Interpretable multi-view attention network for drug-drug interaction prediction

Xuan Lin[†,‡], Qi Wen[†], Sijie Yang[†], Zu-Guo Yu[‡], Yahui Long[#,*] and Xiangxiang Zeng[⊤]
[†] College of Computer Science, Xiangtan University, China
[‡] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, China
[#] Singapore Immunology Network (SIgN), Agency for Science, Technology and Research(A*STAR), Singapore
[⊤] College of Information Science and Engineering, Hunan University, China
[*]Corresponding authors
long_yahui@immunol.a-star.edu.sg

***Abstract*—Drug-drug interaction (DDI) plays an increasingly crucial role in drug discovery. Predicting potential DDI is also essential for clinical research. Given the high cost and risk of wet-lab experiments, *in-silico* DDI prediction is an alternative choice. Recently, deep learning methods have been developed for DDI prediction. However, most of existing methods focus on feature extraction from either molecular SMILES sequences or drug interactive networks, ignoring the valuable complementary information that can be derived from these two views. In this paper, we propose a novel interpretable Multi-View Attention network (MVA-DDI) for DDI prediction. MVA-DDI can effectively extracts drug representations from different perspectives to improve DDI prediction. Specifically, for a given drug, we design a transformer-based encoder and a graph convolutional network-based encoder to learn sequence and graph representations from SMILES sequence and molecular graph, respectively. To fully exploit the complementary information between the sequence and molecular views, an attention mechanism is further adopted to adaptively aggregate the sequence and graph representations by taking the importance of different views into accounts, generating the final drug representations. Comparison experiments demonstrated that our MVA-DDI[1] model achieved superior performance to state-of-the-art models on DDI prediction.**

***Index Terms*—Multi-view learning, Contrastive learning, Interpretable attention network, Drug-drug interaction prediction**

I. Introduction

Drug-drug interaction (DDI) refers to the situation where the administration of one drug affects another or multiple drugs in the human body. Such interactions can be synergistic or they can produce completely new effects [1]. Because few diseases can be cured by a single drug, combination therapy with multiple drugs are more effective than monotherapy for most diseases, which indirectly lead to the emergence of DDI with side effect. Therefore, DDI prediction is of vital importance in drug discovery and clinical research.

Traditionally, DDI prediction is performed by extensive biological or pharmacological trials, which is time-consuming and labor-intensive. With the availability of large-scale biomedical datasets and advancements in artificial intelligence technology, deep learning models have been an emerging paradigm for a wide range of cheminformatics and bioinformatics fields [2]–[4]. Meanwhile, there has been a surge in deep learning methods in DDI prediction, it can serve as a low-cost but effective alternative to predict potential DDIs by extracting drug features from various data formats. Previous methods often concentrate on obtaining the similarity features between drug SMILES and other attribute profiles [5]. With the broad success of graph neural networks (GNNs) [6], there has been a trend toward extracting the topological features of molecular graphs or substructures in recent years [7]. One line along this trend, researchers propose to perform DDI prediction based on chemical structure data of drugs such as CASTER [8]. On the other hand, in KGNN [9], Lin *et al*. integrated the idea of graph convolutional network (GCN) to extract both high-order structures and semantic relations of knowledge graphs.

In most existing methods, researchers propose a single network to learn molecular representations from a single perspective, either similarity-based or graph (network)-based methods [10], [11]. However, these methods will suffer from the issue of inadequate feature encoding due to the single-view learning under different types of DDI tasks. The challenge is that multiple perspectives may interfere with each other [12], especially when the learned features from different views of the same drug can complement each other.

To alleviate the challenge, we propose an interpretable multi-view attention network for DDI prediction (MVA-DDI). In the proposed model, we use SMILES sequences as the single input while taking into account both sequence and molecular graph from the same drug molecule. Specifically, we first adopt the sequence encoder (i.e., Transformer) with Explainable Substructure Partition Fingerprint (ESPF) [13] encoding by fully considering the substructures between drug SMILES. Second, we use the graph encoder (i.e., GCN) to extract the structural features by learning the molecular graph transformed from SMILES via RDKit. Meanwhile, we use self-supervised learning to obtain more informative initial drug features by designing the contrastive loss function between positive and negative samples of molecular graphs. After that, we integrate the sequence and graph features for each drug to obtain the final feature using an interpretable self-attention mechanism, which is used to predict the potential interaction probability of drug pairs via decoder. We evaluate our

[1]https://github.com/Luminous-wq/MVA-DDI

proposed model on real-world dataset, and the experimental results demonstrate that our MVA-DDI outperforms baseline methods.

The main contributions of this paper are summarized as the following:

(i) We propose a novel multi-view attention model named MVA-DDI for predicting DDI, which is conducive to learning high-quality features by taking into account SMILES sequence and molecular structural graph.
(ii) MVA-DDI designs an interpretable attention network to adaptively combine with the representations learned from different views, which contributes to improving the prediction performance of DDI.
(iii) Experimental results on real-world datasets demonstrate the superiority of MVA-DDI to state-of-the-art methods.

II. Related works

A. Single-view learning

Graph embedding-based method Complex data structures, such as the heterogeneous network, is a naturally high-dimensional space and mining effective information from such intricate structures is regarded as an alternative for improving the model performance [10]. To address this challenge, graph embedding-based methods have been proposed to adopt popular network embedding algorithms to capture the underlying structure of the network and derive potentially effective network-based features. These methods can be roughly categorized into matrix factorization-based [14], random walk-based [15] and neural network-based methods [16]. However, this line of work only focuses on the connection between nodes but ignores the node attributes and the types of edges.

Knowledge graph-based method To overcome the lack of knowledge brought by graph embedding-based methods, knowledge graph (KG)-based methods have been gaining increasing attention owing to its powerful expression capabilities for heterogeneous data. KG-DDI [17] is the first specialized for DDI prediction that embeds the nodes in the constructed KG using various embedding approaches. Recently, with the popularity of graph neural network, some novel KG methods integrate the idea of graph convolutional network (GCN) to extract both high-order structures and the neighborhood relations of KG, such as KGNN [9]. These methods achieve promising results while they are easy to ignore the structural information of the drug molecule.

Molecular graph-based method Drug molecules can be naturally encoded by a graph with atoms as nodes and chemical bonds as edges. The emergence of GNN has sparked exploration into the molecular graph representation in DDI prediction. In particular, Graph Convolutional Networks (GCN), specifically designed for graph-structured data, have been widely employed for spatial feature extraction in drug research. More recently, identifying key substructures that contribute most to the DDI prediction is a challenge for GNNs. Substructure-based GNNs are designed to capture important substructures based on the chemical functional groups in molecules, such as CASTER [8]. Molecular graph-based methods show promising performance on various datasets. However, these methods only consider molecular graphs or substructures as fixed size and therefore they use GNNs with predetermined node features to capture structural information.

B. Multi-view learning

Different from the aforementioned methods that model DDI tasks in a single perspective, multi-view learning methods are proposed to combine with two or multiple perspectives in an efficient pattern, such as GoGNN [18] and MUFFIN [19]. In recent years, contrastive learning has been successfully applied in gene regulatory interactions [20] and drug-target interactions [21], and few of them have been applied to DDI prediction. For instance, MIRACLE [22] combines GCN with contrastive self-supervision by treating the DDI network as a multi-view graph. AMDE [12] adopts MPNN [23] and Transformer [24] to learn both graph and sequence features of drugs, respectively. Different from these multi-view learning models, MVA-DDI introduces a novel attention network that can effectively leverage different perspectives of drugs by both adopting sequence and graph encoder to improve performance.

Among the research on deep learning for DDI prediction, ADME [12] is the most relevant to our work. Compared with ADME, the proposed framework adopts the pre-training strategy on large-scale datasets to obtain more generalized node embedding of GNN by contrastive learning, while the existing methods typically initialize or customize the node embedding, which lacks good generalization ability. Moreover, our work also suggests an effective self-attention mechanism to fully consider the weights of sequence and graph features, since different dimensions of features obtained from multi-perspective have varying impacts on DDI prediction.

III. Method

In this section, we present the technical details of the proposed MVA-DDI. Here we define the DDI prediction as a binary classification task. Specifically, given a drug-drug pair d_x and d_y ($d_x, d_y \in D$) with D denoting the set of known DDI, we aim to learn a prediction function $f : D \times D \times I \rightarrow \{0, 1\}$ to infer the probability of the pair to be a real drug-drug interaction.

A. Overview

The overall framework of the proposed MVA-DDI is shown in Figure 1. First, with the drug SMILES sequence as input, we design two feature extraction modules to extract the sequence and graph features from the SMILES sequence and the molecular graph, respectively. Specifically, the sub-sequences are generated by ESPF and fed into *Transformer-based sequence encoder* to obtain sequence representation. Meanwhile, taking molecular graph obtained by converting SMILES sequence with RDKit as input, we utilize a *GCN-based graph encoder* to derive graph representation. To learn more informative representation, a *self-supervised pre-training* strategy is introduced. Second, a *self-attentive feature aggregation* module is

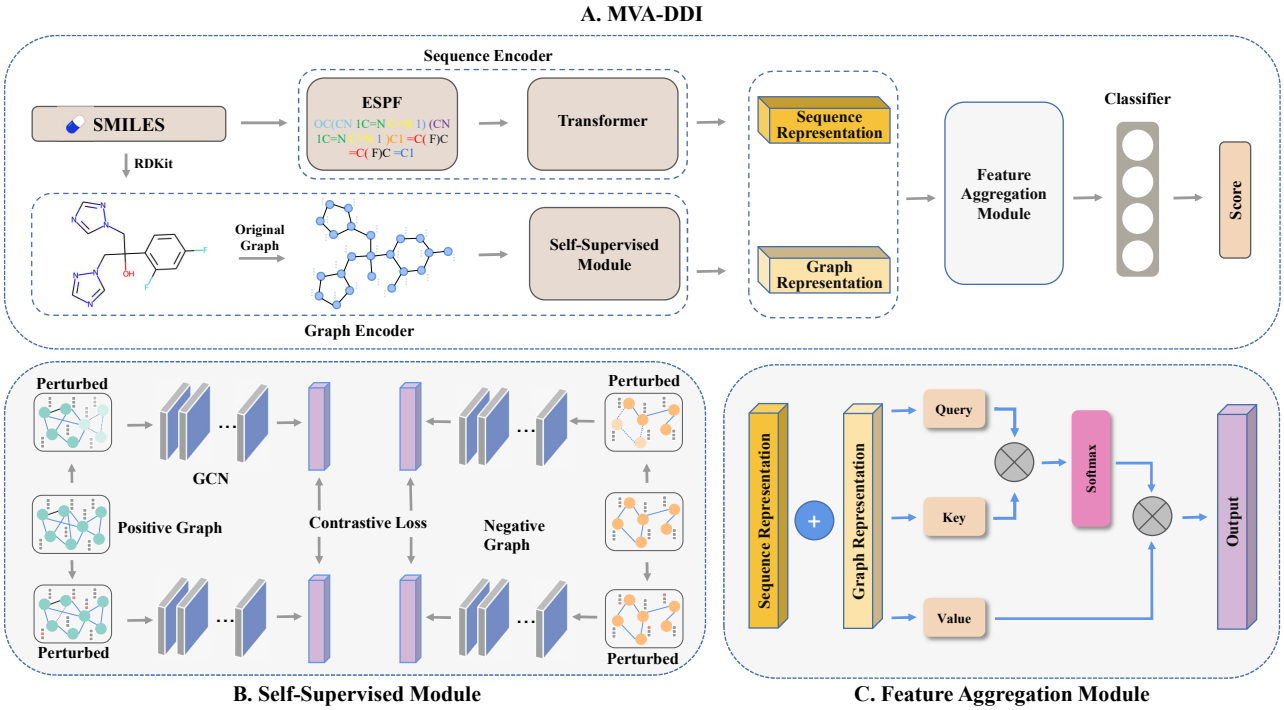

Fig. 1. The overall architecture of the proposed MVA-DDI model. A) Overview of MVA-DDI. B) Self-supervised module. C) Feature aggregation module.

designed to obtain the integrated drug representation. Finally, with the integrated drug representation as input, the classifier outputs the interaction probability of DDI.

B. Transformer-based sequence encoder

We design a transformer-based sequence encoder to extract the chemical context of SMILES sequences. Transformer is an appropriate choice as it is widely used for sequence encoding in natural language processing [24]. Specifically, we employ the ESPF algorithm [13] to facilitate the process of sequence encoding. This algorithm can break down the SMILES sequences into smaller sub-sequences or atomic symbols. These substructures are then mapped to corresponding embedding vectors through predefined dictionary. Thus we obtain the word embedding WE and position embedding PE as the input of the Transformer model.

Generally, the input of Transformer is the word embedding WE, and the position embedding PE of SMILES sequence. Different from the traditional Transformer that takes these two vectors separately, we adopt an *Embedding Layer* to generate a vector X with the same dimension size by adding WE and PE together, and then feed it into the Transformer, which primarily relies on multi-head self-attention and feed-forward neural network. The self-attention mechanism is a core component of the Transformer model. Specifically, it can be formulated as follows.

$$(Q, K, V) = X * (W_Q, W_K, W_V) \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^\top}{\sqrt{d_K}})V \tag{2}$$

where Q, K, and V represent the query, key, and value vectors, respectively, W represents the corresponding weights that are initialized by neural network, and d_K represents a dimension of feature vector K. We obtain the contextual representation x for each position by performing self-attention calculations of input sequence. The feed-forward neural network consists of two fully connected layers and an activation function, which are used to perform nonlinear transformations and to map the representations of each position. Finally, the sequence feature F_s is obtained as follows.

$$F_s = softmax(ReLU(xW_1 + b_1)W_2 + b_2) \tag{3}$$

C. GCN-based graph encoder

We design a GCN-based graph encoder to extract the structural features of molecular graphs. Here we use graphs to model the relationships between atoms and their chemical bonds. GCN can effectively encode the structural information of a graph. Let's denote the feature and adjacency matrices as $X \in R^{N\times d}$ and $A \in R^{N\times N}$, respectively, where d and N represent the dimension of feature and the number of nodes. Each entry $A_{i,j}$ indicates whether there is a connection between node i and j. Taking the molecular graph as input, the encoder outputs graph representation by smoothing features of neighbors. Specifically, as shown in Figure 1, we first use RDKit to convert SMILES sequence into a 2-dimension structure of molecular graph where nodes and edges denote atoms and relationships between them. A feature vector for each atomic is defined, as shown in Table I.

The specific propagation of GCN is as follows:

TABLE I
THE LIST OF PREDEFINED ATOM FEATURES.

Atom Feature	Size	Description
Atomic symbol	44	[C, N, O, S, F, Si, P, Cl, Br, Mg, Na, Ca, Fe, As, Al, I, B, V, K, Tl, Yb, Sb, Sn, Ag, Pd, Co, Se, Ti, Zn, H, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb, Unknow] (One-hot)
Atomic degrees	11	Degree of atoms in a drug [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10] (One-hot)
Implicit value	7	Implicit valence of atoms [0, 1, 2, 3, 4, 5, 6]
Formal charge	1	The formal charge of the atom, which usually ranges from -3 to +3
Radical electrons	1	The number of free radical electrons of an atom, which usually ranges from 0 to 2
Hybridization	5	The atomic hybridization mode [SP, SP2, SP3, SP3D, SP3D2] (One-hot)
Atomic aromaticity	1	Whether the atom is aromatic or not [0/1]
Total hydrogen atoms	5	Total number of hydrogen atoms in the atom [0, 1, 2, 3, 4] (One-hot)

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{4}$$

where $H^{(l)}$ represents the feature matrix at the $l-th$ layer, and we set H^0 to X. To incorporate with the self-features of all nodes in A, we have $\hat{A} = A+I$, and I is the identity matrix, D is the degree matrix of A, and $\tilde{D} = \sum \tilde{A}_{i,j}$. σ is a non-linear activation function such as $ReLU$. In the last layer, $softmax$ is used for classification prediction. $W^{(l)}$ is the trainable parameter matrix for the convolution transformation of the current layer. The convolution in GCN is computed based on the product of the adjacency matrix and the feature vector. This operation allows each node to propagate information with its neighboring nodes, which results in updating the feature representation of each node. Finally, we obtain the graph feature F_g vectors from the last layer as follows.

$$F_g = softmax(H) \tag{5}$$

D. Self-supervised pre-training

The key contribution of this paper lies in the design of self-supervised pre-training module. It consists of data preprocessing, data argumentation, and contrastive loss function.

Data preprocessing. To fully consider the property of drug-like in pre-training, we extracted 4,194 FDA-approved small molecule drugs and 5,805 candidate molecules in preclinical phases from the ChEMBL (Version 32) database. Meanwhile, we downloaded 1,704 drugs from the DrugBank database and obtained a total of 11,703 small-molecule drugs. Then we removed small molecules without SMILES, duplicated, and that could not be parsed by RDKit, and finally obtained a total of 8,722 small molecule drugs for pre-training.

Data argumentation. To facilitate contrastive learning, it is necessary to obtain more samples. The data argumentation method is designed to generate the corresponding positive and negative samples of a given SMILES. The method is based on two graph-specific operations, including random node dropout and edge modification, which result in random perturbations to the graph features and assist the model in learning additional information. Specifically, the positive sample pair is generated by implementing five random node dropouts and five random edge modifications to the current SMILES. The specific operation is to randomly select five columns of data in the node features matrix and set them to 0, and randomly select five pairs of data in the adjacency matrix for value (i.e., 0 or 1) exchange (similar to the deletion and increase of edges). Then the positive sample pair is formed with oneself and labeled as 1. The negative sample pair is obtained by randomly selecting a sample that is different from the current SMILES to perform five random node dropouts and five random edge modifications, then the negative sample pair is formed and labeled as 0, The above operation of the negative sample is repeated 5 times.

Contrastive loss function. The GCN-based contrastive self-supervised module is designed to enhance the discriminative representation of drug features. The proposed model is optimized by using a contrastive loss. For a pair of samples (d_i, d_j), their assigned label is denoted by y_{ij} ($y_{ij} \in \{0, 1\}$), where $y_{ij} = 0$ indicates they belong to a negative sample pair, otherwise the opposite is positive. We denote their corresponding output vectors as o_i and o_j, respectively, and the contrastive loss between them is defined as follows:

$$L_{ij} = \frac{1}{2}\left((1-y_{ij})d(o_i,o_j)^2 + y_{ij}\max(0, m - d(o_i,o_j))^2\right) \tag{6}$$

where $d(oi, oj)$ represents the distance between the two vectors, and we have $d(oi, oj) = \|o_i - o_j\|$. m is a margin parameter that controls the distance between sample pairs. The first and second terms correspond to the loss for negative and positive sample pairs, respectively. The final contrastive loss function is averagely obtained by summing up the loss for each sample pair as follows:

$$L_{cl} = \frac{1}{N}\sum_{i=1}^{N} L_{cl}^i \tag{7}$$

where N represents the number of sample pairs, and we have $L_{cl}^i = L_{ij}$. The entire propagation process of contrastive self-supervised pre-training model is as follows. In the process of forward propagation, the inputs are the output vectors o_i and o_j, along with their labels y_{ij}. The forward function first calculates the L_2 distance between the two vectors, referred to as euclidean distance. Then it computes the contrastive loss based on the formula. Finally, the contrastive loss is returned as the feedback signal for backpropagation to optimize the network parameters.

E. Feature aggregation

To obtain the final embedding of drug representation, the effective feature aggregation module is employed to combine the sequence and graph feature vectors into a more expressive feature vector. For the sequence features F_s and graph features F_g, three feature aggregation methods are designed as follows.

Sum. This is the simple but effective method to obtain the final representation of drug features. The operation is implemented by element-wise adding up two types of features. The computation is obtained by:

$$Sum(F_s, F_g) = F_s + F_g \tag{8}$$

Cat. The concatenate operation, abbreviated as cat, is implemented by splicing multiple features along a certain axis. This method preserves all the information from the individual features, the increase in dimensions of feature vector may affect model complexity.

$$Cat(F_s, F_g) = [F_s, F_g] \tag{9}$$

Attention. The attention operation is a method based on weighted averaging. It involves assigning weights to different feature vectors based on their importance and then combining them through a weighted sum to obtain the final feature vector. This approach allows for dynamic adjustment of the weights W assigned to each feature vector, which can enable a more refined fusion effect by emphasizing or de-emphasizing specific features based on their relevance.

$$Attention(F_s, F_g) = W_1 F_s + W_2 F_g \tag{10}$$

F. Drug-drug interaction prediction

Given a DDI tuple (dx, dy, r), the DDI prediction can be expressed as the joint probability as follows:

$$P(d_x, d_y, r) = \sigma(W_{xy}[(F_s^x \odot F_g^x), (F_s^y \odot F_g^y)] \cdot U_r) \tag{11}$$

where σ is the sigmoid function, W_{xy} and U_r are learnable representations of interaction type r, and $\odot$ represents the *Attention* feature aggregation used in this work. The learning process of the proposed MVA-DDI model can be achieved by minimizing the cross-entropy loss function as follows.

$$L_c = -\sum P log\hat{P} - \lambda(1-P)log(1-\hat{P}) \tag{12}$$

Backpropagation propagates from the output layer to each preceding layer. The end-to-end approach is used to train all trainable parameters in the model.

IV. Results

In this section, we first introduce the experimental setups and then demonstrate the performance of the proposed model MVA-DDI through comparison with baseline methods.

A. Datasets

The known DDI pairs used in our experiment were downloaded from the DrugBank (v5.1.9) database, where we extracted 124,725 drug-drug pairs between 1,704 approved small molecule drugs. Besides, we collected the SMILES sequences and category information for these 1,704 drugs.

B. Baseline methods

To validate the performance of the model, we compare MVA-DDI with seven baseline methods. Next, we simply introduce each method as follows:

- **DeepWalk [15]** is a graph-based method proposed for representation learning.
- **GraRep [14]** learns low dimensional representation by integrating global structure information of the graph into the learning progress.
- **GAE [16]** is a graph autoencoder model that is proposed to learn low dimensional representation of the graph.
- **DeepDDI [5]** is a deep neural network method developed to predict DDI.
- **CASTER [8]** is a deep learning method that encodes the functional substructures of drugs for DDI prediction.
- **KGNN [9]** proposes a knowledge graph neural network to predict DDI by incorporating prior knowledge graph.
- **AMDE [12]** presents a multi-view deep learning for DDI prediction, which simultaneously models SMILES sequence and atom graph to learn drug representation.

We compare MVA-DDI with baseline methods on the same DrugBank datasets. For fair comparison, all baseline methods use the default parameters.

C. Metrics

We use four well-known measurement metrics to evaluate the performance, i.e., accuracy (ACC), area under ROC curve (AUROC), area under the precision-recall curve (AUPR), and F1 score.

D. Experimental settings

In this work, we conduct standard 5-fold cross-validation (CV) to evaluate the performance of the model. In particular, we randomly split known drug-drug pairs into five groups. For each round, we select in turn one group of drug-drug pairs for model testing, while the remaining four groups are used for model training. During pre-training, the ratio of positive and negative sample pairs is set to 1:5. For the sequence encoder, we set the number of attention heads to 8. The length of sequence features is set to 50. The output dimension in the GCN encoder is 75. We use the Adam algorithm for model optimization. The learning rate is set to 1e-4 and the training epoch is 50.

E. Performance evaluation

In this section, we compare the performance of our MVA-DDI model with baseline methods. Table II shows the comparison results. It can be observed that MVA-DDI consistently outperforms baseline methods in terms of four metrics. Specifically, MVA-DDI achieves an average ACC of 94.49%, an average AUROC of 98.43%, an average AUPR of 98.45%, and an average F1-score of 94.55%, which improves the second-best method AMDE by 2.57%, 1.32%, 1.68%, and 1.17%, respectively. Besides, we can uncover from Table II that multi-view learning models (i.e., AMDE and MVA-DDI) have better performance than single-view models (e.g., DeepDDI, and KGNN), indicating that learning representations from multiple perspectives helps to improve the performance of the model.

While both MVA-DDI and AMDE use molecule sequence and graph information to learn representation, there are two main advantages of MVA-DDI compared to AMDE. First, MVA-DDI is able to learn more informative graph representations than AMDE. MVA-DDI introduces a pre-training mechanism to learn graph features using a large amount of prior drug data instead of the training drug data. Second, when integrating representations from different views, MVA-DDI adopts an attention mechanism that allows the model

TABLE II
PERFORMANCE COMPARISON BETWEEN MVA-DDI AND BASELINES.

Model	ACC	AUROC	AUPR	F1
DeepWalk	83.44±0.07	91.76±0.04	90.64±0.05	83.53±0.08
GraRep	84.45±0.06	92.30±0.14	91.15±0.07	84.64±0.12
GAE	81.35±0.46	88.82±0.32	85.95±0.42	82.53±0.40
DeepDDI	82.81±0.21	88.38±0.53	89.23±0.84	83.53±0.21
CASTER	84.38±0.11	91.57±0.09	91.60±0.16	85.60±0.12
KGNN	89.53±0.16	93.64±0.27	92.86±0.28	90.06±0.11
AMDE	91.92±0.46	97.11±0.29	96.77±0.34	93.38±0.37
MVA-DDI	**94.49±0.44**	**98.43±0.34**	**98.45±0.33**	**94.55±0.39**

to adaptively aggregate view-specific representations, while AMDE utilizes simple sum and concatenation methods.

V. DISCUSSION AND CONCLUSION

In this paper, we propose a novel interpretable multi-view attention network, named MVA-DDI, for DDI prediction. MVA-DDI is a dual-channel representation learning framework that fully exploits both the SMILES sequence and the molecular structure information. In particular, for the SMILES sequence, we design a transformer-based encoder to learn the sequence representation by taking the chemical context of atoms into account. For the molecule graph, we use a self-supervised graph contrastive learning encoder to learn the graph representation. This self-supervised learning encoder allows the model to extract informative graph representations by pre-training on a large number of drug molecules. To combine the representations obtained from both views effectively, an attention mechanism is introduced to adaptively fuse both sequence and graph representations by considering the importance of different views to each drug. Extensive experimental results demonstrated that our MVA-DDI outperformed seven state-of-the-art methods in predicting drug-drug interactions.

VI. ACKNOWLEDGMENTS

The work is supported in part by the National Natural Science Foundation of China (No. 62202413, 62122025, U22A2037, 62250028), the Science and Technology Innovation Program of Hunan Province of China (No. 2022WK2009), the Hunan Provincial Natural Science Foundation of China (No. 2021JJ10020,).

REFERENCES

[1] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, and Nicholas P Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, 9(9):2147–2163, 2014.

[2] Xuan Lin, Lichang Dai Dai, Yafang Zhou, Zu-Guo Yu, Wen Zhang, Jian-Yu Shi, Cao Dong-Sheng, Li Zeng, Haowen Chen, Bosheng Song, Philip S Yu, and Xiangxiang Zeng. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. *Briefings in Bioinformatics*, 2023.

[3] Yizheng Wang, Yixiao Zhai, Yijie Ding, and Quan Zou. Sbsm-pro: Support bio-sequence machine for proteins. *arXiv preprint arXiv:2308.10275*, 2023.

[4] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[5] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115(18):E4304–E4311, 2018.

[6] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31th Advances in Neural Information Processing Systems*, volume 30, pages 1–11, 2017.

[7] Xiaoqin Pan, Xuan Lin, Dongsheng Cao, Xiangxiang Zeng, Philip S Yu, Lifang He, Ruth Nussinov, and Feixiong Cheng. Deep learning for drug repurposing: methods, databases, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(4):e1597, 2022.

[8] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. CASTER: Predicting drug interactions with chemical substructure representation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 702–709, 2020.

[9] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. KGNN: Knowledge graph neural network for drug-drug interaction prediction. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, volume 380, pages 2739–2745, 2020.

[10] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[11] Tengfei Ma, Xuan Lin, Bosheng Song, S Yu Philip, and Xiangxiang Zeng. KG-MTL: Knowledge graph enhanced multi-task learning for molecular interaction. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7068–7081, 2023.

[12] Shanchen Pang, Ying Zhang, Tao Song, Xudong Zhang, Xun Wang, and Alfonso Rodriguez-Patón. AMDE: a novel attention-mechanism-based multidimensional feature encoder for drug–drug interaction prediction. *Briefings in Bioinformatics*, 23(1):bbab545, 2022.

[13] Kexin Huang, Cao Xiao, Lucas Glass, and Jimeng Sun. Explainable substructure partition fingerprint for protein, drug, and more. In *NeurIPS Learning Meaningful Representation of Life Workshop*, 2019.

[14] Shaosheng Cao, Wei Lu, and Qiongkai Xu. GraRep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 891–900, 2015.

[15] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[16] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[17] Md Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network. In *PrProceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 113–123, 2019.

[18] Hanchen Wang, Defu Lian, Ying Zhang, Lu Qin, and Xuemin Lin. GoGNN: Graph of graphs neural network for predicting structured entity interactions. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 1317–1323, 2020.

[19] Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. MUFFIN: multi-scale feature fusion for drug-drug interaction prediction. *Bioinformatics*, 37(17):2651–2658, 2021.

[20] Lujing Zheng, Zhenhuan Liu, Yang Yang, and Hong-Bin Shen. Accurate inference of gene regulatory interactions from spatial gene expression with deep contrastive learning. *Bioinformatics*, 38(3):746–753, 2022.

[21] Yang Li, Guanyu Qiao, Xin Gao, and Guohua Wang. Supervised graph co-contrastive learning for drug-target interaction prediction. *Bioinformatics*, 38(10):2847–2854, 2022.

[22] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the Web Conference*, pages 2921–2933, 2021.

[23] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1263–1272, 2017.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31th Advances in Neural Information Processing Systems*, volume 30, 2017.