International Journal of Research in Interdisciplinary Studies
Volume 1, Issue 2, October 2023
https://www.ijris.com | ISSN (Online): 2584-1017

Comparison of Embedding Techniques used in Prediction of Drug Target Binding Affinity

Neha*
Assistant Professor, Department of Electronics and Communication Engineering, Maharaja Surajmal Institute of Technology, Delhi, India

Abstract— The process of identifying of drug-target (DT) interactions is an integral part of the drug discovery process. Drug discovery is a process which focuses on determining new compounds that can be used to cure and treat diseases. Embedding is the process of converting high-dimensional data to low-dimensional data by representing it in the form of a vector. Drug Target sequences need to be transformed into a matrix before they can be fed into a deep learning model. Since the performance of embedding techniques directly affects the quality of the deep learning models and thus the accuracy of the predicted values of drug target binding affinity, we compare the performance and effect of various embedding techniques used for target embedding on deep learning models. Here, the embedding techniques whose performance has been compared are PLUSRNN Embedder, ProtTransBertBFD Embedder, Beppler Embedder, CPCProt Embedder and SeqVec Embedder.

***Index Terms*— drug target binding affinity, drug target interaction, drug discovery, embedding techniques, graph attention network.**

1. Introduction

The strength of the binding (interaction) of a ligand and its receptor is described by affinity. When discussing protein–drug interactions, the binding affinity could be a measurement of how well the drug binds to the protein. Binding affinity suggests the strength that drug-target pairs interaction holds. With binding, drugs can have a positive or negative influence affecting the disease conditions, on functions applied by proteins. By interpretation of drug-target binding affinity, it's attainable to seek out possible drugs that are ready to inhibit the target/protein and benefit many other bio informatics applications. Drugs elicit their desired remedial effects by interacting with particular disease-related targets. Once a target is discovered, drug molecules are screened against the target to spot those who interact with the target and modulate its activity. Having the ability to accurately predict drug–target interactions (DTIs) and DTBA is crucial in early drug development and through drug repurposing endeavors. Here, we glance at some recent research developments during this space. As a result, DTA prediction has received such lots attention within the past few years. Drug target interaction may be defined as the binding of a drug to a target location because of which a change in its behaviour/function are often observed. Drugs work by interacting with target proteins to activate or inhibit the process of the targets. Thus, identifying novel drug-target interactions (DTIs) is a necessary step within the drug discovery field. A drug or medicine is defined as any chemical substance that brings a few physiological changes within the chassis when it's consumed, injected or absorbed. The development of new drugs that overcome resistance mutations requires an understanding of the factors that contribute to resistance. In some cases, an examination of a crystal structure of the drug-bound molecular target is enough to envision how the mutation leads to resistance. However, Gabel et al. [1] showed that RF-score failed in virtual screening and docking tests, speculating that using features like co-occurrence of atom-pairs over-simplified the outline of the protein-ligand complex and led to the loss of knowledge that the raw interaction complex could provide round the same time this study was published, deep learning began to become a preferred architecture powered by the rise in data and high capacity computing machines challenging machine learning methods.

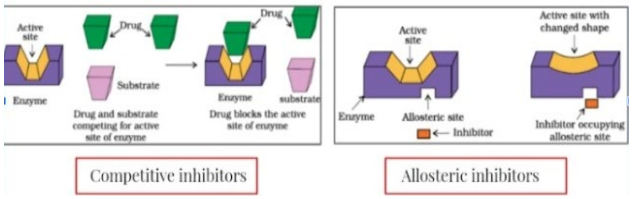


Fig. 1.

2. Literature Review

The drug development is difficult, expensive and success rates are low. Thus, drug target interactions and their binding affinities identification and predication has become an important step in the initial stages of drug discovery. In the meantime, many computational methods have been discovered to minimize the errors of potential DTIs approaches. AI/ML/DL-based Drug discovery and interaction methods have been earlier used in many research projects and have shown good results. However, none of these models are error-free, all of them face a few limitations that could require changes/improvements.

These computational approaches focus on scaling down the search space for DTIs and enlighten drug-target operational condition. This observation accentuated the requirement for:

*Corresponding author: neha.khokhar@msit.in

new, more appropriate drug targets, which will upgrade the capabilities of the drug discovery and screen an outsized number of medications within the very initial phase of drug discovery process, thus guiding toward those drugs that will reveal better efficacy. Due to this, procedures that predict DTIs and most importantly, drug-target binding affinities (DTBA) are highly popular.

Over the last thirty years, various papers that predict DTIs or their affinities have emerged varying from ligand/receptor-based methods (Cheng et al.; Wang et al.) [2], [3] to gene ontology-based (Mutowo et al.) [4], [5], text-mining-based methods (Zhu et al.) [6], and reverse virtual screening techniques (inverse-docking) (Lee et al., 2016; Vallone et al.; Wang et al.) [7]-[9]. Development of such methods continues to be ongoing as each method suffers from different kinds of limitations. More newer approaches established AI, network analysis, and graph mining (Emig et al.; Ba-Alawi et al.; Luo et al.; Olayan et al.,) [10]-[12], and ML and DL techniques (Liu Y. et al.; Rayhan et al.; Zong et al.; Tsubaki et al.) [13], [14] to develop prediction models for DTIs and Drug discovery.

Halogen bond (Zhijian Xu, Zhuo Yang) [15], [16] has attracted a decent deal of attention in the past few years for lead generation optimization aiming at improving drug-target binding affinity. Therefore, we recommend that albeit halogenation may be an important approach for improving ligand bioactivity, it should be given more recognition in the near future to the appliance of the halogen bond for ligand ADME/T property optimization.

Thus, several exhaustive latest reviews summed up the various studies that predict DTIs using various techniques and AI/ML-based methods as presented in Liu Y. et al. (2016) [17], Ezzat et al. [18], Rayhan et al. [19], Trosset and Cavé, and Wan et al. [20], [21].

This approach suffers from two major limitations including: (1) the shortcoming to differentiate between true negative interactions and instances where the dearth of knowledge or missing values impede predicting an interaction, and (2) it doesn't reflect how tightly the drug binds to the target which reflects the potential efficacy of the drug.

Experimental results show that our Deep Learning model can give us a better DTI prediction performance if a few changes are considered. In the future, we would like to work on more protein compounds and hope for better target-interaction results.

We have tried to work on this project to reduce the time, cost and probability of failure of drug discovery process, To propose a method that takes into consideration both local chemical context and topological structure of both targets and ligands to calculate drug target affinity, To perform a comparison of different embedding techniques, To perform feature extraction and selection on matrices obtained after applying embedding, Predicting the binding affinity value based on the embeddings of the drug and the target.

3. Methodology

A. Dataset

We have used the Davis dataset [22], [23] which contains linkage of 442 kinases with 72 kinase inhibitors covering >80% of the human catalytic protein kinome.

Kinase is a kind of catalyst (a protein that paces up substance responses in the body) that adds synthetics called phosphates to different atoms, like sugars or proteins. This might make different particles in the cell become either dormant or active.

Unlike most research papers that treat drug-target binding affinity as a binary problem, in this study we are treating it as a Regression problem. We have split the dataset into 3 parts-training dataset, validation dataset and test dataset for the purpose of this study. As of writing this paper the Davis dataset contains 25,772 DTI pairs, 68 drugs, 379 proteins. It has fields for ‘Drug_ID’, ‘Drug’ containing SMILES notation of the compound, ‘Target_ID’, ‘Target’ and ‘Y’ which is the value of drug target binding affinity.

B. Embedding

Embedding is the process of converting high-dimensional data to low-dimensional data by representing it in the form of a vector. We transformed each protein target sequence into a matrix so that it can be fed into a deep learning model such that each row gives the embedding of a protein. Then, the matrix is passed as input to a fully connected network of molecular graphs containing 2 layers of 1024 and 128 neurons to extract features from target proteins.

The novel idea behind this paper is comparing different embedding techniques suggested to embed proteins and compare their performances with a GAT network as the embedding techniques greatly influence the quality of deep learning models.

Here, the embedding techniques whose performance has been compared are PLUSRNN Embedder [24], ProtTransBertBFD Embedder [25], Beppler Embedder [26], CPCProt Embedder [27] and SeqVec Embedder [28].

C. Encoding

The Drug Molecules have been represented through the SMILES sequence (Simplified Molecular-Input Line-Entry System) which is a system that uses line notation to represent the structure of chemical compounds [29]. A SMILES string could be further converted into different formats before feeding into the machine learning or deep learning model such as molecular fingerprint, one-hot encoding, or word embedding.

Since the focus of this paper is comparing the performance of embedding techniques used on target molecules, we only use one-hot encoding to encode drug molecules.

In one hot encoding, each atom is distinguished from the other using a unique integer allotted to each atom. The result will contain ‘1’ in a cell used to uniquely identify that atom and ‘0’ in others.

D. Model

1) Graph Attention Networks

The attention mechanism in machine learning/deep learning

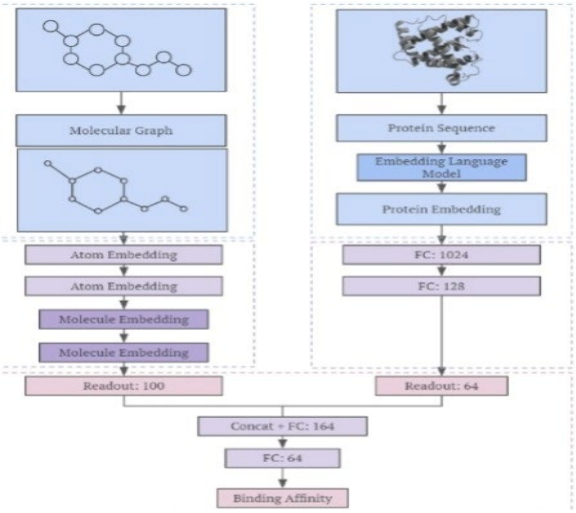
[30] allows the neural network to focus on more important or more relevant parts of input. This further helps us to achieve more accurate predictions.

Graph attention networks [31] operate on graph structured data and take into account features of neighbouring nodes, assigning different importance to each neighbour's contribution, which makes GATs better than GNNs. They work in an anisotropic manner without requiring to compute any costly matrix operation. Thus, here we use the GATs as in a molecule each atom has varying degrees of affect on its local atoms depending on its chemical properties and spatial arrangement. The equations involved in the model are as below:

$$e_{vu} = \text{leaky\_relu}\ (W.\ [h_p\ ,\ h_u]) \quad (1)$$

$$a_{vu} = \text{softmax}(e_{vu}) = \exp(e_{vu})/\Sigma_{u\in N(v)}\exp(e_{uv}) \quad (2)$$

$$C_v = \text{elu}[\ \Sigma_{u\in N(v)}\ a_{vu}.W.h_u] \quad (3)$$



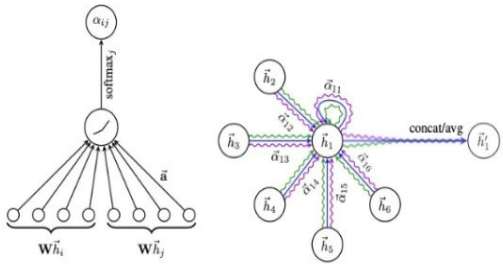


Fig. 2.

2) Early stopping

Early stopping [32] was used to prevent overfitting and also reduce training time consumption. Here, we set the mode as 'lower' i.e. lower metrics will suggest a better model and 'patience' as 20 which means that if the performance metrics does not improve for 20 consecutive iterations, the process will be terminated.

3) Evaluation metrics

Here, we have used three evaluation metrics that are usually used in regression tasks to evaluate the performance of various embedding techniques. These include Mean Square Error (MSE), Root Mean Square Error (RMSE) and R squared.

MSE: Mean Square Error (MSE) is the mean of the difference between the actual value and the predicted value squared.

$$\text{MSE} = 1/n\Sigma_{i=1}(y_i - y'_i)^2 \quad (4)$$

RMSE: Root Mean Square Error can be defined as the estimate of how well a regression line fits the data points.

$$\text{RMSE} = \sqrt{1/n\Sigma_{i=1}(y_i - y'_i)^2} \quad (5)$$

R Squared: R squared, also called the coefficient of determination demonstrates the proximity of the line plotted by plotting the predicted values to the actual values.

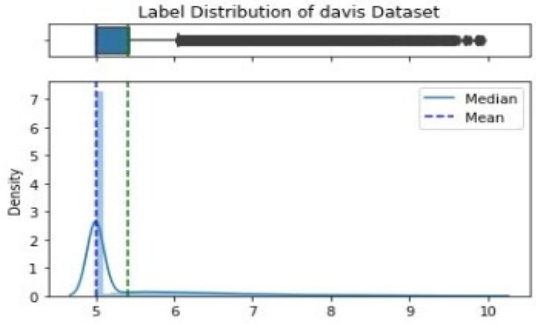


Fig. 3. Data analysis

4. Discussion and Conclusion

Here, we have compared various embedding techniques used for embedding target sequences in machine learning and deep learning approaches. As we can see, the SeqVec Embedder approach gives the best performance in terms of MSE and RMSE.

A. Limitations and Future Studies

One major drawback of predicting drug target binding affinity using deep learning is the lack of data for negative cases of drug target binding affinity. Another loophole is the large number of unknown/ undiscovered drugs.

While this paper compares different protein embedding techniques and their effect on deep learning models. In future, this work can be extended to comparing different embedding techniques used for drug molecules and their effect on deep

Table 1
Result

Embedding Technique	MSE	RMSE	R2
PLUSRNN Embedder	0.36 239239396586964	0.6019903603595905	0.44108237816154794
ProtTransBertBFD Embedder	0.37213957438783835	0.6100324371603844	0.4260493062985934
Beppler Embedder	0.3516710651477847	0.5930186043858866	0.45761787864591097
CPCProt Embedder	0.4231	0.65046137471	0.49235786421545652
SeqVec Embedder	0.3146	0.56089214649	0.4346547891356645

learning/machine learning models.

References

[1] Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., et al. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75.

[2] Chou, T. C., and Talalay, P. (1984). Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Adv. Enzyme Regul.* 22, 27–55.

[3] Lee, A., Lee, K., and Kim, D. (2016). Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov.* 11, 707–715.

[4] Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci.* 11, 320–328.

[5] Li, Q., and Shah, S. (2017). Structure-Based Virtual Screening. *Methods Mol. Biol.* 1558, 111–124.

[6] Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., et al. (2013). Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE* 8:e60618.

[7] Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., and Philbrick, K. (2017). Toolkits and libraries for deep learning. *J. Digit. Imaging* 30, 400–405.

[8] Ezzat, A., Wu, M., Li, X., and Kwoh, C. K. (2019). Computational prediction of drug-target interactions via ensemble learning. *Methods Mol Biol.* 1903, 239–254.

[9] Zong, N., Kim, H., Ngo, V., and Harismendy, O. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics* 33, 2337–2344.

[10] Zong, N., Wong, R. S. N., and Ngo, V. (2019). Tripartite network-based repurposing method using deep learning to compute similarities for drug-target prediction. *Methods Mol. Biol.* 1903, 317–328.

[11] Zhengdan Zhu, Zhijian Xu, Weiliang Zhu. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *Journal of Chemical Information and Modeling* 2020, *60* (6), 2683-2696.

[12] [Zhuo Yang, Zhijian Xu, Yingtao Liu, Jinan Wang, Jiye Shi, Kaixian Chen, and Weiliang Zhu. Unstable, Metastable, or Stable Halogen Bonding Interaction Involving Negatively Charged Donors? A Statistical and Computational Chemistry Study. *The Journal of Physical Chemistry B* 2014, *118* (49), 14223-14233.

[13] F. Chollet et al., Keras, 2015.

[14] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 3642–3649, 2012.

[15] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar. Predicting drug–target interactions using probabilistic matrix factorization. Journal of chemical information and modeling, 53(12):3399–3409, 2013.

[16] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1):30–42, 2012.

[17] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. Nature biotechnology, 29(11):1046–1051, 2011.

[18] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In ICML, pages 647–655, 2014.

[19] A. M. Edwards, R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and H. Y. Frank. Too many roads not taken. Nature, 470(7333):163, 2011.

[20] O. Fedorov, S. M¨uller, and S. Knapp. The (un) targeted cancer kinome. Nature chemical biology, 6(3):166, 2010.

[21] Davis, M., Hunt, J., Herrgard, S. et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29, 1046–1051 (2011).

[22] Huang, Kexin, et al. "DeepPurpose: a Deep Learning Library for Drug-Target Interaction Prediction," Bioinformatics.

[23] Yao, Y.; Rosasco, L.; Caponnetto, A. On Early Stopping in Gradient Descent Learning Constr. Approx. 2007, 26 (2), 289−315.

[24] Seonwoo Min, Seunghyun Park, Siwon Kim, Hyun-Soo Choi, Sungroh Yoon, "Pre-Training of Deep Bidirectional Protein Sequence Representations with Structural Information."

[25] Elnaggar, Ahmed, et al. "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High-Performance Computing," 2020.

[26] Bepler, Tristan, and Bonnie Berger. "Learning protein sequence embeddings using information from structure." 2019.

[27] Lu, Amy X., et al. "Self-supervised contrastive learning of protein representations by mutual information maximization," 2020.

[28] Heinzinger, Michael, et al. "Modeling aspects of the language of life through transfer-learning protein sequences." BMC bioinformatics 20.1 (2019): 723.

[29] David Weininger, 'Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules', Journal of chemical information and computer sciences, 28(1), 31–36, (1988).

[30] Bahdanau et al. 'Neural Machine Translation by Jointly Learning to Align and Translate'.

[31] Veličković et al. 'Graph Attention Networks,' 2017.