MolBERT: Molecular Structure Pre-training Model for Molecular Property Prediction Tasks

Anonymous

Dataset

Task-related Datasets

The benchmark datasets for model comparisons are all from the MoleculeNet benchmark[10]. The detailed information of the 9 downstream molecular property prediction datastes is listed as follows:

- **BBBP**. The Blood-brain barrier penetration (BBBP) dataset focuses on the modeling of the barrier permeability. It contains 1513 compounds.
- **BACE**. Qualitative binding results for a set of inhibitors of human β-secretase 1. This dataset contains 1513 compounds.
- **Tox21**. The ”Toxicology in the 21st Century” (Tox21) project created this dataset which measures the toxicity of compounds. It includes qualitative toxicity property for 8014 molecules on 12 distinct targets.
- **ToxCast**. ToxCast is the data collection providing toxicology measures for compounds. The ToxCast dataset contains quanlitative toxicology measures of 8615 compounds on 617 different targets.
- **ClinTox**. Qualitative data classifying drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.
- **Sider**. Database of marketed drugs and adverse drug reactions (ADR), grouped into 27 system organ classes.
- **HIV**. Experimentally measured abilities to inhibit HIV replication (HIV).
- **ESOL**. A small dataset which consists of water solubility data for 1128 compounds.
- **Freesolv**. The Free Solvation Database (FreeSolv) provides experimental and calculated hydration free energy of small molecules in water. This dataset only contains 642 molecules.
- **Lipo**. The dataset, which is curated from ChEMBL database, provides experimental results of octanol/water distribution coefficient (log D at pH 7.4) of 4200 compounds.

Table 1: A summary of task-related datasets.

Dataset	Type	#Mol.	#Tasks
BBBP	Classification	2039	1
BACE	Classification	1513	1
Tox21	Classification	7831	12
ToxCast	Classification	8675	617
ClinTox	Classification	1478	2
Sider	Classification	1427	27
HIV	Classification	41127	1
ESOL	Regression	1128	1
Freesolv	Regression	642	1
Lipo	Regression	4200	1

Dataset Splitting

For molecular property prediction tasks, following the research work conducted by [10], we cluster compounds by scaffold (molecular graph substructure), and reassemble the clusters by placing the most common scaffolds in the training datasets, generating validation and test datasets which include structurally different compounds. Previous research works have demonstrated that this special *scaffold splitting* is more approximate to the ideal chronological split, which can stimulate the real-world application scenarios better. The model evaluated in the setting of *scaffold splitting* is convinced to give more reliable predictions. Therefore, in our research work, we use the special *scaffold splitting* to split the downstream benchmark datasets. The split ratio for train, validation and test datasets are typically 80%, 10% and 10%, respectively.

Experimental Setup

Hyperparameter Configuration of fine-tuned downstream tasks

Table 2 shows the configuration for the fine-tuned shallow-layered neural networks. The notation $[256, 1]$ denotes a neural network layer with ReLU activation function and dropout of 0.1 whose input size is 256 and output size is 1. That the output size is greater than 1 means the multi-tasking nature of the dataset.

Table 2: Hyperparameter Configuration of Our MolBERT architecture.

Dataset	Fine-tuned Network Architecture
BBBP	[256, 1]
BACE	[256, 1]
Tox21	[256, 12]
ToxCast	[768, 768], [768, 615]
ClinTox	[256, 2]
Sider	[256, 27]

More about Ablation Study

Effect of MolBERT depth.

We studied the influence of the total number of the Transformer blocks in our MolBERT architecture. We refer to this number as MolBERT depth. We show the associated experimental results in Table 3. The token radius is set as 1. We can observe that various downstream molecular machine learning tasks benefit from more Transformer blocks because predictive performance often peaks #Blocks=4 or #Blocks=6.At the same time, we note that overly large radius can make the predictive performance decrease. It may be due to the fact that the research objects involved in molecular machine learning tasks are often small molecules. The size of small molecule substances is so finite that we need not use overly broad neighborhood information during the process of pretraining and fine-tuning.

Experiments on DDI Dataset

Dataset

- **BinaryDDI**. BinaryDDI is a dataset. It contains 548 drugs, 48,548 pairwise DDIs as well as multiple types of pairwise similarity information about these drug pairs. At the stage of data preprocessing, we remove the molecules and associated similarity information if the SMILES strings of those molecules cannot be converted into specified graph objects successfully by the RDKit tool.
- **MultiDDI**. MultiDDI is released by [7]. The dataset contains 86 distinct interaction labels, and each drug is represented as a canonical SMILES string. Our data preprocessing procedure also removes the data items that cannot be converted into graph objects by the RDKit tool. The final dataset contains 1704 molecules and 191,400 interacting pairs.

Setups

Baseline

- **Nearest Neighobr**[9]. This model used the combination of known pairwise interactions and similarity derived from substructure [9] to conduct DDI prediction. We refer to the model as NN for convenience.
- **Label Propagation**[5]. This model turned to the label propagation(LP) algorithm to build three similarity-based predictive models. The pairwise similarity information is computed based on substructures, side effects and off-label side effects, respectively. We named the model as LP-Sub, LP-SE, and LP-OSE, respectively.
- **Multi-Feature Ensemble**[12]. Multi-Feature Ensenmble is built on top of the combination of three different algorithms: neighbor recommendation(NR), label propagation(LP), and matrix disturb(MD) algorithms. We refer to the model as Ens.
- **SSP-MLP**[7]. Ref. [7] use the sequential combination of structural similarity profile(SSP) and multi-layer perceptron to conduct the classification. We refer to the model as SSP-MLP.
- **Mol2vec**[3]. This model uses the pre-trained reprsentations provided by Mol2vec [12] as inputs. The vector representations of a drug pair are fed into a feed-forward neural network.
- **Graph Autoencoder**[6]. Ref. [6] developed an attention mechanism to integrate multiple types of drug features, which will be passed into a GCN-style graph autoencoder to learn the embedding for individual drug node. We refer to the model as GA.
- **NFP**[2]. Neural fingerprint developed by [18] is the first graph neural network tailored for molecular property prediction. We substitute our siamese graph neural network with neural fingerprint. We name the model as NFP for short.
- **GIN**[11]. Graph Isomorphism Network(GIN) is the state-of-the-art graph neural network. We change our siamese GNN encoder with the graph isomorphism network and name the model as GIN.
- **GCN-BMP**[1]. A special variant of GCN with bond-aware message passing mechanism and global attention-based graph pooling is used to predict drug-drug interaction.

Evaluatioin Setups

Setups for BinaryDDI Dataset We divide the whole dataset into the training set, validation set, and test set with the ratio 8:1:1. Note that we have only reliable positive drug pairs in the dataset, we regard the same number of randomly sampled negative drug pairs as the negative training samples for simplicity. As for the valid set and test set, we keep it the same as the original situation. We implement our model with Chainer[8]. We use the Adam optimizer, set the initial learning rate as 0.001. We also exploit the exponential shift strategy with a ratio of 0.5 every 10 epochs. The batch size is 32. For the hyper-parameters of the GNN encoder, we set the dimension of node-level hidden states as 32, that of the whole graph as 16. The total number of graph convolutional layers is 8, deeper than the usual graph neural networks such as GCN[4]. Since the tasks conducted on BinaryDDI dataset is binary classification, we select three metrics: *area under ROC curve(AUC)*, *area under PRC curve(PRC)* and *F1* to evaluate the model performances. For the sake of reliability, we report the mean and standard deviation of the four metrics over 20 repetitions in Table I. The experimental results are obtained on the test set.

Table 3: Abalation study of the number of Transformer blocks in our MolBERT

#Blocks	BBBP	BACE	Tox21	ToxCast	ClinTox	Sider	HIV	Esol	Freesolv	Lipo
2	89.09	84.81	**78.99**	67.15	85.26	61.89	79.26	1.251	3.119	0.7842
4	**90.37**	**85.99**	78.27	**70.80**	85.30	**62.27**	**80.02**	**1.216**	**3.018**	**0.7648**
6	89.09	84.50	78.77	69.23	**85.77**	61.12	78.18	1.312	3.152	0.7928
8	89.78	83.14	78.80	68.75	85.29	59.84	77.29	1.351	3.208	0.8011

Table 4: Model comparison on the BinaryDDI dataset

Model Name	**Performance**		
	AUROC	***AUPRC***	***F1***
NN	67.81 ± 0.25	52.61 ± 0.27	49.84 ± 0.43
LP-Sub	93.70 ± 0.13	90.36 ± 0.18	76.41 ± 0.28
LP-SE	93.79 ± 0.28	90.53 ± 0.39	78.48 ± 0.50
LP-OSE	93.88 ± 0.14	90.63 ± 0.36	79.44 ± 0.43
Ens	95.54 ± 0.13	92.75 ± 0.33	83.81 ± 0.39
SSP-MLP	93.09 ± 0.34	88.58 ± 0.51	78.38 ± 0.57
GA	93.84 ± 0.61	90.27 ± 0.66	54.84 ± 0.47
Mol2vec	93.63 ± 0.14	88.74 ± 0.32	81.03 ± 0.26
NFP	81.82 ± 0.13	68.89 ± 0.21	60.93 ± 0.24
GIN	61.63 ± 0.26	48.28 ± 0.31	56.00 ± 0.56
GCN-BMP	96.66 ± 0.09	94.02 ± 0.12	85.00 ± 0.17
Our	**96.83 ± 0.10**	**94.18 ± 0.10**	**86.10 ± 0.15**

Table 5: Model comparison on the MultiDDI dataset

Model Name	**Performance**		
	AUROC	***AUPRC***	***F1***
NN	98.52 ± 0.17	63.04 ± 0.29	13.74 ± 0.36
NR-Sub	99.01 ± 0.09	66.18 ± 0.24	47.89 ± 0.92
SSP-MLP	74.30 ± 0.21	61.20 ± 0.32	53.56 ± 0.54
NFP	98.95 ± 0.16	68.32 ± 0.17	62.56 ± 0.33
GIN	87.67 ± 0.18	14.58 ± 0.13	10.42 ± 0.24
GCN-BMP	99.01 ± 0.09	80.18 ± 0.30	67.31 ± 0.23
Our	**99.25 ± 0.13**	**82.29 ± 0.11**	**70.21 ± 0.51**

Setups for MultiDDI Dataset We use the same dataset as DeepDDI[7]. The hyperparameter configuration is the same as the binary classification task performed on BinaryDDI dataset. Since the task conducted on the MultiDDI dataset is in essence multi-label classification, we select *macro AUC*, *macro PRC* and *macro F1* to measure the model performance.

Comparison Results

Table 4 shows the performance comparison among our proposed model and baseline approaches. Our model achieve the best performance.

Table 5 demonstrates the model performance comparison results obtained on the MultiDDI dataset. Our MolBERT model can achieve the best performance.

References

[1] Chen, X.; Liu, X.; and Wu, J. 2020. GCN-BMP: Investigating graph representation learning for DDI prediction task. *Methods* 179: 47–54.

[2] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, 2224–2232.

[3] Jaeger, S.; Fulle, S.; and Turk, S. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58(1): 27–35.

[4] Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* .

[5] Li, P.; Huang, C.; Fu, Y.; Wang, J.; Wu, Z.; Ru, J.; Zheng, C.; Guo, Z.; Chen, X.; Zhou, W.; et al. 2015. Large-scale exploration and analysis of drug combinations. *Bioinformatics* 31(12): 2007–2016.

[6] Ma, T.; Xiao, C.; Zhou, J.; and Wang, F. 2018. Drug similarity integration through attentive multi-view graph auto-encoders. *arXiv preprint arXiv:1804.10850* .

[7] Ryu, J. Y.; Kim, H. U.; and Lee, S. Y. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences* 115(18): E4304–E4311.

[8] Tokui, S.; Oono, K.; Hido, S.; and Clayton, J. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, 1–6.

[9] Vilar, S.; Harpaz, R.; Uriarte, E.; Santana, L.; Rabadan, R.; and Friedman, C. 2012. Drug—drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association* 19(6): 1066–1074.

[10] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9(2): 513–530.

[11] Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* .

[12] Zhang, W.; Chen, Y.; Liu, F.; Luo, F.; Tian, G.; and Li, X. 2017. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics* 18(1): 18.