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: tensorboardX>=2.0 in /usr/local/lib/python3.10/dist-packages (from chemprop) (2.6.
    Requirement already satisfied: torch>=1.4.0 in /usr/local/lib/python3.10/dist-packages (from chemprop) (2.1.0+cu1
    Requirement already satisfied: tqdm>=4.45.0 in /usr/local/lib/python3.10/dist-packages (from chemprop) (4.66.2)
    Requirement already satisfied: typed-argument-parser>=1.6.1 in /usr/local/lib/python3.10/dist-packages (from chem
    Requirement already satisfied: rdkit>=2020.03.1.0 in /usr/local/lib/python3.10/dist-packages (from chemprop) (202
    Requirement already satisfied: Werkzeug>=2.2.2 in /usr/local/lib/python3.10/dist-packages (from flask>=1.1.2->che
    Requirement already satisfied: Jinja2>=3.0 in /usr/local/lib/python3.10/dist-packages (from flask>=1.1.2->chempro
    Requirement already satisfied: itsdangerous>=2.0 in /usr/local/lib/python3.10/dist-packages (from flask>=1.1.2->c
    Requirement already satisfied: click>=8.0 in /usr/local/lib/python3.10/dist-packages (from flask>=1.1.2->chemprop
    Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->chemprop)
    Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->chemprop) (1
    Requirement already satisfied: networkx>=2.2 in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->ch
    Requirement already satisfied: future in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->chemprop)
    Requirement already satisfied: cloudpickle in /usr/local/lib/python3.10/dist-packages (from hyperopt>=0.2.3->chem
    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->chemprop
    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
    Requirement already satisfied: typing-inspect>=0.7.1 in /usr/local/lib/python3.10/dist-packages (from typed-argum
    Requirement already satisfied: docstring-parser>=0.15 in /usr/local/lib/python3.10/dist-packages (from typed-argu
    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.
    Requirement already satisfied: mypy-extensions>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from typing-ins
    Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.4.0-
    Requirement already satisfied: rdkit-pypi in /usr/local/lib/python3.10/dist-packages (2022.9.5)
    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)
```

```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

```
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	41127.000000
mean	0.035086
std	0.184001
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]
```

```
unique_values = hiv_df['smiles'].unique()
print(f"Unique values in 'smiles': {unique_values}")
print(f"length of uniqe value: {len(unique_values)}")
```

```
Unique values in 'smiles': ['CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC)CC(CC)=[O+]2'
 'C(=Cc1ccccc1)C1=[O+][Cu-3]2([O+]=C(C=Cc3ccccc3)CC(c3ccccc3)=[O+]2)[O+]=C(c2ccccc2)C1'
 'CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21' ...
 'Cc1ccc(N2C(=O)C3c4[nH]c5ccccc5c4C4CCC(C(C)(C)C)CC4C3C2=O)cc1'
 'Cc1cccc(N2C(=O)C3c4[nH]c5ccccc5c4C4CCC(C(C)(C)C)CC4C3C2=O)c1'
 'CCCCCC=C(c1cc(Cl)c(OC)c(-c2nc(C)no2)c1)c1cc(Cl)c(OC)c(-c2nc(C)no2)c1']
length of uniqe value: 41127
```

```
# 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

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

```
                                                    smiles  activity  HIV_active
11                O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1        CM           1
16                                   NNP(=S)(NN)c1ccccc1        CM           1
80                                   O=Nc1ccc(O)c(N=O)c1O        CM           1
203                Oc1ccc(Cl)cc1C(c1cc(Cl)ccc1O)C(Cl)(Cl)Cl     CM           1
234                  NNC(=O)c1ccccc1SSc1ccccc1C(=O)NN          CM           1
...                                                   ...       ...         ...
41090          Cc1cn(COCCCOCC(=O)c2ccccc2)c(=O)[nH]c1=O        CM           1
41092   Cc1cn(C2CC3C(COC(CCC[Se]c4ccccc4)N3O)O2)c(=O)[...      CM           1
41093  Cc1cn(C2CC3C(COC(CCCC[Se]c4ccccc4)N3O)O2)c(=O)...       CM           1
41098   Cc1cn(C2CC3C(COC(CC[Se]C#N)N3O)O2)c(=O)[nH]c1=O        CM           1
41099      C[Se]CCC1OCC2OC(n3cc(C)c(=O)[nH]c3=O)CC2N1O        CA           1

1443 rows × 3 columns
```

Next steps: View recommended plots

```
# Filter rows where 'target_column' is equal to 1h
hiv_df_filtered_inactive = hiv_df[hiv_df['HIV_active'] == 0]
hiv_df_filtered_inactive = hiv_df_filtered_inactive.sample(n=1500, axis=0, replace=True)
hiv_df_filtered_inactive
```

```
                                                     smiles  activity  HIV_active
26461  CON=C(C(=O)OC(=NC1CCCCC1)NC1CCCCC1)c1csc(NC(c2...        CI           0
40599                          O=C(C=CC(=O)c1cccs1)c1cccs1        CI           0
3248                       N=c1c2c(ncn1N)CCN(Cc1ccccc1)C2        CI           0
14536              COC(=O)C12C=CC(=O)C3CC(C(C(C)C)C1)C32OC        CI           0
39045       N#CCCN(CCC#N)c1ccc(C=C2N=C(c3ccccc3)N(c3ccc(C(...       CI           0
...                                                     ...       ...         ...
16582          CCC(C)c1cccc(C)c1NC(=O)C(=O)Cc1nc2ccccc2s1        CI           0
6013                             Cn1c2ccccc2c2nn3cnnc3nc21        CI           0
14727                       CCOC(=O)CSc1c([N+](=O)[O-])ncn1C        CI           0
27104       COc1cccc(C=[N+]2[N-]C(c3ccncc3)=[O+][Co-4]2(O)...       CI           0
19994          CC(=O)NC(CCCNC(=O)N(C)N=O)C(=O)NCc1ccccc1        CI           0

1500 rows × 3 columns
```

Next steps: View recommended plots

```
hiv_df_sampled = pd.concat([hiv_df_filtered_active, hiv_df_filtered_inactive], axis=0, ignore_index=True)
hiv_df_sampled
```

	smiles	activity	HIV_active
0	O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1	CM	1
1	NNP(=S)(NN)c1ccccc1	CM	1
2	O=Nc1ccc(O)c(N=O)c1O	CM	1
3	Oc1ccc(Cl)cc1C(c1cc(Cl)ccc1O)C(Cl)(Cl)Cl	CM	1
4	NNC(=O)c1ccccc1SSc1ccccc1C(=O)NN	CM	1
...	...	...	...
2938	CCC(C)c1cccc(C)c1NC(=O)C(=O)Cc1nc2ccccc2s1	CI	0
2939	Cn1c2ccccc2c2nn3cnnc3nc21	CI	0
2940	CCOC(=O)CSc1c([N+](=O)[O-])ncn1C	CI	0
2941	COc1cccc(C=[N+]2[N-]C(c3ccncc3)=[O+][Co-4]2(O)...	CI	0
2942	CC(=O)NC(CCCNC(=O)N(C)N=O)C(=O)NCc1ccccc1	CI	0

2943 rows × 3 columns

Next steps: View recommended plots

```
hiv_df_sampled.to_csv('HIV_2.csv', index=False)
# .drop(['activity'], axis=1).
hiv_df_sampled_2 = pd.read_csv("HIV_2.csv")
hiv_df_sampled_2.head()
hiv_df_sampled_2.tail()
```

	smiles	activity	HIV_active
0	O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1	CM	1
1	NNP(=S)(NN)c1ccccc1	CM	1
2	O=Nc1ccc(O)c(N=O)c1O	CM	1
3	Oc1ccc(Cl)cc1C(c1cc(Cl)ccc1O)C(Cl)(Cl)Cl	CM	1
4	NNC(=O)c1ccccc1SSc1ccccc1C(=O)NN	CM	1

	smiles	activity	HIV_active
2938	CCC(C)c1cccc(C)c1NC(=O)C(=O)Cc1nc2ccccc2s1	CI	0
2939	Cn1c2ccccc2c2nn3cnnc3nc21	CI	0
2940	CCOC(=O)CSc1c([N+](=O)[O-])ncn1C	CI	0
2941	COc1cccc(C=[N+]2[N-]C(c3ccncc3)=[O+][Co-4]2(O)...	CI	0
2942	CC(=O)NC(CCCNC(=O)N(C)N=O)C(=O)NCc1ccccc1	CI	0

```
arguments = [
    '--data_path', 'HIV_2.csv',
    '--dataset_type', 'classification',
    '--save_dir', 'test_checkpoints_multimolecule',
    '--epochs', '30',
    '--save_smiles_splits',
    '--quiet',
    '--batch_size', '64',
    '--ignore_columns', 'activity',
    '--depth', '5',
    '--hidden_size', '300'
]

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

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

```
 92%|█████████▏| 34/37 [00:03<00:00,  9.35it/s]
 95%|█████████▍| 35/37 [00:03<00:00,  9.49it/s]
100%|██████████| 37/37 [00:03<00:00, 10.29it/s]

  0%|          | 0/5 [00:00<?, ?it/s]
 40%|████      | 2/5 [00:00<00:00, 10.97it/s]
 80%|████████  | 4/5 [00:00<00:00, 11.15it/s]
 97%|█████████▋| 29/30 [02:18<00:04,  4.77s/it]
  0%|          | 0/37 [00:00<?, ?it/s]
  3%|▎         | 1/37 [00:00<00:03,  9.40it/s]
  5%|▌         | 2/37 [00:00<00:03,  9.04it/s]
 11%|█         | 4/37 [00:00<00:03, 10.15it/s]
 14%|█▎        | 5/37 [00:00<00:03,  9.91it/s]
 16%|█▌        | 6/37 [00:00<00:03,  9.71it/s]
 19%|█▉        | 7/37 [00:00<00:03,  9.18it/s]
 22%|██▏       | 8/37 [00:00<00:03,  9.18it/s]
 24%|██▍       | 9/37 [00:00<00:03,  9.04it/s]
 27%|██▋       | 10/37 [00:01<00:02,  9.01it/s]
 30%|██▉       | 11/37 [00:01<00:02,  9.15it/s]
 32%|███▏      | 12/37 [00:01<00:02,  9.35it/s]
 35%|███▌      | 13/37 [00:01<00:02,  9.13it/s]
 38%|███▊      | 14/37 [00:01<00:02,  9.25it/s]
 41%|████      | 15/37 [00:01<00:02,  9.39it/s]
 43%|████▎     | 16/37 [00:01<00:02,  9.20it/s]
 46%|████▌     | 17/37 [00:01<00:02,  9.10it/s]
 49%|████▊     | 18/37 [00:01<00:02,  9.24it/s]
 51%|█████▏    | 19/37 [00:02<00:01,  9.43it/s]
 54%|█████▍    | 20/37 [00:02<00:01,  9.08it/s]
 57%|█████▋    | 21/37 [00:02<00:01,  9.14it/s]
 62%|██████▏   | 23/37 [00:02<00:01,  9.95it/s]
 65%|██████▍   | 24/37 [00:02<00:01,  9.73it/s]
 68%|██████▊   | 25/37 [00:02<00:01,  9.78it/s]
 70%|███████   | 26/37 [00:02<00:01,  9.35it/s]
 76%|███████▌  | 28/37 [00:02<00:00,  9.51it/s]
 78%|███████▊  | 29/37 [00:03<00:00,  9.55it/s]
 81%|████████  | 30/37 [00:03<00:00,  9.56it/s]
 84%|████████▍ | 31/37 [00:03<00:00,  9.58it/s]
 89%|████████▉ | 33/37 [00:03<00:00, 10.12it/s]
 92%|█████████▏| 34/37 [00:03<00:00,  9.90it/s]
 95%|█████████▍| 35/37 [00:03<00:00,  9.73it/s]
 97%|█████████▋| 36/37 [00:03<00:00,  9.47it/s]

  0%|          | 0/5 [00:00<?, ?it/s]
 40%|████      | 2/5 [00:00<00:00, 11.86it/s]
 80%|████████  | 4/5 [00:00<00:00, 11.65it/s]
100%|██████████| 30/30 [02:23<00:00,  4.78s/it]
Model 0 best validation auc = 0.841157 on epoch 19
Model 0 test auc = 0.805663
Ensemble test auc = 0.805663
1-fold cross validation
        Seed 0 ==> test auc = 0.805663
Overall test auc = 0.805663 +/- 0.000000
Elapsed time = 0:02:25
```

```
mean_score, std_score
```

```
(0.8056628056628057, 0.0)
```

```
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.tail()
```

	num	name	p_np	smiles
2045	2049	licostinel	1	C1=C(Cl)C(=C(C2=C1NC(=O)C(N2)=O)[N+](=O)[O-])Cl
2046	2050	ademetionine(adenosyl-methionine)	1	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...
2047	2051	mesocarb	1	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...
2048	2052	[illegible]	1	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)

```
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()
bp_df_2.tail()
```

	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)...

	smiles
2045	C1=C(Cl)C(=C(C2=C1NC(=O)C(N2)=O)[N+](=O)[O-])Cl
2046	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...
2047	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...
2048	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)[NH-])C3=CC...
2049	[N+](=NCC(=O)N[C@@H]([C@H](O)C1=CC=C([N+]([O-]...

```
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, 120747.70it/s]
100%|██████████| 2050/2050 [00:00<00:00, 92757.30it/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 = 2,039
  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%|          | 0/41 [00:00<?, ?it/s]
  2%|▏         | 1/41 [00:04<03:04,  4.60s/it]
  7%|▋         | 3/41 [00:04<00:50,  1.33s/it]
 12%|█▏        | 5/41 [00:05<00:25,  1.44it/s]
 17%|█▋        | 7/41 [00:05<00:14,  2.29it/s]
 20%|█▉        | 8/41 [00:05<00:12,  2.63it/s]
 22%|██▏       | 9/41 [00:07<00:28,  1.14it/s]
 27%|██▋       | 11/41 [00:08<00:17,  1.74it/s]
 32%|███▏      | 13/41 [00:08<00:11,  2.53it/s]
 37%|███▋      | 15/41 [00:08<00:07,  3.62it/s]
 41%|████▏     | 17/41 [00:09<00:06,  3.73it/s]
 46%|████▋     | 19/41 [00:09<00:04,  5.04it/s]
 51%|█████     | 21/41 [00:09<00:03,  6.13it/s]
 56%|█████▋    | 23/41 [00:09<00:02,  7.02it/s]
 61%|██████    | 25/41 [00:09<00:02,  5.82it/s]
 68%|██████▊   | 28/41 [00:10<00:01,  8.43it/s]
 78%|███████▊  | 32/41 [00:10<00:00, 12.69it/s]
 85%|████████▌ | 35/41 [00:10<00:00, 14.16it/s]
 93%|█████████▎| 38/41 [00:10<00:00, 16.91it/s]
100%|██████████| 1/1 [00:11<00:00, 11.22s/it]Saving predictions to BBBP_preds.csv
Elapsed time = 0:00:12
```

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

	smiles	HIV_active
0	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.08685699850320816
1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	0.03052304871380329
2	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	0.6467068791389465
3	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	0.06845816969871521
4	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	0.4430862367153168

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	0.2801685929298401
2046	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...	0.15042510628700256
2047	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...	0.5994424819946289
2048	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)[NH-])C3=CC...	0.30857348442077637
2049	[N+](=NCC(=O)N[C@@H]([C@H](O)C1=CC=C([N+]([O-]...	0.4408627152442932

```
bp_preds_df.describe()
```

	smiles	HIV_active
count	2050	2050
unique	2050	2004
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	2003
top	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.13198897242546082
freq	1	3

```
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 > 0.8 else 0)
bp_preds_df.head()
```

	smiles	HIV_active	HIV_active_2
0	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	0.086857	0
1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	0.030523	0
2	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	0.646707	0
3	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	0.068458	0
4	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	0.443086	0

Next steps: View recommended plots

```
bp_preds_df.describe()
```

	HIV_active	HIV_active_2
count	2039.000000	2039.000000
mean	0.322916	0.078960
std	0.255242	0.269743
min	0.000220	0.000000
25%	0.126423	0.000000
50%	0.242825	0.000000
75%	0.457704	0.000000
max	0.996994	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
```

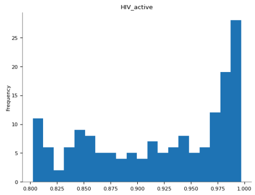
1 to 25 of 161 entries Filter

index	smiles
5	CCN1CCN(C(=O)N[C@@H](C(=O)N[C@H]2[C@H]3SCC(=C(N3C2=O)C(O)=O)CSc4nnnn4C)c5ccc(O)cc5)C(=O)C1=O
6	CN(C)[C@H]1[C@@H]2C[C@H]3C(=C(O)c4c(O)cccc4[C@@]3(C)O)C(=O)[C@]2(O)C(=O)\C(=C(/O)NCN5CCCC5)C1=O
11	CC1=CN([C@H]2C[C@H](F)[C@@H](CO)O2)C(=O)NC1=O
30	CCCC(C)C1(CC)C(=O)NC(=O)NC1=O
47	Cn1nnnc1SCC2=C(N3[C@H](SC2)[C@H](NC(=O)[C@H](O)c4ccccc4)C3=O)C(O)=O
69	[Na+].CO\N=C(C(=O)N[C@@H]1[C@@H]2SCC(=C(N2C1=O)C([O-])=O)COC(C)=O)\c3csc(N)n3
76	CO/N=C(C(=O)N[C@H]1[C@H]2SCC(=C(N2C1=O)C(O)=O)COC(N)=O)/c3occc3
86	CO/N=C(C(=O)N[C@H]1[C@H]2SCC(=C(N2C1=O)C(O)=O)CSC3=NC(=O)C(=O)NN3C)/c4csc(N)n4
96	CO[C@]1(NC(=O)Cc2sccc2)[C@H]3SCC(=C(N3C1=O)C(O)=O)COC(N)=O
116	CCc1cc(ccn1)C(N)=S
127	[Cl-].CN(C)[C@H]1[C@@H]2C[C@H]3C(=C(O)c4c(O)ccc(Cl)c4[C@@]3(C)O)C(=O)[C@]2(O)C(=O)\C(=C(N)/O)C1=O.[H+]
132	[Na+].Cc1sc(SCC2=C(N3[C@H](SC2)[C@H](NC(=O)Cn4cnnn4)C3=O)C([O-])=O)nn1
133	CC(=O)OCC1=C(N2[C@H](SC1)[C@H](NC(=O)CSc3ccncc3)C2=O)C(O)=O
152	CN(C)[C@H]1[C@@H]2C[C@H]3C(=C(O)c4c(O)ccc(Cl)c4[C@@]3(C)O)C(=O)[C@]2(O)C(=O)\C(=C(N)/O)C1=O
170	CO[C@@H]([C@@H]1Cc2cc3cc(O[C@H]4C[C@@H](O[C@H]5C[C@@H](O)[C@H](O)[C@@H](C)O5)[C@H](O)[C@@H](C)O4)c(C)c(O)c3c(O)c2C(=O)[C@H]1O[C@H]6C[C@@H](O[C@H]7C[C@@H](O[C@H]8C[C@](C)(O)[C@H](O)[C@@H](C)O8)[C@H](O)[C@@H](C)O7)[C@H](O)[C@@H](C)O6)C(=O)[C@@H](O)[C@@H](C)O
176	NC1[C@H]2CN(C[C@@H]12)c3nc4N(C=C(C(O)=O)C(=O)c4cc3F)c5ccc(F)cc5F
204	c12[C@]34[C@@]56[C@H]([N@@](CC7CC7)CC4)Cc2ccc(c1O[C@H]3[C@](OC)([C@H](C5)[C@](C(C)(C)C)(C)O)CC6)O
207	O.C[C@@H]1[C@H]2[C@H](O)[C@H]3[C@H](N(C)C)C(=O)C(=C(N)/O)/C(=O)[C@@]3(O)C(=O)C2=C(O)c4c(O)cccc14
226	CC1=C(N2[C@H](SC1)[C@H](NC(=O)[C@H](N)C3=CCC=CC3)C2=O)C(O)=O
235	OC[C@H]1O[C@H](C[C@@H]1O)N2C=C(F)C(=O)NC2=O
237	FC1=CNC(=O)NC1=O
284	CN(C)C1C2C(O)C3C(=C)c4c(Cl)ccc(O)c4C(=C3C(=O)C2(O)C(=O)\C(=C(N)/O)C1=O)O
289	OC[C@@H]1CC[C@@H](O1)n2cnc3C(=O)N=CNc23
293	C1[C@@H]2[C@](C(=C3C(c4c(ccc(c4C[C@@H]13)N(C)C)O)=O)O)(C(C(C(N)=O)=C([C@H]2N(C)C)O)=O)O
319	CC1=CN([C@@H]2O[C@H](CO)C=C2)C(=O)NC1=O

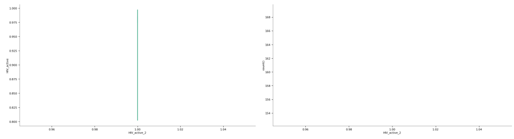
Show 25 per page 1 2 3 4 5 6 7

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

Distributions



Time series



Values

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, 56800.19it/s]
100%|██████████| 51/51 [00:00<00:00, 39041.71it/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,  2.44it/s]
100%|██████████| 1/1 [00:01<00:00,  1.37s/it]Saving predictions to substances_preds.csv
Elapsed time = 0:00:02
```

```
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, 161716.84it/s]
100%|██████████| 892/892 [00:00<00:00, 130929.80it/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 = 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 [00:01<00:26,  1.57s/it]
 11%|█         | 2/18 [00:01<00:13,  1.21it/s]
 22%|██▏       | 4/18 [00:02<00:06,  2.16it/s]
 50%|█████     | 9/18 [00:02<00:01,  6.20it/s]
 67%|██████▋   | 12/18 [00:02<00:00,  8.62it/s]
100%|██████████| 1/1 [00:03<00:00,  3.08s/it]Saving predictions to fda_approved_preds.csv
Elapsed time = 0:00:03
```

```
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	0.009933
1	ZINC000003807804	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	0.596011
2	ZINC000000120286	Nc1nc(N)c2nc(-c3ccccc3)c(N)nc2n1	0.106520
3	ZINC000242548690	C[C@H]1O[C@@H](O[C@H]2[C@@H](O)C[C@H](O[C@H]3[...	0.734897
4	ZINC000000008492	Oc1cccc2cccnc12	0.092682

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 > 0.8 else 0)
fda_preds_df.head()
```

	HIV_active	HIV_active_2
count	892.000000	892.000000
mean	0.297628	0.223094
std	0.232897	0.416555
min	0.002699	0.000000
25%	0.109553	0.000000
50%	0.230911	0.000000
75%	0.421475	0.000000
max	0.996994	1.000000

	zinc_id	smiles	HIV_active	HIV_active_2
0	ZINC000001530427	C[C@@H]1O[C@@H]1P(=O)(O)O	0.009933	0
1	ZINC000003807804	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	0.596011	0
2	ZINC000000120286	Nc1nc(N)c2nc(-c3ccccc3)c(N)nc2n1	0.106520	0
		C[C@H]1O[C@@H]		

Next steps: View recommended plots 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
```

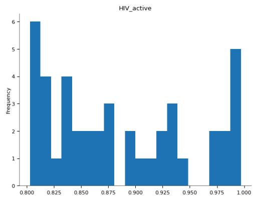
1 to 25 of 43 entries Filter

index	zinc_id	smiles
47	ZINC000003813010	O=c1[nH]c(=O)n([C@H]2C[C@H](O)[C@@H](CO)O2)cc1F
55	ZINC000000592419	O=C(Nc1c(Cl)cncc1Cl)c1ccc(OC(F)F)c(OCC2CC2)c1
81	ZINC000003818726	O=C(/C=C/c1cccc(S(=O)(=O)Nc2ccccc2)c1)NO
153	ZINC000245204949	C[N@+]1(CC2CC2)CC[C@]23c4c5ccc(O)c4O[C@H]2C(=O)CC[C@@]3(O)[C@H]1C5
158	ZINC000003830391	CC1=C(C(=O)O)N2C(=O)[C@@H](NC(=O)[C@H](N)c3ccc(O)cc3)[C@H]2SC1
197	ZINC000000005423	Cc1nc(-c2ccc(OCC(C)C)c(C#N)c2)sc1C(=O)O
228	ZINC000009302239	NC(=O)[C@@H]1CC[C@@H]2CN1C(=O)N2OS(=O)(=O)O
247	ZINC000003922770	C[C@@H](O)[C@H]1C(=O)N2C(C(=O)O)=C(S[C@@H]3CN[C@H](CNS(N)(=O)=O)C3)[C@H](C)[C@H]12
264	ZINC000040899447	CS(=O)(=O)c1ccc(C(=O)Nc2ccc(Cl)c(-c3ccccn3)c2)c(Cl)c1
283	ZINC000003955219	CC(C)CN(C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)O[C@H]1CO[C@H]2OCC[C@@H]12)S(=O)(=O)c1ccc(N)cc1
313	ZINC000014210457	CC(C)(C)NC(=O)N[C@H](C(=O)N1C[C@H]2[C@@H]([C@H]1C(=O)N[C@H](CC1CCC1)C(=O)C(N)=O)C2(C)C)C(C)(C)C
314	ZINC000014879992	CN(C)c1ccc(O)c2c1C[C@H]1C[C@H]3[C@H](N(C)C)C(O)=C(C(N)=O)C(=O)[C@@]3(O)C(O)=C1C2=O
321	ZINC000013597823	O=c1[nH]cnc2c1ncn2[C@H]1CC[C@@H](CO)O1
324	ZINC000019632618	Cc1ccc(NC(=O)c2ccc(CN3CCN(C)CC3)cc2)cc1Nc1nccc(-c2cccnc2)n1
340	ZINC000001530621	CCN[C@H]1C[C@H](C)S(=O)(=O)c2sc(S(N)(=O)=O)cc21
410	ZINC000004097225	C[C@@H](O)[C@H]1C(=O)N2C(C(=O)O)=C(SCCNC=N)C[C@H]12
443	ZINC000000643114	C[C@@H]1Cc2ccccc2N1NC(=O)c1ccc(Cl)c(S(N)(=O)=O)c1
479	ZINC000003830264	C[C@H]1[C@H](NC(=O)/C(=N\OC(C)(C)C(=O)O)c2csc(N)n2)C(=O)N1S(=O)(=O)O
513	ZINC000058581064	C[C@@H]1CCO[C@H]2Cn3cc(C(=O)NCc4ccc(F)cc4F)c(=O)c(O)c3C(=O)N21
612	ZINC000003929508	CCOC(=O)C1=C[C@@H](OC(CC)CC)[C@H](NC(C)=O)[C@@H](N)C1
641	ZINC000038212689	O=c1[nH]cc(F)c(=O)[nH]1
662	ZINC000035342787	CCN(CC)C(=O)/C(C#N)=C/c1cc(O)c(O)c([N+](=O)[O-])c1
677	ZINC000000012346	Nc1ccn([C@@H]2CS[C@H](CO)O2)c(=O)n1
690	ZINC000000601305	C[C@H]1Cc2ccccc2N1NC(=O)c1ccc(Cl)c(S(N)(=O)=O)c1
712	ZINC000004095696	CC1(C)C[C@@H]1C(=O)N/C(=C\CCCCSC[C@H](N)C(=O)O)C(=O)O

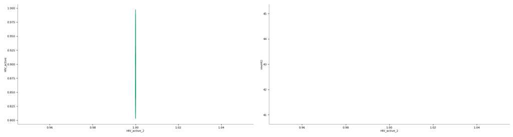
Show 25 per page 1 2

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

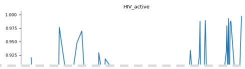
Distributions



Time series



Values



Next steps: View recommended plots

```
!wget https://zinc15.docking.org/substances/subsets/named.csv
```

```
--2024-03-10 05:58:28--  https://zinc15.docking.org/substances/subsets/named.csv
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.1’

named.csv.1             [ <=>                ]   9.28K  --.-KB/s    in 0.04s

2024-03-10 05:58:29 (242 KB/s) - ‘named.csv.1’ saved [9499]
```

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

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
<ipython-input-123-5bc884f87412> in <cell line: 1>()
----> 1 zinc_df = pd.read_csv("named.csv")
      2 zinc_df.head()
      3 zinc_df.tail()
```

6 frames

```
/usr/local/lib/python3.10/dist-packages/pandas/io/common.py in get_handle(path_or_buf, mode, encoding,
compression, memory_map, is_text, errors, storage_options)
    854         if ioargs.encoding and "b" not in ioargs.mode:
    855             # Encoding
--> 856             handle = open(
    857                 handle,
    858                 ioargs.mode,
```