SCIENCE CHINA
Information Sciences

• LETTER •

February 2024, Vol. 67, Iss. 2, 129101:1–129101:2
https://doi.org/10.1007/s11432-022-3793-7

Integrating sequence and graph information for enhanced drug-target affinity prediction

Haohuai HE†, Guanxing CHEN† & Calvin Yu-Chian CHEN*

Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107*, China*

Received 1 December 2022/Revised 17 May 2023/Accepted 1 June 2023/Published online 25 January 2024

Drug discovery is a pivotal discipline involving the identification and development of innovative pharmaceuticals designed to combat a wide array of illnesses. It plays a crucial role in advancing global health outcomes and improving the standard of living for individuals worldwide. However, drug discovery often relies on laborious in vitro experiments, which consume substantial time and human resources [1]. Consequently, large-scale prediction of drug-target affinity (DTA) becomes paramount in this field [2]. Fortunately, machine learning techniques have expedited the process of large-scale DTA prediction and provided highly precise forecasts, revolutionizing the landscape of drug discovery [3].

In recent years, deep learning methods have emerged as the preferred approach for constructing DTA prediction model frameworks. These methods fall broadly into two categories: sequence-based and graph-based. The sequence-based method treats the simplified molecular input line entry system (SMILES) of drugs and the amino acid sequences of proteins as languages, employing natural language processing techniques to extract drug and protein features. For example, FusionDTA [4] is a sequence-based approach. Conversely, the graph-based method, exemplified by MgraphDTA [5], considers the structural characteristics of drugs and proteins, viewing the atoms and amino acids of drugs and proteins as nodes in a graph. The edges connecting these nodes represent chemical bonds, and the graph neural network (GNN) technique is then employed to extract the features of drug-protein (DP) pairs.

Despite their effectiveness, these methods have their limitations. The sequence-based method does not account for the structural attributes of the drug and protein, whereas the graph-based method overlooks the significance of sequence information in feature extraction. Furthermore, it is infeasible to construct DP pairs in a single graph as the protein and drug molecule graphs exist at different hierarchical levels.

Methods. To overcome these limitations, we propose a novel model that combines the strengths of both sequence-based and graph-based methods while mitigating their shortcomings. As shown in Figure 1(a), our method uses a sequence-based model for feature extraction and a graph-based method to represent the internal structure of DP pairs. We tokenize and extract features from the SMILES of the drug and the amino acid sequence of the protein using tokenizers and long short-term memory networks (LSTMs). We then construct separate graph structures for the drug and protein and connect them using a virtual node to create a hybrid graph. Node features are derived from the sequence-based model. The hybrid graph is then passed into a GNN for DTA prediction. The details of our method are provided in Appendix A.

Results. We validated our model using the Davis and KIBA datasets. Our evaluation metrics were the mean squared error (MSE), concordance index (CI), and regression toward the mean index (r_m^2). Detailed information on these datasets and metrics can be found in Appendix C.

As depicted in Figure 1(b), our experimental results show that our model outperforms in all three metrics for both datasets. This suggests that combining graph-based and sequence-based methods is not only feasible but also outperforms the state-of-the-art method in two categories.

Moreover, we developed a 3D graph model based on GraphDTA. We hypothesized that using 3D structures of drugs and proteins could improve the accuracy of prediction. However, our experiments on the Davis dataset showed that this 3D GNN-based model did not outperform GraphDTA with a 2D GNN (Table E1). In fact, the 3D GNN-based approach performed worse than our model, which combines sequence-based and graph-based methods. This might be due to the 3D GNN's inability to effectively simulate the DTA binding pocket and model chemical bonds. While 3D GNNs have their advantages in other tasks, they may not be the optimal approach for DTA prediction.

Figure 1(c) presents the results of an error analysis performed on subsets of different drugs and protein sizes in the Davis and KIBA datasets to analyze factors affecting the prediction accuracy of our model. We observed an increasing trend in MSE with longer SMILES lengths and protein sequence lengths under 1500. However, intervals exceeding 1500 exhibited random MSE, likely due to limited data. Our error analysis underscores the challenge of handling larger inputs in DTA models, which require more message passing for feature interactions in GNNs. Detailed information on this analysis can be found in Appendix D.

To verify the interpretability of our model, we examined known biological findings such as the catalytic triad

* Corresponding author (email: chenyuchian@mail.sysu.edu.cn)
† He H H and Chen G X have the same contribution to this work.

© Science China Press 2024

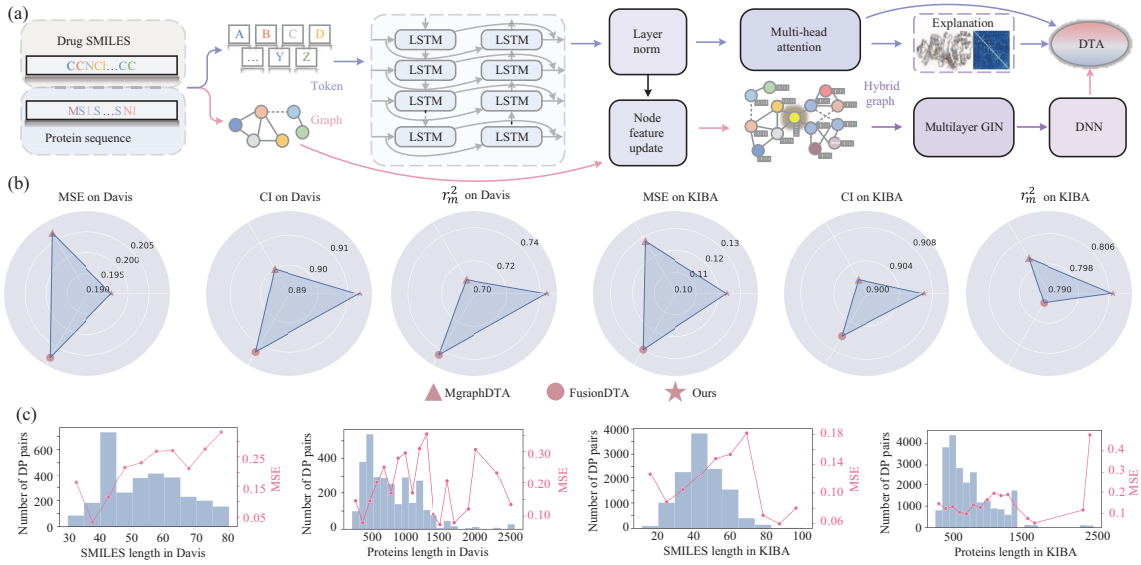

Figure 1 (Color online) Overview of our method and its performance evaluation. (a) The framework of our method: utilization of sequence and structural information of drugs and proteins to enhance the predictive accuracy of DTA; (b) comparative analysis: performance comparison of our method with other models; (c) error analysis results: a depiction of error trends in our model's predictions.

in hydrolase and transferase proteins. Specifically, we investigated examples of Hydrolase and Transfer (PDB ID: 4EY7 and 1ATP) proteins by inputting them along with their respective ligands into our model to explore the attention weight ranking of trimers. Residues His447, Glu334, and Ser203 of 4EY7 were ranked 33, 2, and 245 respectively out of a total of 530, while residues Asp166, Lys168, and Glu91 of 1ATP were ranked 5, 17, and 8 respectively out of a total of 336. Except for Ser203, all other residues were within the top 10%. The coverage rate is 83%. These cases demonstrate the model's ability to capture key residues.

Discussion. In conclusion, our study analyzed the classification and inherent problems of deep learning-based methods for DTA prediction. While both graph-based and sequence-based methods have their limitations, they also have distinct advantages. Consequently, we proposed a novel DTA prediction method that integrates the benefits of both methods to avoid their respective shortcomings. Our approach demonstrated superior performance compared to previous methods on both the Davis and KIBA datasets. Furthermore, we investigated the impact of drug and protein size on prediction accuracy in DTA tasks and examined the performance of 3D graph methods for DTA tasks. Finally, we validated the interpretability of our method through the analysis of trimeric structures.

There is still ample room for improvement in the current methods. For instance, there are many ways to combine protein and drug graphs, and large-scale pre-training models can be used for feature extraction. However, this study has demonstrated the feasibility of using graph-based methods to extract the structural information of DP pairs and using sequence-based methods to extract features from both. In the future, we aim to further explore the details of combining these two methods to improve the accuracy and interpretability of DTA prediction. We will also investigate the use of curriculum learning to mitigate the impact of protein and drug sizes on accuracy.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62176272), Research and Development Program of Guangzhou Science and Technology Bureau (Grant No. 2023B01J1016), Key-Area Research and Development Program of Guangdong Province (Grant No. 2020B1111100001).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Smietana K, Siatkowski M, Møller M. Trends in clinical success rates. Nat Rev Drug Discov, 2016, 15: 379–380
2 Zheng S J, Li Y J, Chen S, et al. Predicting drug-protein interaction using quasi-visual question answering system. Nat Mach Intell, 2020, 2: 134–140
3 Pandey M, Fernandez M, Gentile F, et al. The transformational role of GPU computing and deep learning in drug discovery. Nat Mach Intell, 2022, 4: 211–221
4 Yuan W N, Chen G X, Chen C Y C. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. Briefings BioInf, 2022, 23: bbab506
5 Yang Z D, Zhong W H, Zhao L, et al. MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction. Chem Sci, 2022, 13: 816–833