFSGTIQGGLQDGLQITVNGTVLSSSGTRFA...
336 MLEICLKLVGCKSKKGLSSSSCYLEEALQRPVASFEPQGLSEAA...
```

```
Mol \
141 <img data-content="rdkit/molecule" src="data:i...
336 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
141 (2246728737, 3975275337, 864674487, 1978844073...
336 (2246728737, 3824063894, 847961216, 828793177,...
```

```
mol2vec Clusters
141 (300,) dimensional vector 29
336 (300,) dimensional vector 29
First 2 entries for cluster: 30
```

Drug \

51 Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc...
 111 CCCCNC(=S)N[C@@H]1[C@@H](O)[C@H](O[C@@H]2[C@@H]...

Target \

51 MRGARGAWDFLCVLLLLLRVQTGSSQPSVSPGEPSPPSIHPGKSDL...
 111 MADNFSLHDALSGSGNPNPQGWPAGWGNQPAGAGGYPGASYPGAYP...

Mol \

51 <img data-content="rdkit/molecule" src="data:i...
 111 <img data-content="rdkit/molecule" src="data:i...

Sentence \

51 (2246728737, 422715066, 3217380708, 3124581743...
 111 (2246728737, 3542456614, 2245384272, 117312591...

mol2vec Clusters

51 (300,) dimensional vector 30
 111 (300,) dimensional vector 30
 First 2 entries for cluster: 31

Drug \

13 NS(=O)(=O)c1ccc(C(=O)NCc2cc(F)c(F)c(F)c2)cc1
 422 CC(C)(C)c1cnc(CSc2cnc(NC(=O)C3CCNCC3)s2)o1

Target \

13 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKPL...
 422 MAKQYDSVECPFCDEVSKYEKLAKIGQGTFGEVFKARHRKTGQKVA...

Mol \

13 <img data-content="rdkit/molecule" src="data:i...
 422 <img data-content="rdkit/molecule" src="data:i...

Sentence \

13 (847957139, 3275683399, 999334238, 3532053313,...
 422 (2246728737, 3537123720, 2245277810, 994237755...

mol2vec Clusters

13 (300,) dimensional vector 31
 422 (300,) dimensional vector 31
 First 2 entries for cluster: 32

Drug \

16 NS(=O)(=O)c1ccc(C(=O)NCCOCCOCCN(CC(=O)[O-])CC(...
 275 NC(=O)[C@@H](CCCN=C(N)N)NC(=O)[C@@H](CCCN=C(N)...

Target \

16 MASPDWGYDDKNGPEQWSKLYPIANGNNQSPVDIKTSETKHDTSLK...
275 MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDR...

Mol \

16 <img data-content="rdkit/molecule" src="data:i...
275 <img data-content="rdkit/molecule" src="data:i...

Sentence \

16 (847957139, 3275683399, 999334238, 3532053313,...
275 (847957139, 2591432844, 2246699815, 3304582332...

mol2vec Clusters

16 (300,) dimensional vector 32
275 (300,) dimensional vector 32
First 2 entries for cluster: 33

Drug \

72 COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...
258 NC(=O) [C@H] (CCCN=C(N)N)NC(=O)CCCCNC(=O) [C@H]...

Target \

72 HHSTVADGLITTLHYPAPKRNKPTVYGVSPPNYDKWEMERTDITMKH...
258 MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDR...

Mol \

72 <img data-content="rdkit/molecule" src="data:i...
258 <img data-content="rdkit/molecule" src="data:i...

Sentence \

72 (2246728737, 3975275337, 864674487, 2076190208...
258 (847957139, 2591432844, 2246699815, 3304582332...

mol2vec Clusters

72 (300,) dimensional vector 33
258 (300,) dimensional vector 33
First 2 entries for cluster: 34

Drug \

170 CC(C)NCC(O)COc1cccc2ccccc12
371 COc1cc2c(Nc3ccc(Br)cc3F)ncnc2cc1OCC1CCN(C)CC1

Target \

170 MGAGVLVLGASEPGLSSAAPLPDGAATAARLLVPASPPASLLPPA...
371 MSSPRAVVQLGKAQPAGEELATANQTAQQPSSPAMRRLTVDDFEIG...

Mol \

170 <img data-content="rdkit/molecule" src="data:i...
371 <img data-content="rdkit/molecule" src="data:i...

Sentence \

170 (2246728737, 3537119515, 2245273601, 263530394...

371 (2246728737, 3975275337, 864674487, 2076190208...

mol2vec Clusters

170 (300,) dimensional vector 34

371 (300,) dimensional vector 34

First 2 entries for cluster: 35

Drug \

88 COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...

285 C[C@@H](NC(=O)CCCCNC(=O)[C@H]1O[C@@H](n2cnc3c...

Target \

88 MRGARGAWDFLCVLLLLLRVQTGSSQPSVSPGEPSPPSIHPGKSDL...

285 MSRPPTGKMPGAPETAPGDGAGASRQRKLEALIRDPRSPINVESL...

Mol \

88 <img data-content="rdkit/molecule" src="data:i...

285 <img data-content="rdkit/molecule" src="data:i...

Sentence \

88 (2246728737, 3975275337, 864674487, 2076190208...

285 (2246728737, 3537119515, 2245273601, 263743996...

mol2vec Clusters

88 (300,) dimensional vector 35

285 (300,) dimensional vector 35

First 2 entries for cluster: 36

Drug \

107 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

365 Cc1ccc(Nc2nc(N(C)c3ccc4c(C)n(C)nc4c3)n2)cc1S...

Target \

107 MACGLVASNLNLKPGECLRVGEVAPDAKSFVLNLGKDSNNLCLHF...

365 MSSPRAVVQLGKAQPAGEELATANQTAQQPSSPAMRRLTVDDFEIG...

Mol \

107 <img data-content="rdkit/molecule" src="data:i...

365 <img data-content="rdkit/molecule" src="data:i...

Sentence \

107 (2246728737, 3975275337, 864674487, 1978844073...

365 (2246728737, 422715066, 3217380708, 3124581743...

mol2vec Clusters

107 (300,) dimensional vector 36

365 (300,) dimensional vector 36

First 2 entries for cluster: 37

Drug \

255 O=S(=O)(NCCNC/C=C/c1ccc(Br)cc1)c1cccc2cnccc12

341 CN1CCN(c2ccc3nc(-c4c(N)c5c(F)cccc5[nH]c4=O)[nH...

Target \

255 MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDR...

341 MGAIGLLWLLPLLLSTAAVSGSGMGTGQRAGSPAAGPPLQPREPLSY...

Mol \

255 <img data-content="rdkit/molecule" src="data:i...

341 <img data-content="rdkit/molecule" src="data:i...

Sentence \

255 (864942730, 2228063684, 999334238, 288315894, ...

341 (2246728737, 3657471097, 2092489639, 202181243...

mol2vec Clusters

255 (300,) dimensional vector 37

341 (300,) dimensional vector 37

First 2 entries for cluster: 38

Drug \

115 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

355 CCN(CC)CCNC(=O)c1c(C)[nH]c(/C=C2\C(=O)Nc3ccc(F...

Target \

115 MADNFSLHDALSGSGNPNPQGWPGAWGNQPAGAGGYPGASYPGAYP...

355 MDRSKENCISGPVKATAPVGGPKRVLVTQQFPCQNPLPVNSGQAQR...

Mol \

115 <img data-content="rdkit/molecule" src="data:i...

355 <img data-content="rdkit/molecule" src="data:i...

Sentence \

115 (2246728737, 3975275337, 864674487, 1978844073...

355 (2246728737, 3542456614, 2245384272, 592593828...

mol2vec Clusters

115 (300,) dimensional vector 38

355 (300,) dimensional vector 38

First 2 entries for cluster: 39

Drug \

23 CN[C@@H](C)C(=O)N[C@H](C(=O)N1CCC[C@H]1C(=O)N...

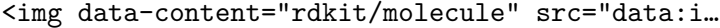
319 COCCNC(=O)[C@@H]1CCCN1C(=O)[C@H](N)Cc1ccc(Cl)cc1

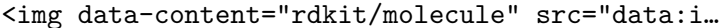
Target \

23 MSDAVSSDRNFNPSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...

319 MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLRVRRANT...

Mol \

23  rdkit/molecule" src="data:i...

319  rdkit/molecule" src="data:i...

Sentence \

23 (2246728737, 3824063894, 847961216, 1362425715...

319 (2246728737, 3975275337, 864674487, 2222621677...

mol2vec Clusters

23 (300,) dimensional vector 39

319 (300,) dimensional vector 39

First 2 entries for cluster: 40

Drug \

76 COc1cc2c(N3CCN(C(=O)Nc4ccc(OC(C)C)cc4)CC3)ncnc...

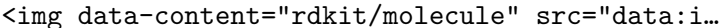
398 Cc1ccc(-n2nc(C(C)(C)C)cc2NC(=O)Nc2ccc(OCCN3CCO...

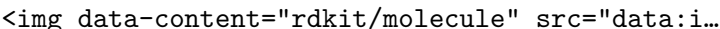
Target \

76 HHSTVADGLITTLHYPAKRNKPTVYGVSPNYDKWEMERTDITMKH...

398 MDYDFKAKLAAERERVEDLFHEYEGCKVGRGTYGHVYKARRKDGKDE...

Mol \

76  rdkit/molecule" src="data:i...

398  rdkit/molecule" src="data:i...

Sentence \

76 (2246728737, 3975275337, 864674487, 2076190208...

398 (2246728737, 422715066, 3217380708, 3207567135...

mol2vec Clusters

76 (300,) dimensional vector 40

398 (300,) dimensional vector 40

First 2 entries for cluster: 41

Drug \

124 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...

397 CCN1CCN(Cc2ccc(NC(=O)Nc3ccc(Oc4cc(NC)ncn4)cc3)...

Target \

124 MSNVPKSSSLPEGIRPGTVLRIRGLVPPNASRFHVNLLCGEEQGSD...

397 MDYDFKAKLAAERERVEDLFHEYEGCKVGRGTYGHVYKARRKDGKDE...

Mol \

124 <img data-content="rdkit/molecule" src="data:i...
397 <img data-content="rdkit/molecule" src="data:i...

Sentence \

124 (2246728737, 3975275337, 864674487, 1978844073...
397 (2246728737, 3542456614, 2245384272, 225184566...

mol2vec Clusters

124 (300,) dimensional vector 41
397 (300,) dimensional vector 41
First 2 entries for cluster: 42

Drug \

102 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...
281 NC(=O)[C@@H](CCCN=C(N)N)NC(=O)CCCCNC(=O)[C@H]...

Target \

102 MACGLVASNLNLKPGECLRVGEVAPDAKSFVLNLGKDSNNLCLHF...
281 MSRPPTGKMPGAPETAPGDGAGASRQRKLEALIRDPRSPINVESL...

Mol \

102 <img data-content="rdkit/molecule" src="data:i...
281 <img data-content="rdkit/molecule" src="data:i...

Sentence \

102 (2246728737, 3975275337, 864674487, 1978844073...
281 (847957139, 2591432844, 2246699815, 3304582332...

mol2vec Clusters

102 (300,) dimensional vector 42
281 (300,) dimensional vector 42
First 2 entries for cluster: 43

Drug \

103 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...
273 NCCCC[C@@H](NC(=O)CCCCNC(=O)[C@H]1O[C@@H](n2c...

Target \

103 MACGLVASNLNLKPGECLRVGEVAPDAKSFVLNLGKDSNNLCLHF...
273 MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDR...

Mol \

103 <img data-content="rdkit/molecule" src="data:i...
273 <img data-content="rdkit/molecule" src="data:i...

Sentence \

103 (2246728737, 3975275337, 864674487, 1978844073...

273 (847957139, 2592785365, 2245384272, 787341104,...

```
mol2vec Clusters
103 (300,) dimensional vector 43
273 (300,) dimensional vector 43
First 2 entries for cluster: 44
```

```
Drug \
204 CC(C)(C)NCC(O)C0c1cccc2[nH]c(=O)[nH]c12
346 CC[C@H](CO)Nc1nc(NCc2cccc2)c2ncn(C(C)C)c2n1
```

```
Target \
204 MAPWPHENSSLAPWPDLPNTANTSGLPVPWEAALAGALLAL...
346 MGAIGLLWLLPLLLSTA AVGSGMGTGQRAGSPAAGPPLQPREPLSY...
```

```
Mol \
204 <img data-content="rdkit/molecule" src="data:i...
346 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
204 (2246728737, 3537123720, 2245277810, 144568923...
346 (2246728737, 3542456614, 2245384272, 150656359...
```

```
mol2vec Clusters
204 (300,) dimensional vector 44
346 (300,) dimensional vector 44
First 2 entries for cluster: 45
```

```
Drug \
224 CCC1(C)CC(CCN)(c2ccc(OC)cc2)CCO1
301 C0c1ccc2c(c1)c(CC(=O)NCCc1cccc(Cl)c1)c(C)n2C(=...
```

```
Target \
224 MSAVALPRVSGGHDEHGHLEEFRTDPIGLMQRVRDECGDVGTFQLA...
301 MFIEAIVLALTALILYSVYSVKSFNTRPTDPPVYPVTVPLGHIV...
```

```
Mol \
224 <img data-content="rdkit/molecule" src="data:i...
301 <img data-content="rdkit/molecule" src="data:i...
```

```
Sentence \
224 (2246728737, 3542456614, 2245384272, 359304122...
301 (2246728737, 3975275337, 864674487, 2076190208...
```

```
mol2vec Clusters
224 (300,) dimensional vector 45
301 (300,) dimensional vector 45
First 2 entries for cluster: 46
```

Drug \

144 CS[C@@H]1O[C@H](CO)[C@H](O)[C@H](NC(=O)c2ccc(C...
173 Cc1ccc(Cl)c(OCC(O)CNC(C)(C)C)c1

Target \

144 MADNFSLHDALSGSGNPNPQGWPAGWGNQPAGAGGYPGASYPGAYP...
173 MGAGVLVLGASEPGNLSSAAPLPDGAATAARLLVPASPPASLLPPA...

Mol \

144 <img data-content="rdkit/molecule" src="data:i...
173 <img data-content="rdkit/molecule" src="data:i...

Sentence \

144 (2246728737, 2592252298, 1026928756, 39590684,...
173 (2246728737, 422715066, 3217380708, 3207567135...

mol2vec Clusters

144 (300,) dimensional vector 46
173 (300,) dimensional vector 46
First 2 entries for cluster: 47

Drug \

12 NS(=O)(=O)c1ccc(C(=O)NCc2ccc(F)c(F)c2F)cc1
326 N[C@@H](C(=O)N1CCC[C@H]1C(=O)NCc1cccs1)C1CCCCC1

Target \

12 SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDHTAKYDPSLKPL...
326 MAHVRGLQLPGCLALALCSLVHSQHVFLAPQQARSLLQRRVRRANT...

Mol \

12 <img data-content="rdkit/molecule" src="data:i...
326 <img data-content="rdkit/molecule" src="data:i...

Sentence \

12 (847957139, 3275683399, 999334238, 3532053313,...
326 (847957139, 2599973650, 2245273601, 1675508538...

mol2vec Clusters

12 (300,) dimensional vector 47
326 (300,) dimensional vector 47
First 2 entries for cluster: 48

Drug \

25 C[C@@H](C(=O)N[C@H](C(=O)N1C[C@@H](Oc2ccccc2)C...
348 CN[C@H]1C[C@@H]2O[C@](C)([C@H]1OC)n1c3ccccc3c3...

Target \

25 MSDAVSSDRNFPNSTNLPRNPSMADYEARIFTFGTWIYSVNKEQLA...
348 MGAIGLLWLLPLLLSTAAGSGMGTGQRAGSPAAGPPLQPREPLSY...

Mol \

25 <img data-content="rdkit/molecule" src="data:i...
348 <img data-content="rdkit/molecule" src="data:i...

Sentence \

25 (2246728737, 3537119515, 2245273601, 900670219...
348 (2246728737, 3824063894, 847961216, 1732628033...

mol2vec Clusters

25 (300,) dimensional vector 48
348 (300,) dimensional vector 48

First 2 entries for cluster: 49

Drug \

104 CO[C@@H]1O[C@H](CO)[C@@H](O[C@@H]2O[C@H](CO)[C...
333 C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1

Target \

104 MACGLVASNLNLKPGECLRVGEVAPDAKSFVLNLGKDSNNLCLHF...
333 MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAA...

Mol \

104 <img data-content="rdkit/molecule" src="data:i...
333 <img data-content="rdkit/molecule" src="data:i...

Sentence \

104 (2246728737, 3975275337, 864674487, 1978844073...
333 (2245897107, 3492293582, 2245900962, 189948681...

mol2vec Clusters

104 (300,) dimensional vector 49
333 (300,) dimensional vector 49

[]: