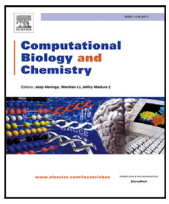
Computational Biology and Chemistry 115 (2025) 108310

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Research Article

Improving binding affinity prediction by emphasizing local features of drug and protein

Daejin Choi [a,*], Sangjun Park [b]

[a] Department of Computer Science and Engineering, Incheon National University, Incheon, Republic of Korea
[b] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Dataset link: https://github.com/Koreaj9u7n/LLF

Keywords:
Binding affinity prediction
Graph neural network
Convolutional neural network

ABSTRACT

Binding affinity prediction has been considered as a fundamental task in drug discovery. Despite much effort to improve accuracy of binding affinity prediction, the prior work considered only macro-level features that can represent the characteristics of the whole architecture of a drug and a target protein, and the features from local structure of the drug and the protein tend to be lost. In this paper, we propose a deep learning model that can comprehensively extract the local features of both a drug and a target protein for accurate binding affinity prediction. The proposed model consists of two components named as Multi-Stream CNN and Multi-Stream GCN, each of which is responsible for capturing micro-level characteristics or local features from subsequences of a target protein sequence and subgraph of a drug molecule, respectively. Having multiple streams consisting of different numbers of layers, both the components can compute and preserve the local features with a stream consisting of a single layer. Our evaluation with two popular datasets, Davis and KIBA, demonstrates that the proposed model outperforms all the baseline models using the global features, implying that local features play significant roles of binding affinity prediction.

1. Introduction

Binding affinity prediction, which anticipates the strength of interaction between a given pair of a small drug-like molecule and a target protein, has been considered as a fundamental task of virtual screening in drug discovery (Nguyen et al., 2019). That is, by providing the information to filter out drug candidates that tend less to be combined with the target protein, predicting binding affinity can help to achieve a primary goal of virtual screening, which is to find drug-like compounds that can be combined with target proteins. However, exhaustive search for drug molecules with human labors is costly, time-consuming, and even infeasible (Ma et al., 2017; Mukherjee et al., 2022), the computational approach has been leveraged as an alternative method for binding affinity prediction. Indeed, it has been reported that virtual screening methods can not only explore more number of drug-like molecules to interact with a target protein, but also reduce human resources on practical experiment of docking drugs with target protein (He et al., 2017; Mukherjee et al., 2022; Nguyen et al., 2019). The growing need and importance of the computational approach for binding affinity prediction have attracted the research community to develop machine-learning models for drug–target affinity (DTA) prediction. This endeavor manifests in the form of two primary categories: feature-based (He et al., 2017; Cichonska et al., 2017, 2018) and deep learning-based (Tsubaki et al., 2018; Öztürk et al., 2018; Ma et al., 2017; Mukherjee et al., 2022; Nguyen et al., 2019). The feature-based models extract the hand-crafted features (e.g., object-based features or network metrics in drug–drug, target–target, or drug–target networks), which are fed into the derived models based on traditional machine learning methods such as gradient boosting method (He et al., 2017), or regularized least squares (Cichonska et al., 2017, 2018).

With the great upsurge and improvement of deep learning techniques and their success in a wide range of predicting applications, a few models adopting popular deep learning techniques such as Convolutional Neural Network (CNN) (Öztürk et al., 2018; Ma et al., 2017), Long Short-Term Memory (LSTM) (Mukherjee et al., 2022), Generative Adversarial Network (GAN) (Zhao et al., 2020), attention mechanism (Zhao et al., 2023) have been proposed. However, despite the improved performance by capturing comprehensive features from drugs and proteins, these models were restricted by the lack of representation of the structures of both molecules and proteins as only string representation through SMILES was employed.

In recent years, the graph neural network (GNN) has been proposed as a solution to overcome the limitation of prior methods. Instead of using a string-based representation for the structure of a drug molecule

* Corresponding author.
E-mail addresses: djchoi@inu.ac.kr (D. Choi), wedm2401@korea.ac.kr (S. Park).

https://doi.org/10.1016/j.compbiolchem.2024.108310
Received 21 March 2024; Received in revised form 10 October 2024; Accepted 4 December 2024
Available online 11 December 2024

(e.g., SMILE), the GNN-based approaches model the structure of the drugs as a graph and directly computing feature vectors from the constructed graph by considering the relations among the nodes in the graph. By extracting comprehensive features of a structure of a drug, the GNN-based models, such as DeepGS (Ma et al., 2017), GraphGTA (Nguyen et al., 2019), or DeepGLSTM (Mukherjee et al., 2022), have shown the enhanced performance for binding affinity predictions.

Unfortunately, despite such significant advance, the suggested models have been restricted to preserve the features of *local structures* of a given drug molecule. That is, since the information of a broader range of neighbors (e.g., neighbors of the neighbors) for a given node can be considered by increasing the number of GNN layers, all the prior GNN models employed a single pipeline with multiple GNN layers to compute the embeddings of a given graph. However, the increase of the number of GNN layers may lose the information of the closer neighbors of a node (e.g., the directly communicating atoms) or even the node itself. In other word, passing a graph through multiple layers may cause the loss of the micro-level (or local) information of a node and its closer network while considering larger network, as reported in the prior studies (Hamilton et al., 2017; Pan et al., 2023). Therefore, the GNN-based models in the prior studies are unlikely to preserve the local features of a drug molecule. Similarly, the characteristics of subsequences (e.g., 3- or 5-consecutive amino acids) of protein has been less considered. That is, the recent approaches have popularly used RNNs or LSTMs for feature extraction of a given protein sequences, which puts a whole protein sequence to compute the feature vectors of the given protein. Thus, the models consider only the macro-level characteristics of a protein, instead of the information made from subsequences.

We claim that the local features (i.e., the characteristics of the subsequences of a protein and sub-graphs of a molecule) should be more considered for binding affinity prediction since the local features are the fundamental blocks to indicate the characteristics of the protein and the drug, as reported by the recent studies (Jiang et al., 2020; Pan et al., 2023). Assuming that local features play significant roles in binding affinity prediction, in this paper, we propose a deep learning model that can accurately estimate binding affinity between a drug and a target protein, by using emphasized local features of the drug and the protein sequence. In particular, we design two components named as Multi-Stream GCN (MS-GCN) and Multi-Stream CNN (MS-CNN). Consisting of multiple flows with different numbers of GCN and CNN layers, MS-GCN and MS-CNN compute and keep the local features of the given pair of a molecule and a target protein, which indicate the characteristics of local relations among the atoms in a part of the molecule and subsequences of a protein, respectively. Our evaluation with two public datasets demonstrates that the proposed model shows competitive performance on both datasets. Throughout several experiments for the proposed model with different settings of the modules, we also show that the features extracted from local structure play significant roles in binding affinity prediction. We highlight the contributions of this paper as follows.

- We propose a deep learning model that captures and uses the comprehensive local features of the given protein and a drug molecule. In particular, the proposed model consists of two components, each of which uses multiple streams with different numbers of CNN or GCN layers. By preserving local information of the given protein and the drug molecule, the proposed model outperforms all the baseline models in the experiment with two popular datasets.
- We also conduct the ablation study for the proposed model, which shows that the models using multiple streams outperform the other models with a single streamline. Note that this result implies that extracting more comprehensive and diverse information from the local structure can improve binding affinity prediction.

The remainder of this paper is organized as follows. We introduce the model for binding affinity prediction and the evaluation methodology in Section 2. We then describe the experimental results in Section 3. After discussing the implications and the limitations of this paper in Section 4, we conclude the paper in Section 5.

2. Methods

2.1. Proposed model

2.1.1. Overall architecture

Fig. 1 depicts the overall architecture of the proposed model. The proposed model accepts a pair of a protein sequence and a drug with SMILES representation that indicates a protein sequence as a string of ASCII characters, each of which is an amino acid. The pair of the protein sequence and the SMILES code of the target drug is then forwarded through a separate path for feature extraction. The given protein sequence is converted into an embedding vector computed as a concatenation of one-hot encoding of each element of the sequence. The embedding vector is then fed into multi-stream convolutional neural network to capture comprehensive features of the given protein sequence. On the other hand, the SMILES representation of the drug changes into graph representation by RDKit (Landrum, 2016), then are put into multi-stream graph convolutional network, which is designed to extract comprehensive features from the local structure or ego network among atoms. The calculated feature vectors of the given target protein sequence and the drug are fused in the fusion network to estimate a binding affinity score of the given protein–drug pair.

2.1.2. Multi-stream CNN

Focusing on that a protein can be represented as a sequence of amino acids, the deep learning techniques to extract the features from a sequential input, including Long Short-Term Memory (LSTM) or one dimensional (1-D) CNN have popularly been applied to the deep learning models for binding affinity prediction. Although a single streamline consisting of multiple CNN blocks or LSTM cells in these models can compute the macro-level (or global) features of a given protein, the micro-level information of the local blocks (i.e., 3- or 4-consecutive amino acids) of a protein sequence may not be preserved as passing multiple CNN blocks or LSTM cells may lose direct relations of the subsequences. Note that this issue was mentioned by Jiang et al. (2020), which suggested a prediction model using the subsequence blocks of a given protein for binding affinity prediction.

With a hypothesis that the combination of the local features for the protein sequence plays more significant roles in binding affinity prediction than global features, multi-stream CNN of the proposed model not only calculates the comprehensive features of the protein sequence with a flow of three 1-D CNN layers, but also keeps the outputs of a single 1-D CNN layer and two 1-D CNN layers to preserve the local features of the given sequence. The computed features are then concatenated after being passed through max-pooling layer, which is expected that the information of local blocks of a protein sequence is kept. The concatenated vector of the output vectors of each stream is then fed into fusion network.

2.1.3. Multi-stream GCN

Recent studies have demonstrated that indicating a drug as a graph representation and adopting graph neural network (GNN)-based models significantly improve the performance of binding affinity prediction since the graph representation can indicate the topological and relational characteristics among atoms of a drug. In addition, it has been also reported that stacking more GCN layers improves the performance of the binding affinity prediction since it can consider more comprehensive information of multi-hop neighbors (e.g., neighbors of neighbors) for a given node (Hamilton et al., 2017). With these rationales, all

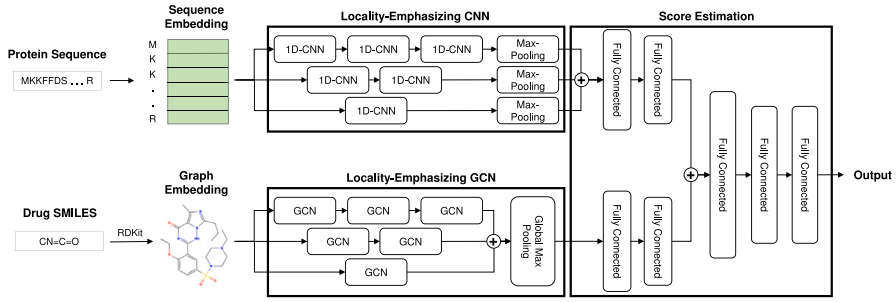


Fig. 1. An illustration of the proposed model is shown.

the prior GNN models employed a single pipeline with multiple GNN layers to compute the embeddings of a given graph. Unfortunately, the increase of the number of GNN layers may lose the information of the closer neighbors of a node (e.g., the directly communicating atoms) or even the node itself. In other word, passing a graph through multiple layers may cause the loss of the micro-level (or local) information of a node and its closer network while considering larger network, as reported in the prior studies (Hamilton et al., 2017; Pan et al., 2023). Therefore, the GNN-based models in the prior studies are unlikely to preserve the local features of a drug molecule.

Inspired by the prior work and not to lose the information of the closer neighbors, the proposed model employs graph convolutional network (GCN), a popular GNN-based model, as a similar form of MS-CNN. That is, the MS-GCN consists of three streamlines, each of which stacks one, two, and three GCN layers. In this way, MS-GCN can capture the comprehensive information in the local structure among nodes without losing the information of the closer neighbors. Since a GCN layer computes the feature vectors of all individual nodes, it is required to convert the vectors into a graph-level feature vector. To this end, the MS-GCN concatenates the vectors of all the nodes of a drug graph for all streamlines and generates a single vector with a global max-pooling layer. The final vector of a given drug graph is then forwarded into fusion network.

2.1.4. Score estimation

Located at the last of the proposed model, the score estimation module accepts the feature vectors of both a given protein sequence and a drug to estimate a binding affinity score for the given pair. In particular, the two fully-connected layers of this module compress the feature vectors of the protein–drug pair as the same dimensional vectors. The compressed vectors are concatenated and passed to three fully-connected layers, which output the final score for binding affinity.

2.1.5. Hyperparameters and implementation detail

We manually explored the best hyperparameters of the proposed model in each experiment and finally set dropout ratio and learning rate to 0.2 and 0.0005, respectively, for all the experiments on both Davis and KIBA datasets. The number of channels and the kernel size of 1-D CNN layers in MS-CNN are 128 and 3, respectively. Note that zero-padding is applied to address the variable sequence lengths of the proteins. The epoch count and the batch size are differently set for two datasets, with the consideration that the amounts of two datasets are different. In particular, the epoch counts (and the batch sizes) for the experiment on Davis and KIBA datasets are 128 (512) and 1000 (2000), respectively. Adam optimizer (Loshchilov and Hutter, 2017), ReLU (Agarap, 2018), and mean squared error (MSE) were used for model training, activation, and a loss function, respectively. The kernel size and stride of 1-D CNN in MS-CNN are set to 3 and 1, respectively. Note that all the proposed models were implemented on a machine of Intel Xeon CPU (@ 2.10 GHz), 256 GB of RAM, and 2-way Tesla V100 graphics cards, using Python 3.6.9 and PyTorch 1.12.1 with CUDA 11.3 toolkit on Ubuntu 18.04. All the libraries used in the implementation and experiments are also reported in the repository.[1]

Table 1
The statistical description of the datasets used for evaluation.

Dataset	#Drugs	#Target proteins	#Drug–Protein pairs	Score range
Davis	68	442	30,056	[5.0, 10.8]
KIBA	2,111	229	118,254	[0.0, 17.2]

2.2. Evaluation

2.2.1. Dataset

The evaluation for the proposed model is conducted with two datasets that have been popularly used for DTA prediction: (i) Davis (Davis et al., 2011) and (ii) KIBA (Tang et al., 2014). The statistical description and the distributions of the binding affinity scores for the datasets are summarized in Table 1 and Fig. 2, respectively.

The proposed model is separately trained and evaluated for individual datasets. We conducted five-fold cross validation with both the datasets. In particular, we randomly split each dataset with 8:2 ratio, each of which is used for training and testing, respectively, and then report the evaluation metrics on average.

2.2.2. Evaluation metric

We choose three well-known metrics, MSE, concordance index (CI), and r_m^2, to evaluate and compare the proposed model with the other baseline models, which are formally defined as follows:

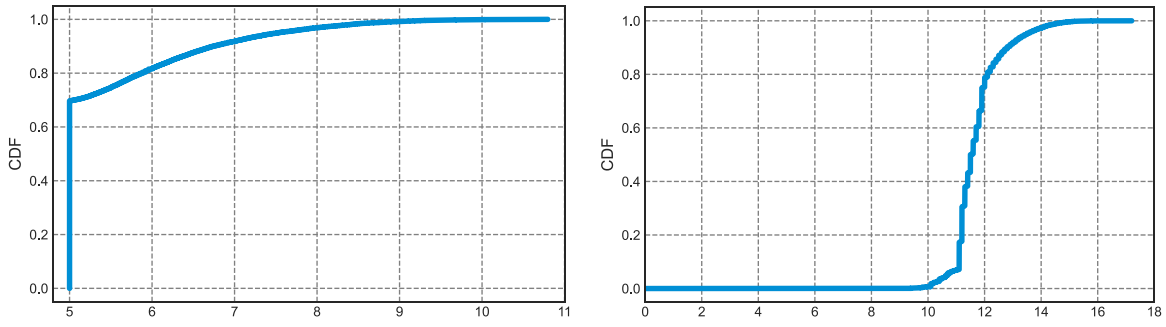
- **Mean Squared Error (MSE):** $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$
- **CI:** $\frac{1}{Z}\sum_{y_i>y_j} h(f_i - f_j)$, where $h(x) = 1, x > 0; 0.5, x = 0; 0, x < 0$ and Z is the number of drug-target pairs with different binding affinity values.
- r_m^2: $r^2 * \left(1 - \sqrt{(r^2 - r_0^2)}\right)$, where r and r_0 are the squared correlation coefficient with and without an intercept, respectively.

3. Results

3.1. Overall performance

Table 2 describes the performance of the proposed model on two benchmark datasets, Davis and KIBA. Here, we also indicate the performance of the other models proposed in the prior studies for a comparison purpose. Note that we borrowed the performance metrics reported in individual studies with the expectation that the best performance was reported in the paper where the model was proposed. The proposed model shows promising performance across all the datasets; the MSE and r_m^2 of the proposed model on Davis (and KIBA) dataset

[1] https://github.com/Koreaj9u7n/LLF



(a) Davis (b) KIBA

Fig. 2. The distributions of the scores for Davis and KIBA datasets are illustrated.

Table 2
Overall performance.

Model	Davis			KIBA		
	MSE	CI	r_m^2	MSE	CI	r_m^2
KronRLS (Cichonska et al., 2017, 2018)	0.379	0.869	0.407	0.411	0.782	0.342
SimBoost (He et al., 2017)	0.282	0.873	0.644	0.222	0.836	0.629
DeepDTA (Öztürk et al., 2018)	0.261	0.878	0.630	0.194	0.863	0.673
MT–DTI (Shin et al., 2019)	0.245	0.887	0.665	0.193	0.882	0.738
DeepCPI (Tsubaki et al., 2018)	0.293	0.867	0.607	0.211	0.852	0.657
WideDTA (Öztürk et al., 2019)	0.262	0.886	0.633	0.179	0.875	0.675
GANsDTA (Zhao et al., 2020)	0.276	0.881	0.653	0.224	0.866	0.775
Attention-DTA (Zhao et al., 2023)	0.245	0.887	0.657	0.162	0.882	0.735
DeepGS (Ma et al., 2017)	0.252	0.882	0.686	0.193	0.860	0.684
GraphDTA (GCN) (Nguyen et al., 2019)	0.254	0.880	–	0.139	0.889	–
GraphDTA (GAT) (Nguyen et al., 2019)	0.232	0.892	0.662	0.179	0.866	0.671
GraphDTA (GIN) (Nguyen et al., 2019)	0.229	0.893	0.649	0.147	0.882	0.684
DeepGLSTM (Mukherjee et al., 2022)	0.232	0.895	0.680	0.133	0.897	0.792
Proposed	**0.217**	**0.894**	**0.710**	**0.131**	**0.902**	**0.795**

Table 3
Performance with the change of streams in MS-CNN.

Settings	MSE	CI	r_m^2
Single Stream: One Layer	0.218	0.900	0.701
Single Stream: Two Layers	0.215	0.899	0.721
Single Stream: Three Layers	0.211	0.895	0.704
Single Stream: Four Layers	0.214	0.895	0.717
Single Stream: Five Layers	0.214	0.903	0.728
Combination of One and Two Layers	0.212	0.901	0.719
Combination of One, Two, Three Layers	0.205	0.903	0.717

are 0.217 (0.131) and 0.710 (0.795), respectively, which outperforms all the other benchmark models. These results not only demonstrate that the proposed model can predict the binding affinity of a pair of a protein sequence and a target drug with high accuracy, but also imply that capturing comprehensive *local* features from both a protein sequence and a target drug can improve the performance of binding affinity prediction. Note that the GNN-based models, GraphDTA and DeepGLSTM, tend to indicate higher performance than other benchmark models, implying that considering the graph structure of drugs plays a significant role in predicting binding affinity scores.

3.2. Performance comparison with differentiating the layers of MS-CNN

We next investigate how MS-CNN and MS-GCN contribute to binding affinity prediction. To this end, we measure the performance differences based on the changes of the number of layers and the combinations of streams consisting of MS-CNN. Here, we use a single fold of KIBA dataset. Table 3 summarize the performance change of the proposed model when one to five streams are used. The MSEs of the models decrease until the number of the layers are three and increase when the number of layers are 4 and 5, which indicates that using more layers does not assure performance improvement. With this result, the combinations of with one to three layers are only tested. The proposed model with the single stream in MS-CNN shows the degraded performance, compared with multiple streams; the MSE values of the models with a single stream are more than 0.211 while the one with three streams is 0.205, meaning that combining multiple streams can improve the performance of binding affinity prediction. The combinations of multiple streams show an improved performance than the models with each stream. This result indicates that the streams complementarily work for the prediction task.

Interestingly, the models of the single stream with different numbers of CNN layers show different performance; the MSE value of the model with a single stream of three 1-D CNN layers is the lowest while the case of two 1-D CNN layers outperforms other models in terms of r_m^2. These results demonstrate the reason why multiple streams should be used. That is, since different streams consisting of different numbers of CNN layers play different roles in binding affinity prediction, the feature vectors generated in the different streams should be preserved and used for the prediction.

3.3. Performance comparison with differentiating locality-emphasizing GCN

We also explore how each stream in MS-GCN affects the model performance by measuring the performance change with differentiating the elements of MS-GCN. Table 4 summarizes the performance of the proposed model with different settings of MS-GCN. Note that MS-CNN was set to have three streams in these experiment. As shown in Table 4, using multiple streams that consist of different numbers of GCN layers improves model performance than the models with single flow in MS-GCN. The MSE values of single stream with one, two, three GCN layers are 0.214, 0.214, and 0.207, respectively, while the ones of the models

Table 4
Performance with the change of streams in locality-emphasizing GCN.

Settings	MSE	CI	r_m^2
Single Stream: One Layer	0.214	0.903	0.715
Single Stream: Two Layers	0.214	0.902	0.729
Single Stream: Three Layers	0.207	0.905	0.726
Single Stream: Four Layers	0.210	0.894	0.708
Single Stream: Five Layers	0.217	0.904	0.688
Combination of One and Two Layers	0.208	0.901	0.707
Combination of One, Two, Three Layers	0.205	0.903	0.717

with two and three streams are 0.208 and 0.205, respectively. These results demonstrate that MS-GCN contributes to improve model performance by combining the features of each stream. Note that the model with a single stream of three layers outperforms other single stream models, indicating that considering more information from multi-hop neighbors at some extent play more important roles in binding affinity prediction.

4. Discussion

In this paper, we proposed a deep learning model that accurately estimate a score of binding affinity for a given pair of a target protein sequence and a drug. We now describe the implications and the limitation of this work.

Importance of local structure of a protein sequence and a target drug for binding affinity prediction

Our evaluation result with two public datasets shows that the proposed model outperforms all the baseline models proposed by the prior studies, which extract and use the global features of the target protein and the drug-like molecule. Such performance improvement may result from the features that the models mainly consider. That is, the prior models adopted the modules like LSTM or multiple 1-D CNN layers to compute the feature vector of a given protein sequence, which can capture macro-level characteristics of a whole protein sequence. Unfortunately, the features of the micro-level subsequences (e.g., 3- or 4-consecutive amino acids) tend to be weaker or lost as more information is compressed into a fixed-length feature vector. As reported by the recent studies (Jiang et al., 2020; Pan et al., 2023), local structures (i.e., subsequences of a protein and sub-graphs of a molecule) are fundamental blocks that construct a whole protein or drug molecule, which may indicate an important characteristics of the protein, and the features extracted from the local structures play a significant role in accurate binding affinity prediction. In line with this, the proposed model preserves the features of the local structures for both a protein and a drug by adopting MS-CNN and MS-GCN modules, respectively. The proposed model outperforms all the other models using a single pipeline of multiple CNN layers or a single LSTM, which demonstrates that using the features of local structures or fundamental blocks of a protein and a drug molecule, which have not been much explored yet, play more significant roles in binding affinity prediction.

Emphasizing local network by multiple streams

The experiment results with changes of the streams in MS-CNN and MS-GCN demonstrate that using multiple streams with different numbers of CNN and GCN layers can improve the performance of binding affinity prediction. A possible reason is in line with the prior studies (Hamilton et al., 2017), which pointed out that outputs passed through multiple layers may lose the information of the input itself. To consider further local features without losing information of input, the MS-CNN and MS-GCN of the proposed model broadcasts the input embeddings to three different streams consisting of different numbers of CNN or GCN layers, which not only preserves the output of a single layer, but also computes the more comprehensive features with the more numbers of the layers. Note that the improved performance of the model with multiple streams verifies that such preserving method helps to estimate a binding affinity score.

Limitation

Despite the valuable insights described above, there are a few limitations in this paper. First, the proposed model was evaluated with two public datasets, Davis and KIBA. Therefore, we cautiously note that (i) the results of the experiments can be affected by the biases in these datasets and (ii) the evaluation is limited to the family of the kinases in these datasets so that the results for a wider set of proteins can be different. However, since the proposed model shows the robust performance in two public datasets, we expect that the tendency will be preserved in other data sources. Second, the proposed model is more complex than the prior studies since two components of the proposed model computes the embeddings of multiple streamlines while the prior models rely on a single streamline. Thus, the proposed model will spend more time and resources for binding affinity prediction and the parallel processing techniques may improve the efficiency from a complexity perspective. Third, although we revealed that capturing the features from local structure of a protein sequence and a target drug is important for binding affinity prediction, further investigations on how to ensemble local and global features and how much the performance can be improved by such ensemble are still remained, which will be explored in future work.

5. Conclusion

In this paper, we proposed a deep learning model to estimate a score of binding affinity for a given pair of a target protein sequence and a drug. With an assumption that local information of both the protein sequence and the drug is important for binding affinity prediction, the proposed model employs MS-CNN and MS-GCN, each of which can use both global and local features without losing local information of subsequences of a target protein sequence and subgraph of a drug, respectively. Based on the evaluation with two popular datasets, Davis and KIBA, we demonstrate that the proposed model outperforms all the baseline models depending on global features, which implies that capturing comprehensive local features can play significant roles of binding affinity prediction. Considering that all the prior studies focused on computing global features of the protein–drug pairs, we expect that the findings and the implications in this paper may provide the researchers of developing machine learning models in drug discovery valuable insights of re-considering local features of the drug and the protein. Despite the contributions, the further investigations on (i) how to fuse local features or (ii) what the detailed role of local features in binding affinity prediction definitely may enhance the proposed models, which will be conducted as future work.

CRediT authorship contribution statement

Daejin Choi: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Sangjun Park:** Writing – review & editing, Validation, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Incheon National University Research Grant in 2023.

Data availability

The dataset and all the implemented source codes that support the findings of this study are available at https://github.com/Koreaj9u7n/LLF.

References

Agarap, A.F., 2018. Deep learning using rectified linear units (relu). CoRR abs/1803.08375, URL: http://arxiv.org/abs/1803.08375, arXiv:1803.08375.

Cichonska, A., Pahikkala, T., Szedmak, S., Julkunen, H., Airola, A., Heinonen, M., Aittokallio, T., Rousu, J., 2018. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics 34, i509–i518.

Cichonska, A., Ravikumar, B., Parri, E., Timonen, S., Pahikkala, T., Airola, A., Wennerberg, K., Rousu, J., Aittokallio, T., 2017. Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. PLoS Comput. Biol. 13, e1005678.

Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Zarrinkar, P.P., 2011. Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046–1051.

Hamilton, W.L., Ying, R., Leskovec, J., 2017. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 1025–1035, URL: https://dl.acm.org/doi/10.5555/3294771.3294869.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., Ester, M., 2017. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Cheminform. 9, 24.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., Wei, Z., 2020. Drug–target affinity prediction using graph neural network and contact maps. RSC Adv. 10, 20701–20712, URL: https://api.semanticscholar.org/CorpusID:219739173.

Landrum, G., 2016. Rdkit: Open-source cheminformatics software. URL: https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

Loshchilov, I., Hutter, F., 2017. Fixing weight decay regularization in adam. CoRR abs/1711.05101, URL: http://arxiv.org/abs/1711.05101.

Ma, W., Qiu, Z., Song, J., Cheng, Q., Ma, C., 2017. Deepgs: Predicting phenotypes from genotypes using deep learning. bioRxiv http://dx.doi.org/10.1101/241414, URL: https://www.biorxiv.org/content/early/2017/12/31/241414, arXiv:https://www.biorxiv.org/content/early/2017/12/31/241414.full.pdf.

Mukherjee, S., Ghosh, M., Basuchowdhuri, P., 2022. Deepglstm: Deep graph convolutional network and LSTM based approach for predicting drug-target binding affinity. In: SDM. SIAM, pp. 729–737.

Nguyen, T., Le, H., Venkatesh, S., 2019. Graphdta: prediction of drug–target binding affinity using graph convolutional networks. bioRxiv http://dx.doi.org/10.1101/684662, URL: https://www.biorxiv.org/content/early/2019/07/02/684662. arXiv:https://www.biorxiv.org/content/early/2019/07/02/684662.full.pdf.

Öztürk, H., Özgür, A., Ozkirimli, E., 2018. Deepdta: deep drug–target binding affinity prediction. Bioinformatics 34, i821–i829. http://dx.doi.org/10.1093/bioinformatics/bty593, URL: arXiv:https://academic.oup.com/bioinformatics/article-pdf/34/17/i821/50582374/bioinformatics_34_17_i821.pdf.

Öztürk, H., Ozkirimli, E., Özgür, A., 2019. Widedta: prediction of drug-target binding affinity. arXiv:1902.04166.

Pan, S., Xia, L., Xu, L., Li, Z., 2023. Submdta: drug target affinity prediction based on substructure extraction and multi-scale features. BMC Bioinformatics 24, http://dx.doi.org/10.1186/s12859-023-05460-4.

Shin, B., Park, S., Kang, K., Ho, J.C., 2019. Self-attention based molecule representation for predicting drug-target interaction. In: Machine Learning for Healthcare Conference. PMLR, pp. 230–248.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., Aittokallio, T., 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J. Chem. Inf. Model. 54, 735–743. http://dx.doi.org/10.1021/ci400709d.

Tsubaki, M., Tomii, K., Sese, J., 2018. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35, 309–318. http://dx.doi.org/10.1093/bioinformatics/bty535, URL: arXiv:https://academic.oup.com/bioinformatics/article-pdf/35/2/309/48963189/bioinformatics_35_2_309.pdf.

Zhao, Q., Duan, G., Yang, M., Cheng, Z., Li, Y., Wang, J., 2023. Attentiondta: Drug-target binding affinity prediction by sequence-based deep learning with attention mechanism. IEEE/ACM Trans. Comput. Biol. Bioinform. 20, 852–863.

Zhao, L., Wang, J., Pang, L., Liu, Y., Zhang, J., 2020. Gansdta: Predicting drug-target binding affinity using gans. Front. Genet. 10, http://dx.doi.org/10.3389/fgene.2019.01243, URL: https://www.frontiersin.org/articles/10.3389/fgene.2019.01243.