Transferable Neural Networks Architecture for Low Data Drug Discovery

Mun-Hwan Lee[1,*], Eung-Hee Kim, Ph.D. [2], Yong-Ju Lee[1], Hong-Gee Kim, Ph.D.[1]

[1] *Biomedical Knowledge Engineering Lab., Seoul National University, Seoul , Republic of Korea*

[2] *Dept. of Global Software Engineering., Sunmoon University, Asan-si, Republic of Korea*

[*]E-mail: munhwanlee@snu.ac.kr

Introduction

Machine learning (ML)'s advances in virtual screening (VS) has made significant contributions to drug discovery accompanied by novel approaches based on deep neural networks (DNN). However, such techniques require a considerable volume of both training and test data, in order to achieve comparable results. While certain models have managed to learn from small datasets[1], they optimize only within the defined range of data which is nontransferable. Thus, the current limitation in ML models necessitates the need of a universal architecture for generalization regardless of dataset size.

In this study, we propose a transferable DNN to overcome such limitations in generalizing ability by incorporating binary response (positive or negative) rather than non-fixed output dimensions depending on a dataset. We demonstrate that the proposed architecture is capable to learn transferable information between varied datasets.

Method

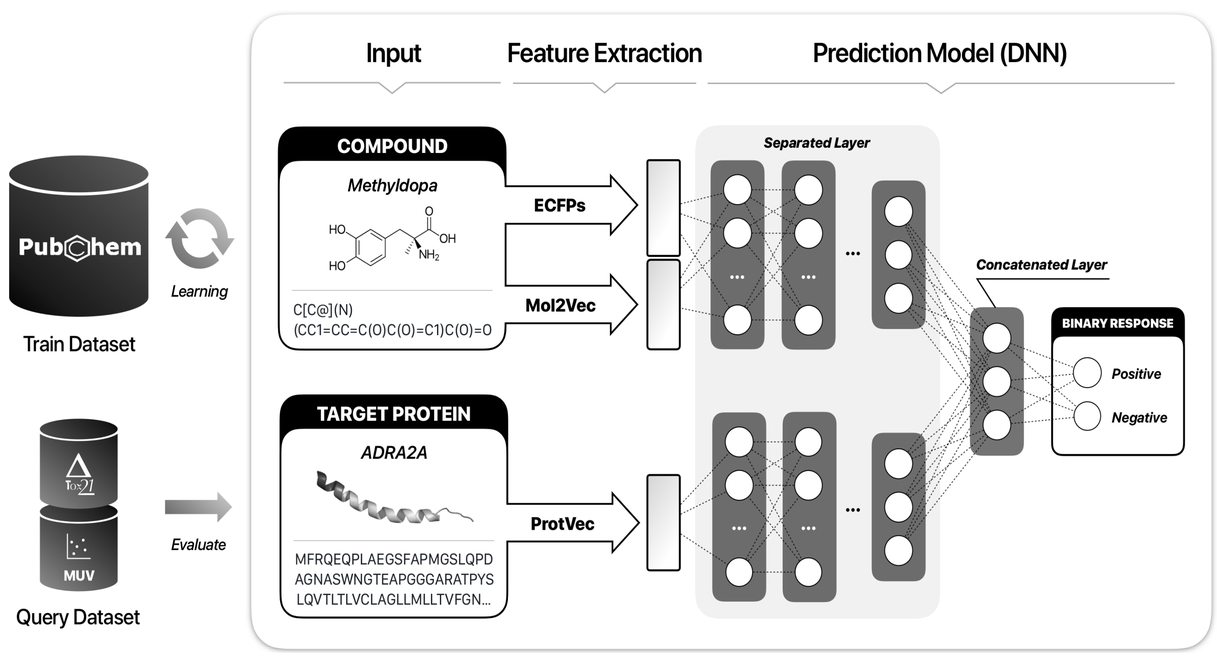

Figure 1. Overview of the transferable DNN

The proposed model consists of feature extraction and prediction model by implementing proteoche-mometric (PCM) approaches, which utilize the additional use of protein information as shown in Figure 1. First, Extended-Connectivity Fingerprints (ECFP) and Mol2Vec were exploited for compound feature extraction and ProtVec for target protein. The dataset comprised both positive and negative examples for each feature pairs, which can be mathematically defined as

$$D = \left\{ \left[Feature_{compound_i}, Feature_{protein_i} \right], y_{i=1}^{|D|} \right\}, where\ y_i \in \{0,1\}.$$

Second, separated layers pass the paired data into concatenated layer for classification. By adjusting the dimensions of hidden nodes through separated layers, the model can prevent a disproportions between feature dimensions (e.g. between 2048-dimension for compounds and 256-dimesion for targets), and avoid bias.

Experiment and Result

The proposed models were trained on PubChem BioAssay (PCBA), and evaluated on Maximum Unbiased Validation (MUV) and Tox21 as a query dataset. To justify the generalization ability of the model, we removed overlapping data from evaluation dataset. In order to avoid learning bias in skewed distribution, the same number of positive and negative pairs were extracted in each compound. This resulted in 361,632 positive examples and 361,632 negative examples. Our proposed networks consist of 5 separated layers and 1 concatenated layer. For training the networks we used stochastic gradient descent with Adaptive Moment Estimation (Adam). We used 50% of dropout on the hidden layers to prevent overfitting of the networks. The proposed model shows better performance over the random-forest baseline as shown in Table 1, Table 2, and Table 3.

Table 1. ROC-AUC Scores of Models on Median Held-out Task for Each Model on Datasets [a]

Models	MUV	Tox21
Proposed Model	**0.730 ± 0.079**	**0.772 ± 0.067**
RF (100 trees)	0.661 ± 0.081	0.709 ± 0.100

[a] Numbers reported are medians and standard deviations.

Table 2. ROC-AUC Scores of Models on Each Tasks on MUV

Models	MUV - 466	MUV - 548	MUV - 600	MUV - 644	MUV - 652	MUV - 689	MUV - 692	MUV - 712	MUV - 713	MUV - 733	MUV - 737	MUV - 810	MUV - 832	MUV - 846	MUV - 852	MUV - 858	MUV - 859
Proposed Model	**0.755**	**0.684**	**0.777**	**0.649**	**0.840**	**0.833**	**0.782**	**0.869**	**0.716**	**0.823**	**0.730**	**0.730**	**0.689**	**0.617**	**0.681**	**0.615**	**0.657**
RF (100 trees)	0.735	0.667	0.576	0.548	0.776	0.768	0.635	0.780	0.630	0.699	0.723	0.702	0.661	0.563	0.622	0.610	0.528

Table 3. ROC-AUC Scores of Models on Each Tasks on Tox21

Models	NR-AR	NR-AR-LBD	NR-AhR	NR-Aromatase	NR-ER	NR-ER-LBD	NR-PPAR-gamma	SR-ARE	SR-ATAD5	SR-HSE	SR-MMP	SR-p53
Proposed Model	**0.636**	**0.762**	**0.741**	**0.782**	**0.628**	**0.714**	**0.721**	**0.783**	**0.800**	**0.807**	**0.826**	**0.829**
RF (100 trees)	0.461	0.528	0.666	0.763	0.583	0.641	0.701	0.749	0.732	0.718	0.746	0.772

Conclusion

We propose a transferable DNN to improve a learning model by transferring information from abundant datasets to small datasets. This allows generalization ability for ML models with various feature extraction methods in a scalable way. We also demonstrated that the model learns transferable ability for various size of datasets especially in small one. In the future work, We are planning to augment transferability to multitask deep learning for robust performance by adopting conventional transfer learning approaches.

Acknowledge

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-00398, Development of drug discovery software based on big data).

Reference

1. Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning ACS Cent. Sci. 2017, 3, 283– 293
2. 2. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J. Chem. Inf. Model., 2018, 58 (1), pp 27–35.
3. 3. Asgari, E.; Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One 2015, 10, e0141287