*Briefings in Bioinformatics*, 2022, 1–11

https://doi.org/10.1093/bib/bbac296

**Problem Solving Protocol**

# Molormer: a lightweight self-attention-based method focused on spatial structure of molecular graph for drug–drug interactions prediction

Xudong Zhang, Gan Wang, Xiangyu Meng, Shuang Wang, Ying Zhang, Alfonso Rodriguez-Paton, Jianmin Wang and Xun Wang

Corresponding authors: Jianmin Wang. E-mail: jmwang113@hotmain.com; Xun Wang. E-mail: wangsyun@upc.edu.cn

## Abstract

Multi-drug combinations for the treatment of complex diseases are gradually becoming an important treatment, and this type of treatment can take advantage of the synergistic effects among drugs. However, drug–drug interactions (DDIs) are not just all beneficial. Accurate and rapid identifications of the DDIs are essential to enhance the effectiveness of combination therapy and avoid unintended side effects. Traditional DDIs prediction methods use only drug sequence information or drug graph information, which ignores information about the position of atoms and edges in the spatial structure. In this paper, we propose Molormer, a method based on a lightweight attention mechanism for DDIs prediction. Molormer takes the two-dimension (2D) structures of drugs as input and encodes the molecular graph with spatial information. Besides, Molormer uses lightweight-based attention mechanism and self-attention distilling to process spatially the encoded molecular graph, which not only retains the multi-headed attention mechanism but also reduces the computational and storage costs. Finally, we use the Siamese network architecture to serve as the architecture of Molormer, which can make full use of the limited data to train the model for better performance and also limit the differences to some extent between networks dealing with drug features. Experiments show that our proposed method outperforms state-of-the-art methods in Accuracy, Precision, Recall and F1 on multi-label DDIs dataset. In the case study section, we used Molormer to make predictions of new interactions for the drugs Aliskiren, Selexipag and Vorapaxar and validated parts of the predictions. Code and models are available at https://github.com/IsXudongZhang/Molormer.

**Keywords:** drug–drug interactions, molecular graph spatial structure, lightweight self-attention, Siamese network architecture

## Introduction

Multi-drug combinations for the treatment of complex diseases are becoming increasingly popular today, and this type of treatment can take advantage of the synergistic effects between drugs [1]. However, drug–drug interactions (DDIs) are not just all beneficial, and when two drugs enter the body at the same time, some unknown side effects or even harmful toxicity may occur [2]. Therefore, accurate and rapid identification of the interactions between a large number of drug pairs is essential to enhance the effectiveness of combination therapy and avoid unintended side effects. Methods for identifying DDIs in the wet lab are very expensive and time-consuming. In recent years, a large number of computationally based methods for predicting DDIs have been proposed.

Computation-based approaches can be broadly classified into two categories: text mining-based and feature learning-based approaches. Text mining-based approaches [3, 4] usually use scientific literature, medical reports and electronic medical records as data sources to identify annotated DDIs by applying natural language

**Xudong Zhang** is a graduate student in the Department of Computer Science and Engineering at the China University of Petroleum, with a research interest in drug repositioning.

**Gan Wang** is a graduate student in the Department of Computer Science and Engineering at the China University of Petroleum, with a research interest in drug repositioning.

**Xiangyu Meng** is a graduate student in the Department of Computer Science and Engineering at the China University of Petroleum, with a research interest in drug repositioning.

**Shuang Wang** is a lecturer in the School of Computer Science and Technology, China University of Petroleum (East China). Her research interests are focusing on drug repositioning, *de novo* drug design and so on.

**Ying Zhang** is a graduate student in the Department of Computer Science and Engineering at the China University of Petroleum, with a research interest in drug repositioning.

**Alfonso Rodríguez-Patãn** received the BS degree in electronic and computational from the University of Santiago de Compostela, Santiago de Compostela, Spain, in 1992, and the PhD degree in computer science from the Polytechnic University of Madrid, Madrid, Spain, in 1999. He is currently an Associate Professor with the Faculty of Informatics, Polytechnic University of Madrid. His current research interests include interplay of computer science, biology and engineering.

**Jianmin Wang** is a graduate student in Integrative Biotechnology & Translational Medicine, Yonsei University, Korea. His current research interests focus on deep learning for drug discovery and protein–protein interactions.

**Xun Wang** received her PhD degree from the University of Tsukuba, Japan, in 2015. She is currently a Full Professor in the China University of Petroleum (East China) and her current research interests include DNA computing, membrane computing and bioinformatics.

**Received:** May 18, 2022. **Revised:** June 20, 2022. **Accepted:** June 20, 2022

processing (NLP) techniques to automate a large corpus of relevant data for processing text [5]. This method identifies DDIs with high accuracy and can be used to build a DDI-related database, but it cannot identify unannotated DDIs [6]. However, there is still a large number of unknown DDIs, and using only text mining-based methods to identify DDIs still has significant limitations.

The feature-based learning approaches [7–9] refer to using the structural feature vector of a drug as input to learning deeper structural knowledge of the drug through machine learning or deep learning methods to predict potentially unknown DDIs. Ryu *et al.* [10] use simplified molecular input line entry system (SMILES) of a drug to extract molecular fingerprint representations and calculate similarities between drugs to construct drug structure feature vector and finally use a deep neural network to predict DDIs. DeepPurpose [11], as an integrated deep learning model, can input SMILES to convolutional neural network (CNN) [12], recurrent neural network (RNN) [13, 14], message passing neural network (MPNN) [15] and transformer [16] to extract structural features to predict DDIs. The method can be used well with the NLP models RNN and transformer for the task of predicting DDIs, especially transformer can achieve good prediction results due to its location information and attention mechanism. However, from the point of view of feature encoding, SMILES, which only stores one-dimensional structural information of drugs, cannot characterize drugs with the spatial structure well.

There are many graph-based deep learning methods, such as graph convolutional neural network (GCN) [17, 18], graph attention network [19–21] and gated graph neural network [22, 23], which have performed well in the fields of social networks and knowledge graphs. The molecular graph of a drug can store not only the atomic information but also the positional and spatial information of atoms. Recently, graph-based models have been applied in the field of drug development and discovery [24]. In the field of DDIs prediction, DeepDrug [25] uses the feature matrix and adjacency matrix of the drug molecule graph and then feeds both into the GCN to learn the deep representation of the features to predict DDIs. DPDDI [26] uses GCN to extract topological relationships of drugs from DDI networks to learn low-dimensional feature representations of drugs. Attention based [27–29] on multi-headed attention mechanism performs very well on various prediction tasks. Graphormer [30] first uses a transformer for processing graph structure, it uses centrality encoding, spatial encoding and edge encoding combined with attention to the encoding graph structure and the model outperforms traditional graph neural network models. Due to the large number of nodes in the drug molecule map and the large computational effort of the transformer itself, the model is computationally and storage expenses and therefore not suitable for DDIs prediction tasks.

In this paper, we propose Molormer, a method based on a lightweight attention mechanism for DDI prediction that can encode the spatial structure of the molecular graph. The contribution of our method is as follows: (i) traditional DDIs prediction methods only use drug sequence or graph information but ignore the spatial structure information of atoms. Molormer uses 2D structure of drugs as input and uses centrality encoding, spatial encoding and edge encoding [30] to encode molecule graph. (ii) Molormer uses a lightweight-based attention mechanism to process spatially the encoded molecular graph, which not only preserves the multi-headed attention mechanism but also alleviates the computational and storage costs of the model. (iii) A Siamese network architecture [31] is used as the design architecture of Molormer, where the two drugs of the input model share the same neural network model weights so that the limited data can be fully utilized to train the model for a better fit.

## Methods

Before inputting drug *A* and drug *B* into Molormer, we need to extract the graph structure in the 2D structural information of the drug download from DrugBank [32] that can be applied to the attention mechanism. We first extracted the atom and bond information of drug using Rdkit [33]. We then feed the generated graphs of the two drugs into Molormer. The architecture diagram of Molormer is shown in Figure 1. In Molormer, the original drug graph is first fed to the embedding layer to be embedded as four forms of features, which jointly can characterize the representation of a drug in spatial structure. Next, the two generated drug features are input to two ProbSparse self-attention blocks to extract to two feature maps, respectively. After concatenating the two feature maps, they are fed to the decoder for dimensionality reduction and final prediction.

### Weight-sharing-based atom feature embedding

The previous methods [7, 25], after obtaining atom and bond information of drug graph, would be directly embedded and input to graph-based models (such as MPNN, GCN) for feature extraction. As for the molecular graph, the importance of each atom is different, which has been neglected in the current calculations of attention. The more edges that exist for an atom, the more associations or interactions between atoms, then the more we consider it to be of interest. In this paper, we represent the importance of an atom by computing its degree centrality and combine it with the features of the atom itself to form the centrality feature of the atom, which can better allow the model to capture the semantic relevance and importance of atom in the attention mechanism [30].

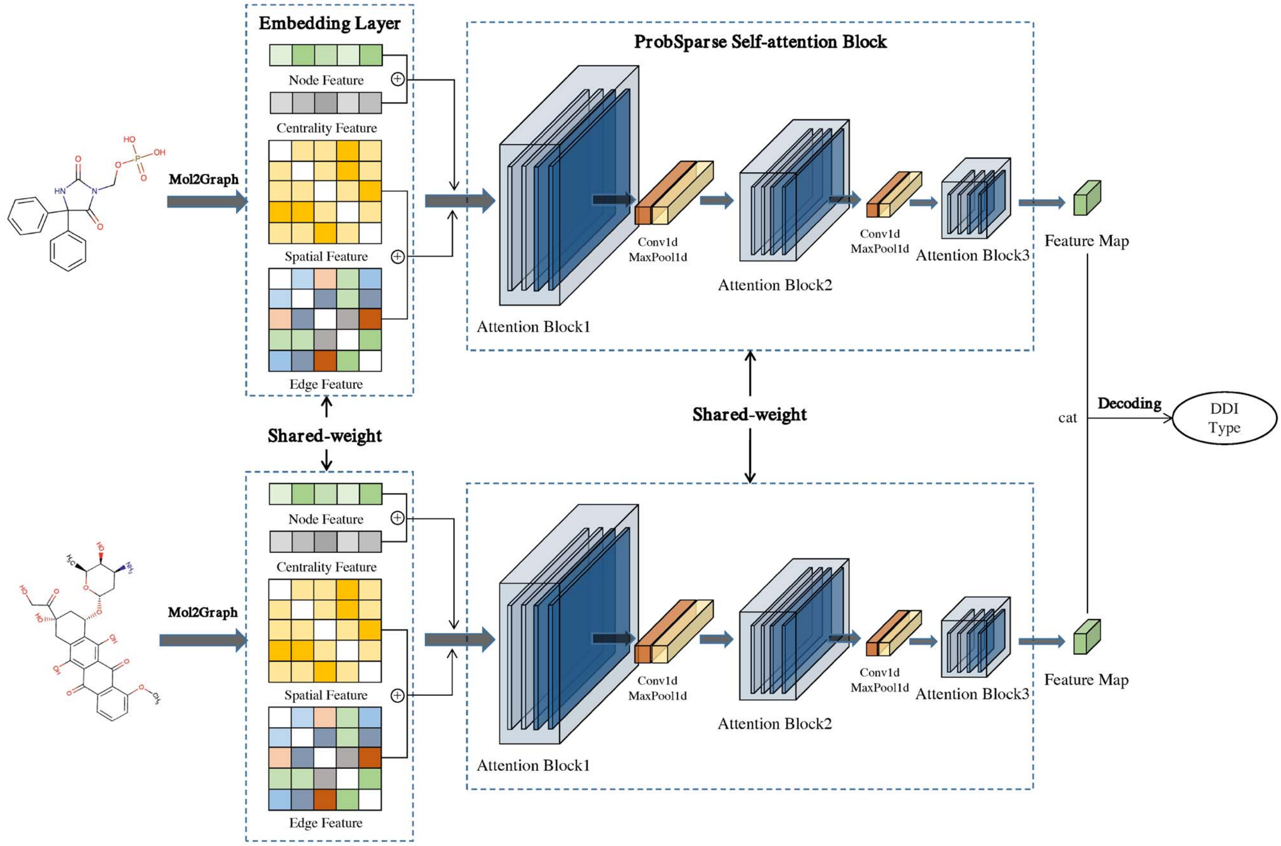

**Figure 1.** The overview of Molormer.

We define the atom *i* in drug *A* as a vector $x_i^A$ comprised of 9 elements, as shown in the following:

$$x_i^A = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9], \tag{1}$$

where $x_i^A$ denotes the digital representation of the drug *A* atom *i*. $\alpha_1$ is the number of the atom in the drug. $\alpha_2$ represents chiral information including unspecified, R-type and S-type. $\alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9$ represent degree (number of bonds involved), formal charge, number of connected hydrogen atoms, number of radical electrons, hybridization, whether it forms an aromatic bond and whether it is in a ring, respectively. All elements in $x_i^A$ can be obtained by Rdkit and are embedded as integers by a dictionary defined in advance.

The centrality of the atom is measured using the degree of the atom. Typically, there is no directionality of bonds in a molecule, so the in and out degrees are equal. For example, if atom *i* in drug *A* has 4 bonds, then the in-degree $I_i^A$ and the out-degree $O_i^A$ of the atom are both 4. In addition, the degree in the definition of atom in Equation (1) is used to represent the atom feature itself, and centrality is used to represent the importance of the atom. In Equations (2) and (3), we use two different learnable shared weight matrices to learn the embedding representation of the atom feature and centrality feature, respectively.

The embedding equations for $E_{\text{atom}_i}^A$ and $E_{\text{atom}_j}^B$ of atom *i* of drug *A* and atom *j* of drug *B* with centrality features are as follows:

$$E_{\text{atom}_i}^A = W_x x_i^A + W_{\text{out}} O_i^A + W_{\text{in}} I_i^A \tag{2}$$

$$E_{\text{atom}_j}^B = W_x x_j^B + W_{\text{out}} O_j^B + W_{\text{in}} I_j^B, \tag{3}$$

where $W_x x_i^A$ denotes the atom feature of drug *A* atom *i*. $O_i^A$ denotes the out-degree centrality vector of drug *A* atom *i* and $I_i^A$ denotes the in-degree centrality vector of drug *A* atom *i*. $W_{\text{out}} O_i^A + W_{\text{in}} I_i^A$ is the centrality feature of drug *A* atom *i*. The same is true for the corresponding parameters of drug *B*. $W_x$,$W_{\text{out}}$ and $W_{\text{in}}$ are learnable shared weight matrices, which are jointly used by drug *A* and drug *B* in the Siamese network architecture to learn features of atoms during parameter updates.

## Weight-sharing-based edge feature embedding

Using only the centrality feature of atom to represent molecule is still not comprehensive enough. The atom spatial position and bond level of the molecular graph all have critical effects on the property of molecule, so how to accurately encode the bond feature is also a crucial issue. In addition, the atoms in a molecular graph are distributed in a multi-dimensional space, so we need to encode the position feature of atoms and edges with the help of the spatial association information between atom pairs. The spatial location encoding of atom is unlike sequential data where absolute location embedding [16] or relative location embedding [34, 35] can be performed for each token. Ying *et al.* [30] proposed to solve this problem with a learnable spatial encoding in an attention mechanism, where the spatial information between two nodes is the shortest path distance between them. However, the combination of feature with huge dimensionality and computationally intensive attention mechanism can lead to dimensional explosion. To solve this problem, we first use a matrix of shared weights to learn the spatial feature embedding of atom, which can be unnecessarily repeated training for the model. And apply ProbSparse self-attention [36] in Molormer, we will introduce this computationally reduced attention mechanism in the next section.

For an adjacent node (atom) pair $(\mathrm{a}, \mathrm{b})$ with edges (bonds), its edge $e_{(\mathrm{a,b})}$ is defined as follows:

$$e_{(\mathrm{a,b})} = [\beta_1, \beta_2, \beta_3], \tag{4}$$

where $\beta_1$ represents the bond type, $\beta_2$ represents the stereochemical bond and $\beta_3$ represents whether or not the bond is conjugated. $\beta_1, \beta_2, \beta_3$ can be obtained by Rdkit and are embedded as integers by a dictionary defined in advance.

Next, we define the spatial structure information. For a node pair $(i, j)$, we first find the shortest path $P = (e_1, e_2, \ldots, e_k)$ from $x_i$ to $x_j$ and then use its shortest path distance to represent the relationship between the positions of two nodes on the space structure. If the node pair $(i, j)$ is connected (can be non-adjacent), then we take the shortest path distance as the spatial location $s_{(i,j)}$ of the edge $e_{(i,j)}$. If the node pair $(i, j)$ is unconnected, then we set the spatial location $s_{(i,j)}$ of its edge $e_{(i,j)}$ to $-1$.

After obtaining the digital representations of edge and spatial structure, the edge $e_{(i,j)}$ of drug $A$ and edge $e_{(m,n)}$ of drug $B$ with spatial structure information are embedded as follows:

$$E^A_{\mathbf{edge}_{(i,j)}} = \frac{1}{k}\sum_{l=1}^{k} P^A_l \mathrm{W}_{\mathbf{edge}} + \mathrm{W}_{\mathbf{spat}} s^A_{(i,j)} \tag{5}$$

$$E^B_{\mathbf{edge}_{(m,n)}} = \frac{1}{k}\sum_{h=1}^{k} P^B_h \mathrm{W}_{\mathbf{edge}} + \mathrm{W}_{\mathbf{spat}} s^B_{(m,n)}, \tag{6}$$

where $\frac{1}{k}\sum_{l=1}^{k} P^A_l \mathrm{W}_{\mathbf{edge}}$ is the embedding of edge$e^A_{(i,j)}$ in drug $A$, which is obtained by averaging the dot-products of edge feature $P^A_l$ and a learnable shared weight matrix $\mathrm{W}_{\mathbf{edge}}$ in the shortest path $P^A$ of the atom pair $(i, j)$ in drug A. $P^A_l$ is the $l$th edge in the shortest path $P^A$ of the atom pair $(i, j)$ in drug A and k is the number of edges in the shortest path $P^A$. $\mathrm{W}_{\mathbf{spat}} s^A_{(i,j)}$ is the embedding of spatial structure feature of atom pair $(i, j)$ in drug $A$. The same is true for the corresponding parameters of drug $B$. $\mathrm{W}_{\mathbf{edge}}$ and $\mathrm{W}_{\mathbf{spat}}$ are two learnable shared weight matrices, which are jointly used by drug $A$ and drug $B$ in the Siamese network architecture to learn features of atoms during parameter updates.

## ProbSparse self-attention stacked encoder

After we get the node features and edge features of the molecular graph, there are many limitations if we process them with traditional attention mechanism [16]. The large number of nodes of molecules and the large dimensionality of the feature vectors can lead to multi-headed attention hardly determining which information is important. And it also suffers from a series of problems such as high time complexity, high memory overhead, and can even lead to dimensional explosion. Therefore, we propose a stacked block molecular graph encoder based on ProbSparse self-attention [36] to filter out the most important queries and reduce computational complexity, while using self-attention distilling to reduce feature dimensionality and the number of parameters of network. Encoder includes three ProbSparse self-attention blocks, and the first two blocks are followed by CNN to do distilling. The molecular graph of drug $A$ and drug $B$ is extracted by encoder and two feature maps are obtained.

Traditional self-attention requires $O(L_Q L_k)$ memory and a quadratic dot product computational cost, which is the main drawback of its predictive power. It is found that a few dot products of sparse self-attention contribute to the main attention and other dot product pairs can be ignored. ProbSparse self-attention can be calculated by the following equation:

$$\mathbf{Attn}\,(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathbf{Softmax}\left(\frac{\overline{\mathrm{Q}}K^T}{\sqrt{d}}\right)\mathrm{V}, \tag{7}$$

where $\overline{\mathrm{Q}}$ is the same sparse matrix as the query size and it contains only top-u queries. We compute and rank all queries in Q according to the KL scatter-based sparsity measurement proposed by Zhou *et al.* [36] and then use top-u queries (in this paper u is 25) to form $\overline{\mathrm{Q}}$ take the place of Q. In this way, the dot product computation time complexity of ProbSparse self-attention is $O(\mathbf{ln} L_Q)$, and the memory occupation per query-key lookup and per block is $\mathrm{O}(L_K \mathbf{ln} L_Q)$ [36].

After inputting the features of embedded drug $A$ and drug $B$ into encoder, the formula for calculating ProbSparse self-attention is as follows:

$$\mathbf{Attn}^A_{(i,j)} = \mathbf{Softmax}\left(\frac{\left(E^A_{\mathbf{atom}_i}\mathrm{W}_{\overline{\mathrm{Q}}}\right)\left(E^A_{\mathbf{atom}_i}\mathrm{W}_K\right)^T}{\sqrt{d}} + E^A_{\mathbf{edge}_{(i,j)}}\right)$$
$$\left(E^A_{\mathbf{atom}_i}\mathrm{W}_V\right) \tag{8}$$

$$\mathrm{Attn}^{B}_{(m,n)} = \mathbf{Softmax}\left(\frac{\left(E^{B}_{\mathbf{atom}_m} W_{\overline{Q}}\right)\left(E^{B}_{\mathbf{atom}_m} W_{K}\right)^{T}}{\sqrt{d}} + E^{B}_{\mathbf{edge}_{(m,n)}}\right)$$

$$\left(E^{B}_{\mathbf{atom}_m} W_{V}\right), \tag{9}$$

where $W_{\overline{Q}}$, $W_K$ and $W_V$ are the learnable shared weight matrices. All three attention blocks in Encoder are calculated based on the above formula.

In addition, each attention block is followed by a distilling operation to privilege the high-level mapping with dominant features and generate a focused feature map at the next level. The specific calculation formula is as follows:

$$X_{j+1} = \mathbf{MaxPool}\left(ELU\left(Conv1d\left(\left[X_j\right]_{op}\right)\right)\right), \tag{10}$$

where $[X_j]_{op}$ denotes the ProbSparse self-attention and necessary operations performed by the input to the $j$th attention block. $Conv1d()$ is a 1D convolution layer and $ELU()$ is activation function [37].

### Decoder and prediction

After the embedded drug A and drug B with spatial structure information are input to the encoder, two low-dimensional feature maps representing the drug features are output. Then, we concatenate the two feature maps and input them into decoder. The decoder for DDIs prediction consists of the CNN and DNN. In the decoder, the concatenated feature map is first input to a layer of CNN, and several order-invariant local convolution filters in the CNN can capture and aggregate the interaction of nearby regions, so the CNN is often used in interaction encoders [38]. After the extracted interaction feature is flattened, they are then input into a fully connected layer consisting of three layers and finally input into the softmax function to predict the DDI type.

## Results and Discussion

The deep learning model for this experiment is based on the Pytorch framework, and the conversion of drug standard delay format (SDF) file into the molecular graph is done using Rdkit [33]. For Parameter setting, the learning rate is 1e−4, batch is 32, epoch is 200, attention block layer is 3, attention head is 8 and hidden dimension of drug is 256. Also, we use adam optimizer and cross entropy loss function to achieve the best performance of Molormer.

### Dataset and evaluation metrics

The dataset of DDIs proposed by Ryu *et al.* [10] is used as the baseline dataset for this paper. This dataset includes 1710 drugs with 192 284 DDIs. All the DDIs are divided into a total of 86 common DDI types, each described by a common sentence. We downloaded SDF files storing 2D structure information of drugs in DrugBank [39] in order to obtain the 2D structure information of the drugs, which will be inputted into Molormer. We randomly divide the dataset into the training set, test set and validation set in the ratio of 7:2:1 to train and evaluate Molormer.

We evaluate the performance of model using metrics that are commonly used in classification experiments, including: accuracy, macro precision, macro recall and macro F1 score. These evaluation metrics are as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{cases} 1 & \text{if } y_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$\text{Macro Precision} = \frac{1}{l}\sum_{i=1}^{1} \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$\text{Macro Recall} = \frac{1}{l}\sum_{i=1}^{1} \frac{TP_i}{TP_i + FP_i} \tag{13}$$

$$\text{Macro } F1 = \frac{2 \times \text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}}, \tag{14}$$

where $n$ and $l$ indicate the number of samples and DDI types, respectively. In equation 9, $x_i$ is the predicted value of true DDI type in dataset of sample $i$. And TP, TN, FN and FN are true positive, true negative, false positive and false negative, respectively.

### Baselines

(i) MPNN uses message passing neural network [15] to extract drug feature. It can extract embedding vector to molecular graph-level by embedding vector for each atom and edge. A multi-layer perceptron decoder is used to predict DDIs.

(ii) LSTM is a special kind of RNN, which is mainly designed to solve the gradient disappearance and gradient explosion problems during training of long sequences [40]. We use frequent consecutive sub-sequence mining [38] as the word embedding algorithm to predict DDIs after getting the embedding features of drugs input to the LSTM.

(iii) Morgan-CNN [11] inputs Morgan Fingerprint of drug into a multi-layer perceptron to extract low-dimensional feature, and after concatenation features of two drugs, finally inputs them into a multilayer 1D CNN for prediction of DDIs.

(iv) MolTrans [38] uses frequent consecutive sub-sequence mining algorithm to extract sub-structures of SMILES, and transformer for enhanced representation of sub-structures. The model is currently a more advanced DTI prediction method, but it is equally applicable to the DDIs task, and we use the hyperparameters from the original paper for our experiments.

(v) Hyper-AttentionDTI [41] is currently the most advanced method for predicting DTIs, and it is also applicable to the DDI prediction task. The model uses a deep CNN to learn the feature matrix of drugs. To model complex non-covalent molecular interactions

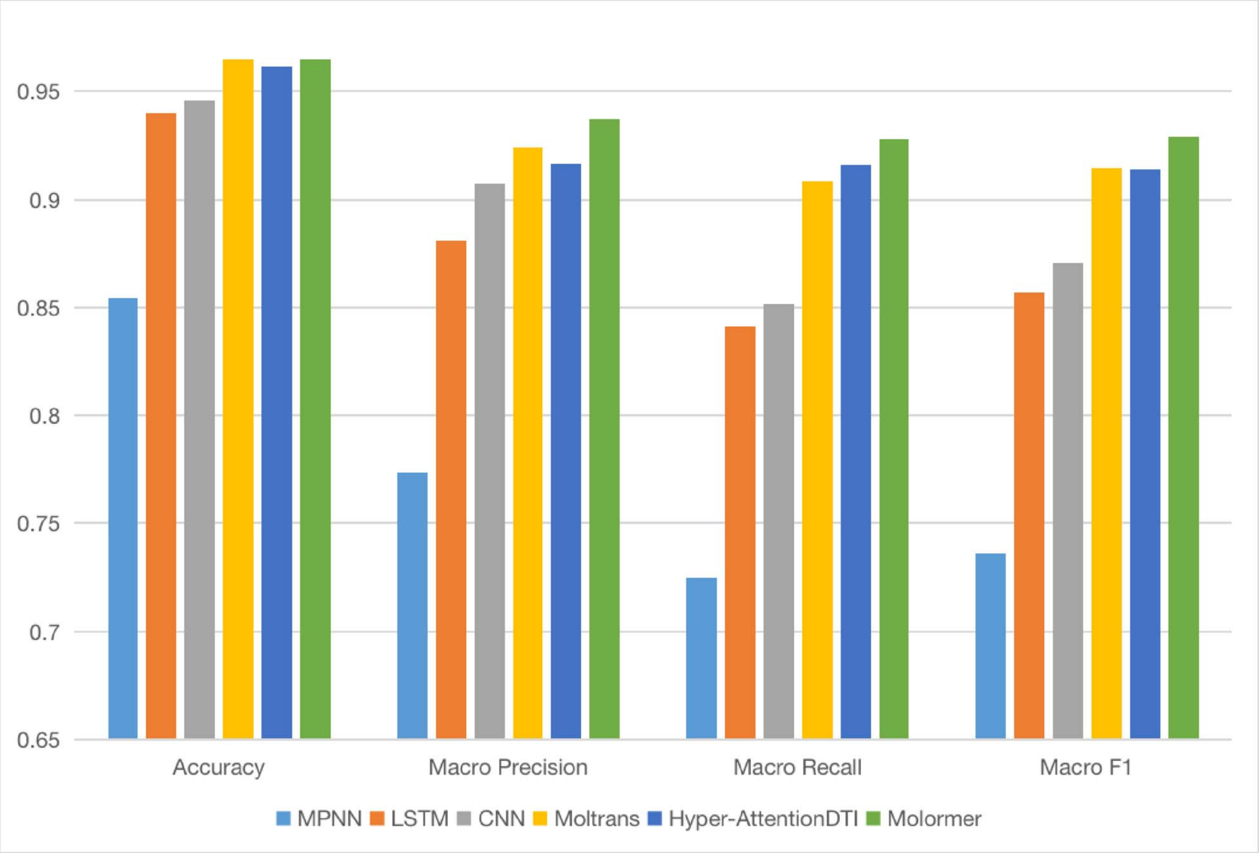

**Figure 2.** Molormer achieved the best prediction performance in comparison with other advanced methods (the average of five random executions).

between atoms and atoms, the model uses the attention mechanism of the feature matrix to assign an attention vector to each atom.

## Comparison with state-of-the-art methods

To evaluate the performance of Molormer on the multi-classes DDIs prediction task, we first conducted experiments comparing it with other state-of-the-art methods, and the experimental results are shown in Figure 2. According to Figure 2, we can see that Molormer achieved the best performance in all evaluation metrics. Among them, accuracy of MolTrans is tied with Molormer for first place, but all other metrics are lower than Molormer. In the multi-class DDIs prediction task, the number of samples per class is extremely unbalanced and accuracy does not fully reflect the performance of model on unbalanced dataset. Note that each sample contains a pair of drugs and a label meaning which type of interactions they can interact with each other. Therefore, we need to combine the other three metrics to evaluate model, which can reduce the influence of the unbalanced categories.

As shown in Table 1, we can see that the macro precision, macro recall and macro F1 of Molormer are 2.13, 2.16 and 2.09% higher than MolTrans, respectively, and 1.6, 2.73 and 2.09% higher than Hyper-AttentionDTI, respectively. In summary, we believe that Molormer has the best performance. While some previous works [7, 11, 25] have treated the prediction of DDIs as a binary classes problem, in this paper, we perform multi-classes prediction of DDIs.

To further compare the prediction accuracy of Molormer with other models for each category, we calculated the prediction accuracy for each category, as shown in Figure 3. The result is calculated based on the test dataset, and some samples are small, so the prediction accuracy will be lower, but Molormer still performs better on the prediction of these categories.

## Ablation studies

To further evaluate Molormer as well as to validate the performance of the model under different settings, we design four ablation studies: (i) with or without centrality encoding and spatial encoding, (ii) with or without ProbSparse self-attention, (iii) with or without CNNs for self-attention distilling and decoding and (iv) with or without weight-sharing Siamese network architecture.

### With or without centrality encoding and spatial encoding

The impact of how drug is encoded on predicted outcome of the final DDIs is crucial, and how to accurately encode drug is a challenge for current research. Frequent consecutive subsequence (FCS) mining algorithm [38] is similar

**Table 1.** Molormer achieved the best prediction performance in comparison with other advanced methods (five random executions)

| Method | Accuracy | Macro precision | Macro recall | Macro F1 |
|---|---|---|---|---|
| MPNN | 0.8554 ± 0.001 | 0.7734 ± 0.002 | 0.7284 ± 0.006 | 0.7397 ± 0.006 |
| LSTM | 0.9393 ± 0.001 | 0.8961 ± 0.011 | 0.8637 ± 0.017 | 0.8722 ± 0.011 |
| CNN | 0.9457 ± 0.002 | 0.9048 ± 0.002 | 0.8796 ± 0.022 | 0.8848 ± 0.012 |
| MolTrans | 0.9649 ± 0.001 | 0.9206 ± 0.003 | 0.9054 ± 0.007 | 0.9102 ± 0.006 |
| Hyper-AttentionDTI | 0.9548 ± 0.002 | 0.9259 ± 0.004 | 0.8997 ± 0.004 | 0.9078 ± 0.003 |
| Molormer | 0.9667 ± 0.002 | 0.9419 ± 0.004 | 0.9270 ± 0.002 | 0.9311 ± 0.002 |

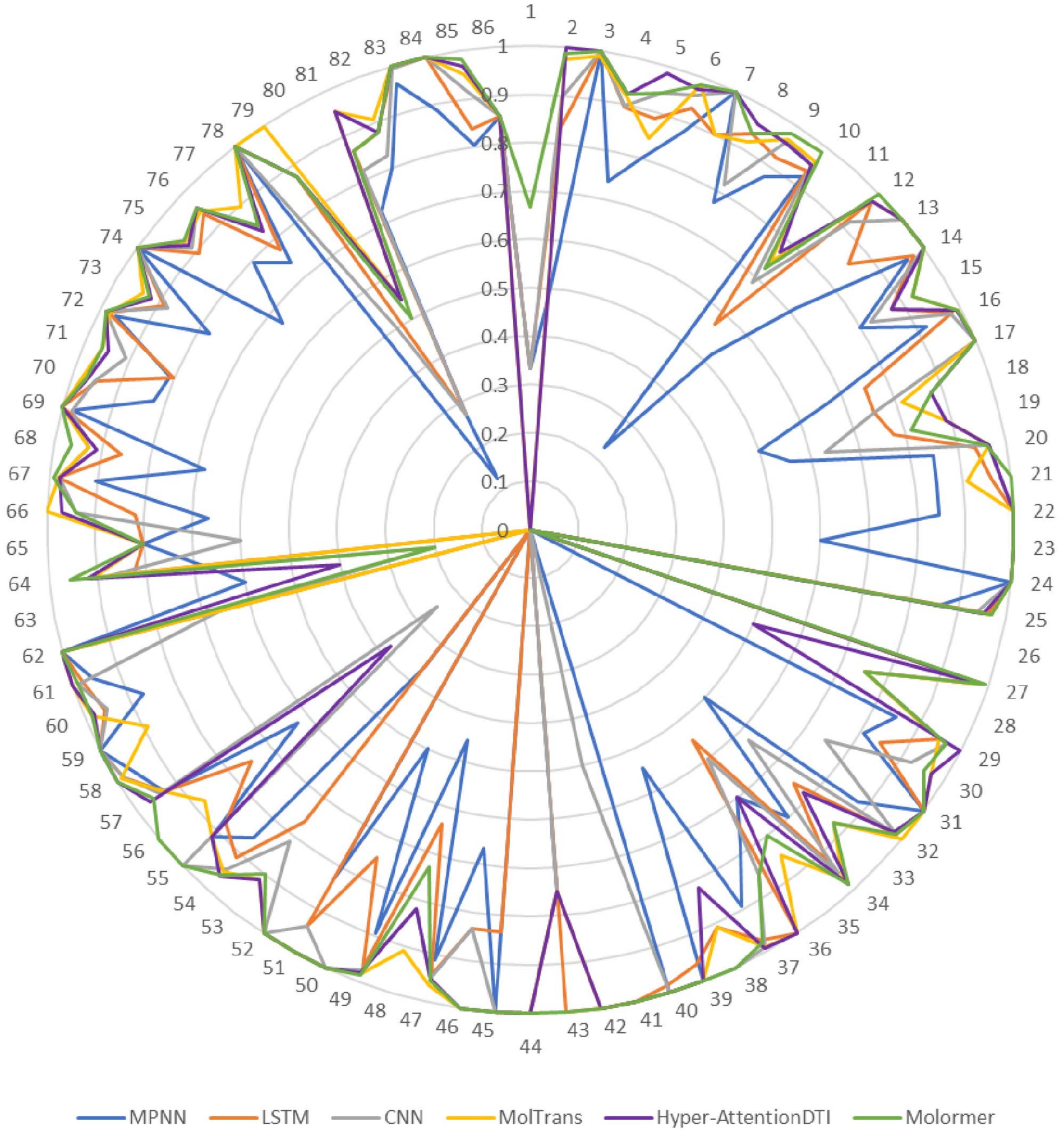

**Figure 3.** Molormer performs best on all predicted types of DDIs.

**Table 2.** Comparison of models with different encoding (five random executions)

| Method | Accuracy | Macro precision | Macro recall | Macro F1 |
|---|---|---|---|---|
| Without cent&spat encoding | 0.9634 ± 0.004 | 0.9383 ± 0.004 | 0.9194 ± 0.004 | 0.9268 ± 0.004 |
| Without centrality encoding | 0.9596 ± 0.002 | 0.9370 ± 0.006 | 0.9229 ± 0.004 | 0.9283 ± 0.003 |
| Without spatial encoding | 0.9626 ± 0.002 | 0.9297 ± 0.006 | 0.9191 ± 0.004 | 0.9226 ± 0.004 |
| FCS encoding | 0.9631 ± 0.004 | 0.9266 ± 0.012 | 0.9108 ± 0.015 | 0.9135 ± 0.012 |
| With cent&spat encoding | 0.9667 ± 0.002 | 0.9419 ± 0.004 | 0.9270 ± 0.002 | 0.9311 ± 0.002 |

**Table 3.** Comparison of models with and without ProbSparse self-attention (five random executions)

| Method | Accuracy | Macro precision | Macro recall | Macro F1 | Time | Memory |
|---|---|---|---|---|---|---|
| Self-attention | 0.9585 ± 0.002 | 0.9322 ± 0.002 | 0.8976 ± 0.004 | 0.9101 ± 0.002 | 2.3 ± 0.1326 | 21,937 MiB |
| ProbSparse self-attention | 0.9667 ± 0.002 | 0.9419 ± 0.004 | 0.9270 ± 0.002 | 0.9311 ± 0.002 | 1.3 ± 0.1010 | 13,201 MiB |

to the word embedding method in NLP, which is based on SMILES and generates a set of subsequences according to frequency of occurrence of sub-strings. In this ablation study, we compare FCS encoding, without centrality encoding and spatial encoding (cent&spat), without centrality encoding, without spatial encoding and with centrality encoding and spatial encoding (cent&spat). The experimental results are shown in Table 2.

According to Table 2, we can see that the model with cent&spat encoding preforms the best on all evaluation metrics. We believe that the reason for poor effect of FCS encoding is that it only uses drug sequence structure information, which ignores the spatial structure information of molecule. In addition, when the model without cent&spat encoding but only atomic and bond information of the molecular map is adopted, it is preferable to using centrality encoding or spatial encoding individually. Therefore, we conclude that these two encoding schemes must be used simultaneously to better characterize the spatial structure of the molecular graph.

## With or without ProbSparse self-attention

Self-attention can assign different attention weights to different content, and self-attention-based approaches have achieved amazing success in many fields. However, the computational effort of the self-attention mechanism is very large, and the computational power and memory capacity of current GPUs are limited. Therefore, this paper uses a lightweight ProbSparse self-attention for the DDIs prediction task. To validate the performance of ProbSparse self-attention in DDIs prediction task, we evaluated not only the prediction performance of the model but also the time and GPU memory used for each epoch during training, and experimental results are shown in Table 3.

According to Table 3, we can see that the accuracy, macro precision, macro recall and macro F1 of the model using ProbSparse self-attention are 0.82, 0.97, 2.94 and 2.1% higher than those of the model using self-attention, respectively. In terms of time and memory, for each epoch in training, the model using self-attention takes 2.3 h and 21 937 MiB of memory, while the model using ProbSparse self-attention takes only 1.3 h, saving almost half the time and memory.

## With or without CNNs for self-attention distilling and decoding

In Molormer, we use CNNs in two places for self-attention distilling and dimensionality reduction, respectively. First, we apply three attention blocks in Molormer and add the self-attention distilling module comprised of CNNs between the first two attention blocks to reduce final interaction feature dimension and the number of parameters of Molormer. Second, we use a CNN to reduce the concatenated feature map and then use the fully connected layer for the final prediction. To verify the effect of CNNs for self-attention distilling and decoding on model performance, we conduct an ablation study without CNNs for distilling and decoding, and the experimental results are shown in Table 4.

In terms of prediction performance, we can see that the model with CNNs performs the best on all evaluation metrics. Molormer concatenate and flatten the two feature maps output from the third attention block to generate interaction feature. The interaction feature dimensions generated by the model without CNNs for self-attention distilling is 8192, without CNN for decoding is 33 280, while the interaction feature dimensions generated by the model with self-attention distilling is 2048. Moreover, the number of model parameters without CNNs for self-attention distilling, without CNN for decoding and with CNNs are 11.76, 25.39 and 9.40 M, respectively. The amount of model parameters is reduced drastically. In summary, it can be seen that using CNNs in Molormer not only improves the overall prediction performance of the model but also is a more lightweight model.

## With or without weight-sharing Siamese network architecture

Deep learning models for DDIs prediction usually need to design two channels to process the features of two drugs,

**Table 4.** Comparison of models with or without CNNs (five random executions)

| Method | Accuracy | Macro precision | Macro recall | Macro F1 | Dimensions | Parameters |
|---|---|---|---|---|---|---|
| Without CNNs for distilling | 0.9455 ± 0.004 | 0.9268 ± 0.004 | 0.8865 ± 0.004 | 0.8993 ± 0.004 | 8192 | 11.76 M |
| Without CNN for decoding | 0.9645 ± 0.004 | 0.9364 ± 0.003 | 0.9205 ± 0.004 | 0.9262 ± 0.004 | 33,280 | 25.39 M |
| With CNNs | 0.9667 ± 0.002 | 0.9419 ± 0.004 | 0.9270 ± 0.002 | 0.9311 ± 0.002 | 2048 | 9.40 M |

**Table 5.** Comparison of models with or without weight-sharing Siamese network architecture (five random executions)

| Method | Accuracy | Macro precision | Macro recall | Macro F1 |
|---|---|---|---|---|
| Without weight-sharing | 0.9608 ± 0.002 | 0.9295 ± 0.011 | 0.9118 ± 0.014 | 0.9204 ± 0.09 |
| With weight-sharing | 0.9667 ± 0.002 | 0.9419 ± 0.004 | 0.9270 ± 0.002 | 0.9311 ± 0.002 |

**Table 6.** Predicted DDI types of drug pairs

| Drug A | Drug B | DDI type (DrugBank v 5.0) | DDI type (new prediction) | Reference |
|---|---|---|---|---|
| Aliskiren | Selexipag | 60 | 49 | Unknown |
| Aliskiren | Furosemide | 75 | 49 | Unknown |
| Aliskiren | Ketoconazole | 73 | 47 | Unknown |
| Selexipag | Sulfamethoxazole | 47 | 20 | Unknown |
| Selexipag | Epoprostenol | 6 | 60 | DrugBank v 5.0 |
| Selexipag | Fluconazole | 47 | 73 | Unknown |
| Vorapaxar | Nimesulide | 6 | 73 | Unknown |
| Vorapaxar | Amiodarone | 73 | 40 | DrugBank v 5.0 |
| Vorapaxar | Argatroban | 49 | 6 | DrugBank v 5.0 |
| Vorapaxar | Clopidogrel | 6 | 47 | Unknown |

6: The metabolism of Drug_B can be decreased when combined with Drug_A. 20: Drug_A may increase the hepatotoxic activities of Drug_B. 40: Drug_A may decrease the sedative activities of Drug_B. 47: Drug_A may increase the antihypertensive activities of Drug_B. 49: Drug_A may increase the antipsychotic activities of Drug_B. 60: Drug_A may increase the hyperglycemic activities of Drug_B. 73: Drug_A may increase the neuromuscular blocking activities of Drug_B. 75: Drug_A may increase the photosensitizing activities of Drug_B.

and the limited number of drugs in the DDIs dataset may result in under-fitting during the training. Siamese network can be used to process features with small differences between the two inputs, and the differences between the two networks are limited to some extent because the weights are shared between them [42]. Based on this property, Siamese network architecture in the DDIs prediction model allows both networks to learn the feature of drug accurately. To further validate the effect of Siamese network on Molormer prediction performance, we designed an ablation study with or without weight-sharing Siamese network architecture, and experimental results are shown in Table 5. According to Table 5, we can observe that the accuracy, macro precision, macro recall and F1 of the model using weight-sharing Siamese network architecture are higher than that without.

## Case study

In the above experiments, we verified that the prediction performance of Molormer is superior and stable, so we use Molormer in this section to predict DDIs for three drugs with a wide range of applications, Aliskiren, Selexipag and Vorapaxar. Aliskiren is the first drug in the renin inhibitor drug class and is used for the treatment of hypertension [43]. Selexipag act as agonists of the prostacyclin receptor to increase vasodilation in the pulmonary circulation and decrease elevated pressure in the blood vessels supplying blood to the lungs [32]. Vorapaxar acts as in the secondary prevention of cardiovascular events in stable patients with peripheral arterial disease or a history of myocardial infarction [44]. The results of the new DDI type predictions are shown in Table 6, where Aliskiren, Selexipag and Vorapaxar predicted three, three and four new interactions, respectively. Among them, the original DDI types in the dataset can be retrieved in DrugBank, three of the predicted DDI types can be validated by DrugBank and the rest of the new DDI types we consider as the results of Molormer recommendations. In summary, we can see that some of the DDI results predicted by Molormer can be verified by the literature, which once again proves the accuracy of the model prediction. Finally, for the completely new DDIs predicted by Molormer, we hope to be provided with some guidance for research and validation in the future.

## Conclusion

In this paper, we propose Molormer, a method based on a lightweight attention mechanism for DDI prediction that can encode the spatial structure of the molecular graph. We use spatial encoding with molecular spatial structure to encoding the molecular graph. Besides, a lightweight-based attention mechanism ProbSparse Self-attention is introduced to process the spatially encoded molecular graph. Finally, we take the Siamese network as the architecture of Molormer to share a weight between two networks processing drug features. Experiments show that our proposed method outperforms other state-of-the-art methods. Our current work still has many potential limitations. The model in this paper uses the 2D spatial structure of the drugs, which leads to excessive time and memory consumption in the data preprocessing part. We suggest that the future solution to the limitations of Molormer can start from two directions: the first is to design lightweight deep learning models to reduce the computation and running time. The second is to design application-specific integrated circuit for running DDIs prediction models to achieve fast computation. We will do further work in the future to overcome these limitations.

**Key Points**

- Accurate and rapid identification of the drug–drug interactions is essential to enhance the effectiveness of combination therapy and avoid unintended side effects.
- We propose Molormer, a deep learning-based method for DDIs prediction. Molormer encodes the molecular graph with spatial structure and uses a lightweight-based attention mechanism to process spatially encoded molecular graph. Finally, Siamese network architecture is used to serve as the architecture of Molormer.
- Experiments show that our proposed method outperforms other state-of-the-art methods in terms of Accuracy, Precision, Recall and F1.
- We designed four ablation studies to verify that spatial encoding, ProbSparse self-attention, self-attention distilling and Siamese network architecture contribute to the overall prediction performance of the model from multiple perspectives.
- In the case study section, we used Molormer to make predictions of new interactions for the drugs Aliskiren, Selexipag and Vorapaxar and validated parts of the predictions.

## Data availability

We provide the Python source code of Molormer model training, which is freely available at https://github.com/IsXudongZhang/Molormer.

## Authors' contributions statement

Conceptualization, X.Z. and X.W.; methodology, X.Z.; software, X.Z.; validation, G.W., X.M. and A.R.-P.; formal analysis, S.W., Y.Z., M.W.; investigation, X.Z and X.W.; resources, X.Z.; data curation, G.W. and X.M.; writing—original draft, X.Z.; writing—review and editing, A.R.-P., S.W. and M.W.; visualization, X.Z.; supervision, S.W. and S.W.; project administration, X.Z.; funding acquisition, A.R.-P. All authors have read and agreed to the published version of the manuscript.

## Acknowledgements

We thank the anonymous reviewers for their valuable suggestions.

## Funding

National Key Research and Development Project of China (2021YFA1000102,2021YFA1000103); Natural Science Foundation of China (Grant Nos. 61873280, 61972416); Taishan Scholarship (tsqn201812029); Foundation of Science and Technology Development of Jinan (201907116); Shandong Provincial Natural Science Foundation (ZR2021QF023); Fundamental Research Funds for the Central Universities (21CX06018A); Spanish project (PID2019-106960GB-I00); Juan de la Cierva (IJC2018-038539-I).

## References

1. Han K, Jeng EE, Hess GT, *et al.* Synergistic drug combinations for cancer identified in a crispr screen for pairwise genetic interactions. *Nat Biotechnol* 2017;**35**(5):463–74.
2. Li Z, Wang R-S, Zhang X-S, *et al.* Detecting drug targets with minimum side effects in metabolic networks. *IET Syst Biol* 2009;**3**(6): 523–33.
3. Zhou D, Miao L, He Y. Position-aware deep multi-task learning for drug–drug interaction extraction. *Artificial intel ligence in medicine* 2018;**87**:1–8.
4. Liu S, Tang B, Chen Q, *et al.* Drug-drug interaction extraction via convolutional neural networks. *Comput Math Methods Med* 2016;**2016**:6146901–8.
5. Hong L, Lin J, Li S, *et al.* A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nat Mach Intell* 2020;**2**(6):347–55.
6. Takeda T, Hao M, Cheng T, *et al.* Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *J Chem* 2017;**9**(1):1–9.
7. Pang S, Zhang Y, Song T, *et al.* Amde: a novel attention-mechanism-based multidimensional feature encoder for drug–drug interaction prediction. *Brief Bioinform* 2021;(1):1.
8. Song T, Zhang X, Mao D, *et al.* DeepFusion: a deep learning based multi-scale feature fusion method for predicting drug-target interactions. *Methods* 2022.
9. Wang S, Jiang M, Zhang S, *et al.* MCN-CPI: multiscale convolutional network for compound-protein interaction prediction. *Biomolecules* 2021;**11**(8):1119.
10. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc Natl Acad Sci* 2018;**115**(18):E4304–11.
11. Huang K, Tianfan F, Glass LM, *et al.* Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* 2020;**36**(22–23):5545–7.
12. Alex K, Ilya S Geoffrey EH. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, Vol. **25**. San Francisco, CA, United States:

Morgan Kaufmann Publishers Inc., Lake Tahoe, Nevada, United States; 2012.

13. Cho K, Van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014.
14. Graves A. Long short-term memory. In: *Supervised Sequence Labelling with Recurrent Neural Networks*. New York, NY, United States: Springer; 2012, 37–45.
15. Justin G, Samuel SS, Patrick FR, *et al.* Neural message passing for quantum chemistry. In: *International Conference on Machine Learning, PMLR*. Massachusetts, United States: JMLR.org, Sydney, NSW, Australia; 2017, 1263–72.
16. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems* 30. San Francisco, CA, United States: Morgan Kaufmann Publishers Inc., Long Beach, California, United States; 2017.
17. Rex Y, Ruining H, Kaifeng C, *et al.* Graph convolutional neural networks for web-scale recommender systems. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, 974–83.
18. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **33**. Palo Alto, CA, United States: AAAI Press, Hawaii, United States; 2019. 7370–7.
19. Xiang W, Xiangnan H, Yixin C, *et al.* Kgat: knowledge graph attention network for recommendation. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, United States: Association for Computing Machinery, Anchorage Alaska, United States; 2019. 950–8.
20. Petar VĆC, Cucurull G, Casanova A, *et al.* Graph attention networks. *Stat* 2017;**1050**:20.
21. Tao Z, Wei Y, Wang X, *et al.* Mgat: multimodal graph attention network for recommendation. *Inf Process Manag* 2020;**57**(5):102277.
22. Li Y, Tarlow D, Brockschmidt M, *et al.* Gated graph sequence neural networks. *arXiv preprint arXiv:151105493* 2015.
23. Beck D, Haffari G, Cohn T. Graph- to-sequence learning using gated graph neural networks. *arXiv preprint arXiv:180609835* 2018.
24. Sun M, Zhao S, Gilvary C, *et al.* Graph convolutional networks for computational drug development and discovery. *Brief Bioinform* 2020;**21**(3):919–35.
25. Cao X, Fan R, Zeng W. Deepdrug: a general graph-based deep learning framework for drug relation prediction. *biorxiv* 2020.
26. Lee G, Park C, Ahn J. Novel deep learning model for more accurate prediction of drug- drug interaction effects. *BMC Bioinformatics* 2019;**20**(1):1–8.
27. Meng X, Wang X, Zhang X, *et al.* A novel attention-mechanism based cox survival model by exploiting pan-cancer empirical genomic information. *Cel ls* 2022;**11**(9):1421.
28. Wang G, Zhang X, Zheng P, *et al.* Multi-transdti: transformer for drug–target interaction prediction based on simple universal dictionaries with multi-view strategy. *Biomolecules* 2022;**12**(5):644.
29. Shenggeng L, Yanjing W, Lingfeng Z, *et al.* MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Brief Bioinform* 2022;**23**(1):bbab421.
30. Ying C, Cai T, Luo S, *et al.* Do transformers really perform badly for graph representation? In: *Advances in Neural Information Processing Systems*, Vol. **34**. San Francisco,CA, United States: Morgan Kaufmann Publishers Inc.; 2021.
31. Hughes LH, Schmitt M, Mou L, *et al.* Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN. *IEEE Geosci Remote Sens Lett* 2018;**15**(5):784–8.
32. Wishart DS, Feunang YD, Guo AC, *et al.* Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.
33. Landrum G, *et al.* Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2013.
34. Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Mach Learn Res* 2020;**21**(140):1–67.
35. Shaw P, Uszkoreit J, Vaswani A. Self- attention with relative position representations. arXiv preprint arXiv:1803.02155. 2018.
36. Haoyi Z, Shanghang Z, Jieqi P, *et al.* Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of AAAI*. Palo Alto, CA, United States: AAAI Press, Vancouver, Canada; 2021.
37. Clevert D-A´ᴇ, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289. 2015.
38. Huang K, Xiao C, Glass LM, *et al.* Moltrans: molecular interaction transformer for drug– target interaction prediction. *Bioinformatics* 2021;**37**(6):830–6.
39. Wishart DS, Knox C, Guo AC, *et al.* Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**(suppl 1):D901–6.
40. Yong Y, Si X, Changhua H, *et al.* A review of recurrent neural networks: Lstm cells and network architectures. *Neural Comput* 2019;**31**(7):1235–70.
41. Zhao Q, Zhao H, Zheng K, *et al.* Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* 2022;**38**(3):655–62.
42. Bromley J, Guyon I, LeCun Y, *et al.* Signature verification using a siamese time delay neural network. In: *Advances in Neural Information Processing Systems*, Vol. **7**. San Francisco,CA, United States: Morgan Kaufmann Publishers Inc.; 1993.
43. Byung-Hee O. Aliskiren, the first in a new class of direct renin inhibitors for hypertension: present and future perspectives. *Expert Opin Pharmacother* 2007;**8**(16):2839–49.
44. Cheng JWM, Colucci V, Howard PA, *et al.* Vincent Colucci, Patricia a Howard, Jean M Nappi, and Sarah a Vorapaxar in atherosclerotic disease management. *Ann Pharmacother* 2015;**49**(5):599–606.