Electronic Supplementary Information

MoleculeNet: A Benchmark for Molecular Machine Learning

Zhenqin Wu,[a‡] Bharath Ramsundar,[b‡] Evan N. Feinberg,[c¶] Joseph Gomes,[a¶] Caleb Geniesse,[c] Aneesh S. Pappu,[b] Karl Leswing,[d] and Vijay Pande*[a]

[a] *Department of Chemistry, Stanford University, Stanford, CA 94305, USA. E-mail: pande@stanford.edu*
[b] *Department of Computer Science, Stanford University, Stanford, CA 94305, USA*
[c] *Program in Biophysics, Stanford School of Medicine, Stanford, CA 94305, USA*
[d] *Schrodinger Inc., USA*
‡ Joint First Authorship
¶ Joint Second Authorship

1 Model Training and Hyperparameter Optimization

All models were trained on Stanford's GPU clusters via DeepChem. No model was allowed to train for more than 10 hours(time profile in Table S1. Users can reproduce benchmarks locally by following directions from DeepChem.

Hyperparameters were determined using Gaussian Process Optimization via pyGPGO(https://github.com/hawk31/pyGPGO), with max number of iterations set to 20. Optimized hyperparameters for each model are listed, detailed hyperparameters can be found on Deepchem.

1.1 Logistic Regression (Logreg)

- Learning rate
- L2 regularization
- Batch size

1.2 Support Vector Classification (KernelSVM)

- Penalty parameter C
- Kernel coefficient gamma for radial basis function

1.3 Kernel Ridge Regression (KRR)

- Penalty parameter

1.4 Random Forest (RF)

- Number of trees in the forest: 500

1.5 Gradient Boosting (XGBoost)

- Maximum tree depth
- Learning rate
- Number of boosted tree

1.6 Multitask/Singletask Networks

- Layer size
- Weight - initial standard deviation
- Bias - initial constant
- Learning rate
- L2 regularization
- Batch size

1.7 Bypass Networks

- Layer size(main layer and bypass layer)
- Weight - initial standard deviation(main layer and bypass layer)
- Bias - initial constant(main layer and bypass layer)
- Learning rate
- L2 regularization
- Batch size

1.8 Influence Relevance Voting (IRV)

- K(number of nearest neighbors)
- Learning rate
- Batch size

1.9 Graph Convolutional models (GC)

- Layer size of convolutional layers
- Layer size of fully-connected layer
- Learning rate
- Batch size

1.10 Weave models

- Length of output features(layer size) of convolutional layers
- Learning rate
- Batch size

1.11 Deep Tensor Neural Networks (DTNN)

- Length of atom embedding(features)
- Size of distance bin(from -1Å to 19Å)
- Learning rate
- Batch size

1.12 Directed Acyclic Graph models (DAG)

- Length of features in the convolutional layer
- Maximum number of propagation of a graph
- Learning rate
- Batch size

1.13 Message Passing Neural Networks (MPNN)

- Number of message passing phases
- Number of steps(iterations) in readout phase
- Learning rate
- Batch size

1.14 ANI-1

- Layer size
- Length of radial and angular symmetry functions
- Learning rate
- Batch size

All final performances were run three times with different fixed numerical seeds on the best-performing hyperparameters, and data splitting methods have been set to maintain determistic behavior. These settings control most randomness in learning process, but benchmark runs(on the same seed) may vary on the order of 1% due to other sources of nondeterminism. Mean and standard deviations of all results are presented in the Performances section of Appendix.

We measured model running time of Tox21, MUV, QM8 and Lipophicility on a single node in Stanford's GPU clusters(CPU: Intel Xeon E5-2640 v3 @2.60 GHz, GPU: NVIDIA Tesla K80), results listed below:

Table S1 Time Profile for Tox21, MUV, QM8 and Lipophilicity(second)

Model	Tox21	MUV	QM8	Lipophilicity
Logreg	93	522		
KernelSVM	2574	2231		
KRR			3390/5153*	24
RF	24273			186
XGBoost	2082	2418		410
Multitask/Singletask	22	858	275/701*	21
Bypass	31	938		
IRV	58	2674		
GC	246	2320	512	131
Weave	323	4593		255
DAG				5142
DTNN			940	
MPNN			3383	1626

* ECFP/Coulomb Matrix

2 Performances

Table S2 PCBA, MUV, HIV and BACE Performances: AUC-PRC for PCBA and MUV, AUC-ROC for HIV and BACE

Model	Model	Training	Validation	Test
PCBA	Logreg	0.166 ± 0.001	0.130 ± 0.004	0.129 ± 0.003
	Multitask	0.100 ± 0.003	0.097 ± 0.000	0.100 ± 0.006
	Bypass	0.121 ± 0.001	0.111 ± 0.003	0.112 ± 0.002
	GC	0.151 ± 0.001	$\mathbf{0.136 \pm 0.003}$	$\mathbf{0.136 \pm 0.004}$
MUV	Logreg	0.238 ± 0.010	0.036 ± 0.009	0.070 ± 0.009
	KernelSVM	0.922 ± 0.034	0.113 ± 0.039	0.137 ± 0.033
	XGBoost	0.159 ± 0.018	0.066 ± 0.053	0.086 ± 0.033
	IRV	0.043 ± 0.006	0.069 ± 0.008	0.087 ± 0.025
	Multitask	0.385 ± 0.014	$\mathbf{0.202 \pm 0.032}$	$\mathbf{0.184 \pm 0.020}$
	Bypass	0.317 ± 0.027	0.166 ± 0.043	0.148 ± 0.069
	GC	0.040 ± 0.013	0.049 ± 0.023	0.046 ± 0.031
	Weave	0.060 ± 0.030	0.127 ± 0.028	0.109 ± 0.028
HIV	Logreg	0.834 ± 0.004	0.788 ± 0.016	0.702 ± 0.018
	KernelSVM	0.999 ± 0.000	0.837 ± 0.000	$\mathbf{0.792 \pm 0.000}$
	XGBoost	0.942 ± 0.000	$\mathbf{0.841 \pm 0.000}$	0.756 ± 0.000
	IRV	0.849 ± 0.000	0.818 ± 0.000	0.737 ± 0.000
	Multitask	0.753 ± 0.012	0.711 ± 0.027	0.698 ± 0.037
	Bypass	0.736 ± 0.017	0.719 ± 0.012	0.693 ± 0.026
	GC	0.903 ± 0.004	0.792 ± 0.014	0.763 ± 0.016
	Weave	0.725 ± 0.004	0.742 ± 0.040	0.703 ± 0.039
BACE	Logreg	0.960 ± 0.001	0.719 ± 0.003	0.781 ± 0.010
	KernelSVM	0.986 ± 0.000	0.739 ± 0.000	0.862 ± 0.000
	XGBoost	0.933 ± 0.000	$\mathbf{0.756 \pm 0.000}$	0.850 ± 0.000
	RF	0.999 ± 0.000	0.728 ± 0.004	$\mathbf{0.867 \pm 0.008}$
	IRV	0.887 ± 0.000	0.715 ± 0.001	0.838 ± 0.000
	Multitask	0.863 ± 0.034	0.696 ± 0.037	0.824 ± 0.006
	Bypass	0.931 ± 0.001	0.745 ± 0.017	0.829 ± 0.006
	GC	0.852 ± 0.046	0.627 ± 0.015	0.783 ± 0.014
	Weave	0.862 ± 0.009	0.638 ± 0.014	0.806 ± 0.002

Table S3 BBBP, Tox21, ToxCast, SIDER, ClinTox Performances (AUC-ROC)

Model	Model	Training	Validation	Test
BBBP	Logreg	0.986 ± 0.001	0.958 ± 0.003	0.699 ± 0.002
	KernelSVM	0.995 ± 0.000	0.964 ± 0.000	$\mathbf{0.729 \pm 0.000}$
	XGBoost	0.987 ± 0.000	0.956 ± 0.000	0.696 ± 0.000
	RF	1.000 ± 0.000	0.956 ± 0.002	0.714 ± 0.000
	IRV	0.915 ± 0.000	$\mathbf{0.964 \pm 0.000}$	0.700 ± 0.000
	Multitask	0.908 ± 0.019	0.955 ± 0.002	0.688 ± 0.005
	Bypass	0.950 ± 0.005	0.960 ± 0.003	0.702 ± 0.006
	GC	0.956 ± 0.004	0.943 ± 0.002	0.690 ± 0.009
	Weave	0.873 ± 0.010	0.951 ± 0.005	0.671 ± 0.014
Tox21	Logreg	0.910 ± 0.002	0.772 ± 0.011	0.794 ± 0.015
	KernelSVM	0.998 ± 0.000	0.818 ± 0.010	0.822 ± 0.006
	XGBoost	0.899 ± 0.011	0.775 ± 0.018	0.794 ± 0.014
	RF	0.999 ± 0.000	0.763 ± 0.002	0.769 ± 0.015
	IRV	0.805 ± 0.003	0.807 ± 0.006	0.799 ± 0.006
	Multitask	0.884 ± 0.001	0.795 ± 0.017	0.803 ± 0.012
	Bypass	0.938 ± 0.001	0.800 ± 0.008	0.810 ± 0.013
	GC	0.905 ± 0.004	0.825 ± 0.013	$\mathbf{0.829 \pm 0.006}$
	Weave	0.875 ± 0.004	$\mathbf{0.828 \pm 0.008}$	0.820 ± 0.010
ToxCast	Logreg	0.828 ± 0.016	0.611 ± 0.024	0.605 ± 0.003
	KernelSVM	0.905 ± 0.012	0.674 ± 0.013	0.669 ± 0.014
	XGBoost	0.764 ± 0.004	0.641 ± 0.009	0.640 ± 0.005
	IRV	0.663 ± 0.004	0.660 ± 0.009	0.663 ± 0.015
	Multitask	0.887 ± 0.002	0.705 ± 0.017	0.702 ± 0.013
	Bypass	0.793 ± 0.002	0.684 ± 0.016	0.676 ± 0.005
	GC	0.815 ± 0.003	0.709 ± 0.013	0.716 ± 0.014
	Weave	0.830 ± 0.006	$\mathbf{0.750 \pm 0.007}$	$\mathbf{0.742 \pm 0.003}$
SIDER	Logreg	0.918 ± 0.001	0.635 ± 0.018	0.643 ± 0.011
	KernelSVM	0.984 ± 0.021	0.655 ± 0.030	0.682 ± 0.013
	XGBoost	0.854 ± 0.016	0.645 ± 0.038	0.656 ± 0.027
	RF	1.000 ± 0.000	0.650 ± 0.013	$\mathbf{0.684 \pm 0.009}$
	IRV	0.628 ± 0.004	$\mathbf{0.657 \pm 0.028}$	0.640 ± 0.020
	Multitask	0.790 ± 0.007	0.632 ± 0.040	0.666 ± 0.026
	Bypass	0.852 ± 0.001	0.644 ± 0.035	0.673 ± 0.025
	GC	0.735 ± 0.013	0.609 ± 0.021	0.638 ± 0.012
	Weave	0.647 ± 0.015	0.591 ± 0.031	0.581 ± 0.027
ClinTox	Logreg	0.990 ± 0.001	0.732 ± 0.065	0.722 ± 0.039
	KernelSVM	0.994 ± 0.002	0.614 ± 0.195	0.669 ± 0.092
	XGBoost	0.926 ± 0.008	0.729 ± 0.140	0.799 ± 0.050
	RF	0.996 ± 0.001	0.688 ± 0.107	0.713 ± 0.056
	IRV	0.804 ± 0.004	0.748 ± 0.075	0.770 ± 0.072
	Multitask	0.917 ± 0.002	0.711 ± 0.186	0.778 ± 0.055
	Bypass	0.943 ± 0.004	0.734 ± 0.111	0.827 ± 0.051
	GC	0.962 ± 0.005	$\mathbf{0.920 \pm 0.035}$	0.807 ± 0.047
	Weave	0.948 ± 0.013	0.875 ± 0.066	$\mathbf{0.832 \pm 0.037}$

Table S4 PDBbind Performances (Root-Mean-Square Error)

Model	Model	Training	Validation	Test
PDBbind - core	RF	0.82 ± 0.00	2.02 ± 0.02	2.03 ± 0.01
	RF(grid)	0.73 ± 0.01	1.98 ± 0.01	2.27 ± 0.01
	Multitask	1.62 ± 0.03	$\mathbf{1.86 \pm 0.01}$	2.21 ± 0.02
	Multitask(grid)	1.51 ± 0.05	1.92 ± 0.02	2.20 ± 0.03
	GC	1.42 ± 0.04	2.10 ± 0.05	$\mathbf{1.92 \pm 0.07}$
PDBbind - refined	RF	0.66 ± 0.00	1.48 ± 0.00	1.62 ± 0.00
	RF(grid)	0.51 ± 0.00	$\mathbf{1.37 \pm 0.00}$	$\mathbf{1.38 \pm 0.00}$
	Multitask	1.09 ± 0.01	1.53 ± 0.03	1.66 ± 0.05
	Multitask(grid)	0.55 ± 0.02	1.41 ± 0.02	1.46 ± 0.05
	GC	1.20 ± 0.01	1.55 ± 0.05	1.65 ± 0.03
PDBbind - full	RF	0.66 ± 0.00	1.40 ± 0.00	1.31 ± 0.00
	RF(grid)	0.51 ± 0.00	$\mathbf{1.35 \pm 0.00}$	$\mathbf{1.25 \pm 0.00}$
	Multitask	1.52 ± 0.17	1.42 ± 0.05	1.45 ± 0.14
	Multitask(grid)	0.39 ± 0.01	1.40 ± 0.03	1.28 ± 0.02
	GC	1.65 ± 0.10	1.57 ± 0.20	1.44 ± 0.12

Table S5 ESOL, FreeSolv, Lipophilicity Performances (Root-Mean-Square Error)

Model	Model	Training	Validation	Test
ESOL	RF	0.51 ± 0.01	1.16 ± 0.15	1.07 ± 0.19
	Multitask	0.59 ± 0.04	1.17 ± 0.13	1.12 ± 0.15
	XGBoost	0.51 ± 0.08	1.05 ± 0.10	0.99 ± 0.14
	KRR	0.38 ± 0.01	1.65 ± 0.19	1.53 ± 0.06
	GC	0.43 ± 0.20	1.05 ± 0.15	0.97 ± 0.01
	DAG	0.32 ± 0.03	0.74 ± 0.04	0.82 ± 0.08
	Weave	0.34 ± 0.04	0.57 ± 0.04	0.61 ± 0.07
	MPNN	0.25 ± 0.06	$\mathbf{0.55 \pm 0.02}$	$\mathbf{0.58 \pm 0.03}$
FreeSolv	RF	0.80 ± 0.03	2.12 ± 0.68	2.03 ± 0.22
	Multitask	1.07 ± 0.06	1.95 ± 0.41	1.87 ± 0.07
	XGBoost	0.85 ± 0.12	1.76 ± 0.21	1.74 ± 0.15
	KRR	0.31 ± 0.03	2.10 ± 0.12	2.11 ± 0.07
	GC	0.31 ± 0.09	1.35 ± 0.15	1.40 ± 0.16
	DAG	0.49 ± 0.46	1.48 ± 0.15	1.63 ± 0.18
	Weave	0.32 ± 0.04	$\mathbf{1.19 \pm 0.08}$	1.22 ± 0.28
	MPNN	0.31 ± 0.05	1.20 ± 0.02	$\mathbf{1.15 \pm 0.12}$
Lipophilicity	RF	0.318 ± 0.006	0.835 ± 0.036	0.876 ± 0.040
	Multitask	0.385 ± 0.065	0.852 ± 0.048	0.859 ± 0.013
	XGBoost	0.135 ± 0.012	0.783 ± 0.021	0.799 ± 0.054
	KRR	0.180 ± 0.002	0.889 ± 0.009	0.899 ± 0.043
	GC	0.471 ± 0.001	$\mathbf{0.678 \pm 0.040}$	$\mathbf{0.655 \pm 0.036}$
	DAG	0.173 ± 0.026	0.857 ± 0.050	0.835 ± 0.039
	Weave	0.549 ± 0.051	0.734 ± 0.011	0.715 ± 0.035
	MPNN	0.363 ± 0.043	0.757 ± 0.030	0.719 ± 0.031

Table S6 QM7, QM7b, QM8 and QM9 Performances (Mean Absolute Error)

Model	Model	Training	Validation	Test
QM7	RF	47.1 ± 0.1	124.0 ± 4.6	122.7 ± 4.2
	Multitask	101.8 ± 13.7	121.7 ± 7.5	123.7 ± 15.6
	KRR	65.5 ± 0.3	108.3 ± 5.4	110.3 ± 4.7
	GC	67.8 ± 4.0	77.9 ± 10.0	77.9 ± 2.1
	Multitask(CM)	10.4 ± 1.8	11.0 ± 1.7	10.8 ± 1.3
	KRR(CM)	0.1 ± 0.0	9.9 ± 0.1	10.2 ± 0.3
	DTNN	8.2 ± 3.9	$\mathbf{8.9 \pm 3.7}$	$\mathbf{8.8 \pm 3.5}$
	ANI-1	24.7 ± 1.2	27.7 ± 1.2	27.8 ± 0.7
QM7b	Multitask(CM)	2.95 ± 0.70	2.90 ± 0.82	2.89 ± 0.65
	KRR(CM)	0.01 ± 0.00	$\mathbf{1.08 \pm 0.08}$	$\mathbf{1.05 \pm 0.06}$
	DTNN	1.68 ± 0.18	1.79 ± 0.14	1.77 ± 0.17
QM8	Multitask	0.0081 ± 0.0002	0.0155 ± 0.0005	0.0150 ± 0.0005
	KRR	0.0152 ± 0.0001	0.0197 ± 0.0004	0.0195 ± 0.0003
	GC	0.0123 ± 0.0009	0.0150 ± 0.0006	0.0148 ± 0.0006
	Multitask(CM)	0.0163 ± 0.0010	0.0181 ± 0.0012	0.0179 ± 0.0013
	KRR(CM)	0.0002 ± 0.0000	0.0242 ± 0.0003	0.0238 ± 0.0004
	DTNN	0.0140 ± 0.0009	0.0170 ± 0.0007	0.0169 ± 0.0009
	MPNN	0.0128 ± 0.0010	$\mathbf{0.0146 \pm 0.0010}$	$\mathbf{0.0143 \pm 0.0011}$
QM9	Multitask	15.3 ± 0.2	15.9 ± 0.2	16.0 ± 0.2
	GC	4.6 ± 0.5	4.7 ± 0.5	4.7 ± 0.5
	Multitask(CM)	4.3 ± 1.0	4.3 ± 1.1	4.4 ± 1.0
	DTNN	2.3 ± 1.1	$\mathbf{2.4 \pm 1.1}$	$\mathbf{2.4 \pm 1.1}$
	MPNN	3.2 ± 1.5	3.2 ± 1.5	3.2 ± 1.5

Table S7 QM7b Test Set Performances of All Tasks(Mean Absolute Error)

Task	Multitask(CM)	KRR(CM)	DTNN
Atomization energy - PBE0	36.0	**9.3**	21.5
Excitation energy of maximal optimal absorption - ZINDO	1.31	1.83	**1.26**
Highest absorption - ZINDO	0.086	0.098	**0.074**
HOMO - ZINDO	0.293	0.369	**0.192**
LUMO - ZINDO	0.255	0.361	**0.159**
1st excitation energy - ZINDO	0.368	0.479	**0.296**
Ionization potential - ZINDO	0.305	0.408	**0.214**
Electron Affinity - ZINDO	0.271	0.404	**0.174**
HOMO - KS	0.247	0.272	**0.155**
LUMO - KS	0.187	0.239	**0.129**
HOMO - GW	0.270	0.294	**0.166**
LUMO - GW	0.172	0.236	**0.139**
Polarizability - PBE0	0.335	0.225	**0.173**
Polarizability - SCS	0.317	**0.116**	0.149

Table S8 QM8 Test Set Performances of All Tasks(Mean Absolute Error)

Task	Multitask	GC	KRR	Multitask(CM)	KRR(CM)	DTNN	MPNN
E1 - CC2	0.0088	**0.0074**	0.0115	0.0125	0.0137	0.0092	0.0084
E2 - CC2	0.0098	**0.0085**	0.0116	0.0114	0.0124	0.0092	0.0091
f1 - CC2	**0.0145**	0.0175	0.0202	0.0186	0.0272	0.0182	0.0151
f2 - CC2	0.0320	0.0328	0.0387	0.0358	0.0460	0.0377	**0.0314**
E1 - PBE0	0.0089	**0.0076**	0.0118	0.0126	0.0140	0.0090	0.0083
E2 - PBE0	0.0096	**0.0083**	0.0117	0.0114	0.0122	0.0086	0.0086
f1 - PBE0	**0.0121**	0.0125	0.0189	0.0152	0.0258	0.0155	0.0123
f2 - PBE0	0.0252	0.0246	0.0319	0.0267	0.0376	0.0281	**0.0236**
E1 - CAM	0.0083	**0.0070**	0.0111	0.0119	0.0132	0.0086	0.0079
E2 - CAM	0.0090	**0.0076**	0.0109	0.0106	0.0115	0.0082	0.0082
f1 - CAM	0.0140	0.0153	0.0208	0.0177	0.0304	0.0180	**0.0134**
f2 - CAM	0.0274	0.0285	0.0345	0.0303	0.0417	0.0322	**0.0258**

Table S9 QM9 Test Set Performances of All Tasks(Mean Absolute Error)

Task	Multitask	Multitask(CM)	GC	DTNN	MPNN
mu	0.602	0.519	0.583	**0.244**	0.358
alpha	3.10	**0.85**	1.37	0.95	0.89
HOMO	0.00660	0.00506	0.00716	**0.00388**	0.00541
LUMO	0.00854	0.00645	0.00921	**0.00513**	0.00623
gap	0.0100	0.0086	0.0112	**0.0066**	0.0082
R2	125.7	46.0	35.9	**17.0**	28.5
ZPVE	0.01109	0.00207	0.00299	**0.00172**	0.00216
U0	15.10	2.27	3.41	2.43	**2.05**
U	15.10	2.27	3.41	2.43	**2.00**
H	15.10	2.27	3.41	2.43	**2.02**
G	15.10	2.27	3.41	2.43	**2.02**
Cv	1.77	0.39	0.65	**0.27**	0.42

3 Grid Featurizer

In our implementation, we generate a vector with length 2052 for each pair of ligand and protein. Detailed process listed below:

First, binding pocket atoms of the protein are extracted using a distance cutoff of 4.5 Å. In this process, atom in the protein will be extracted only if it locates within this distance from any atom in the ligand molecule.

Intra-ligand and intra-protein fingerprints are generated (using the ordinary circular fingerprint with radius of 2) respectively on the atoms from the ligand and atoms in the binding pocket of the protein, and then hashed together to form a vector of length 512.

Then we form three different sets of contacting atom pairs between ligand and protein, whose intra-pair distance falls within bins: $0 \sim 2$ Å, $2 \sim 3$ Åand $3 \sim 4.5$ Å. Each set of pairs is hashed into a fixed length fingerprint with length 512.

Finally, salt bridges are counted, hydrogen bonds are counted in three different distance bins, forming the last four digits. In total the fingerprints have length of 2052.

4 ClinTox

The ClinTox dataset addresses clinical drug toxicity by providing a qualitative comparison of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons. We compiled the FDA-approved drug names from annotations in the SWEETLEAD database. We compiled the names of drugs that failed clinical trials for toxicity reasons from the Aggregate Analysis of ClinicalTrials.gov (AACT) database. To identify these drug names, we relied on annotations from the clinical study table titled "clinical_study_noclob.txt" in the AACT database. From this table, we selected clinical trials where the overall status was "terminated," "suspended," or "withdrawn," and the explanation for the status included the terms "adverse," "toxic," or "death."

5 Dataset and model access

Table S10 listed DeepChem commands to load datasets and models in MoleculeNet. For more detailed instructions please refer to the docs and examples. Tutorial for building customized datasets can be found at `https://github.com/deepchem/deepchem/blob/master/examples/notebooks/dataset_preparation.ipynb`

Table S10 DeepChem commands to load MoleculeNet datasets and models

Dataset	Command
QM7	`deepchem.molnet.load_qm7_from_mat`
QM7b	`deepchem.molnet.load_qm7b_from_mat`
QM8	`deepchem.molnet.load_qm8`
QM9	`deepchem.molnet.load_qm9`
ESOL	`deepchem.molnet.load_delaney`
FreeSolv	`deepchem.molnet.load_sampl`
Lipophilicity	`deepchem.molnet.load_lipo`
PCBA	`deepchem.molnet.load_pcba`
MUV	`deepchem.molnet.load_muv`
HIV	`deepchem.molnet.load_hiv`
BACE	`deepchem.molnet.load_bace_classification`
PDBbind	`deepchem.molnet.load_pdbbind_grid`
BBBP	`deepchem.molnet.load_bbbp`
Tox21	`deepchem.molnet.load_tox21`
ToxCast	`deepchem.molnet.load_toxcast`
SIDER	`deepchem.molnet.load_sider`
ClinTox	`deepchem.molnet.load_clintox`
Model	**Command**
Logreg	`deepchem.models.TensorflowLogisticRegression`
KernelSVM[a]	`sklearn.svm.SVC`
KRR[a]	`sklearn.kernel_ridge.KernelRidge`
RF[a]	`sklearn.ensemble.RandomForestClassifier` `sklearn.ensemble.RandomForestRegressor`
XGBoost[b]	`deepchem.models.xgboost_models.XGBoostModel`
Multitask/Singletask	`deepchem.models.TensorflowMultiTaskClassifier` `deepchem.models.TensorflowMultiTaskRegressor`
Bypass	`deepchem.models.RobustMultitaskClassifier`
IRV	`deepchem.models.TensorflowMultiTaskIRVClassifier`
GC[c]	`deepchem.nn.SequentialGraph` `deepchem.models.GraphConvTensorGraph`
Weave[c]	`deepchem.nn.AlternateSequentialWeaveGraph` `deepchem.models.WeaveTensorGraph`
DAG[c]	`deepchem.nn.SequentialDAGGraph` `deepchem.models.DAGTensorGraph`
DTNN[c]	`deepchem.nn.SequentialDTNNGraph` `deepchem.models.DTNNTensorGraph`
MPNN	`deepchem.models.MPNNTensorGraph`

[a] These models are based on scikit-learn package.[1]

[b] XGBoost is based on xgboost package.[2]

[c] These models are implemented in two frameworks, with identical underlying structures and performances. All benchmark numbers are run with `deepchem.nn.SequentialGraph` series model

6 Model validation

MoleculeNet includes multiple models that are previously proposed. To validate our reimplementation, here we compare the performances of our implementation with reported values in previous papers. All model validation scripts and trained

models can be found in DeepChem.

Note that performances of our models might be different from values in the benchmark tables due to no limitation imposed on running time(more epochs), different random splitting patterns, etc.

6.1 Graph Convolutional models

We evaluate the model on ESOL dataset, note that we provide performances based on a 80/10/10 random train, valid, test splitting, while the original paper reported performance under cross validation.[3]

RMSE in logS(log solubility in mol per litre):

- Original result: 0.52 ± 0.07
- Reimplementation: 0.39 for valid subset, 0.31 for test subset

6.2 Directed Acyclic Graph models

We evaluate the model on ESOL dataset with the same splitting pattern, the original paper reported performance under 10-fold cross validation.[4]

RMSE in logS(log solubility in mol per litre):

- Original result: 0.58 ± 0.07
- Reimplementation: 0.68 for valid subset, 0.58 for test subset

6.3 Weave models

We evaluate the model on Tox21 dataset, using 80/10/10 random train, valid, test splitting. The original paper reported performance as median score of 5-fold cross validation.[5]

mean ROC-AUC:

- Original result: $0.846 \sim 0.867$ for different model structure settings.
- Reimplementation: 0.857 for valid subset, 0.843 for test subset

6.4 Deep Tensor Neural Network

We evaluate the model on the atomization energy task of qm9, using 80/10/10 random train, valid, test splitting.(train subset with 106,400 samples) The original paper reported performance using different size of training set.[6]

MAE in kcal/mol:

- Original result: 0.93 ± 0.02 with 2 DTNN layers and 100,000 training samples.
- Reimplementation: 1.15 for valid subset, 1.26 for test subset

6.5 Message Passing Neural Network

We evaluate the model on the HOMO-LUMO gap task of qm9, using 80/10/10 random train, valid, test splitting.(train subset with 106,400 samples) The original paper reported performance with a training set containing 110,462 randomly picked samples.[7] Due to that no hyperparameter is specified for the model, we are not able to fully repeat the results.

Note that the original paper trained a single model for each task in the qm9 dataset. Here we only picked one representative task to compare.

MAE in eV:

- Original result: 0.0544
- Reimplementation: 0.0997 for valid subset, 0.101 for test subset

6.6 Influence Relevance Voting

We evaluate the model on the HIV dataset, using 80/10/10 random train, valid, test splitting. The original paper reported performance under 10-fold cross validation.[8]

ROC-AUC:

- Original result: 0.845
- Reimplementation: 0.840 for valid subset, 0.852 for test subset

References

[1] *scikit-learn: Machine Learning in Python*, `http://scikit-learn.org/stable/`, Accessed: 2017-10-18.

[2] *eXtreme Gradient Boosting*, `https://github.com/dmlc/xgboost`, Accessed: 2017-10-18.

[3] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, *arXiv preprint arXiv:1509.09292*, 2015.

[4] A. Lusci, G. Pollastri and P. Baldi, *Journal of chemical information and modeling*, 2013, **53**, 1563–1575.

[5] S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *arXiv preprint arXiv:1603.00856*, 2016.

[6] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *arXiv preprint arXiv:1609.08259*, 2016.

[7] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *arXiv preprint arXiv:1704.01212*, 2017.

[8] S. J. Swamidass, C.-A. Azencott, T.-W. Lin, H. Gramajo, S.-C. Tsai and P. Baldi, *Journal of chemical information and modeling*, 2009, **49**, 756–766.