SmileGNN: Drug-Drug Interaction Prediction Based on SMILES and Graph Neural Network

Xueting Han
Harbin Institute of Technology (Shenzhen)
Xutao Li
Harbin Institute of Technology (Shenzhen)
Junyi Li (lijunyi@hit.edu.cn)
Harbin Institute of Technology (Shenzhen)

Research Article

Keywords: drug-drug interaction prediction, graph neural network, knowledge graph, structural features, topological features

Posted Date: June 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-598562/v1

Version of Record: A version of this preprint was published at Life on February 21st, 2022. See the published version at https://doi.org/10.3390/life12020319.

SmileGNN: Drug-Drug Interaction Prediction Based on SMILES and Graph Neural Network

Xueting Han, Xutao Li, Junyi Li*

School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China

*To whom correspondence should be addressed. Email: lijunyi@hit.edu.cn.

Abstract

Background：The use of multiple drugs at the same time can lead to unexpected adverse drug reactions. The interaction between drugs can be confirmed by routine in vitro and clinical trials. But it is difficult to test the drug-drug interaction widely and effectively before the drug is put into market. Therefore, the prediction of drug-drug interaction has become an important research in biomedical field.

Results：In recent years, researchers have used deep learning to predict drug-drug interaction by using drug structural features and graph theory, and they have achieved a series of achievements. A drug-drug interaction prediction model SmileGNN is proposed in this paper. The structural features of drugs are constructed by using SMILES data. The topological features of drugs in knowledge graph are obtained by graph neural network. The structural and topological features of drugs are aggregated to predict the interaction of new drug pairs.

Conclusions：The experimental results show that the model proposed in this paper combines a variety of data sources, and has better prediction performance compared with the existing prediction model of drug-drug interaction prediction. The most striking result is that five out of top ten predicted new interaction of drugs are verified from the latest database, which proves the credibility of SmileGNN.

Keywords: drug-drug interaction prediction, graph neural network, knowledge graph, structural features, topological features.

1 Introduction

Drug-drug Interaction (DDI) is one of the focuses of biomedical research. For many diseases with complex pathways of action, single drugs may not be

ideal for treatment. One solution is combination drug therapy, which uses several drugs at the same time. For instance, Venetoclax and Idasanutlin are used together to treat leukemia. Venetoclax inhibits the anti-apoptotic Bcl-2 family protein and Idasanutlin activates the p53 pathway, which are effective [1]. However, concurrent use of multiple drugs may lead to Adverse Drug Events (ADEs) [2,3]. Routine in vitro and clinical trials confirm DDIs. But it is difficult to test DDIs extensively and effectively before they are marketed, because it is impossible to test every two drugs for DDIs considering the large number of drugs and the time and cost of validation. At the same time, due to the fact that ADEs are not always reported and counted in time after the occurrence, there are relatively few documented and verified DDIs compared with the large number of drugs.

At present, DDI prediction methods are mainly divided into two categories: drug stuctural feature based approach and graph-based approach.

The drug stuctural feature based approach assumes that chemically similar drugs have similar DDIs. Ryu et al. [4] proposed DeepDDI model, which is the first model to use deep learning in drug-drug predition. Structural Similarity Profiles (SSP) of pairs of drugs are generated by using SMILES (Simplified Molecular Input Line Entry Specification) data of the drugs, which are then sent into Deep Neural Network (DNN) for classification through PCA dimensification reduction. On the basis of DeepDDI, Lee et al. [5] added two new data with the method similar to the SSP generated by drugs' SMILES data: Target Gene data to generate TSP (Target Similarity Profile) and Gene Ontology (GO) to generate GSP (Gene Ontology Term Similarity Profile). These three feature vectors are reduced in dimension by an improved encoder, and then stitch into a single feature vector for the drug pair, which is put into DNN for training. This improved model uses more data and has higher accuracy. Based on the DeepDDI, a polymorphic deep learning model was proposed by Deng et al. [6], which uses the complete information screened for training. It can use the information related to a variety of drugs to learn more efficiently and has a higher accuracy. The methods based on drug features have high accuracy on known data sets, but they have some limitations. The hypothesis that "drugs with similar chemical structures have similar DDIs" has not been scientifically verified. Thus, there may be a large deviation in the prediction results in actual clinical verification.

In recent years, a series of studies on the application of graph theory in molecular level have achieved successful results, and many researchers are trying to use graph theory to analyze DDI prediction. Marinka et al. [7] proposed the model Decagon, which is a two-layer heterograph. It was constructed to predict the type of polypharmacy side effects of drug pairs whose drug targets are all proteins. In this study, the Graph Neural Network (GNN) was used to train the model by Graph representation learning, and it was shown that the GNN has better performance in predicting DDIs than the traditional shallow Graph structure model and the traditional graph embedding method. Bougiatiotis et al. [8] extracted the three dimensional relationships related to a specific disease from various databases, and expressed them with The Unified Medical Language System (UMLS) to construct multiple knowledge graphs (KG) for specific diseases. The model DDI-BLKG extracts drug features based on its pathways, which has a certain enlightenment for the prediction of DDIs. Lin et al. [9] extracted a large number of drug-related data from the database, and processed data into triples. The triples are encoded to construct a huge KG. The feature vector of the drug is generated through two times of aggregation by GNN. Thus, the vector includes not only the information of the drug itself, but also the information of drug-related entities. The method based on graph models drug action pathway and other data, and uses deep learning and other methods to make training prediction. The graph-based method has a good explanatory ability but sometimes neglects the information contained in the entities.

Graph Neural Network (GNN) extends the convolutional neural network to non-Euclidean space, which provides a more natural and effective method for the modeling of graph structured data [10]. GNN can be regarded as an embedding method, which extracts the embedding vectors of adjacent nodes for updating its own embedding vectors, without the need for manual feature engineering [11].

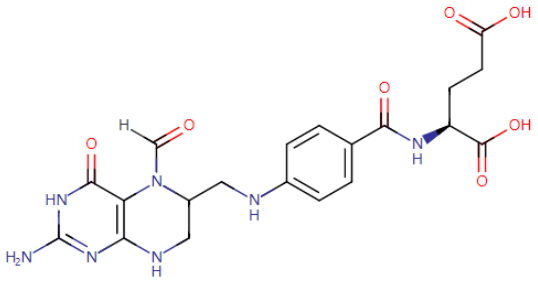
Knowledge Graph (KG), as a knowledge representation and management method, was proposed by Google in 2012. In recent years, KG has become popular in academia and industry, and its use has expanded from the search engine field to all fields involving big data [18]. KG is a kind of data structure based on graph, and is usually represented as triples, $\mathrm{G} = (\mathrm{head}, \mathrm{relation}, \mathrm{tail})$. Head and tail are the head entity and tail entity respectively, which are different entities from the form of web pages. Relation, on the other hand, is

the relation in the knowledge base, which is transformed from hyperlink to web page into semantic relation between entities.

2 Methodology

2.1 Drug Structural Features

The data sources for this topic is DrugBank [13]. DrugBank is a drug knowledge database that describes clinical information on drugs, such as side effects, DDIs, etc. DrugBank also provides data on the molecular level, such as the chemical structure of the drug, the target protein of the drug, etc. SMILES (Simplified Molecular Input Line Entry Specification) is a specification that explicitly describes molecular structures using ASCII strings. SMILES can describe a three-dimensional chemical structure with a string of characters, as shown in Figure 1 is a two-dimensional graph of the drug Leucovorin and its corresponding SMILES. SMILES can be imported by molecular editing software and converted into two-dimensional graphics or three-dimensional models of molecules.



[H]C(=O)N1C(CNC2=CC=C(C=C2)C(=O)N[C@@H](CCC(O)=O)C(O)=O)CNC2=C1C(=O)NC(N)=N2,

Figure 1 Two-dimensional graphs of the drug Leucovorin and its corresponding SMILES

SMILES2Vec [14] method was proposed to apply Seq2seq [15] technology in natural language processing to SMILES string, in which the chemical structure information is used as input variable into the deep neural network to predict the physical properties of compounds. SMILES2Vec removes some of the long (more than 250 letters) SMILES during preprocessing, and conducts

one-hot coding on the remaining SMILES, converting each SMILES into a vector of length 26. According to this pretreatment method, the chemical structure of the drug was pretreated, as shown in Figure 2.

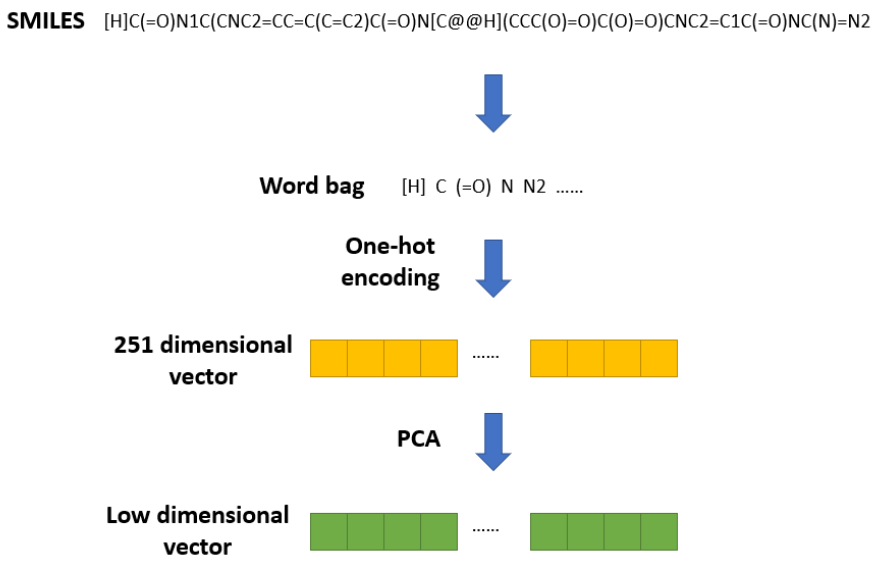


Figure 2 Pretreatment methods of SMILES

All the SMILES stored in DrugBank are converted into a word bag with 251 elements in it. Then one-hot encoding is used to transform them into 251 dimensional vectors. Finally, PCA is used to reduce the 251-dimensional SMILES vectors to a specific dimension, that is, a vector of lower dimension used to represent the structure of a drug.

2.2 Drug Topological Features

Construction of KG. The data from two databases are used to construct KG, which are used to obtain the topological features of the drugs on the corresponding one. Kyoto Encyclopedia of Genes and Genomes (KEGG) [16] is a database resource for understanding advanced functions and utilities of biological systems from molecular level information. There are multiple sub-

databases under KEGG. Wang et al. [17] constructed a large, high-quality heterogeneous map linking Patient, Disease, and Drug (PDD) in Electronic Medical Record (EMR). PDD database extracts key medical entities from MIMIC-III (Medical Information Mart for Intensive Care III) [18] and linked them to current biomedical knowledge graphs (including ICD-9 Ontology and DrugBank). PDD diagrams are accessible on the web through SPARQL endpoints and provide information for medical research and treatment recommendations.

RDF (Resource Description Framework) [19] is a resource description language commonly used as a representation of the KG. Bio2RDF project [20] provides tools to convert data to n-quads or other formats of RDF. Then, the RDFlib library is used to parse these n-quads data and divide them into triples (entity, relationship, entity) in a format that is convenient for subsequent KG to generate embedded features, as shown in Figure 3.

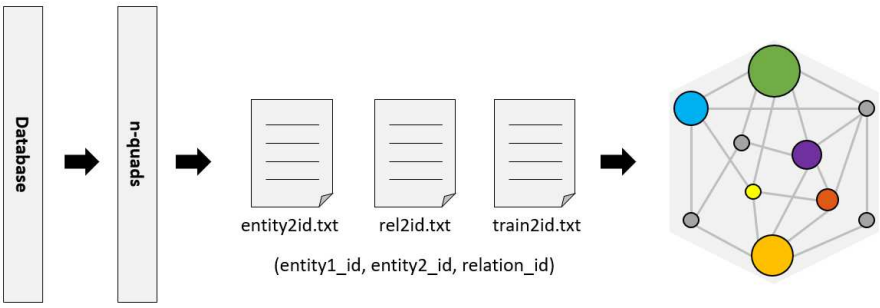


Figure 3 KG Construction

Here we introduce a metric to evaluate the KG. Density is used to describe the density of edge connections between nodes in a graph/network. For a graph G with L edges and N nodes, the density calculation formula is shown in (1) :

$$d(G) = \frac{2L}{N(N-1)} \tag{1}$$

The density of the graph has a certain influence on the results of graph-based research and machine learning. This will be discussed in subsequent experiments.

We construct two KGs by KEGG and PDD respectively. The corresponding data is shown in the following Table 1.

Table 1 Comparison of KEGG KG and PDD KG

	KEGG	PDD
Number of drugs	11,174	1,495
The proportion of drugs with structural records	13.96%	72.37%
The density of graph	4.300 e-5	8.571 e-4
Number of positive samples	56983	36768
Drug interaction subgraph density	9.128 e-4	3.292 e-2

It can be seen from the table that there are more types of drugs in KEGG data set, but the graph itself is relatively sparse, and the proportion of drugs with structure records is relatively low. PDD dataset has fewer drug types, but the graph is more dense, and the proportion of drugs with structure records is higher.

It should be noted that the positive and negative samples in the experiment are not the result of manual labeling, but come from the existing data in the database. For negative samples, this paper believes that there is no DDI between the two drugs in the experiment, but in fact there may be DDI between them, which has not been clinically verified, so it has not been recorded in the database.

Extraction of topological features. Generally, the models that use KG to predict DDIs can only capture data information in a small range. In order to expand the receptive field, obtain the rich entity information in the KG, and explore the potential correlation between drugs and other entities, KGNN [9] model was proposed. KGNN extracts the higher-order structure and semantic relations of drugs by GNN, and learns the representation of drugs and

their neighborhoods from the KG. We use KGNN model to calculate the topological features of drugs on the KG, as shown in Figure 4. For each entity, the model extracts several entities from the domain of the entity and aggregates the information of these entities to form the topological feature representation of the entity. There are three kinds of entity aggregation methods: sum aggregation is a superposition operation, concatenate is a concatenate operation, neighbor only considers the neighborhood without considering the information of the node itself. These three aggregation methods are abbreviated as sum, concat and neigh.

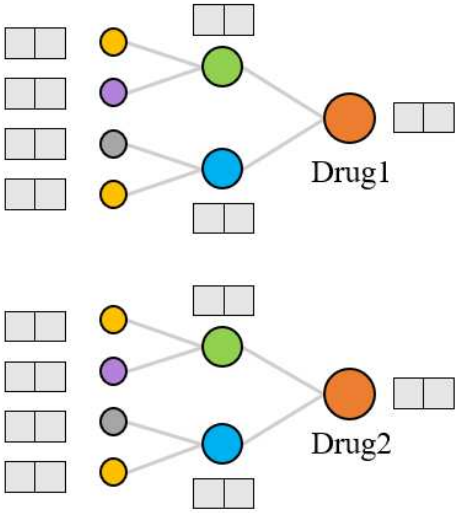


Figure 4 Extraction of topological features

2.3 Drug-drug Interaction Prediction

We consider using GNN to obtain drug topological features on the KG, and fuse drug structural features into the model to study the influence of drug structural features on DDI prediction. Hence, we propose a novel model SmileGNN as shown in Figure 5.

The algorithm can be summarized as follows. The method SMILES2Vec mentioned in section 2.1 is used to calculate the feature vector of drug

structure by using the data of SMILES. The KGNN model is retained to calculate the drug topological features, in which the graph neural network (GNN) is used to aggregate the entity information of the receptive field within two hops of the entity to obtain the drug topological features. Then the two features of the drug are aggregated to obtain a comprehensive drug features including drug topological features and drug structural features. Two algorithms are specifically designed to aggregate drug structural features and drug topological features. See section 3.4 for detailed algorithms and comparative analysis.

On this basis, the interaction value between the two drugs is calculated. After passing the threshold value of 0.5, it is classified as the presence of DDI or the absence of DDI.

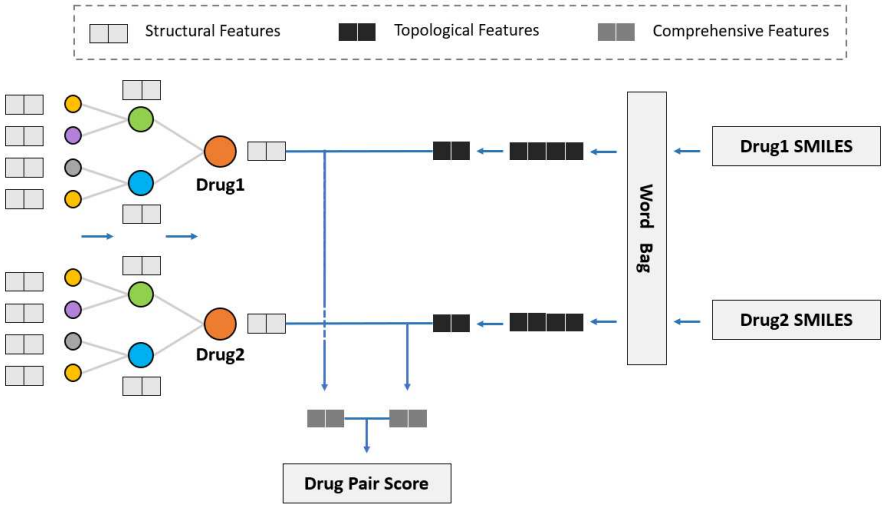


Figure 5 SmileGNN model

This model uses the dichotomous cross loss entropy as the loss function, and its calculation formula is shown in (2) :

$$Loss = \sum_{(i,j)\in Y(i,j\in N_d, j\neq i)} -y_{i,j} log\widehat{y_{i,j}} - (1-y_{i,j})\log(1-\widehat{y_{i,j}}) \quad (2)$$

Where, $\widehat{y_{i,j}}$ represents the predicted value, $y_{i,j}$represents the true value of drug pairs in the data set, and Y represents the set of all drug pairs.

3 Experiment

3.1 Experimental Settings

In this paper, the prediction of DDI is considered to be a binary task. It is not necessary to specifically predict the type of DDI and what side effects the DDI may cause, but only to judge whether there is a possible DDI between the drug pair.

Metrics. ACC (Accuracy) and AUC (Area Under Curve) are mainly used as evaluation indexes for a series of models. In some comparative experiments, F1 is also used as an evaluation index.

Settings. The experiment is conducted on two datasets, KEGG and PDD. See section 2.2 for the construction and data features of the dataset. For the two datasets, a parameter combination that achieves the highest AUC value is adopted through parameter tuning based on grid search. The final parameters to be used are shown in Table 2.

Table 2 Experimental parameters

	KEGG	PDD
batch size	2048	1024
learning rate	2e-2	1e-2
GNN embed dimension	32	64

Baselines. In addition to KGNN, two classic models, DeepDDI and Decagon are compared with the new model proposed in this paper. See Section 1 for a detailed introduction of the models.

- **DeepDDI** [4]: DeepDDI model is based on the drug stuctural feature method, and is the first to use deep neural network to predict DDI. The model was put forward in 2017, and established the Gold Standard Database (Gold Standard Database) of DDIs. DeepDDI is considered as a benchmark among stuctural feature methods.

- **Decagon** [7]: Decagon model is the first model using graph neural network among graph-based methods. This model was proposed in 2018, and is a model with great influence among graph-based methods in recent years.
- **KGNN** [9]: The usage of KG and GNN to predict DDI can mine the potential correlations between drugs and other entities.

3.2 Results and Analysis

The experimental results of these models are compared and analyzed, as shown in Table 3.

Table 3 Comparative analysis of the new model and several classical models

model	The data source	ACC	AUC
DeepDDI	KEGG	0.8217	0.8987
Decagon	STITCH, etc.	--	0.8720
KGNN	KEGG	0.8834	0.9422
SmileGNN	KEGG	**0.8936**	**0.9521**

SmileGNN achieves the best performance among all the models. Compared to the classic DeepDDI and Decagon models, there is a 5.3% and 8.0% improvement in AUC values, respectively. Compared with the KGNN model using drug topological features alone, it also has a certain performance improvement.

According to the experimental results, both DeepDDI model and Decagon model are the pioneer models in the field of DDI prediction. However, the model designs still need to be improved, and their prediction performance is relatively poor. In the graph-based method, both Decagon model and KGNN model only use the topological features of the drug, but KGNN not only considers the topological features of the current node of the drug, but also the topological features of the node in the neighborhood of the drug within a certain range, so more information can be learned from the graph and the

effect is improved more than that of Decagon model. The new model SmileGNN proposed in this paper combines the topological features and structural features of the drug, and has a better performance than the Decagon and KGNN models using topological features alone or the DeepDDI model using structural features alone.

SmileGNN model retains the method of KGNN model in learning drug topological features, and has excellent performance. However, in terms of the learning of drug structural features, the model proposed in this paper deals with SMILES in a relatively independent and rough way. In the future, the feature expression algorithm of drug structural features can be further optimized to improve the prediction ability of the model.

3.3 Ablation Study

SmileGNN adds the use of drug structural features to KGNN, and integrates multi-source information to predict new DDIs. The comparison with the original performance of the KGNN model [9] is an ablation experiment, so as to compare and analyze the influence of the new drug structural features on the model performance.

Experiments are carried out in KEGG and PDD dataset to conduct experiments on the three drug topological feature aggregation types of sum, concat and Neigh respectively. The experimental results are shown in Table 4.

Table 4-1 Comparison of the performance of SmileGNN and KGNN on KEGG dataset

Model	Aggregator type	Average Accuracy	Average AUC	Average F1
KGNN	sum	0.8801	0.9390	0.8851
	concat	**0.8834**	**0.9422**	**0.8881**
	neigh	0.8642	0.9267	0.8690
	Average	0.8759	0.9360	0.8807
SmileGNN	sum	0.8888	0.9467	0.8943

	concat	**0.8936**	**0.9521**	**0.8957**
	neigh	0.8744	0.9329	0.8788
	Average	0.8856	0.9439	0.8896

Table 4-2 Comparison of the performance of SmileGNN and KGNN on PDD dataset

Model	Aggregator type	Average Accuracy	Average AUC	Average F1
KGNN	sum	0.8920	0.9542	0.8947
	concat	**0.8970**	**0.9576**	**0.8995**
	neigh	0.8896	0.9518	0.8919
	Average	0.8929	0.9545	0.8954
SmileGNN	sum	0.9040	0.9618	0.9056
	concat	**0.9065**	**0.9642**	**0.9084**
	neigh	0.9000	0.9613	0.9018
	Average	0.9035	0.9624	0.9053

For both KEGG and PDD datasets, the performance of SmileGNN, which uses drug structural features, is better than that of KGNN in all three kinds of aggregation methods of drug topological features. Conconsistent with the KGNN model, SmileGNN has the best effect when using concat for obtaining drug topological features, with the AUC value reaching 0.9521 and 0.9642 in KEGG and PDD dataset respectively. It proves that the newly added drug structural features can steadily improve the performance of the model.

By comparing Table 4-1 with Table 4-2, it can be found that the performance of both KGNN and SmileGNN models on PDD dataset is better than

that using KEGG dataset. As for the improvement of model performance after adding SMILES, PDD dataset has the same degree of improvement on the original good results, with about 1% improvement in ACC, AUC and F1 value.

According to the comparison of KEGG and PDD datasets in section 2.2, the following conclusions can be basically drawn:

1. On the denser graph, the drug topology information learned from the model is richer and can better represent the drug topological features.

2. In PDD data, there is a higher proportion of drugs corresponding to drug structure, and drug structural features have a greater influence on the model, which is positive.

Due to the limitations of the dataset itself, that is, the drug pairs in the dataset are classified as drug pairs without DDIs, but may also have DDIs. Therefore, the predicted results of the model cannot be infinitely close to 1, and the excellent performance obtained in training and cross-validation does not explain everything. In section 4, special attention is paid to drug pairs that are classified "incorrectly", i.e., those that the datasets records as non-DDIs but the model predicts as DDIs.

3.4 Case Study

- **Influence of drug feature aggregation method.**

Referring to the ways that KGNN designed to aggregate the topological features of multiple nodes together, methods sum and concat are designed to aggregate the structural features and topological features of drugs together by corresponding addition and connection.

Given two matrices as input: drug topological feature matrix A, whose shape is $BatchSize * EmbedDimension_A$; drug structural feature matrix B, whose shape is $BatchSize * EmbedDimension_B$. For the sum method, the weight matrix W of the shape $EmbedDimension_A * EmbedDimension_A$ is designed, and the bias vector b. Notice that the matrices A and B have to have the same shape. Output is shown in formula (3). For concat method, the

weight matrix W of the shape $(EmbedDimension_A + EmbedDimension_B) * EmbedDimension_A$ is designed, and the bias vector b. Output is shown in formula (4).

$$\tanh([A + B] * W + b) \quad (3)$$

$$\tanh([A, B] * W + b) \quad (4)$$

For PDD dataset, when other parameters are unchanged, the drug topological feature dimension is set as 64 dimension, so is the drug structural feature dimension. The two aggregation methods are used to obtain drug features, and the other parameters are consistent. The experimental results are shown in Table 5.

Table 5 Different aggregation methods on PDD dataset

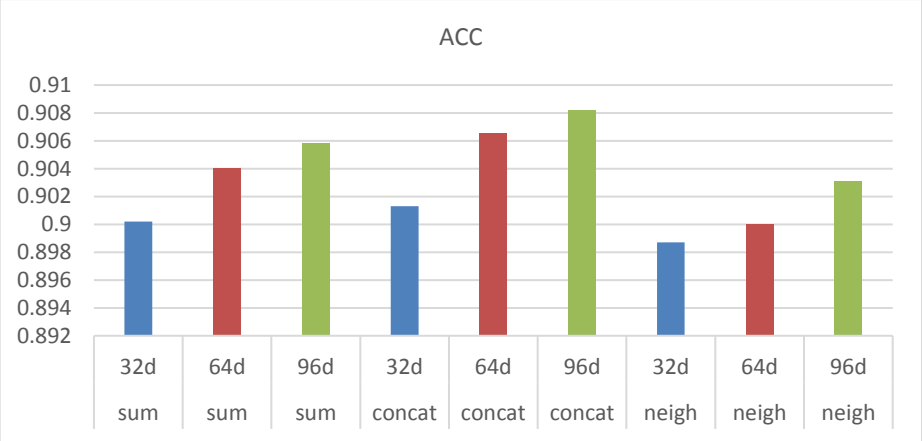
	ACC	AUC	F1
sum	**0.9059**	**0.9647**	**0.9070**
concat	0.9056	0.9618	0.9040

As can be seen from Table 5, when sum and concat are used to aggregate drug topological features and drug structural features, the performance of sum method is slightly better than that of concat method, but the performance gap is not significant. In view of the fact that concat method is more flexible and has no requirement on feature dimension, subsequent experiments all adopted concat method.

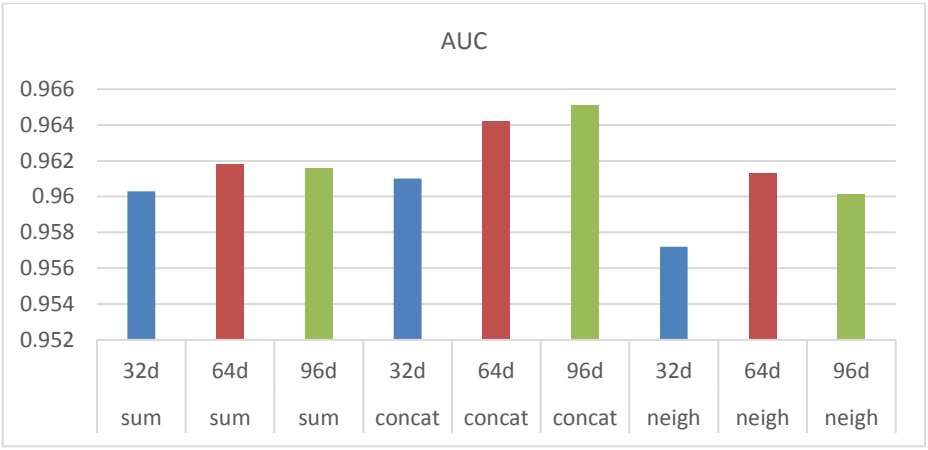
- **Influence of drug structural feature dimension.**

To measure the influence of drug structural feature dimension on the result of model training, and study the loss of PCA dimension reduction method, we conduct the following experiment. The model unified concat method is used to connect drugs topological features and structural features, using the PDD dataset, set the PCA dimension reduction of drug structure dimension respectively 32 d, 64 d, 96 d. Other parameters remain the same.

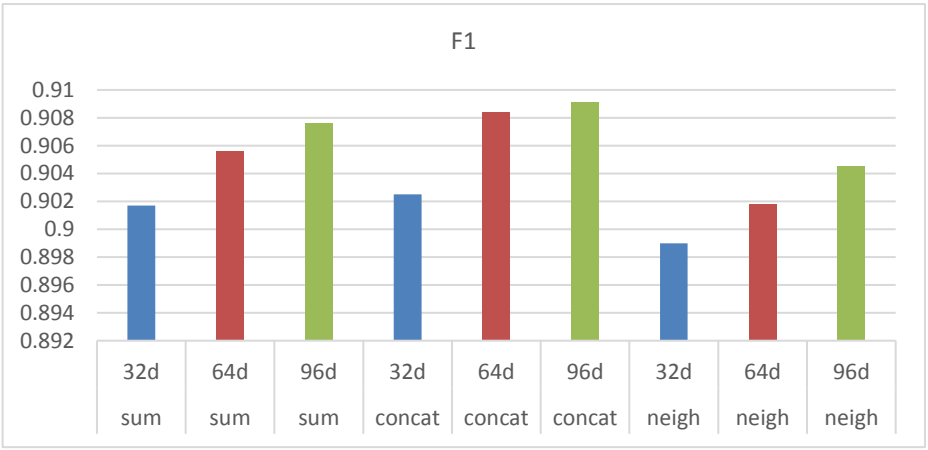
Among them, three methods of sum, concat and neigh are used to obtain drug topological features respectively, in order to observe whether the influence of drug structural feature dimensions is stable and consistent. As shown in Figure 6, with the enhancement of drug structure characteristic dimension from 32 dimension to 64 dimension, the effect of sum, concat and neigh models, three aggregation methods for drug structures, all improve slightly, indicating that the influence of drug structural feature dimension on model performance is stable and consistent. Note that when the drug structure dimension was increased from 64 to 96 dimensions, the performance of the model is not significantly improved.



(a)



(b)



(c)

Figure 6 Influence of drug structure characteristic dimension on model performance

In conclusion, when PCA is used to reduce the dimension of drug structural features, the effect of dimension reduction is better and the information loss is smaller in the process of decreasing from 251 dimension to 64 dimension. When the dimension is further reduced, the representation of drug structural features may be greatly lost, and the performance of the final model will be affected. Considering when using 64-dimensional drug structural features, the model has already had a relatively good performance,

while the use of a higher dimension of drug structural features will occupy more computing resources and storage space, and the performance improvement is not obvious, so the experiments uniformly use 64-dimensional drug structural features.

4 Discussion

Instead of sending the score of the drug pairs into the threshold category of 0.5, the drug pairs with a score over 0.9 are directly printed and ranked from highest to lowest. To get better result, use PDD dataset to get drugs pairs classified as DDIs, and eliminate pairs which are recorded with DDIs in PDD. Get the highest score of the top ten new prediction of DDIs, and send the results to the latest DrugBank database query. The ones that are recorded as DDI in DrugBank are marked as 1, otherwise marked as 0, as shown in Table 6.

The PDD dataset is updated to version 1.3 and was uploaded in October 2018. The DDIs in the PDD dataset were extracted from version 5.1.1 of DrugBank, which was uploaded in July 2018. The latest DrugBank database is version 5.1.8 uploaded in January 2021. So there's a 2.5-year gap, during which many new DDIs are discovered and verified.

Table 6 New DDI

drug1	drug2	score	Whether you can query DDI in DrugBank
DB00437	DB09322	0.999964	0
DB00450	DB00768	0.999917	0
DB00437	DB00959	0.999854	0
DB00660	DB01656	0.999831	1
DB00722	DB01039	0.999817	1
DB00437	DB00633	0.999764	1
DB00346	DB01173	0.999618	1

DB04908	DB05521	0.999571	1
DB00475	DB00820	0.999542	0
DB00040	DB00564	0.999236	0

It can be seen that the five new DDIs shown in Table 6 are the latest ones that have been clinically verified and included in DrugBank database in recent two years, while the remaining five DDIs have not been experimentally verified yet. The model proposed in this paper is reliable for the prediction of novel DDIs, and the experimental results are of great auxiliary significance for clinical trials of novel DDIs.

In the following, two drug pairs are studied separately, and the influence of drug structural features and drug topological features on drug pair interaction prediction is discussed. It can be seen that both drug pairs [DB00437, DB00959] and [DB00437, DB00633] have high scores above 0.99, and both contain drug DB00437.

According to the SSP calculated in DeepDDI [4], it is known that the structural similarity between drug DB00959 and drug DB00633 is only about 35.19%, which does not have a high similarity. However, only 72% of the drugs in the PDD data set have SMILES data, so about 48% of the drug pairs cannot be directly calculated for their structural similarity. In the context of sparse data, 35.19% similarity also has a greater impact on the results.

In the drug targeting data, it is found that both drug DB00959 and drug DB00633 act on Cytochromes P450 group protein enzymes. Due to the similar pathway of action, the model is more inclined to believe that drug DB00959 and drug DB00437 also have DDIs. DDI records in the DrugBank database show that the adverse drug event of drug combination [DB00437, DB00633] is due to competition for the excretory pathway of the kidney [21]. Based on the relevant information in literature and a series of databases, it is believed that the interaction mechanism of this drug pair is not obviously related to the protein enzymes of Cytochromes P450 group [22,23].

Through the study of this example, it is realized that SmileGNN can make good use of the known drug structural information and drug topological information to predict DDIs. However, on the one hand, the model is limited by

insufficient information of drug structure; On the other hand, the learning of topological information in KG is relatively blind and random. In the future, this model still has some room for improvement in learning drug features.

5 Conclusion

In this paper, a new model SmileGNN (model based on SMILES and graph neural network) is proposed to predict drug-drug interactions by comprehensively using drug structural features and drug topological features. We implement the proposed method and conduct experimental comparisons on two datasets. Through experiments, it is verified that SmileGNN has better performance than the classic models and KGNN. According to the latest database, SmileGNN's prediction results are credible.

Declarations

Abbreviations

DDI: Drug-drug interaction; GNN: Graph neual network; KG: Knowledge graph;

ACC: Accuracy; AUC: Area under the curve.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the codes are available at: https://github.com/AshleyHan/SmileGNN

Competing interests.

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' contributions.

JL and XH designed the study. XH performed bioinformatics analysis and drafted the manuscript. XL helped to revise the manuscript. JL conceived of the study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements.

This work was supported by the grants from the National Key Research Program (2017YFC1201201, 2018YFC0910504 and 2017YFC0907503), Shenzhen Science and Technology Program the university stable support program (2021) from JL, NSFC under Grant No.61972111 from XL.

References

1. Pan R, Ruvolo V, Mu H, et al. Synthetic lethality of combined Bcl-2 inhibition and p53 activation in AML: mechanisms and superior antileukemic efficacy[J]. Cancer cell, 2017, 32(6): 748-760. e6.
2. Edwards I R, Aronson J K. Adverse drug reactions: definitions, diagnosis, and management[J]. The lancet, 2000, 356(9237): 1255-1259.
3. Bansal M, Yang J, Karan C, et al. A community computational challenge to predict the activity of pairs of compounds[J]. Nature biotechnology, 2014, 32(12): 1213-1222.
4. Ryu J Y, Kim H U, Lee S Y. Deep learning improves prediction of drug–drug and drug–food interactions[J]. Proceedings of the National Academy of Sciences, 2018, 115(18): E4304-E4311.
5. Lee G , Park C , Ahn J . Novel deep learning model for more accurate prediction of drug-drug interaction effects[J]. BMC Bioinformatics, 2019, 20(1):415.
6. Yifan D , Xinran X , Yang Q , et al. A multimodal deep learning framework for predicting drug–drug interaction events[J]. Bioinformatics, 2020, (15):15.
7. Marinka Z , Monica A , Jure L . Modeling polypharmacy side effects with graph convolutional networks[J]. Bioinformatics, 2018, 34(13): i457-i466.

8. Bougiatiotis K, Aisopos F, Nentidis A, et al. Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph[C]//International Conference on Artificial Intelligence in Medicine. Springer, Cham, 2020: 122-132.

9. Lin X , Quan Z , Wang Z J , et al. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction[C]// Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20. 2020.

10. Defferrard M , Bresson X , Vandergheynst P . Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering[J]. 2016, 30: 3844–3852.

11. Hamilton W L, Ying R, Leskovec J. Representation learning on graphs: Methods and applications[J]. IEEE Data Eng. Bull., 2017, 40, 52–74.

12. Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]//International Semantic Web Conference. Springer, Berlin, Heidelberg, 2013: 542-557.

13. Wishart D S, Feunang Y D, Guo A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J]. Nucleic acids research, 2018, 46(D1): D1074-D1082.

14. Goh G B, Hodas N O, Siegel C, et al. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties[J]. arXiv preprint arXiv:1712.02034, 2017.

15. Xu Z, Wang S, Zhu F, et al. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery[C]//Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics. 2017: 285-294.

16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30.

17. Wang M, Zhang J, Liu J, et al. Pdd graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking[C]//International Semantic Web Conference. Springer, Cham, 2017: 219-227.

18. Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific data, 2016, 3(1): 1-9.

19. Lassila O, Swick R R. Resource description framework (RDF) model and syntax specification[J]. 1998.

20. Belleau F, Nolin M A, Tourigny N, et al. Bio2RDF: towards a mashup to build bioinformatics knowledge systems[J]. Journal of biomedical informatics, 2008, 41(5): 706-716.

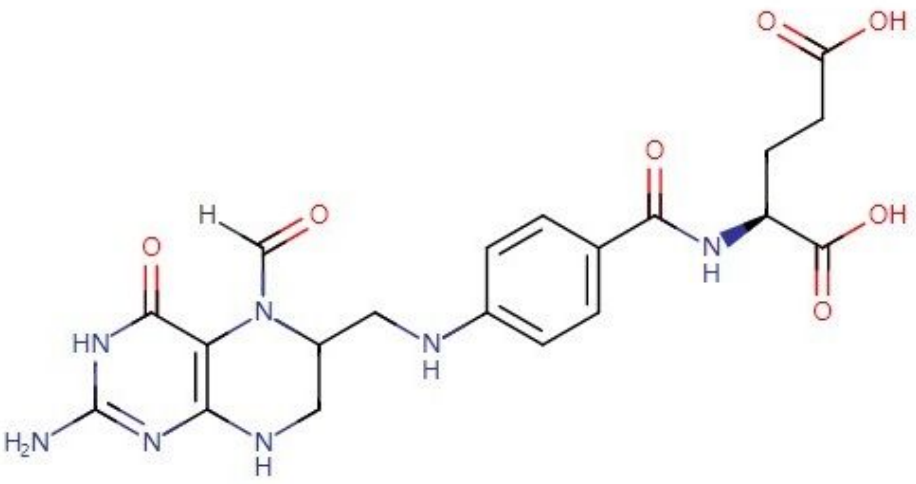
21. Van Ginneken C A M, Russel F G M. Saturable pharmacokinetics in the renal excretion of drugs[J]. Clinical pharmacokinetics, 1989, 16(1): 38-54.

22. Avsaroglu H, Bull S, Maas-Bakker R F, et al. Differences in hepatic cytochrome P450 activity correlate with the strain-specific biotransformation of medetomidine in AX/JU and

IIIVO/JU inbred rabbits[J]. Journal of veterinary pharmacology and therapeutics, 2008, 31(4): 368-377.

23. Duhamel M C, Troncy É, Beaudry F. Metabolic stability and determination of cytochrome P450 isoenzymes' contribution to the metabolism of medetomidine in dog liver microsomes[J]. Biomedical Chromatography, 2010, 24(8): 868-877.

Figures



[H]C(=O)N1C(CNC2=CC=C(C=C2)C(=O)N[C@@H](CCC(O)=O)C(O)=O)CNC2=C1C(=O)NC(N)=N2

Figure 1

Two-dimensional graphs of the drug Leucovorin and its corresponding SMILES

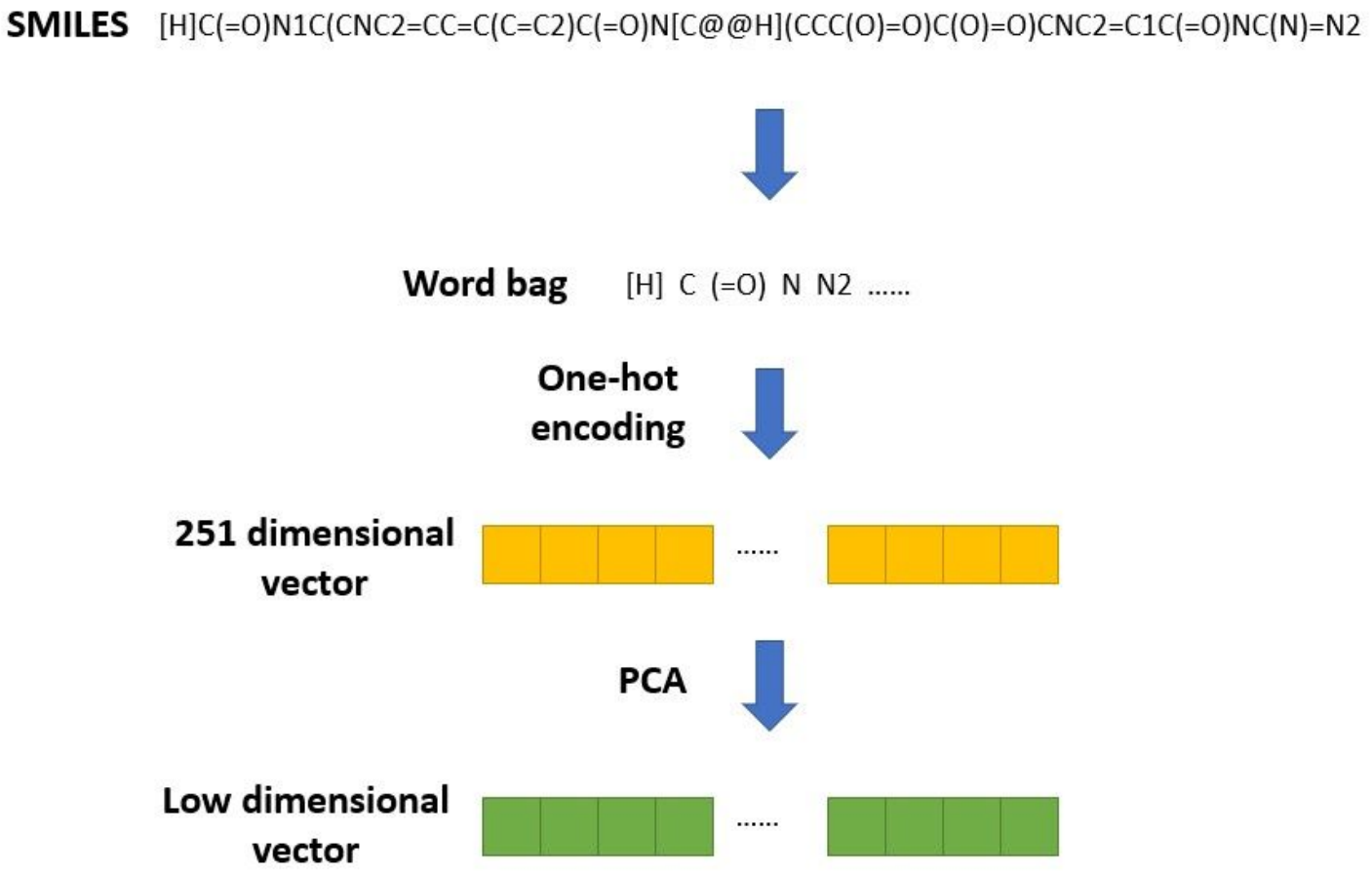


Figure 2

Pretreatment methods of SMILES

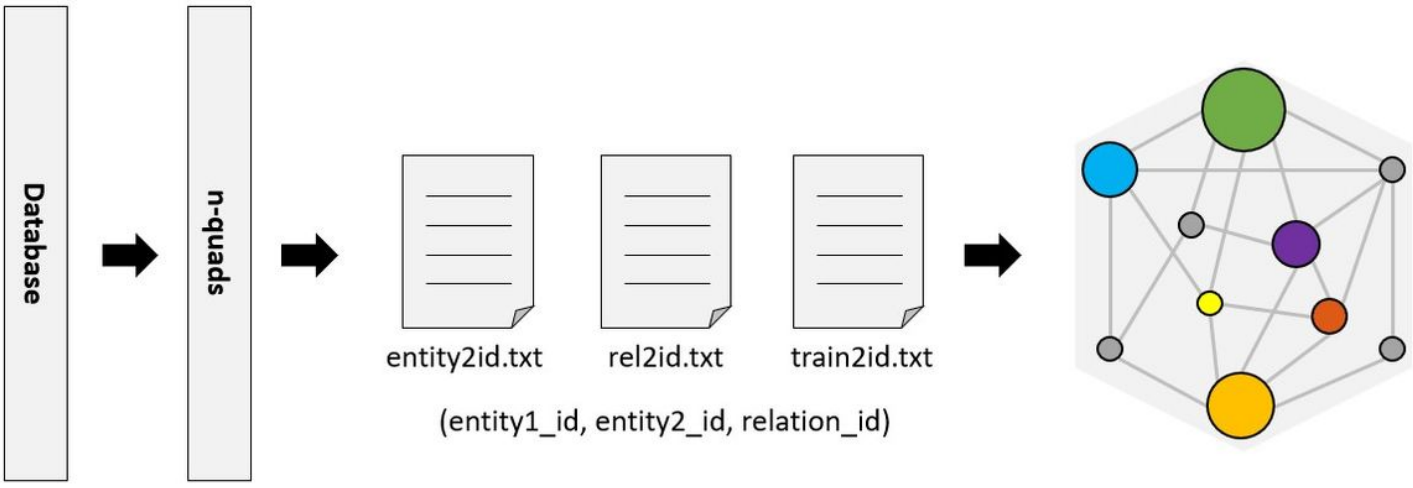


Figure 3

KG Construction

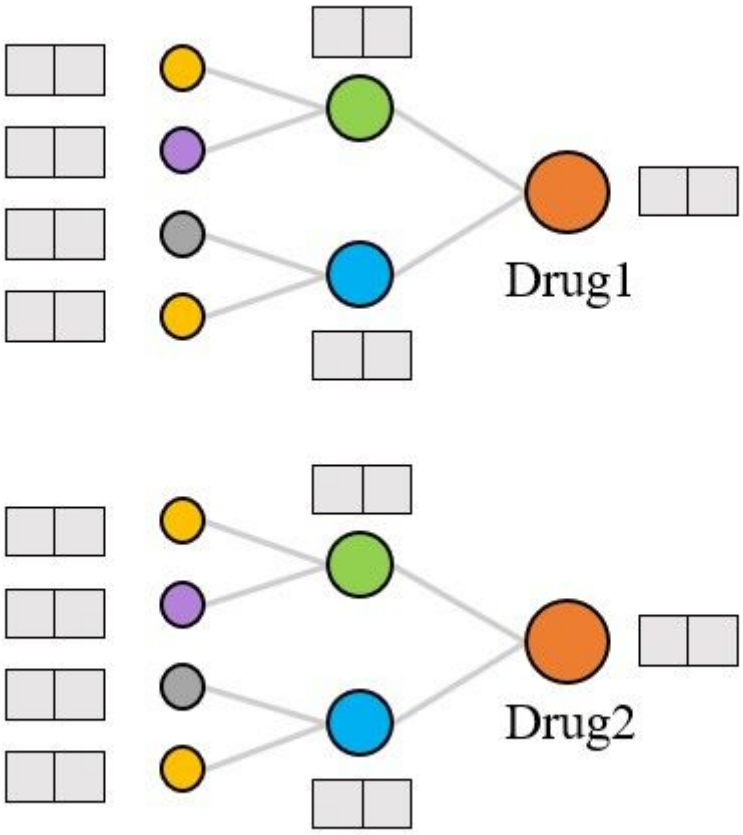


Figure 4

Extraction of topological features

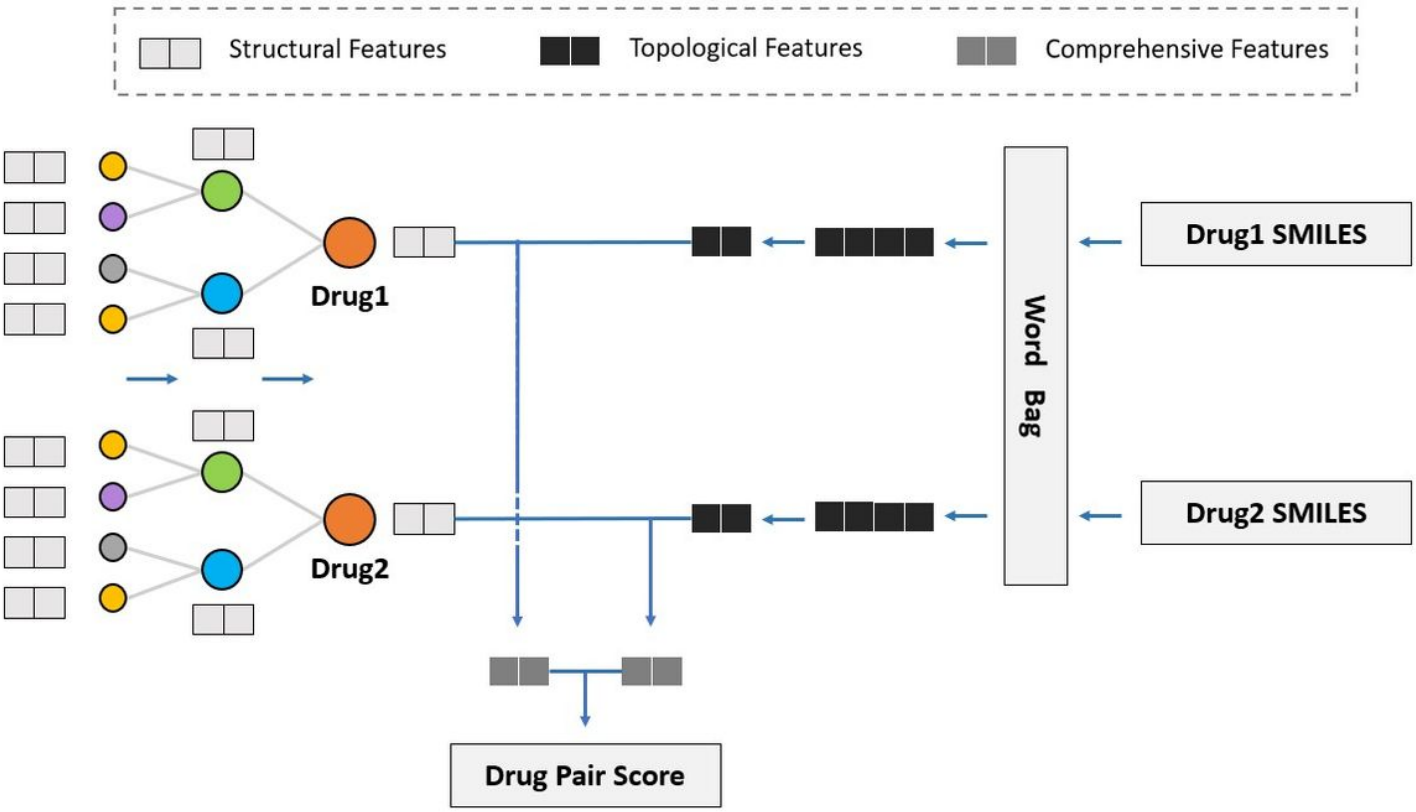
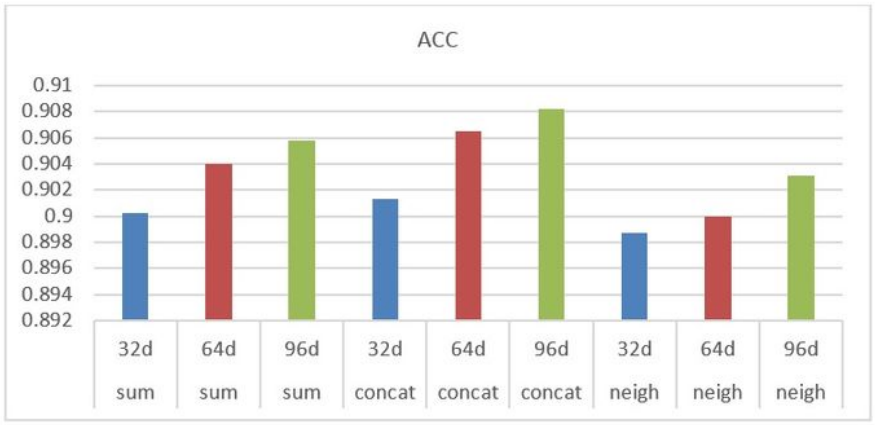
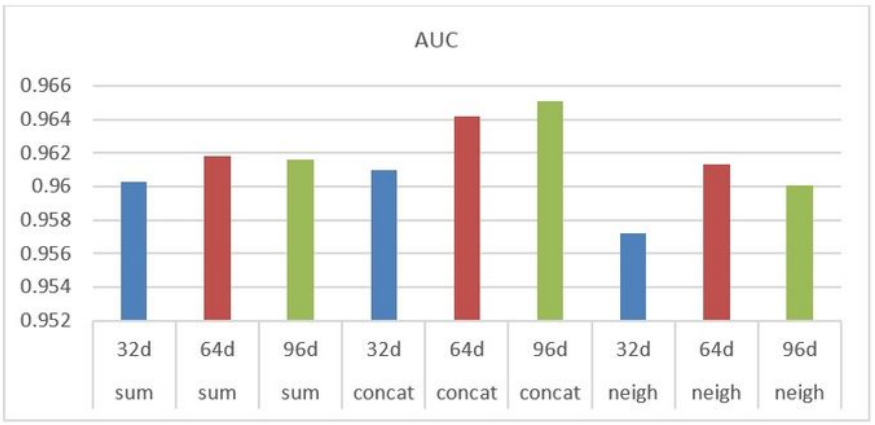


Figure 5

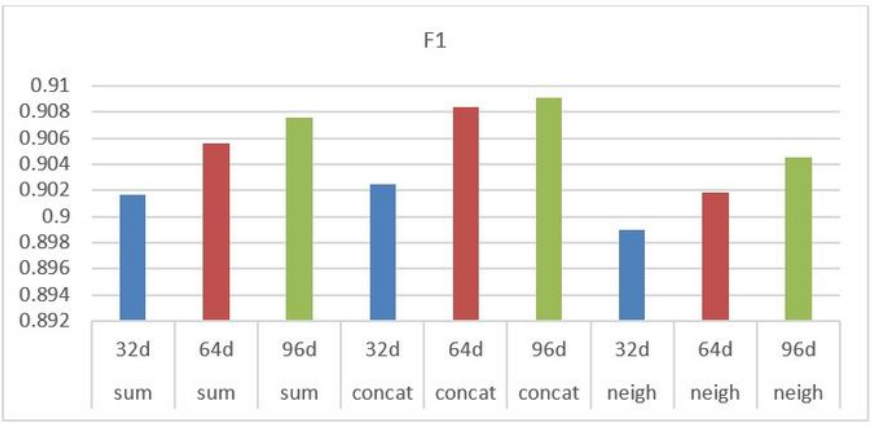
SmileGNN model



(a)



(b)



(c)

Figure 6

Influence of drug structure characteristic dimension on model performance