Setup

```
!pip install chemprop
!pip install rdkit-pypi  # should be included in above after Chemprop v1.6 release

import chemprop
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.offsetbox import AnchoredText
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.decomposition import PCA
```

```
Requirement already satisfied: py4j in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->chemprop) (
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1.
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1.3->c
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1.3
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1.3->
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.1.
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.0.3->chemp
Requirement already satisfied: xarray in /usr/local/lib/python3.10/dist-packages (from pandas-flavor>=0.2.0->chem
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.22.
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn
Requirement already satisfied: sphinxcontrib-applehelp in /usr/local/lib/python3.10/dist-packages (from sphinx>=3
Requirement already satisfied: sphinxcontrib-devhelp in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1
Requirement already satisfied: sphinxcontrib-jsmath in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.
Requirement already satisfied: sphinxcontrib-htmlhelp>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from sph
Requirement already satisfied: sphinxcontrib-serializinghtml>=1.1.5 in /usr/local/lib/python3.10/dist-packages (f
Requirement already satisfied: sphinxcontrib-qthelp in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.
Requirement already satisfied: Pygments>=2.0 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.2->chem
Requirement already satisfied: docutils<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.
Requirement already satisfied: snowballstemmer>=1.1 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.
Requirement already satisfied: babel>=1.3 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.2->chempro
Requirement already satisfied: alabaster<0.8,>=0.7 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.2
Requirement already satisfied: imagesize in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.2->chemprop
Requirement already satisfied: requests>=2.5.0 in /usr/local/lib/python3.10/dist-packages (from sphinx>=3.1.2->ch
Requirement already satisfied: protobuf>=3.20 in /usr/local/lib/python3.10/dist-packages (from tensorboardX>=2.0-
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from torch>=1.4.0->chemprop)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from torch>=1.4.0->c
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.4.0->chemprop) (1.
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from torch>=1.4.0->chemprop) (2
Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dist-packages (from torch>=1.4.0->chemp
Collecting typing-inspect>=0.7.1 (from typed-argument-parser>=1.6.1->chemprop)
  Downloading typing_inspect-0.9.0-py3-none-any.whl (8.8 kB)
Collecting docstring-parser>=0.15 (from typed-argument-parser>=1.6.1->chemprop)
  Downloading docstring_parser-0.15-py3-none-any.whl (36 kB)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=3.0->flas
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests>=2.5.0->sph
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests>=2.5.
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests>=2.5.
Collecting mypy-extensions>=0.3.0 (from typing-inspect>=0.7.1->typed-argument-parser>=1.6.1->chemprop)
  Downloading mypy_extensions-1.0.0-py3-none-any.whl (4.7 kB)
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.4.0-
Building wheels for collected packages: typed-argument-parser
  Building wheel for typed-argument-parser (setup.py) ... done
  Created wheel for typed-argument-parser: filename=typed_argument_parser-1.9.0-py3-none-any.whl size=25615 sha25
  Stored in directory: /root/.cache/pip/wheels/f0/94/0f/9539f578bed7e1bd423c702e403712f5ee8989f831a71db000
Successfully built typed-argument-parser
Installing collected packages: tensorboardX, rdkit, mypy-extensions, docstring-parser, typing-inspect, typed-argu
Successfully installed chemprop-1.6.1 docstring-parser-0.15 mypy-extensions-1.0.0 pandas-flavor-0.6.0 rdkit-2023.
Collecting rdkit-pypi
  Downloading rdkit_pypi-2022.9.5-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (29.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 29.4/29.4 MB 10.0 MB/s eta 0:00:00
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from rdkit-pypi) (1.25.2)
Requirement already satisfied: Pillow in /usr/local/lib/python3.10/dist-packages (from rdkit-pypi) (9.4.0)
Installing collected packages: rdkit-pypi
Successfully installed rdkit-pypi-2022.9.5
```

```
hiv_df = pd.read_csv("HIV.csv")
hiv_df.head()
```

	smiles	activity	HIV_active
0	CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC)CC(CC)...	CI	0
1	C(=Cc1ccccc1)C1=[O+][Cu-3]2([O+]=C(C=Cc3ccccc3...	CI	0
2	CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21	CI	0
3	Nc1ccc(C=Cc2ccc(N)cc2S(=O)(=O)O)c(S(=O)(=O)O)c1	CI	0
4	O=S(=O)(O)CCS(=O)(=O)O	CI	0

Next steps: View recommended plots

```
hiv_df.describe()
```

	HIV_active
count	20261.000000
mean	0.029219
std	0.168423
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

```
unique_values = hiv_df['HIV_active'].unique()
print(f"Unique values in 'HIV_active': {unique_values}")
```

```
    Unique values in 'HIV_active': [0 1]
```

```
# Filter rows where 'your_column' is not equal to 1 or 0
filtered_df = hiv_df[(hiv_df['HIV_active'] != 1) & (hiv_df['HIV_active'] != 0)]
filtered_df
```

smiles	activity	HIV_active

```
hiv_df.drop(['activity'], axis=1).to_csv('HIV_2.csv', index=False)
```

```
hiv_df_2 = pd.read_csv("HIV_2.csv")
```

```
hiv_df_2.head()
```

	smiles	HIV_active
0	CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC)CC(CC)...	0
1	C(=Cc1ccccc1)C1=[O+][Cu-3]2([O+]=C(C=Cc3ccccc3...	0
2	CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21	0
3	Nc1ccc(C=Cc2ccc(N)cc2S(=O)(=O)O)c(S(=O)(=O)O)c1	0
4	O=S(=O)(O)CCS(=O)(=O)O	0

Next steps: View recommended plots

```
# Filter rows where 'target_column' is equal to 1h
hiv_df_filtered = hiv_df_2[hiv_df_2['HIV_active'] == 1]
hiv_df_filtered
```

1 to 50 of 592 entries Filter

index	smiles	HIV_active
11	O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1	1
16	NNP(=S)(NN)c1ccccc1	1
80	O=Nc1ccc(O)c(N=O)c1O	1
203	Oc1ccc(Cl)cc1C(c1cc(Cl)ccc1O)C(Cl)(Cl)Cl	1
234	NNC(=O)c1ccccc1SSc1ccccc1C(=O)NN	1
235	O=C(NN=Cc1ccc(Cl)cc1Cl)c1ccccc1SSc1ccccc1C(=O)NN=Cc1ccc(Cl)cc1Cl	1
244	S=c1[nH][nH]c(=S)s1	1
271	Cc1cc(-c2ccc(N=Nc3cc(S(=O)(=O)O)c4ccccc4c3N)c(C)c2)ccc1N=Nc1cc(S(=O)(=O)O)c2ccccc2c1N	1
279	Nc1ccc2ccccc2c1N=Nc1ccc(C=Cc2ccc(N=Nc3c(N)ccc4ccccc34)cc2S(=O)(=O)O)c(S(=O)(=O)O)c1	1
326	Cc1c(C(=O)O)c(O)cc2c1C(=O)c1c(O)c(OC3OC(CO)C(O)C(O)C3O)c(O)c(O)c1C2=O	1
352	O=S(=O)(O)c1cc2ccc1ccc1ccc(cc1S(=O)(=O)O)nnc1ccc(ccc3ccc(cc3S(=O)(=O)O)nn2)c(S(=O)(=O)O)c1	1
353	Nc1c(N=Nc2ccc(-c3ccc(N=Nc4cc(S(=O)(=O)O)c5ccccc5c4O)cc3)cc2)cc(S(=O)(=O)O)c2ccccc12	1
361	O=S(c1ccccc1O)S(=O)c1ccccc1O	1
384	O=S1(=O)OC(c2cc(Cl)c(O)c(Br)c2)(c2cc(Cl)c(O)c(Br)c2)c2ccccc21	1
387	Cc1c(C2(c3cc(C(C)C)c(O)c(Br)c3C)OS(=O)(=O)c3ccccc32)cc(C(C)C)c(O)c1Br	1
429	COC1C(OC(=O)C=CC=CC=CC=CC(=O)O)CCC2(CO2)C1C1(C)OC1CC=C(C)C	1
434	CC1OC(OCC2OC(Oc3c(-c4ccc(O)cc4)oc4cc(OC5OC(C)C(O)C(O)C5O)cc(O)c4c3=O)C(O)C(O)C2O)C(O)C(O)C1O	1
443	COc1cc(-c2ccc(N=Nc3ccc4c(S(=O)(=O)O)cc(S(=O)(=O)O)c(N)c4c3O)c(OC)c2)ccc1N=Nc1ccc2c(S(=O)(=O)O)cc(S(=O)(=O)O)c(N)c2c1O	1
498	Nc1c(S(=O)(=O)O)cc2cc(S(=O)(=O)O)ccc2c1N=Nc1ccc(-c2ccc(N=Nc3c(N)c(S(=O)(=O)O)cc4cc(S(=O)(=O)O)ccc34)c(S(=O)(=O)O)c2)cc1	1
499	Cc1cc(-c2ccc(N=Nc3c(S(=O)(=O)O)cc4cc(S(=O)(=O)O)cc(N)c4c3O)c(C)c2)ccc1N=Nc1c(S(=O)(=O)O)cc2cc(S(=O)(=O)O)cc(N)c2c1O	1
654	C#CC1(O)CCC2C3CCC4=C(CCC(=O)C4)C3CCC21C	1
676	c1ccc2c3c(ccc2c1)O[Fe-4]12(Oc4ccc5ccccc5c4N=[O+]1)(Oc1ccc4ccccc4c1N=[O+]2)[O+]=N3	1
699	O=S1c2ccccc2Sc2ccccc21	1
740	COc1cc2nncc(S)c2cc1OC	1
818	CC(=O)C1=C(C)C2=Cc3c(C(C)=O)c(C)c4n3[Fe-3]35(Cl)n6c(c(C)c(CCC(=O)O)c6=CC6=[N+]3C(=C4)C(C)=C6CCC(=O)O)=CC1=[N+]25	1
869	N=C(N)Nc1ccccc1SSc1ccccc1NC(=N)N.O=S(=O)(O)O	1
879	COC(=S)SSC(=S)NCCN(CCNC(=S)SSC(=S)OC)C(=S)SSC(=S)OC	1
978	O=c1[nH]c2cc(Cl)c(Br)cc2o1	1
996	COc1ccc(C)cc1S(=O)(=O)c1c(Cl)cccc1[N+](=O)[O-]	1
1004	Cc1c(N=Nc2cc(N=Nc3ccc(N=Nc4ccc(O)c(C(=O)O)c4)cc3)c(N)cc2N)cccc1S(=O)(=O)O	1
1059	NC(=O)CSSCC(N)=O	1
1153	O=[N+]([O-])c1ccc(C=Cc2ccc(N=Nc3ccc(C=Cc4ccc([N+](=O)[O-])cc4S(=O)(=O)O)c(S(=O)(=O)O)c3)cc2S(=O)(=O)O)c(S(=O)(=O)O)c1	1
1195	O=C(O)c1cc(N=Nc2ccc(C=Cc3ccc(N=Nc4ccc(O)c(C(=O)O)c4)cc3S(=O)(=O)O)c(S(=O)(=O)O)c2)ccc1O	1
1196	O=S(=O)(O)c1cc(N=Nc2ccc(O)c3ccccc23)ccc1C=Cc1ccc(N=Nc2ccc(O)c3ccccc23)cc1S(=O)(=O)O	1
1198	O=S(=O)(O)c1cc(N=Nc2cc(S(=O)(=O)O)c3ccccc3c2O)ccc1C=Cc1ccc(N=Nc2cc(S(=O)(=O)O)c3ccccc3c2O)cc1S(=O)(=O)O	1
1199	O=C(O)c1cc2ccccc2c(N=Nc2ccc(C=Cc3ccc(N=Nc4c(O)c(C(=O)O)cc5ccccc45)cc3S(=O)(=O)O)c(S(=O)(=O)O)c2)c1O	1
1200	Cc1ccc(C(=O)Nc2ccc(S(=O)(=O)O)c3cc(S(=O)(=O)O)cc(S(=O)(=O)O)c23)cc1NC(=O)c1cccc(NC(=O)Nc2cccc(C(=O)Nc3cc(C(=O)Nc4ccc(S(=O)(=O)O)c5cc(S(=O)(=O)O)cc(S(=O)(=O)O)c45)ccc3C)c2)c1	1
1212	c1ccc2[nH]c(SSc3nc4ccccc4[nH]3)nc2c1	1
1266	O=C1c2ccccc2C(=O)c2c(S)cccc21	1
1277	Nc1cc(C(=O)O)ccc1SSc1ccc(C(=O)O)cc1N	1
1322	CC(=O)Nc1ccc(S2=Nc3ccccc3S2)cc1	1
1461	Cc1cc(Cc2cc(C)cc(S(=O)(=O)O)c2O)c(O)c(S(=O)(=O)O)c1	1
1482	COc1ccc(C2CC(=O)c3c(O)cc(OC4OC(COC5OC(C)C(O)C(O)C5O)C(O)C(O)C4O)cc3O2)cc1O	1
1510	CN(c1ccc(C(N)=O)cc1)c1c(O)nc2ccccc2c1O	1
1511	COc1ccc(OC)c(N=Nc2c(O)nc3ccccc3c2O)c1	1
1528	O=C(Nc1ccc2c(O)c(N=Nc3ccc4c(O)cc(S(=O)(=O)O)cc4c3)c(S(=O)(=O)O)cc2c1)c1ccccc1	1
1530	O=C(Nc1ccc(N=Nc2ccc(S(=O)(=O)O)cc2)cc1)c1ccc(N=Nc2ccc3c(S(=O)(=O)O)cccc3c2O)cc1	1
1532	O=C(Nc1ccc2c(O)cc(S(=O)(=O)O)cc2c1)Nc1ccc2c(O)c(N=Nc3ccc4cc(S(=O)(=O)O)ccc4c3)c(S(=O)(=O)O)cc2c1	1
1533	CC(=O)Nc1cc2c(O)c(N=Nc3ccc(C(=O)Nc4ccc(N=Nc5ccc(S(=O)(=O)O)cc5)cc4)cc3)c(S(=O)(=O)O)cc2cc1S(=O)(=O)O	1
1534	Cc1cc(NC(=O)c2ccc(N)cc2)ccc1N=Nc1ccc2c(O)c(N=Nc3ccc4cc(S(=O)(=O)O)ccc4c3)c(S(=O)(=O)O)cc2c1	1

Show 50 per page 1 2 10 12

Next steps: View recommended plots

```
bp_df = pd.read_csv("BBBP.csv")
bp_df.head()
```

	num	name	p_np	smiles
0	1	Propanolol	1	[Cl].CC(C)NCC(O)COc1cccc2ccccc12
1	2	Terbutylchlorambucil	1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl
2	3	40730	1	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...
3	4	24	1	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C
4	5	cloxacillin	1	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...

Next steps: View recommended plots

```
bp_df.drop(['num', 'name', 'p_np'], axis=1).to_csv('BBBP_2.csv', index=False)
```

```
bp_df_2 = pd.read_csv("BBBP_2.csv")
bp_df_2.head()
```

	smiles
0	[Cl].CC(C)NCC(O)COc1cccc2ccccc12
1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl
2	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...
3	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C
4	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...

Next steps: View recommended plots

```
arguments = [
    '--data_path', 'HIV_2.csv',
    '--dataset_type', 'classification',
    '--save_dir', 'test_checkpoints_multimolecule',
    '--epochs', '5',
    '--save_smiles_splits',
    '--quiet',
    '--batch_size', '2048'
]

args = chemprop.args.TrainArgs().parse_args(arguments)
```

```
mean_score, std_score = chemprop.train.cross_validate(args=args, train_func=chemprop.train.run_training)
```

```
20261it [00:00, 236491.33it/s]
100%|██████████| 20261/20261 [00:00<00:00, 32268.28it/s]
100%|██████████| 20261/20261 [00:06<00:00, 3342.72it/s]
Fold 0
20261it [00:00, 346024.27it/s]
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will cre
  warnings.warn(_create_warning_msg(
  0%|          | 0/5 [00:00<?, ?it/s]
  0%|          | 0/8 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557:
  warnings.warn(_create_warning_msg(

 12%|█▎        | 1/8 [00:53<06:13, 53.38s/it]
 25%|██▌       | 2/8 [00:55<02:19, 23.20s/it]
 38%|███▊      | 3/8 [00:59<01:12, 14.42s/it]
 50%|█████     | 4/8 [01:01<00:38,  9.62s/it]
 62%|██████▎   | 5/8 [01:04<00:21,  7.14s/it]
 75%|███████▌  | 6/8 [01:08<00:11,  5.94s/it]
 88%|████████▊ | 7/8 [01:10<00:04,  4.77s/it]
100%|██████████| 8/8 [01:12<00:00,  3.88s/it]

  0%|          | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:07<00:00,  7.08s/it]
 20%|██        | 1/5 [01:20<05:22, 80.56s/it]
  0%|          | 0/8 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557:
  warnings.warn(_create_warning_msg(

 12%|█▎        | 1/8 [00:52<06:04, 52.02s/it]
 25%|██▌       | 2/8 [00:55<02:21, 23.64s/it]
 38%|███▊      | 3/8 [00:58<01:10, 14.04s/it]
 50%|█████     | 4/8 [01:05<00:44, 11.18s/it]
 62%|██████▎   | 5/8 [01:07<00:23,  7.98s/it]
 75%|███████▌  | 6/8 [01:09<00:12,  6.05s/it]
 88%|████████▊ | 7/8 [01:12<00:04,  4.83s/it]
100%|██████████| 8/8 [01:14<00:00,  3.92s/it]

  0%|          | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:06<00:00,  6.53s/it]
 40%|████      | 2/5 [02:42<04:04, 81.50s/it]
  0%|          | 0/8 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557:
  warnings.warn(_create_warning_msg(

 12%|█▎        | 1/8 [00:46<05:28, 46.94s/it]
 25%|██▌       | 2/8 [00:55<02:27, 24.57s/it]
 38%|███▊      | 3/8 [00:59<01:16, 15.21s/it]
 50%|█████     | 4/8 [01:02<00:40, 10.11s/it]
 62%|██████▎   | 5/8 [01:04<00:22,  7.42s/it]
 75%|███████▌  | 6/8 [01:06<00:11,  5.61s/it]
 88%|████████▊ | 7/8 [01:09<00:04,  4.54s/it]
100%|██████████| 8/8 [01:11<00:00,  3.72s/it]

  0%|          | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:07<00:00,  7.00s/it]
 60%|██████    | 3/5 [04:02<02:41, 80.80s/it]
  0%|          | 0/8 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557:
  warnings.warn(_create_warning_msg(

 12%|█▎        | 1/8 [00:49<05:43, 49.10s/it]
 25%|██▌       | 2/8 [00:54<02:21, 23.54s/it]
```

```
mean_score, std_score
```

```
(0.5501948822233519, 0.0)
```

```
arguments = [
    '--test_path', 'BBBP_2.csv',
    '--preds_path', 'BBBP_preds.csv',
    '--checkpoint_dir', 'test_checkpoints_multimolecule'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args)
```

```
Loading training args
Setting molecule featurization parameters to default.
Loading data
2050it [00:00, 131105.98it/s]
100%|██████████| 2050/2050 [00:00<00:00, 100628.74it/s]Validating SMILES

[15:12:04] Explicit valence for atom # 1 N, 4, is greater than permitted
[15:12:04] WARNING: not removing hydrogen atom without neighbors
[15:12:04] Explicit valence for atom # 6 N, 4, is greater than permitted
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] Explicit valence for atom # 6 N, 4, is greater than permitted
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] Explicit valence for atom # 11 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 12 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] Explicit valence for atom # 5 N, 4, is greater than permitted
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
```

```
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
[15:12:05] WARNING: not removing hydrogen atom without neighbors
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will cre
```

```
smiles = [['CCC'], ['CCCC'], ['OCC']]
arguments = [
    '--test_path', '/dev/null',
    '--preds_path', '/dev/null',
    '--checkpoint_dir', 'test_checkpoints_multimolecule'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args, smiles=smiles)
```

```
Loading training args
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will cre
  warnings.warn(_create_warning_msg(
Setting molecule featurization parameters to default.
Loading data
Validating SMILES
Test size = 3
  0%|          | 0/1 [00:00<?, ?it/s]Loading pretrained parameter "encoder.encoder.0.cached_zero_vector".
Loading pretrained parameter "encoder.encoder.0.W_i.weight".
Loading pretrained parameter "encoder.encoder.0.W_h.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.bias".
Loading pretrained parameter "readout.1.weight".
Loading pretrained parameter "readout.1.bias".
Loading pretrained parameter "readout.4.weight".
Loading pretrained parameter "readout.4.bias".
Moving model to cuda

  0%|          | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:00<00:00,  1.95it/s]
100%|██████████| 1/1 [00:01<00:00,  1.23s/it]Saving predictions to /dev/null
Elapsed time = 0:00:02
```

```
preds
```

```
[[0.11991178244352341], [0.09433130919933319], [0.13542072474956512]]
```

```
smiles = [['CCCC(=O)Nc1ccc(OCC(O)CNC(C)C)c(c1)C(C)=O'], ['Nc1nc(NC2CC2)c2ncn([C@@H]3C[C@H](CO)C=C3)c2n1 |c:18|'], ['(
arguments = [
    '--test_path', '/dev/null',
    '--preds_path', '/dev/null',
    '--checkpoint_dir', 'test_checkpoints_multimolecule'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args, smiles=smiles)
print(preds)
```

```
Loading training args
Setting molecule featurization parameters to default.
Loading data
Validating SMILES
Test size = 5
  0%|          | 0/1 [00:00<?, ?it/s]Loading pretrained parameter "encoder.encoder.0.cached_zero_vector".
Loading pretrained parameter "encoder.encoder.0.W_i.weight".
Loading pretrained parameter "encoder.encoder.0.W_h.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.bias".
Loading pretrained parameter "readout.1.weight".
Loading pretrained parameter "readout.1.bias".
Loading pretrained parameter "readout.4.weight".
Loading pretrained parameter "readout.4.bias".
```

```
Moving model to cuda

  0%|          | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:00<00:00,  2.12it/s]
100%|██████████| 1/1 [00:00<00:00,  1.00it/s]Saving predictions to /dev/null
Elapsed time = 0:00:01
[[0.046882860362529755], [0.029342399910092354], [0.03106580302119255], [0.034708183258771896], [0.03785863891243
```

```
bp_preds_df = pd.read_csv("BBBP_preds.csv")
bp_preds_df.head()
```

	smiles	HIV_active
0	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.04202035814523697
1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	0.042415913194417953
2	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	0.021459585055708885
3	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	0.038869235664606094
4	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	0.021893160417675972

Next steps: View recommended plots

```
bp_preds_df.tail()
```

	smiles	HIV_active
2045	C1=C(Cl)C(=C(C2=C1NC(=O)C(N2)=O)[N+](=O)[O-])Cl	0.030196895822882652
2046	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...	0.02806824818253517
2047	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...	0.039688270539045334
2048	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)[NH-])C3=CC...	0.030639704316854477

```
bp_preds_df.describe()
```

	smiles	HIV_active
count	2050	2050
unique	2050	1997
top	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	Invalid SMILES
freq	1	11

```
bp_preds_df = bp_preds_df[bp_preds_df['HIV_active'] != "Invalid SMILES"]
bp_preds_df.describe()
```

	smiles	HIV_active
count	2039	2039
unique	2039	1996
top	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.049100425094366074
freq	1	3

```
bp_preds_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2039 entries, 0 to 2049
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   smiles      2039 non-null   object
 1   HIV_active  2039 non-null   object
dtypes: object(2)
memory usage: 47.8+ KB
```

```
bp_preds_df['HIV_active'] = bp_preds_df['HIV_active'].astype(float)
```

```
bp_preds_df['HIV_active_2'] = bp_preds_df['HIV_active'].apply(lambda x: 1 if x > 0.1 else 0)
bp_preds_df.head()
```

	smiles	HIV_active	HIV_active_2
0	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.042020	0
1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	0.042416	0
2	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	0.021460	0
3	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	0.038869	0
4	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	0.021893	0

Next steps: View recommended plots

```
bp_preds_df.describe()
```

	HIV_active	HIV_active_2
count	2039.000000	2039.000000
mean	0.030481	0.006376
std	0.014017	0.079612
min	0.006127	0.000000
25%	0.023413	0.000000
50%	0.028977	0.000000
75%	0.035188	0.000000
max	0.212985	1.000000

```
# Filter rows where 'target_column' is equal to 1
bp_preds_df_filtered = bp_preds_df[bp_preds_df['HIV_active_2'] == 1]
bp_preds_df_filtered
```

1 to 13 of 13 entries Filter ?

index	smiles	HIV_active	HIV_active_2
12	C(Cl)Cl	0.1286938637495041	1
120	[Na+].[Na+].[Na+].[O-]C(=O)[P]([O-])([O-])=O	0.1037321612238884	1
177	NC(N)=O	0.10228211432695389	1
433	ClCCl	0.1286938637495041	1
435	C=COC=C	0.11848479509353638	1
454	CCO	0.13927006721496582	1
509	CCCO	0.10898642987012863	1
510	CC(C)=O	0.10306501388549805	1
1203	C(C)Cl	0.1303427666425705	1
1204	C=C	0.21298544108867645	1
1502	[CH2-][CH-]C	0.13932982087135315	1
1658	C(=C)OC=C	0.11848479509353638	1
2044	[N+](=[N-])=O	0.13185620307922363	1

Show 50 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Next steps: View recommended plots

```
sub_df = pd.read_csv("substances.csv")
sub_df.head()
```

	zinc_id	smiles
0	ZINC000000000027	N[C@@H](CCc1ccc(N(CCCl)CCCl)cc1)C(=O)O
1	ZINC000016090786	N[C@H](CCc1ccc(N(CCCl)CCCl)cc1)C(=O)O
2	ZINC000001763088	N[C@H](CCCc1ccc(N(CCCl)CCCl)cc1)C(=O)O
3	ZINC000002033385	N[C@@H](CCCc1ccc(N(CCCl)CCCl)cc1)C(=O)O
4	ZINC000000001673	N[C@@H](Cc1ccc(N(CCCl)CCCl)cc1)C(=O)O

Next steps: View recommended plots

```
sub_df.tail()
```

	zinc_id	smiles
46	ZINC000196349655	O=C(O)CCSc1ccc(N(CCCl)CCCl)cc1
47	ZINC000064454242	N=NCCCc1ccc(N(CCCl)CCCl)cc1
48	ZINC000005161807	O=C(O)C/C=C/c1ccc(N(CCCl)CCCl)cc1
49	ZINC000001682294	O=C(O)CCOc1ccc(N(CCCl)CCCl)cc1
50	ZINC000079564304	O=C(O)CNC(=O)c1ccc(N(CCCl)CCCl)cc1

```
sub_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   zinc_id  51 non-null     object
 1   smiles   51 non-null     object
dtypes: object(2)
memory usage: 944.0+ bytes
```

```
arguments = [
    '--test_path', 'substances.csv',
    '--preds_path', 'substances_preds.csv',
    '--checkpoint_dir', 'test_checkpoints_multimolecule',
    '--smiles_columns', 'smiles'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args)
```

```
Loading training args
Setting molecule featurization parameters to default.
Loading data
51it [00:00, 45580.55it/s]
100%|██████████| 51/51 [00:00<00:00, 64998.33it/s]
/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will cre
  warnings.warn(_create_warning_msg(
Validating SMILES
Test size = 51
  0%|          | 0/1 [00:00<?, ?it/s]Loading pretrained parameter "encoder.encoder.0.cached_zero_vector".
Loading pretrained parameter "encoder.encoder.0.W_i.weight".
Loading pretrained parameter "encoder.encoder.0.W_h.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.bias".
Loading pretrained parameter "readout.1.weight".
Loading pretrained parameter "readout.1.bias".
Loading pretrained parameter "readout.4.weight".
Loading pretrained parameter "readout.4.bias".
Moving model to cuda

  0%|          | 0/2 [00:00<?, ?it/s]
 50%|█████     | 1/2 [00:00<00:00,  1.68it/s]
100%|██████████| 1/1 [00:01<00:00,  1.06s/it]Saving predictions to substances_preds.csv
Elapsed time = 0:00:01
```

```
fda_df = pd.read_csv("fda_approved.csv")
fda_df.head()
```

	zinc_id	smiles
0	ZINC000001530427	C[C@@H]1O[C@@H]1P(=O)(O)O
1	ZINC000003807804	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1
2	ZINC000000120286	Nc1nc(N)c2nc(-c3ccccc3)c(N)nc2n1
3	ZINC000242548690	C[C@H]1O[C@@H](O[C@H]2[C@@H](O)C[C@H](O[C@H]3[...
4	ZINC000000008492	Oc1cccc2cccnc12

Next steps: View recommended plots

```
arguments = [
    '--test_path', 'fda_approved.csv',
    '--preds_path', 'fda_approved_preds.csv',
    '--checkpoint_dir', 'test_checkpoints_multimolecule',
    '--smiles_columns', 'smiles'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args)
```

```
Loading training args
Setting molecule featurization parameters to default.
Loading data
892it [00:00, 160296.45it/s]
100%|██████████| 892/892 [00:00<00:00, 131565.19it/s]Validating SMILES

/usr/local/lib/python3.10/dist-packages/torch/utils/data/dataloader.py:557: UserWarning: This DataLoader will cre
  warnings.warn(_create_warning_msg(
Test size = 892
  0%|          | 0/1 [00:00<?, ?it/s]Loading pretrained parameter "encoder.encoder.0.cached_zero_vector".
Loading pretrained parameter "encoder.encoder.0.W_i.weight".
Loading pretrained parameter "encoder.encoder.0.W_h.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.weight".
Loading pretrained parameter "encoder.encoder.0.W_o.bias".
Loading pretrained parameter "readout.1.weight".
Loading pretrained parameter "readout.1.bias".
Loading pretrained parameter "readout.4.weight".
Loading pretrained parameter "readout.4.bias".
Moving model to cuda

  0%|          | 0/18 [00:00<?, ?it/s]
  6%|▌         | 1/18 [00:02<00:50,  2.96s/it]
 22%|██▏       | 4/18 [00:03<00:08,  1.62it/s]
 44%|████▍     | 8/18 [00:03<00:02,  3.84it/s]
 61%|██████    | 11/18 [00:03<00:01,  4.67it/s]
 94%|█████████▍| 17/18 [00:03<00:00,  8.70it/s]
100%|██████████| 1/1 [00:04<00:00,  4.51s/it]Saving predictions to fda_approved_preds.csv
Elapsed time = 0:00:05
```

```
fda_preds_df = pd.read_csv("fda_approved_preds.csv")
fda_preds_df.head()
```

	zinc_id	smiles	HIV_active
0	ZINC000001530427	C[C@@H]1O[C@@H]1P(=O)(O)O	0.028843
1	ZINC000003807804	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	0.027752
2	ZINC000000120286	Nc1nc(N)c2nc(-c3ccccc3)c(N)nc2n1	0.029231
3	ZINC000242548690	C[C@H]1O[C@@H](O[C@H]2[C@@H](O)C[C@H](O[C@H]3[...	0.013390
4	ZINC000000008492	Oc1cccc2cccnc12	0.034843

Next steps: View recommended plots

```
fda_preds_df = fda_preds_df[fda_preds_df['HIV_active'] != "Invalid SMILES"]
fda_preds_df.describe()
fda_preds_df['HIV_active'] = fda_preds_df['HIV_active'].astype(float)
fda_preds_df['HIV_active_2'] = fda_preds_df['HIV_active'].apply(lambda x: 1 if x > 0.1 else 0)
fda_preds_df.head()
```

	zinc_id	smiles	HIV_active	HIV_active_2
0	ZINC000001530427	C[C@@H]1O[C@@H]1P(=O)(O)O	0.028843	0
1	ZINC000003807804	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	0.027752	0
2	ZINC000000120286	Nc1nc(N)c2nc(-c3ccccc3)c(N)nc2n1	0.029231	0
		C[C@H]1O[C@@H]		

Next steps: View recommended plots

```
# Filter rows where 'target_column' is equal to 1
fda_preds_df_filtered = fda_preds_df[fda_preds_df['HIV_active_2'] == 1]
fda_preds_df_filtered
```

	zinc_id	smiles	HIV_active	HIV_active_2
598	ZINC000006827693	O=C(O)O	0.10556	1

```
!wget https://zinc15.docking.org/substances/subsets/named.csv?count=all
```

```
--2024-03-08 15:56:47--  https://zinc15.docking.org/substances/subsets/named.csv?count=all
Resolving zinc15.docking.org (zinc15.docking.org)... 169.230.75.4
Connecting to zinc15.docking.org (zinc15.docking.org)|169.230.75.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: ‘named.csv?count=all’

named.csv?count=all     [    <=>            ]   3.01M  28.5KB/s    in 1m 59s

2024-03-08 15:58:48 (25.8 KB/s) - Read error at byte 3160219 (Success).Retrying.

--2024-03-08 15:58:49--  (try: 2)  https://zinc15.docking.org/substances/subsets/named.csv?count=all
Connecting to zinc15.docking.org (zinc15.docking.org)|169.230.75.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: ‘named.csv?count=all’

named.csv?count=all     [         <=>       ]   3.03M  24.2KB/s    in 2m 0s

2024-03-08 16:00:49 (25.9 KB/s) - Read error at byte 3179558 (Success).Retrying.

--2024-03-08 16:00:51--  (try: 3)  https://zinc15.docking.org/substances/subsets/named.csv?count=all
Connecting to zinc15.docking.org (zinc15.docking.org)|169.230.75.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: ‘named.csv?count=all’

named.csv?count=all     [           <=>     ] 205.58K  28.7KB/s               ^C
```

```
zinc_df = pd.read_csv("named.csv")
zinc_df.head()
```

	zinc_id	smiles
0	ZINC000030727788	C=C[C@]1(C)C[C@@H](OC(=O)CSC(C)(C)CNC(=O)[C@H]...
1	ZINC000150377216	CCCCCC/C=C\C/C=C\CCCCCCCC(=O)OC[C@H](COCCCCCCC...
2	ZINC000100780125	CC(=O)O[C@H]1C[C@](C)(O)[C@@H]2CC=C(C)[C@@H]2[...
3	ZINC000006580536	O=C(O)[C@H](Cc1ccccc1)N(CCCl)CCCl
4	ZINC000150351802	O=C1C[C@H](c2ccc(O)c(O)c2)Oc2c1c(O)cc(O[C@H]1O...

Next steps: View recommended plots

```
arguments = [
    '--test_path', 'named.csv',
    '--preds_path', 'named_preds.csv',
    '--checkpoint_dir', 'test_checkpoints_multimolecule',
    '--smiles_columns', 'smiles'
]

args = chemprop.args.PredictArgs().parse_args(arguments)
preds = chemprop.train.make_predictions(args=args)
```

```
  5%|▌         | 37/697 [00:12<05:44,  1.91it/s]
  6%|▌         | 41/697 [00:12<03:10,  3.45it/s]
  6%|▋         | 44/697 [00:13<03:10,  3.43it/s]
  6%|▋         | 45/697 [00:13<02:54,  3.74it/s]
  7%|▋         | 49/697 [00:14<02:16,  4.76it/s]
  7%|▋         | 52/697 [00:15<02:23,  4.50it/s]
  8%|▊         | 53/697 [00:15<02:35,  4.15it/s]
  8%|▊         | 55/697 [00:15<02:01,  5.30it/s]
  8%|▊         | 57/697 [00:17<04:05,  2.60it/s]
  8%|▊         | 59/697 [00:17<03:07,  3.41it/s]
  9%|▊         | 60/697 [00:18<04:02,  2.63it/s]
  9%|▉         | 61/697 [00:18<03:44,  2.84it/s]
  9%|▉         | 64/697 [00:18<02:12,  4.77it/s]
  9%|▉         | 66/697 [00:19<03:06,  3.38it/s]
 10%|▉         | 68/697 [00:19<02:46,  3.78it/s]
 10%|▉         | 69/697 [00:20<03:35,  2.92it/s]
 10%|█         | 73/697 [00:21<02:19,  4.48it/s]
 11%|█         | 76/697 [00:21<02:31,  4.11it/s]
 11%|█         | 77/697 [00:22<02:25,  4.27it/s]
 11%|█▏        | 80/697 [00:22<01:37,  6.31it/s]
 12%|█▏        | 82/697 [00:23<02:32,  4.03it/s]
 12%|█▏        | 83/697 [00:23<02:18,  4.44it/s]
 12%|█▏        | 84/697 [00:24<03:19,  3.07it/s]
 12%|█▏        | 85/697 [00:24<03:33,  2.87it/s]
 12%|█▏        | 87/697 [00:24<02:25,  4.20it/s]
 13%|█▎        | 89/697 [00:25<02:37,  3.85it/s]
 13%|█▎        | 92/697 [00:25<02:31,  4.00it/s]
 13%|█▎        | 93/697 [00:26<02:24,  4.17it/s]
 14%|█▍        | 97/697 [00:26<01:55,  5.21it/s]
 14%|█▍        | 100/697 [00:27<02:07,  4.67it/s]
 14%|█▍        | 101/697 [00:27<02:36,  3.80it/s]
 15%|█▍        | 104/697 [00:28<01:46,  5.58it/s]
 15%|█▌        | 106/697 [00:28<01:55,  5.12it/s]
 15%|█▌        | 108/697 [00:29<02:20,  4.18it/s]
 16%|█▌        | 109/697 [00:29<02:13,  4.39it/s]
 16%|█▌        | 113/697 [00:29<01:47,  5.44it/s]
 16%|█▋        | 115/697 [00:30<01:27,  6.68it/s]
 17%|█▋        | 117/697 [00:30<02:06,  4.60it/s]
 17%|█▋        | 120/697 [00:31<01:29,  6.46it/s]
 18%|█▊        | 122/697 [00:31<01:44,  5.50it/s]
 18%|█▊        | 124/697 [00:32<02:03,  4.66it/s]
 18%|█▊        | 125/697 [00:32<01:55,  4.94it/s]
 19%|█▊        | 129/697 [00:32<01:35,  5.92it/s]
 19%|█▉        | 131/697 [00:32<01:19,  7.13it/s]
 19%|█▉        | 133/697 [00:34<02:19,  4.05it/s]
 20%|█▉        | 137/697 [00:34<01:49,  5.10it/s]
 20%|█▉        | 139/697 [00:34<01:30,  6.16it/s]
 20%|██        | 141/697 [00:35<02:40,  3.47it/s]
 21%|██        | 143/697 [00:36<02:04,  4.43it/s]
 21%|██        | 145/697 [00:37<02:41,  3.41it/s]
 21%|██        | 146/697 [00:37<02:28,  3.70it/s]
 21%|██        | 147/697 [00:37<02:16,  4.02it/s]
 21%|██        | 148/697 [00:38<03:41,  2.48it/s]
 21%|██▏       | 149/697 [00:38<03:20,  2.73it/s]
 22%|██▏       | 152/697 [00:38<01:49,  4.96it/s]
 22%|██▏       | 154/697 [00:39<02:09,  4.20it/s]
 22%|██▏       | 156/697 [00:40<02:31,  3.58it/s]
 23%|██▎       | 157/697 [00:40<02:22,  3.80it/s]
```

	zinc_id	smiles	HIV_active
0	ZINC000030727788	C=C[C@]1(C)C[C@@H](OC(=O)CSC(C)(C)CNC(=O)[C@H]...	0.018594
1	ZINC000150377216	CCCCCC/C=C\C/C=C\CCCCCCCC(=O)OC[C@H](COCCCCCCCC...	0.049322
2	ZINC000100780125	CC(=O)O[C@H]1C[C@](C)(O)[C@@H]2CC=C(C)[C@@H]2[...	0.019618
3	ZINC000006580536	O=C(O)[C@H](Cc1ccccc1)N(CCCl)CCCl	0.043474

Next steps: View recommended plots

```
zinc_preds_df = pd.read_csv("named_preds.csv")
zinc_preds_df.head()
zinc_preds_df = zinc_preds_df[zinc_preds_df['HIV_active'] != "Invalid SMILES"]
zinc_preds_df.describe()
zinc_preds_df['HIV_active'] = zinc_preds_df['HIV_active'].astype(float)
zinc_preds_df['HIV_active_2'] = zinc_preds_df['HIV_active'].apply(lambda x: 1 if x > 0.1 else 0)
zinc_preds_df.head()
```

	zinc_id	smiles	HIV_active	HIV_ac
0	ZINC000030727788	C=C[C@]1(C)C[C@@H](OC(=O)CSC(C)(C)CNC(=O)[C@H]...	0.018594	
1	ZINC000150377216	CCCCCC/C=C\C/C=C\CCCCCCCC(=O)OC[C@H](COCCCCCCC...	0.049322	
2	ZINC000100780125	CC(=O)O[C@H]1C[C@](C)(O)[C@@H]2CC=C(C)[C@@H]2[...	0.019618	
3	ZINC000006580536	O=C(O)[C@H](Cc1ccccc1)N(CCCl)CCCl	0.043474	

Next steps: View recommended plots

```
# Filter rows where 'target_column' is equal to 1
zinc_preds_df_filtered = zinc_preds_df[zinc_preds_df['HIV_active_2'] == 1]
zinc_preds_df_filtered
```

	zinc_id	smiles	HIV_active	HIV_active_2
920	ZINC000004978447	OCCBr	0.110170	1
1479	ZINC000030728765	O=CCBr	0.124223	1
1694	ZINC000070667158	COCCl	0.105435	1
2015	ZINC000012153803	CC/C=C\CC=O	0.104406	1
3924	ZINC000005224354	OCCO	0.115865	1
4057	ZINC000001530493	O=C/C=C/Cl	0.118323	1
6038	ZINC000008214610	CNC=O	0.124328	1
8339	ZINC000001530491	OC/C=C/Cl	0.120551	1
9856	ZINC000001658755	CSCSSC	0.100051	1
10845	ZINC000000897143	C=CC=O	0.129973	1
12691	ZINC000000901212	C=CC#N	0.113243	1
13076	ZINC000004429713	CS[S@](C)=O	0.104607	1
13156	ZINC000017856915	CCN=O	0.120628	1
13450	ZINC000058593509	CCON	0.105297	1
14716	ZINC000002038899	C=CCCC=O	0.104293	1
14838	ZINC000064624118	COOC	0.110023	1
15179	ZINC000008830539	CN=C=S	0.114074	1
15989	ZINC000005132894	C/C=C/SC	0.126821	1
17540	ZINC000002031662	CCNC	0.102240	1
20217	[illegible]	[illegible]	[illegible]	1
21806	ZINC000008220697	OCCCl	0.109488	1

Next

View recommended plots

```
from google.colab import drive
drive.mount('/content/drive')
```

```
    Mounted at /content/drive
```

```
!mkdir '/content/drive/My Drive/Chemprop_Backup/'
```

```
!pwd
```

```
    /content
```

```
!ls -al
```

```
    total 10564
    drwxr-xr-x 1 root root    4096 Mar  8 16:13 .
```