Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 370–376

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Towards a more general drug target interaction prediction model using transfer learning

Derwin Suhartono[a], Muhammad Rizki Nur Majiid[a,*], Alif Tri Handoyo[a], Pandu Wicaksono[a], Henry Lucky[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

The topic of Drug-Target Interaction (DTI) topic has emerged nowadays since the COVID-19 outbreaks. DTI is one of the stages of finding a new cure for a recent disease. It determines whether a chemical compound would affect a particular protein, known as binding affinity. Recently, significant efforts have been devoted to artificial intelligence (AI) powered DTI. However, the use of transfer learning in DTI has not been explored extensively. This paper aims to make a more general DTI model by investigating DTI prediction method using Transfer learning. Three popular models will be tested and observed: CNN, RNN, and Transformer. Those models combined in several scenarios involving two extensive public datasets on DTI (BindingDB and DAVIS) to find the most optimum architecture. In our finding, combining the CNN model and BindingDB as the source data became the most recommended pre-trained model for real DTI cases. This conclusion was proved with the 6% AUPRC increase after fine-tuning the BindingDB pre-trained model to DAVIS dataset than without pre-training the model first.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Computer Science and Computational Intelligence 2022

Keywords: drug-target interaction; transfer learning; drug discovery; deep learning; SMILES

1. Introduction

Identifying drug interactions and the target drug (protein) is the most crucial in drug discovery. The main goal of drug discovery is to find safe and effective drugs to cure diseases [1]. All drug development processes, from target

* Corresponding author:
E-mail address: muhammad.majiid@binus.ac.id

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Computer Science and Computational Intelligence 2022
10.1016/j.procs.2022.12.148

identification to clinical trials, are costly and time-consuming at each stage. Drug Target Interaction (DTI) helps describe the level of attachment of drug compounds to target proteins [2]. Due to the high cost of the experiment, it is necessary to develop a suitable computational tool that can accurately detect interactions between drug compounds and target proteins.

Recently, significant efforts have been devoted to artificial intelligence (AI) powered DTI, enabling cost-effective development strategies in academia and the pharmaceutical industry. Leveraging deep neural networks can provide an advantage in understanding complex biological spatial contexts. Due to the remarkable success of deep learning in computer vision, speech recognition, and Natural Language Processing (NLP), deep learning models are also widely used for DTI prediction. In predicting DTI, there are at least two main parts to do, namely encoding methods and DTI prediction methods [3]. The encoding method will label the input data with a character-based dictionary. Like in DeepDTA [4], with a dictionary like {'C':1,'H':2,'N':3,...,'=':63}, drugs with sequences like 'CN=C=O' are labeled as [1 363 1 63 5] using the simplified molecular input line entry system (SMILES). Models using SMILES can be used to predict chemical reactions. DTI prediction methods are seen as binary classification problem [1]. This approach uses abinary dataset with labels of ones and zeros for known and unknown interactions.

Several learning-based algorithms have been developed to predict drug-target interactions, such as KronRLS [5] to predict DTI based on chemical structure and protein sequences and SimBoost [6] to predict drug-target binding affinities. When it comes to DTI modeling, the most frequently used prediction method [7-10], is by using 1D descriptors for each drug and target. However, just like most tasks in deep learning, the task of DTI sometimes is treated as benchmarking with specific datasets and researchers are competing to get higher accuracies in the scoreboard. This can be misleading,as the model, which is trained with specific dataset, may not be able to perform well in real life testing as different datasets can have different characteristics. For the model to be able to perform well in real life testing, it needs to be more general in nature.

Due to the ability of deep learning to learn features automatically, several deep learning methods are starting to be applied to DTI, such as transfer learning methods. Transfer learning is a method to use a model that has been trained on a data to other data that is different but still in the same clump [1]. Thus, this method will lead the model to be more general. Recently, transfer learning methods have attracted attention in the field of bioinformatics. Parts of this research include degeneration of biological systems [11], single-cell Ribonucleic Acid (RNA) sequencing [12], drug sensitivity prediction [13], andpatient response estimation [14]. In drug discovery, transfer learning is applied mainly in three aspects: molecular property and activity prediction (including DTI), molecule generation, and structure-based virtual screening [15]. Transfer learning is an essential method in machine learning development to solve the fundamental problem of the lack of data for training [15]. For example, molecular generative models typically pre-train the model on a large dataset,such as Chemical European Molecular Biology (ChEMBL) [16], and then train the model on a specific target dataset using smaller data to generate a particular functional molecule. Then the second model was initialized with the parameters obtained from the first model [17]. Transfer learning can fix data lack problems by fine-tuning pre-trained models that have been trained previously usinglarge datasets [17].

The model used for DTI prediction can be combined with the transfer learning method. Bonggun et al. [18] importeda molecular representation model learned from the PubChem database [19] and applied it to the DTI model to improve performance. Transfer learning methods also can be used in the de novo stage [20]. However, building and evaluating deep learning models to make predictions requires programming skills and an in-depth understanding of chemical drugs. Therefore, integrating data preparation with modeling becomes more difficult [21]. It is for this reason that many data scientists prefer to use specialized frameworks because of the ease with which they develop models. One of the frameworks is DeepPurpose [22].

To make a more general DTI model, in this paper we investigate DTI prediction method using Transfer learning. Transfer learning concepts can formulate the binary DTI problem in a more realistic situation. Even in the drug discovery process (especially when designing new drugs/proteins), some known interactions are insufficient to create a good model. Therefore, transferring knowledge from the related source dataset to the target dataset would be helpful. We use publicly available datasets for our experiment, namely BindingDB [2] and DAVIS [3].

2. Method

2.1. Dataset

In this paper, we use two datasets, BindingDB [2] and DAVIS [3]. BindingDB is a publicly available dataset containingmeasured binding affinities focusing on drug and target (protein) interactions. The DAVIS dataset is fully populated with the observed binding affinities for all 68 drugs and 442 target pairs as measured by Kd (kinase dissociation constant) values. Both datasets use SMILES as their drug representation, which is easier to process, since SMILES representation is string. To make the transfer learning easier, we binarize the datasets using default threshold from DeepPurpose framework [22], resulting in two labels for each drug-target pair: "*1"* or "*0"*. A "*1"* label means that the drug-target pair has strong interaction, which means that the drug could significantly change the protein. Meanwhile, the *"0"* label means that the drug-target pair has weak interaction, which means there will be no significant change to the protein.

2.2. Experiment setting

In this paper There are two commonly used transfer learning approaches: (1) the develop model approach and (2) the pre-trained model approach. The developed model approach consists of four stages as seen in Fig. 1. The first step is selecting the source model; this stage determines the model architecture and datasets with an enormous number that can be used as source datasets. The second step is developing the source model; at this stage, the predetermined architecture is trained using the source dataset until the model can perform well with the data. The next step is reusing the model; this stage continues the model's training process using a dataset that will be used as a parameter for evaluating the model's performance. Model tuning, this last stage aims to get the best evaluation results by adjusting the values of the hyperparameters.

While in the pre-trained model approach, only three stages are carried out. The stage omitted from this approach is the stage of developing the source model because, in selecting the source model, the chosen architecture is a pre-trained model. Pre-trained model is an architectural model that already stores the weight values from the training results using a public dataset. In addition, the stages of reusing and tuning the model are the same as the developed model approach[23]. This study uses a developed model approach because it wants to determine the source dataset manually.

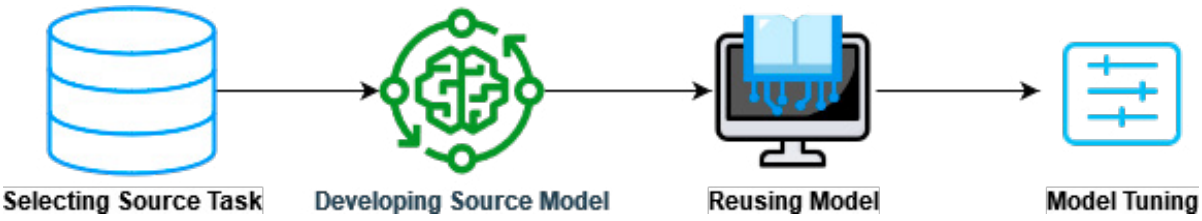


Fig. 1. Transfer Learning Steps with Develop Model Approach

In this case, two datasets will be used: BindingDB and DAVIS. The two datasets will alternately be tested as a source dataset and a fine-tuning dataset. The models are built using a combination of Convolutional Neural Network (CNN), Convolutional Neural Network - Recurrent Neural Network (CNN-RNN), and Transformers using DeepPurpose framework [22]. Table 1 contains details of the combination of experiments that will be carried out. Experiments that do not have a source dataset do not apply transfer learning, which we call the direct training method. All experiments were trained for 100 epochs with a learning rate and batch sizes of 0.001 and 256, respectively.

The value of the hyperparameters itself follows the best results from previous studies[4]. Each dataset will be divided into three sets; 70% as the training set, 10% as the validation set, and 20% for the testing set. Area Under Precision- Recall Curve (AUPRC) was chosen as the evaluation metric because most of the DTI classification cases

use this metric [4].

AUPRC is the percentage of the area under the curve between the precision and recall axes. Precision and recall are metrics that are often used in classification cases because they can reduce evaluation problems in imbalanced data.

In addition, Area Under Receiver-Operating Curve (AUROC) is also used for comparison. The AUROC is calculatedas the area under the Receiver Operating Characteristic (ROC) curve. A ROC curve shows the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across different decision thresholds. Each dataset will be observed eight times with various scenarios. So, in total, there will be 16 trials summarized in Table 1. All the observations already covered the combinations of three popular models: Transformer, CNN, and RNN. The drug and target (protein) data in every scenario will be processed with separate models using the ensemble method. All trials served in the table will already have been fine-tuned.

Table 1. Experimental Scenarios of Datasets and Encoders Combination

No	Source Dataset		Fine-tuning Dataset		Drug Encoder		Target Encoder		
	BindingDB	DAVIS	BindingDB	DAVIS	CNN	CNN-RNN	Transformer	CNN	CNN-RNN
1			✓		✓			✓	
2			✓			✓			✓
3			✓				✓	✓	
4			✓				✓		✓
5				✓	✓			✓	
6				✓		✓			✓
7				✓			✓	✓	
8				✓			✓		✓
9		✓	✓		✓			✓	
10		✓	✓			✓			✓
11		✓	✓				✓	✓	
12		✓	✓				✓		✓
13	✓			✓	✓			✓	
14	✓			✓		✓			✓
15	✓			✓			✓	✓	
16	✓			✓			✓		✓

3. Results and Discussion

In this section, we present our experiment results and discuss them further, starting from all results that are gained from experiment mentioned in section 2.

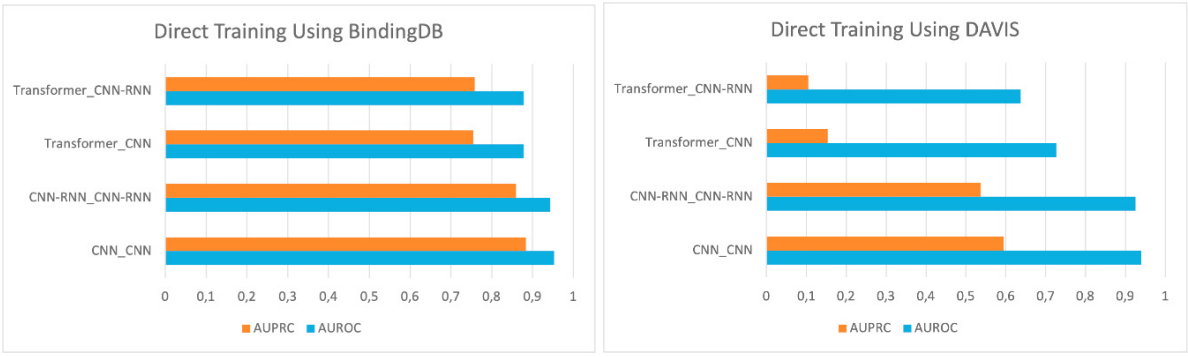


Fig. 2. Results of direct training on BindingDB (left); Results of direct training on DAVIS (right)

The results that are gained using direct training method are shown in Fig. 2. Overall, the models work better on BindingDB dataset than DAVIS dataset. Surprisingly, the best performing model for BindingDB is the CNN-CNN with AUROC of 95.3% and AUPRC of 88.37%, meanwhile the models with Transformer have lower performance scores. As for DAVIS, CNN-CNN also has the best performance score. However, there is a big gap between AUROC and AUPRC (AUROC of 93.95% and AUPRC of 59.52%), which indicates the imbalance level of the dataset is extremely high and the model could not solve the imbalance problem, especially for the Transformer models with almost 60% gap.



Fig. 3. Results of transfer learning with DAVIS as source dataset and BindingDB as Fine-tuning dataset (left); Results of transfer learning with BindingDB as source dataset and DAVIS as Fine-tuning dataset (right)

Fig. 3 shows the results from the transfer learning method. On the left side, the results from pre-training the models on DAVIS dataset, then fine-tuning them on BindingDB dataset are shown. In comparison to the results from direct training on BindingDB dataset, we can see almost no difference in performance, indicating that the transfer learning method failed to increase the performance for this case. On the contrary, the right side of Fig. 3, which is the results from pre-training the models on BindingDB dataset, then fine-tuning them on DAVIS dataset, shows increment on both AUROC and AUPRC for all models, especially the Transformer CNN-RNN. This difference on performance increment shows that DAVIS is not a good dataset to be used as source dataset, as it is imbalanced. Instead, BindingDB can be a good source dataset as it increases the performance for the models that are fine-tuned with DAVIS.

For more thorough discussion, we also present the results in three aspects: (1) the models, (2) the datasets, and (3) the transfer learning method. The comparison between the models used to see the best performing model. Table 2 summarizes the experimental results of each model with the average AUPRC and time cost. The best score was obtained with a CNN decoder for both the drug and the target. The use of Transformer as a decoder does not seem to meet expectations, even the average AUPRC does not reach 50%. This proves that the use of an encoder designed specifically for sequential data does not seem to be so important in this case.

Table 2. Model Comparison

Drug Encoder	Target Encoder	Average AUPRC	Average AUROC
CNN	CNN	**76.06%**	**94.87%**
CNN-RNN	CNN-RNN	72.57%	93.92%
Transformer	CNN	45.32%	80.82%
Transformer	CNN-RNN	44.04%	78.38%

Table 3 compares the mean AUPRC and time costs of the trials using each of the data. The AUPRC value of the model trained and tested using BindingDB far outperforms the model using DAVIS data. This may be due to BindingDB's much larger amount of data, allowing deep learning models to learn more effectively. Another possible factor is the high false positive rate generated by the DAVIS-trained model, given that the AUPRC is very sensitive to this. The large size of the BindingDB dataset also underlies the reason why the average training time is up to 2.5 times longer than DAVIS.

Table 3. Dataset Comparison

Dataset	Average AUPRC	Average AUROC
BindingDB	**81.43%**	**91.33%**
DAVIS	34.75%	80.75%

The core experiment of this research is to investigate further the transfer learning method for DTI. Table 4 contains the average AUPRC results from experiments that use transfer learning and those that do not. In the experiment using BindingDB as the target dataset, the application of transfer learning from DAVIS did not seem to succeed in increasing the AUPRC value in the test data. In contrast to the DAVIS target dataset, the use of BindingDB as the source dataset succeeded in increasing the AUPRC by up to 6%. This result indicates that BindingDB could act better as source dataset than DAVIS. We hypothesize that the amount of data in BindingDB is larger than DAVIS, thus make it a better option to be a source dataset.

Table 4. Transfer Learning Comparison

Source Dataset	Target Dataset	Average AUPRC	Average AUROC
None	BindingDB	**81.43%**	**91.33%**
DAVIS	BindingDB	81.08%	91.28%
None	DAVIS	34.75%	80.75%
BindingDB	DAVIS	40.72%	84.62%

4. Conclusion

There are some interesting findings in this study. The first is CNN's performance, which consistently outperforms decoders using CNN-RNN and Transformers, which have been known to be effective in dealing with sequential data. Of course, these findings need to be explored further to find the cause. The second is that the performance of the model trained using DAVIS is very far from expectations, in contrast to BindingDB which produces a high AUPRC and AUROC value. Even pre-trained models with DAVIS dataset produce slightly lower AUPRC than non-pre-trained models that are directly trained on BindingDB. This case may be caused by the high imbalance in the dataset. Therefore, it can be concluded that the implementation of transfer learning results in better model performance in this DTI case when the source dataset is good. BindingDB is good enough to be used as a source dataset, this is proven in this study by consistently improving model performance compared to non-pre-trained models. For future works, we can explore on more datasets to be used as source dataset and finetuning dataset, also more model variations. We can also explore how to make an imbalanced dataset like DAVIS to be a better source dataset.

References

[1] Ezzat A, Wu M, Li XL, Kwoh CK. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform [Internet]. 2019 Jul 19;20(4):1337–57. Available from: https://doi.org/10.1093/bib/bby002.

[2] Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007 Jan;35(Database issue): 98-201.

[3] Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011 Oct;29(11):1046–51.

[4] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics [Internet]. 2018 Sep1;34(17):i821–9. Available from: https://doi.org/10.1093/bioinformatics/bty593.

[5] Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, et al. Toward more realistic drug–target interaction predictions. Brief Bioinform [Internet]. 2015 Mar 1;16(2):325–37. Available from: https://doi.org/10.1093/bib/bbu010.

[6] He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform [Internet]. 2017;9(1):24. Available from: https://doi.org/10.1186/s13321-017-0209-z.

[7] Cortés-Ciriano I, Ain QU, Subramanian V, Lenselink EB, Méndez-Lucio O, Ijzerman AP, et al. Polypharmacology modelling using proteochemometrics (PCM): Recent methodological developments, applications to target families, and future prospects. Vol. 6,

MedChemComm. 2015.

[8] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations ofwords and phrases and their compositionality. In:Advances in Neural Information Processing Systems. 2013.

[9] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-Learning-Based Drug-Target Interaction Prediction. J Proteome Res.2017;16(4).

[10] Gao K, Nguyen DD, Sresht V, Mathiowetz AM, Tu M, Wei GW. Are 2D fingerprints still valuable for drug discovery? Phys Chem Chem Phys. 2020;22(16).

[11] Zou N, Zhu Y, Zhu J, Baydogan M, Wang W, Li J. A Transfer Learning Approach for Predictive Modeling of Degenerate Biological Systems. Technometrics [Internet]. 2015 Jul 3;57(3):362–73. Available from: https://doi.org/10.1080/00401706.2015.1044117.

[12] Mieth B, Hockley JRF, Görnitz N, Vidovic MMC, Müller KR, Gutteridge A, et al. Using transfer learning from prior referenceknowledge to improve the clustering of single-cell RNA-Seq data. Sci Rep. 2019;9(1):20353.

[13] Turki T, Wei Z, Wang JTL. Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients. IEEE Access. 2017;5:7381–93.

[14] Mourragui S, Loog M, van de Wiel MA, Reinders MJT, Wessels LFA. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. Bioinformatics [Internet]. 2019 Jul 15;35(14):i510–9. Available from: https://doi.org/10.1093/bioinformatics/btz372.

[15] Pan SJ, Yang Q. A Survey on Transfer Learning - {IEEE} Xplore Document. IEEE Trans Knowl Data Eng. 2009;22(10):1345–59.

[16] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945–54.

[17] Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, et al. Transfer Learning for Drug Discovery. J Med Chem [Internet]. 2020 Aug 27;63(16):8683–94. Available from: https://doi.org/10.1021/acs.jmedchem.9b02147.

[18] Shin B, Park S, Kang K, Ho J. Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction. 2019.

[19] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202–13.

[20] Kotsias PC, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. Nat Mach Intell [Internet]. 2020;2(5):254–65. Available from: https://doi.org/10.1038/s42256- 020-0174-5.

[21] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. JChem Inf Model. 2019;59(8).

[22] Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics [Internet]. 2020 Dec 1;36(22–23):5545–7. Available from: https://doi.org/10.1093/bioinformatics/btaa1005.

[23] Olivas ES, Guerrero JDM, Martinez Sober M, Magdalena Benedito JR, Serrano López AJ. Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. 2009.