Bioinformatics, 38(18), 2022, 4387–4394
https://doi.org/10.1093/bioinformatics/btac538
Advance Access Publication Date: 29 July 2022
Original Paper

Data and text mining

Multidrug representation learning based on pretraining model and molecular graph for drug interaction and combination prediction

Shujie Ren, Liang Yu * and Lin Gao

School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 17, 2022; revised on July 6, 2022; editorial decision on July 26, 2022; accepted on July 27, 2022

Abstract

Motivation: Approaches for the diagnosis and treatment of diseases often adopt the multidrug therapy method because it can increase the efficacy or reduce the toxic side effects of drugs. Using different drugs simultaneously may trigger unexpected pharmacological effects. Therefore, efficient identification of drug interactions is essential for the treatment of complex diseases. Currently proposed calculation methods are often limited by the collection of redundant drug features, a small amount of labeled data and low model generalization capabilities. Meanwhile, there is also a lack of unique methods for multidrug representation learning, which makes it more difficult to take full advantage of the originally scarce data.

Results: Inspired by graph models and pretraining models, we integrated a large amount of unlabeled drug molecular graph information and target information, then designed a pretraining framework, MGP-DR (Molecular Graph Pretraining for Drug Representation), specifically for drug pair representation learning. The model uses self-supervised learning strategies to mine the contextual information within and between drug molecules to predict drug–drug interactions and drug combinations. The results achieved promising performance across multiple metrics compared with other state-of-the-art methods. Our MGP-DR model can be used to provide a reliable candidate set for the combined use of multiple drugs.

Availability and implementation: Code of the model, datasets and results can be downloaded from GitHub (https://github.com/LiangYu-Xidian/MGP-DR).

Contact: lyu@xidian.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

When facing complex diseases, medical staff usually use multiple drugs to act simultaneously because most human diseases are attributed to complex biological processes (Han *et al.*, 2017). Combination therapy refers to the use of two or more drugs to improve the clinical effect and offers many advantages compared with single-agent therapy (Sun *et al.*, 2013). It can manifest the synergistic therapeutic effect of drugs to improve the efficacy, delay or reduce the occurrence of drug resistance or reduce the risk of adverse reactions (Ali *et al.*, 2017). At present, combination therapies have been widely used in the treatment of many complex diseases, such as hypertension (Giles *et al.*, 2014), cancer (Ramzan *et al.*, 2021) and infectious diseases (Zheng *et al.*, 2018). Besides, when a drug is used together with a specific food ingredient, its efficacy may also change significantly. Recent study shows 67% of the elderly in the USA take an average of five or more drugs in their daily lives, including prescription drugs and dietary supplements (Qato *et al.*, 2016). So, even if the patient is extremely self-disciplined, drug interaction seems to be inevitable.

One category of drug interaction prediction methods is traditional dose screening. This method is mainly based on synergy models, such as HSA (Berenbaum, 1989) and Loewe (Loewe, 1953). It works by drawing a dose–response curve to determine the degree and the direction of interactions between drugs, such as SynergyFinder (Ianevski *et al.*, 2017). But because there are certain differences between the theoretical foundations on which the models rely in the early stages of their development, the screening results are quite different (Meyer *et al.*, 2020). So, this type of traditional method is gradually fading from people's field of vision.

Another type of drug interaction prediction model is data-driven calculation methods, which can be divided into three categories:

1. The first type of method, such as the work of Bai *et al.* (2018), using an improved naive Bayes algorithm to predict effective drug combinations is a representative machine learning method. Deep learning-related methods mainly include AuDNNsynergy (Zhang *et al.*, 2021), DEEPDDI (Ryu *et al.*, 2018) and embedding models based on anti-autoencoder and Wasserstein distance (Dai *et al.*, 2021). This type of method is closely related to feature engineering. Data and features determine the upper limit of the learning effect, and models and algorithms only approach this upper limit. Therefore, how to design more efficient features become very worthy of attention.
2. The second type of method aims to construct a large-scale network with drugs as nodes and then uses known network nodes and network structures to predict the possibility of a link between two nodes that do not yet have edges, such as the trigraph information propagation algorithm (Xu *et al.*, 2020) and the graph random walk with neighbor recommendation method (Zhang *et al.*, 2017). However, because the network is constructed in advance, it becomes impossible for the predictor to jump out of the current graph to predict the relationships of new drugs, which greatly increases the limitations of the model.
3. The third type of method is dominated by drug structure. The differences in the chemical structures of individual drugs in drug combinations are significantly related to the synergistic effects of drugs (Liu and Zhao, 2016; Ru *et al.*, 2020). Therefore, to avoid tedious feature collection and redundant feature design, more researchers are focusing on the molecular structure of drugs. Two kinds of representations are generally adopted to describe molecules: text-based representations, such as the Simplified Molecular Input Line Entry Specification (SMILES) (Weininger, 1988) and graph-based representations, such as 2D undirected cyclic graphs. For SMILES strings, related methods include CASTER (Huang *et al.*, 2020) and BERTChem-DDI (Mondal, 2020). For undirected graphs, related methods include drug representation learning based on the Siamese GCN network (Chaudhuri *et al.*, 2019) proposed by Chen *et al.* (2019), GCN-BMP (Chen *et al.*, 2020) and Attentive FP (Xiong *et al.*, 2020), as well as the MG-BERT (Zhang *et al.*, 2021).

Although these proposed methods have achieved good performance, shortcomings remain. To date, among the existing molecular representation methods, there has not been a representation strategy specifically designed for drug molecules. Therefore, to overcome the abovementioned drawbacks, we constructed a multidrug representation learning framework named MGP-DR based on pretraining and molecular graphs. We have designed two strategies for multimolecule graph representation to aggregate information between different drug molecules. Then, two pretraining tasks, Mask Atoms Prediction and S_{AB} Score Prediction, were adopted to mine the hidden structural and target information in the drug pairs. Results illustrate that MGP-DR can generate context-sensitive atomic representations and global representations of drug pairs after pretraining and improve the predictive performance of related downstream tasks on different datasets, outperforming previous state-of-the-art models. Additionally, MGP-DR can also be extended to higher-order drug representation learning to further promote the development of combination therapy.

2 Materials and methods

2.1 Overview of the MGP-DR model

The MGP-DR model was developed to learn multidrug representations and then apply them to multiple drug-related bioinformatics problems. As shown in Figure 1a, when pretraining, a large number of unlabeled drug molecule pairs with target information were used

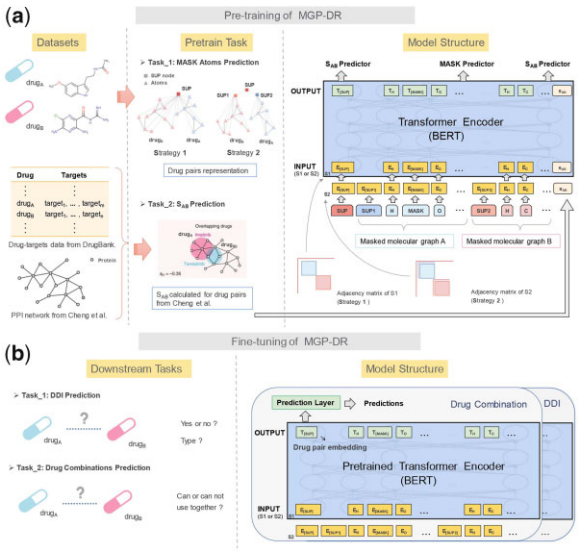

Fig. 1. Overview of the MGP-DR framework. (Top subfigure **a**) Pretraining of MGP-DR. We use the drug structure information and target information to design two training tasks to pretrain our model. The supernode denoted by [SUP] was used to extract the global information of the drug pair. [SUP1] and [SUP2] represent global information of $drug_A$ and $drug_B$, respectively. S_{AB} reflects the overlapping condition of the targets of the drug pair. (Bottom subfigure **b**) Fine-tuning of MGP-DR with the same encoder structure as the model structure in subfigure a. During the fine-tuning stage, the embedding of the supernode was used to perform downstream-related prediction tasks

to mine contextual information in molecules themselves and target overlapping conditions between different drugs through the two pretraining tasks we defined. After learning the potential information, the pretrained transformer encoder in our MGP-DR model was transferred, as shown in Figure 1b. The encoder layers were kept, and the pretraining task-related head was removed. Then, a two-layer fully connected neural network was added to achieve DDI (Drug–Drug Interaction) and drug combination prediction.

2.2 Data resources

In the pretraining stage, we collected the target information and the structure information of 2675 approved drugs from the DrugBank (Wishart *et al.*, 2018). After removing those without SMILE strings or SMILE strings that could not be converted, remaining 2501 drugs. To calculate the overlapping degree of drug targets, we used a human protein–protein interaction network from Cheng *et al.* (2019). Ten two pretraining datasets were designed, named the large-scale dataset and small-scale dataset. Drug pairs in the large-scale set reached 3.1 million, which only contains structural information. In the small-scale set, drug pairs are about 1.3 million, including both target and structure information. To verify the pretraining performance, we randomly reserved 10% of them for evaluation. **In the fine-tuning stage**, for the binary DDI prediction task, we used two datasets, one named BIOSNAP (Zitnik *et al.*, 2018) and the other from Zhang *et al.* (2017). BIOSNAP consists of 1322 approved drugs with 41 520 labeled DDIs. Zhang *et al.* contain 548 drugs and 48 584 pairwise DDIs. For the multi-DDI prediction task, we used the datasets collected by Ryu *et al.* named DEEPDDI. It is composed of 1710 drugs and 86 different interaction types, capturing 192 284 drug–drug pairs as samples. For the drug combination prediction task, the datasets were collected from the DCDB (Liu *et al.*, 2014).

2.3 Pretraining of MGP-DR

2.3.1 Pretraining tasks

The first task is named Mask Atoms Prediction. Molecular graph has a natural graph structure, in which the nodes represent atoms, and the edges represent whether there are chemical bonds between

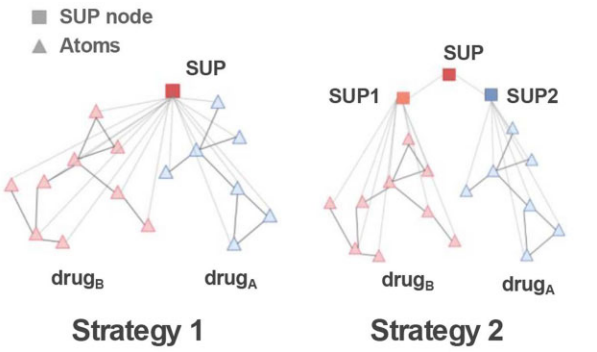

Fig. 2. Two strategies for generate drug pair molecular graph. The super-node denoted by [SUP], represented by square. Atoms are represented by triangle

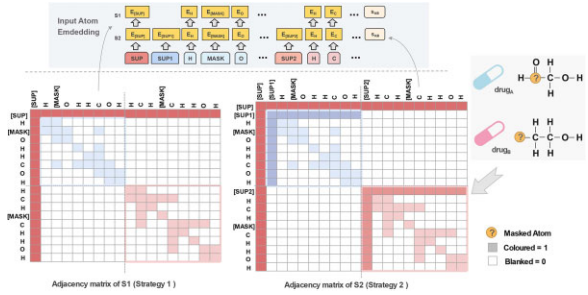

Fig. 3. Input representations of the pretraining stage. The adjacency matrix can be viewed as a block diagonal symmetric matrix. The adjacency matrix of drug A and drug B falls on its diagonal. The colored squares in the matrix indicate that two nodes have connected edges, which is represented as 1, and the blank squares indicate that there are no connected edges between two nodes, which is represented as 0

the atoms. To generate a molecular graph for multiple drugs, we proposed two strategies. As shown in Figure 2, a supernode (SUP) is defined to collect information from all drugs. In Strategy 1, the supernode is connected with each atom in both drugs. In Strategy 2, supernode 1 (SUP1) connects with all atoms in drug$_A$, and supernode 2 (SUP2) connects with all atoms in drug$_B$. After aggregating the information of each drug, SUP1 and SUP2 were aggregated by the node SUP. For the whole drug pair molecular graph, we took advantage of a mask strategy proposed by MG-BERT. Fifteen percent of the atoms will be randomly selected. For a graph with only a few atoms, at least one atom would be masked. Supernodes would not be permitted to be selected. Each selected atom has an 80% probability of being replaced by the [MASK] mark, a 10% probability of being randomly replaced by other atoms and a 10% probability of remaining unchanged.

The second task named S_{AB} Score Prediction. Inspired by the work from Cheng *et al.*, which utilizes a recently introduced separation measure (Menche *et al.*, 2015):

$$s_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \tag{1}$$

to describe the network proximity of drug-target modules A and B. It compares the mean shortest distance between the target proteins within each drug, $\langle d_{AA} \rangle$and $\langle d_{BB} \rangle$, to the mean shortest distance $\langle d_{AB} \rangle$ between A and B target pairs. When $S_{AB} < 0$, the two drug targets overlap, while for $S_{AB} \geq 0$, the two drug targets are topologically separated. It was found that compared to random drug pairs, FDA (Katz, 2004) approved drug combinations have lower S_{AB}, which confirms that S_{AB} can provide a reliable way to measure the drug–drug relationship within the human interactome. We introduce this feature into our pretraining mission and design a regression task to predict the S_{AB} score for each drug pair. The precalculated scores are used as the ground truth to train the model. Before training, we use two different normalization methods to preprocess the S_{AB} value, namely the min-max method and the z-score method:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2}$$

$$X^* = \frac{X - \mu}{\sigma} \tag{3}$$

where X^* represents the normalized result of the X value, $\min(X)$ and $\max(X)$ represent the minimum and maximum values of X, μ represents the mean of all X values and σ represents the variance of all X values.

2.3.2 Input representations

The numbers of each atom type appearing in the molecules are very unbalanced. After statistical analysis of drugs in our dataset, 13 atomic types with the highest frequency were included in the dictionary, and the other rarely encountered atom types were uniformly denoted by [UNK]. The supernode is denoted by the mark [SUP]. The masked tokens selected are denoted by [MASK]. In summary, the dictionary contains the following tokens: [H], [C], [N], [O], [F], [S], [Cl], [P], [Br], [I], [Na], [Fe], [Mg], [UNK], [SUP] and [MASK]. In addition, the degree of overlapping target information between drug pairs is represented by S_{AB}.

Before training, each molecule was converted into a 2D undirected graph by RDKit (Landrum, 2013). The graph is stored by an adjacency matrix and a list of atoms. Then, two atom lists and two adjacency matrices are concatenated as input (Fig. 3). If two atoms have connected edge, the corresponding position of matrix is 1, otherwise is 0.

2.3.3 Model architecture and training process

MGP-DR consists of three components: an embedding layer, several Transformer encoder layers (Vaswani *et al.*, 2017) and a task-related output layer. After being inspired by the ideas of BERT (Devlin *et al.*, 2018) and MG-BERT, we made some modifications to the model to make it more suitable for the potential feature learning of drug pairs. In the embedding layers, each token in our dictionary represents a node from the graph. The order of the drug pairs does not affect the overall topology of the graph, drug pair (A, B) or drug pair (B, A) could be represented by the same adjacency matrix and atom list. In the Transformer encoder layers, every atom token exchanges information with each other through a global attention mechanism, then a local attention mechanism based on chemical bonds is added to the attention score calculation. The adjacency matrix is used to control local information exchange, which makes atoms more related with the nodes that connect with themselves. The supernodes are used to solve long-distance dependence. Task-related output layer is generated by a fully connected neural network to perform specific tasks. For the first pretraining task, we obtain an embedding vector with the same dimension as the dictionary for each token. Then, the cross-entropy loss is calculated only at the masked atoms to optimize the parameters of the model:

$$L_C = -\sum_{i=1}^{m} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{4}$$

L_C stands for the classification loss, where m is the number of samples, y_i is the label (same as the masked atom or not) of sample i and p_i is the probability that sample i is predicted to be in the positive (same). For the second pretraining task, we convert the supernode embedding vector from encoder layers into one dimension through the output layer and calculate the mean square error loss with the precalculated S_{AB} score:

$$L_R = \frac{1}{m} \sum_{i=1}^{m} (y_i - y'_i)^2 \tag{5}$$

L_R stands for the regression loss, where m is the number of samples, y_i is the label (precalculated S_{AB} score) of sample i and

y'_i is the output predicted score. The overall loss of the MGP-DR model is:

$$L_{TOTAL} = L_C + L_R \quad (6)$$

2.4 Fine-tuning of MGP-DR

The architecture of the MGP-DR model was slightly adjusted during the pretraining phase. The embedding layers and Transformer layers are shared with the pretraining model. We remove the last layer for the pretraining stages, which is called the pretraining head, and replace it with a downstream task-related layer called the prediction head. According to the different downstream tasks, different prediction heads need to be built corresponding to the supernode. For binary DDI prediction and drug combination prediction, the prediction head is a fully connected neural network with two neurons in the output layer. For multi-DDI prediction, the prediction head is a fully connected neural network with 86 neurons in the output layer, which is the same number of categories of DDI. The input representation of the model is the same as what is shown in Figure 3, except that it does not contain S_{AB} information and does not mask the atoms.

2.5 Evaluation metrics and parameter settings

To verify the pretraining performance, we calculate the recovery accuracy of the masked atoms in the mask atom prediction task. That is, the proportion of masked atoms that are correctly predicted. For the fine-tuning stage, the evaluation metrics included AUC (ROC-AUC), AUPR (PR-AUC), Precision, Recall, ACC and F-score (F1). These metrics have been utilized to assess the performance of MGP-DR with other state-of-the-art methods. The metrics definitions and parameter settings for MGP-DR model could be found in Supplementary Material.

3 Results and discussion

3.1 Pretraining stage

3.1.1 Choice of model structure

To determine a better model structure for both pretraining and downstream tasks, we designed and compared three model structures, named Small, Medium and Large. In this section, we use the large-scale dataset to pretrain our model and the BIOSNAP as the fine-tuning dataset. Pretraining recovery accuracy and the averaged fine-tuning performance were used as the evaluation metrics.

As listed in Table 1, because of the small embedding size and feedforward network size, the small model is inferior to the other two. There is only a slight difference between the Medium and Large models. Therefore, we further evaluate these two structures through the downstream performance, as shown in Table 2. Compared with the medium model, the large MGP-DR model performs better on the pretraining task while performing slightly worse on DDI prediction tasks. This phenomenon may be caused by the fact that the large model has an overfitting risk. Meanwhile, a large number of parameters makes the training speed of the model very slow. The time to train one epoch for a medium-sized model is ~13.63 h; for a large-sized model, is about 27.94 h. After the evaluation of various factors, the structure of the medium-sized MGP-DR model was finally adopted for subsequent experiments.

3.1.2 Influence of pretraining strategies

In this section, a controlled experiment was conducted to explore the effects of different aggregation strategies and the presence or absence of hydrogen atoms in molecules. We use the large-scale pretraining dataset and choose the BIOSNAP dataset as the downstream task.

As illustrated in Supplementary Table S1, the pretraining accuracy with hydrogens can reach 0.9818, whereas that of the hydrogen-free model can only reach 0.9019 under Strategy 1. There is also a similar trend for Strategy 2. This phenomenon can be attributed to the processing method of adjacent information. The initial adjacency matrix does not distinguish the chemical bonds between atoms. Under this circumstance, the hydrogen atom can play an auxiliary role in helping to determine the chemical bond. The retention of hydrogen atoms can make the contextual information more abundant, which is critical for predicting the type of masked atoms during pretraining. For the two different information aggregation strategies, the results show that there is only a slight difference between them. Therefore, in subsequent experiments, we retained the hydrogen atoms in the drug pair molecules and analyzed the experimental results of both strategies.

Table 1. Pretraining performance between the three structures

Name	Layers	Heads	Embedding size	FFN size	Recovery accuracy
Small	3	2	64	128	0.9797
Medium	3	2	128	256	0.9818*
Large	3	2	256	512	0.9823*

*The underlined value represents the first and second largest value of the current column.

Table 2. Fine-tuning performance under two excellent structures

Name	ROC-AUC	PR-AUC	F1	ACC	# Pre-train parameters
Medium	0.9941*	0.9927*	0.9646*	0.9635*	418 577
Large	0.9829	0.9784	0.9449	0.9430	1 656 337

*The underlined value represents the largest value of the current column.

3.1.3 Influence of target information

In this section, we use a small-scale dataset to pretrain our MGP-DR model. The total loss contains the classification loss and the regression loss. As shown in Supplementary Table S2, the 'S_{AB}' column represents the preprocessing method for the target overlap score. The word 'original' means keeping the initial calculated S_{AB} score without any operation, while 'z-score' and 'min-max' represent the transformation of the scores using the corresponding normalization method.

According to the results in Supplementary Table S2, it can be found that 'min-max' obviously performs better than the other two methods. For the DDI data, the difference among results is not very significant, but for the drug combination data, the difference is more apparent. In the fourth column, the difference between the maximum value and the minimum value is ~4%, while in the fifth column, the difference can reach 16%. We summarize the reasons for these results as follows. The first is the impact of the data size. There are 83 040 samples in the BIOSNAP dataset, but only 1417 samples in the DCDB dataset, and the scale gap between them can exceed a factor of 50. So, compared with the latter, the former can fit the model better, and the fluctuation of each performance is smaller, which is maintained at ~0.1%. Second, the definition and calculation method of S_{AB} are based on drug combination data, and its biological significance is also inferred from this kind of data. As described in the article by Cheng *et al.*, 'We find that FDA approved drug combinations have lower S_{AB} compared to random drug pairs, confirming that it offers a reliable measure of drug–drug relationships within the human interactome'. As a result, the perturbation of the S_{AB} score is more likely to exert a great impact on the performance of the DCDB dataset.

3.2 Fine-tuning stage

3.2.1 Pretraining indeed succeeded

We compared the performance of the pretrained and non-pretrained MGP-DR models on different downstream datasets. For the following four datasets, the exact same pretraining data, fine-tuning scheme and hyperparameters are set between pretrained and non-

pretrained models. ROC-AUC was chosen as the performance metric, excluding the DEEPDDI dataset. In the multiclassification problem, the calculation of ROC-AUC can be divided into two types: macroaverage and microaverage. At the same time, the categories in the DEEPDDI dataset are seriously unbalanced, resulting in obvious fluctuations in this value. Therefore, here, we choose ACC as the evaluation metric.

Supplementary Table S3 shows a general performance improvement across all datasets after pretraining. For the DDI prediction task, the ROC-AUC value of the binary classification can be improved by 3.37%, and the ACC value of the multiclassification can be improved by 5.61%. For the drug combination prediction task, the ROC-AUC improvement under the DCDB dataset reached 7.9%. In addition, when pretraining is not introduced, the related indicators decrease, and the fluctuation increases. It is clearly shown that the pretrained MGP-DR model can outperform the non-pretrained model, clearly demonstrating the effectiveness of the pretraining strategy and the excellent generalization ability of our model.

3.2.2 Comparison with other state-of-the-art methods

CASTER (Chemic*Al* Substruc*Tur*E *R*epresentation) (Huang *et al.*, 2020): CASTER is a framework that predicts DDIs through given chemical structures of drugs. We use the BIOSNAP dataset provided in this article and adopt the same data splitting method. For each experiment, three independent runs with different random splits were conducted, and early stopping was used based on the ROC-AUC on the validation set. The dropout rate was set to 0.5, and the total epoch number was 100. Classification results are shown in Table 3, and other comparison methods involved include: LR (Logistic Regression), Nat.Prot (Vilar *et al.*, 2014), Mol2Vec (Jaeger *et al.*, 2018), MolVAE (Gómez-Bombarelli *et al.*, 2018) and DeepDDI (Ryu *et al.*, 2018). It can be observed that our method has an absolute lead in these three metrics, with an average improvement of more than 10% compared to the previous best method, CASTER. In addition, MGP-DR contains fewer intermediate parameters, which is beneficial to model training and storage. Besides, we do not need to spend much time constructing chemical substructures and can obtain better results in less time.

GCN-BMP (Graph Convolutional *N*etwork with *B*ond-aware *M*essage *P*ropagation) (Chen *et al.*, 2020) **and Zhang *et al.*** (Zhang *et al.*, 2017): GCN-BMP is a graph-based method that uses the Siamese GCN network as a generic encoder. Another method proposed by Zhang *et al.* is a link prediction model based on multifeature fusion. We take advantage of the dataset provided by Zhang *et al.*, which was also used in the GCN-BMP paper. The ratio of the training set, validation set and test set is set to 8:1:1. The dropout rate is 0, and the total epoch number is 250. For each experiment, three independent runs with different random splits were conducted, and early stopping was introduced, too. The classification results are shown in Table 4, and other comparison methods involved include: NN (Vilar *et al.*, 2012), LP-Sub/SE/OSE (Li *et al.*, 2015), SSP-MLP (Ryu *et al.*, 2018), GA (Ma *et al.*, 2018), Mol2Vec (Jaeger *et al.*, 2018), NFP (Duvenaud *et al.*, 2015) and GIN (Xu *et al.*, 2018). It can be concluded from the table that compared with other methods, our MGP-DR method is absolutely ahead in terms of the PR-AUC and F1 score and has a slight advantage in terms of the ROC-AUC. Moreover, the results under the MGP-DR model are more stable, and the standard deviation of each indicator is below 0.001.

DEEPDDI (*Deep* learning for *D*rug–*D*rug *I*nteractions) (Ryu *et al.*, 2018) **and Dai *et al.*** (Dai *et al.*, 2021): the DEEPDDI model collected a multiclassification dataset that integrates 86 DDI types from DrugBank. Both two methods use this dataset. The dropout rate is 0, the total epoch number is 250 and an early stopping strategy is also introduced. The DEEPDDI model showed reasonably accurate performance, attaining 84.8–93.2% in their article. Our MGP-DR model can achieve 87.53% average accuracy, but as shown in Table 5, both ROC-AUC and PR-AUC values are above 90%. It indicates that our prediction results can be very stable in an unbalanced sample set. The comparison results with Dai's and other baseline methods are listed in Table 5, include: ComplEx (Trouillon *et al.*, 2016), KBGAN (Cai and Wang, 2017), SimplE (Kazemi and Poole, 2018) and RotatE (Sun *et al.*, 2019). The MGP-DR model offers obvious advantages in both indicators, especially PR-AUC, which exceeds other methods by more than 10%.

Table 3. MGP-DR provides more accurate DDI prediction than other strong baselines on the BIOSNAP dataset

Model	ROC-AUC	PR-AUC	F1	# Parameters
LR	0.802 ± 0.001	0.779 ± 0.001	0.741 ± 0.002	1723
Nat.Prot	0.853 ± 0.001	0.848 ± 0.001	0.714 ± 0.001	N/A*
Mol2Vec	0.879 ± 0.006	0.861 ± 0.005	0.798 ± 0.007	8 061 953
MolVAE	0.892 ± 0.009	0.877 ± 0.009	0.788 ± 0.033	8 012 292
DeepDDI	0.886 ± 0.007	0.871 ± 0.007	0.817 ± 0.007	8 517 633
CASTER	0.910 ± 0.005	0.887 ± 0.008	0.843 ± 0.005	7 813 429
MGP-DR	0.994 ± 0.001	0.993 ± 0.002	0.965 ± 0.001	416 257

The first and second-largest values in each column are underlined. The baselines' performances are taken from CASTER.

*"N/A" represents parameter quantities that cannot be evaluated.

Table 4. MGP-DR provides more accurate DDI prediction than other strong baselines on Zhang *et al.*'s dataset

Model	ROC-AUC	PR-AUC	F1
NN	0.678 ± 0.250	0.526 ± 0.270	0.498 ± 0.430
LP-Sub	0.937 ± 0.130	0.904 ± 0.180	0.764 ± 0.280
LP-SE	0.938 ± 0.280	0.905 ± 0.390	0.785 ± 0.500
LP-OSE	0.939 ± 0.140	0.906 ± 0.360	0.794 ± 0.430
SSP-MLP	0.931 ± 0.340	0.886 ± 0.510	0.784 ± 0.570
GA	0.938 ± 0.610	0.903 ± 0.660	0.548 ± 0.470
Mol2Vec	0.936 ± 0.140	0.887 ± 0.320	0.810 ± 0.260
NFP	0.818 ± 0.130	0.689 ± 0.210	0.609 ± 0.240
GIN	0.616 ± 0.260	0.483 ± 0.310	0.560 ± 0.560
GCN-BMP	0.967 ± 0.090	0.940 ± 0.120	0.850 ± 0.170
Zhang *et al.*	0.957 ± 0.002	0.841 ± 0.024	0.751 ± 0.020
MGP-DR	0.968 ± 0.001	0.966 ± 0.001	0.912 ± 0.001

The first and second largest values in each column are underlined. The baselines' performances are taken from GCN-BMP and Zhang *et al.*

Table 5. MGP-DR provides more accurate DDI prediction than other strong baselines on the DEEPDDI dataset

Model	ROC-AUC	PR-AUC
ComplEx	0.9355	0.7419
KBGAN	0.9436	0.7562
SimplE	0.9310	0.7499
RotatE	0.9348	0.7676
Dai (ComplEx)	0.9527	0.7615
Dai (SimplE)	0.9431	0.7693
Dai (RotatE)	0.9480	0.7899
MGP-DR	0.9781	0.9129

The first and second-largest values in each column are underlined. The baselines' performances are taken from Dai *et al.*

3.2.3 Validation of the top-ranked novel predictions

In this section, we conduct experiments to demonstrate the ability of MGP-DR to predict novel DDIs. We validate the prediction results

of the binary classification task: BIOSNAP, Zhang *et al.* and DCDB. For the first two datasets, the sample sizes in the test set are 19 016 and 16 608. Therefore, among the samples with the top 1000 Pred_score, the drug pairs whose original label is 0 are selected. For the last dataset, the sample size is only 142, so we choose top 10. BIOSNAP dataset is divided in advance; thus, no seed is needed. For the remaining two datasets, due to expensive training overhead, we cannot use cross-validation to estimate each sample; consequently, different seeds are chosen to control the different divisions.

As shown in Supplementary Table S4, under the BIOSNAP and Zhang *et al.* datasets, there are 1 and 2∼3 novel drug pairs in the top 1000. Under the DCDB dataset, there are 1–2 novel drug pairs in the top 10. Of the 13 new drug pairs shown in Supplementary Table S4, 6 can be validated (identified by √) in multiple ways. The detailed valid information for each pair could be checked in Supplementary Material. Drugs marked by * in the table are minimally annotated in major drug-related databases, such as DrugBank, CTD (Comparative Toxicogenomics Database) (Davis *et al.*, 2021), pubChem (Kim *et al.*, 2019), etc., and only a small amount of drug information exists.

3.3 Analysis of learned embeddings and case study

3.3.1 Representation visualization

MGP-DR takes a drug pair as input and generates a two-drug ensemble representation. To intuitively prove that our method is a graph-level strategy, we visualized the representation embeddings of SUP nodes extracted by the pretrained models during the training stage for different datasets. Figure 4 shows the t-SNE (van der Maaten and Hinton, 2008) visualization of drug pair graph features for the positive drug pair samples and the negative drug pair samples. We observed that for each dataset, the embedding vectors learned by our proposed model can easily separate the interacting drug molecules and the non-interacting ones, especially for BIOSNAP. This means that the MGP-DR model can learn discriminative but stable data-driven molecular representations on both the DDI prediction task and the drug combination prediction task.

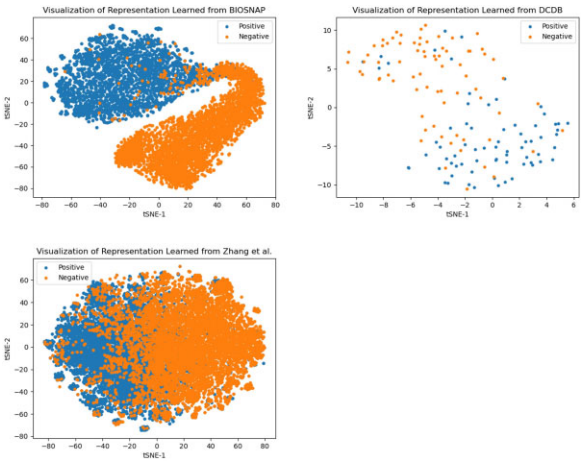

Fig. 4. Representation embeddings of SUP nodes for different datasets

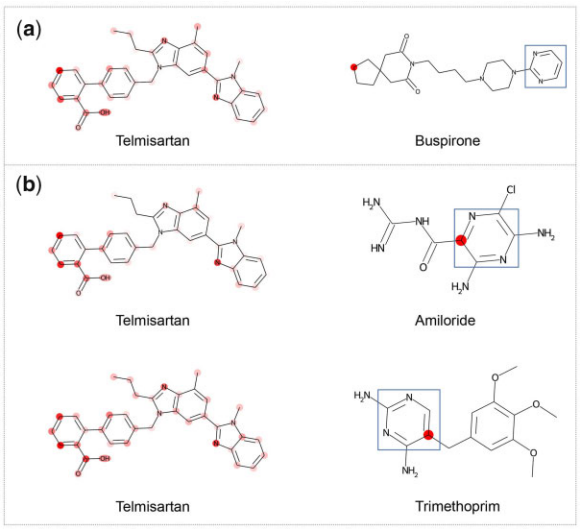

Fig. 5. The molecular graphs of interacting drugs involving with telmisartan. (**a**) Predicted interaction by our model. (**b**) Reported interaction from the training set. The darker the color is, the higher the attention scores are

3.3.2 Case study

We chose a certain pair of predicted new drugs, named telmisartan and buspirone, as well as two reported telmisartan-related drugs amiloride and trimethoprim to further explain the experimental results. Detailed information for the following five aspects is listed in Supplementary Material.

Attention distribution. We visualize the molecular graphs of these three pairs of drugs and label the atoms with different colors according to the attention scores output by the Transformer encoder. As shown in Figure 5, the darker the atom color is, the higher the attention scores are. This result suggested that for the same drug molecule, when it forms drug pair graphs with drug molecules that have similar interacting structures, the obtained attention scores are concentrated on some specific atoms or substructures, although the scores fluctuate.

Similar substructure. For similar molecular structures, similar parts often have specific molecular functions. Among the three telmisartan-related drugs listed in Figure 5, substructure $C_4H_4N_2$, shown in the blue box, is named pyrimidine. Many studies have shown that pyrimidine analogs can result in decreased activity of some disease-related proteins, such as AKR1C1 (Brozic *et al.*, 2009). We collected all diseases related to pyrimidine from CTD, and there were 11 diseases in total. After querying, nine of these diseases were also strongly associated with two of the three drugs mentioned above in the CTD record, and the specific statistical results are shown in Supplementary Table S5. This indicates that the properties and function of the drug substructures are closely related to the function of the whole drug and may affect the overall efficacy of the drug.

Pathway enrichment. We collected the genes related to the two drugs as annotated gene sets. Then, GO (Harris *et al.*, 2008) (Gene Ontology) and KEGG (Kanehisa and Goto, 2000) (Kyoto Encyclopedia of Genes and Genomes) enrichment analyses were performed. The enrichment results are shown in Supplementary Figure S1 (left: buspirone; right: telmisartan). It can be observed that the annotated gene sets enriched by the two drugs have a strong overlapping effect at the BP level. We also performed GO enrichment analysis at the molecular function level and KEGG enrichment analysis of the annotated gene set. Results were also closely related to lipids, and the pathways involved included fatty acid binding, long-chain fatty acid binding and fatty acid ligase activity.

Disease enrichment. The DO (Schriml *et al.*, 2012) (Disease Ontology) enrichment results are shown in Supplementary Figure S2 (left: buspirone; right: telmisartan). It is shown that genes associated with the two drugs are broadly enriched in many of the same diseases (marked by the red box), including hepatitis, obesity, kidney disease, urinary system disease and nutrition disease. Meanwhile, some diseases, such as obesity, overnutrition, lipid storage disease and fatty liver, were all strongly related to the results of pathway enrichment. This means that the drug pairs we predicted can exhibit a certain degree of connections under different verification aspects. In addition, there is something in common between these connections.

Exploration of target overlap patterns. We selected one disease related to both drugs, hypertension, and used the method proposed by Cheng *et al.* to explore the overlapping mode of the targets among the three based on the human protein interaction network. We denote the annotated gene sets of telmisartan and buspirone as

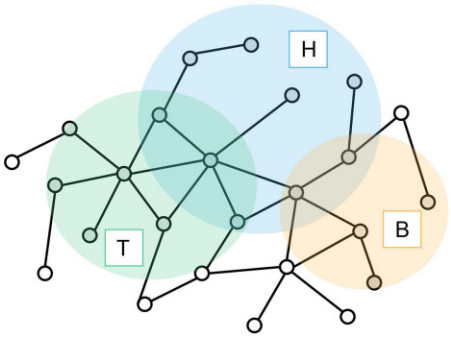

Fig. 6. The target overlap pattern of the hypertension module (H), telmisartan module (T) and buspirone module (B)

*T*and *B*. The network proximity between the disease module and the drug module is Z_{HT}, Z_{HB}.The network separation between the drug modules is S_{TB}. The calculation results are shown in Figure 6, and this target overlap pattern conforms to 'Complementary exposure', that is, satisfying $Z_{HT} < 0$, $Z_{HB} < 0$, $S_{TB} \geq 0$. The validation result also coincides with the most significant cotreatment target overlap pattern proposed in the article.

4 Conclusions

This article developed a novel end-to-end computational framework called MGP-DR, which is used for multidrug representation learning based on a pretraining model and molecular graphs to predict DDIs and drug combinations. Extensive experiments show that our model can achieve advanced prediction performance compared with other state-of-the-art methods. The validation of the experimental results strongly demonstrates the predictive ability of MGP-DR for new drug pairs. The distribution of drug pair embeddings and attention scores for molecular graphs indicate that the trained strategies we choose can effectively capture potential intradrug and interdrug associations. The representation of the molecule is consistent with the domain knowledge insights. There is still room for further improvement. In this study, we only perform a simple linear combination of losses corresponding to different drug pretraining tasks. Meanwhile, with the development of the multitask learning method, it would be a promising direction to combine different pretraining tasks using the well-established strategy of multitask learning in the training phase. We believe that this fusion may help to learn different aspects based on a large number of unlabeled drugs to further advance research related to drug discovery.

Acknowledgements

Thanks to all those who maintain excellent databases and to all experimentalists who enabled this work by making their data publicly available.

Funding

This work was supported by the National Natural Science Foundation of China [62072353 and 62132015].

Conflict of 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.

Data availability

The data underlying this article are available in https://github.com/LiangYu-Xidian/MGP-DR.

References

Ali,M.A. *et al.* (2017) Trends in the market for antihypertensive drugs. *Nat. Rev. Drug Discov.*, **16**, 309–310.

Bai,L.Y. *et al.* (2018) Prediction of effective drug combinations by an improved naive Bayesian algorithm. *Int. J. Mol. Sci.*, **19**, 14.

Berenbaum,M.C. (1989) What is synergy? *Pharmacol. Rev.*, **41**, 93–141.

Brozic,P. *et al.* (2009) Derivatives of pyrimidine, phthalimide and anthranilic acid as inhibitors of human hydroxysteroid dehydrogenase AKR1C1. *Chem. Biol. Interact.*, **178**, 158–164.

Cai,L. and Wang,W. (2017) *KBGAN: adversarial learning for knowledge graph embeddings*. arXiv preprint arXiv:1711.04071.

Chaudhuri,U. *et al.* (2019) Siamese graph convolutional network for content based remote sensing image retrieval. *Comput. Vis. Image Underst.*, **184**, 22–30.

Chen,X. *et al.* (2019) Drug-drug interaction prediction with graph representation learning. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA*. IEEE, pp. 354–361.

Chen,X. *et al.* (2020) GCN-BMP: investigating graph representation learning for DDI prediction task. *Methods*, **179**, 47–54.

Cheng,F. *et al.* (2019) Network-based prediction of drug combinations. *Nat. Commun.*, **10**, 1197.

Dai,Y. *et al.* (2021) Drug-drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Brief. Bioinform.*, **22**, bbaa256.

Davis,A.P. *et al.* (2021) Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res.*, **49**, D1138–D1143.

Devlin,J. *et al.* (2018) *BERT: pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Duvenaud,D.K. *et al.* (2015) Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural. Inf. Process. Syst.* arXiv:1509.09292.

Giles,T.D. *et al.* (2014) Efficacy and safety of nebivolol and valsartan as fixed-dose combination in hypertension: a randomised, multicentre study. *Lancet*, **383**, 1889–1898.

Gómez-Bombarelli,R. *et al.* (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.*, **4**, 268–276.

Han,K. *et al.* (2017) Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.*, **35**, 463–474.

Harris,M.A. *et al.* (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Huang,K. *et al.* (2020) CASTER: predicting drug interactions with chemical substructure representation. *Proc. Conf. AAAI Artif. Intell.*, **34**, 702–709.

Ianevski,A. *et al.* (2017) SynergyFinder: a web application for analyzing drug combination dose-response matrix data. *Bioinformatics*, **33**, 2413–2415.

Jaeger,S. *et al.* (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.*, **58**, 27–35.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Katz,R.J.N. (2004) FDA: evidentiary standards for drug development and approval. *NeuroRx*, **1**, 307–316.

Kazemi,S.M. and Poole,D. (2018) Simple embedding for link prediction in knowledge graphs. *Adv. Neural. Inf. Process. Syst.*, **31**.

Kim,S. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.

Landrum,G. (2013) *RDKit Documentation, Release, 2013*. Vol. **1**. p. 4. https://buildmedia.readthedocs.org/media/pdf/rdkit/latest/rdkit.pdf.

Li,P. *et al.* (2015) Large-scale exploration and analysis of drug combinations. *Bioinformatics*, **31**, 2007–2016.

Liu,Y. *et al.* (2014). DCDB 2.0: a major update of the drug combination database. *Database (Oxford)*, **2014**, bau124.

Liu,Y.Y. and Zhao,H.Y. (2016) Predicting synergistic effects between compounds through their structural similarity and effects on transcriptomes. *Bioinformatics*, **32**, 3782–3789.

Loewe,S. (1953) The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung*, **3**, 285–290.

Ma,T. *et al.* (2018) *Drug similarity integration through attentive multi-view graph auto-encoders*. arXiv preprint arXiv:1804.10850.

Menche,J. *et al.* (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.

Meyer,C.T. *et al.* (2020) Charting the fragmented landscape of drug synergy. *Trends Pharmacol. Sci.*, **41**, 266–280.

Mondal,I. (2020) *BERTChem-DDI: Improved Drug-Drug Interaction Prediction from Text Using Chemical Structure Information*. arXiv preprint arXiv:2012.11599.

Qato,D.M. *et al.* (2016) Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the United States, 2005 vs 2011. *JAMA Intern. Med.*, **176**, 473–482.

Ramzan,Z. *et al.* (2021) A machine learning-based self-risk assessment technique for cervical cancer. *Curr. Bioinform.*, **16**, 315–332.

Ru,X.Q. *et al.* (2020) Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput. Biol. Med.*, **119**, 103660.

Ryu,J.Y. *et al.* (2018) Deep learning improves prediction of drug-drug and drug-food interactions. *Proc. Natl. Acad. Sci. USA*, **115**, E4304–E4311.

Schriml,L.M. *et al.* (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

Sun,X.C. *et al.* (2013) High-throughput methods for combinatorial drug discovery. *Sci. Transl. Med.*, **5**, 205rv1.

Sun,Z. *et al.* (2019) *Rotate: knowledge graph embedding by relational rotation in complex space*. arXiv preprint arXiv:1902.10197.

Trouillon,T. *et al.* (2016) Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA*. pp. 2071–2080.

van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Vaswani,A. *et al.* (2017) Attention is all you need. *Adv. Neural. Inf. Process. Syst.*, **30**.

Vilar,S. *et al.* (2012) Drug-drug interaction through molecular structure similarity analysis. *J. Am. Med. Inform. Assoc.*, **19**, 1066–1074.

Vilar,S. *et al.* (2014) Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.*, **9**, 2147–2163.

Weininger,D.J.J. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

Wishart,D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.

Xiong,Z.P. *et al.* (2020) Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.*, **63**, 8749–8760.

Xu,H. *et al.* (2020) *Tri-graph information propagation for polypharmacy side effect prediction*. arXiv preprint arXiv:2001.10516.

Xu,K. *et al.* (2018) *How powerful are graph neural networks?* arXiv preprint arXiv:1810.00826.

Zhang,T. *et al.* (2021) Synergistic drug combination prediction by integrating multiomics data in deep learning models. *Methods Mol. Biol.*, **2194**, 223–238.

Zhang,W. *et al.* (2017) Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics*, **18**, 18.

Zhang,X.-C. *et al.* (2021) MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Brief. Bioinform.*, **22**, bbab152.

Zheng,W. *et al.* (2018) Drug repurposing screens and synergistic drug-combinations for infectious diseases. *Br. J. Pharmacol.*, **175**, 181–191.

Zitnik,M. *et al.* (2018) *BioSNAP Datasets: Stanford Biomedical Network Dataset Collection*, http://snap. stanford.edu/biodata Cited by 5.1.