

CBSG_DPI — A Deep Learning Method for Drug-target Interactions Based on a Graph Convolutional Neural Network

Yuyu GAN

College of Electronic Information, Guangxi Minzu University, Nanning 530006, China

E-mail: 617440284@qq.com

Yanjia YU

College of Electronic Information, Guangxi Minzu University, Nanning 530006, China

E-mail: 2201077265@qq.com

Yongshang YU

College of Electronic Information, Guangxi Minzu University, Nanning 530006, China

E-mail: BioMedicalJoeYu@163.com

Le ZHONG

College of Electronic Information, Guangxi Minzu University, Nanning 530006, China

E-mail: 735260753@qq.com

Yong LIU*

College of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China

E-mail: liuhyonge@gxmzu.edu.cn

Abstract Drug-protein interaction (DPI) prediction in drug discovery and new drug design plays a key role, but the traditional in vitro experiments would incur significant temporal and financial costs, cannot smoothly advance drug-protein interaction research, so many computer prediction models have emerged, and the current commonly used is based on deep learning method. In this paper, a deep learning model Computer-based Drug-Protein Interaction CBSG_DPI is proposed to predict drug-protein interactions. This model uses the protein features extracted by the Computed Tomography CT and Bert method and the drug features extracted by the SMILES2Vec method and input into the graph convolutional neural network (GCN) to complete the prediction of drug-protein interactions. The obtained results show that the proposed model can not only predict drug-protein interactions more accurately but also train hundreds of times faster than the traditional deep learning model by abandoning the traditional grid search algorithm to find the best parameters.

Keywords deep learning; graph convolutional neural network; drug-target interactions; CBSG_DPI

1 Introduction

In the past few years, drug discovery and new drug design have both been time-consuming and expensive processes, with the prediction of drug-protein interactions (DPI) being a critical aspect of these processes. Traditionally, DPI prediction has relied heavily on a series of in vitro experiments, which are often tedious and incur significant time and financial expenditures^[1–5]. However, the rapid advancement of artificial intelligence, and more specifically, the application of machine learning methods within bioinformatics, has revolutionized this field. Deep learning, a subset of machine learning, has been increasingly applied in the area of Drug-Target Interaction (DTI), effectively addressing the challenges of time and cost associated with traditional methods. Current research suggests that, aside from deep learning, machine learning methods for predicting drug-protein interactions can be broadly categorized into three approaches: 1) molecular docking-based methods, 2) similarity-based methods, and 3) network-based models.

The theoretical and practical significance of this research lies in its ability to accelerate drug discovery and new drug design processes, improve the accuracy of drug-protein interaction prediction, and reduce time and costs. The applications of this research are manifold and substantial, ranging from streamlining drug development, minimizing redundant experimental procedures, to providing more efficient strategies for medical and pharmaceutical research. Moreover, precise prediction of drug-protein interactions is instrumental in elucidating the mechanisms of action of drugs, which supports personalized drug therapy and furthers the advancement of precision medicine.

1.1 Molecular Docking-Based Methods

Molecular docking is a method of drug design by predicting the interaction mode of the receptor and the drug molecule. It represents a theoretical simulation approach to study the intermolecular interactions such as those between ligands and receptors, and to anticipate their binding modes and affinities. Recently, molecular docking methods have gained significant importance in the field of drug-protein interaction research. Figure 1 presents an approximate flowchart of the molecular docking process.

Molecular docking is a simulation-based method using the three-dimensional structural features of molecules (ligands) and proteins (receptors). A key advantage of this method is its independence from prior information beyond structural inputs of drugs and targets. This allows for the identification of potential targets for a given drug and vice versa, potential drugs for a specific target. Despite its popularity in drug development and repurposing, molecular docking faces certain challenges. To enhance prediction accuracy, Luo *et al.*^[6] addressed these challenges by optimizing the target conformation, incorporating solvent considerations, and integrating binders into their methodology.

The recent emergence of the 2019 Novel coronavirus (SARS-CoV-2) and the resultant COVID-19 disease can cause severe or even fatal respiratory infections, but there is no approved treatment or effective treatment to effectively combat the outbreak. This emergency is forcing the

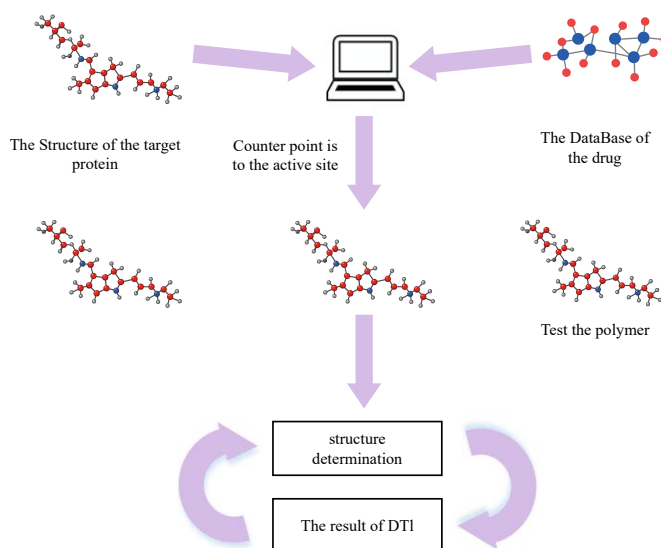


Figure 1 Flow chart of molecular docking

world to respond to the development of novel vaccines or small-molecule therapies for SARS-CoV-2. Through these efforts, the structure of the SARS-CoV2 major protease (Mopar) has been rapidly resolved and made publicly available to facilitate global efforts in the development of new drug candidates. So recently, Ton et al.^[7] developed a new deep learning platform, termed Deep Docking (DD), which provides rapid prediction of Glide (or any other docking program) docking scores, thus enabling virtual screening of billions of available molecules in a short period. In the current study, they applied DD to all 1.3 billion compounds in the ZINC15 library to identify the top 1,000 potential ligands of the SARS-CoV-2 Mopar protein.

1.2 Similarity-Based Approach

The similarity-based approach contains assumptions based on similar drugs sharing similar targets, and opposing assumptions based on similar proteins sharing ligands^[8]. This approach employs two primary similarity matrices: (a) drug similarity (calculated by chemical structure) and (b) target similarity (generated by protein sequence alignment).

In recent years, Xuan et al.^[9] introduced DTIP, a method for drug-protein interaction prediction, which encodes and integrates multiple similarities and associations of drugs and proteins along with neighborhood topological information. This approach enhances the effectiveness of feature extraction for predicting drug-protein interactions. Lin et al.^[10] utilized feature similarity fusion and molecular mapping, considering both drug and target similarities. They employed graphical representation to extract internal structural information and intermolecular interactions between drug and target. Thus, the purpose of drug target mutual use detection is achieved. Wang et al.^[11] proposed a novel network-based method for predicting DTIs, PPDTs. In this work, the authors predict potential DTIs from the local structure of the drug-target

association network, which is different from traditional structure and ligand-based global network similarity methods. Initially, known DTIs are represented as binary networks based on the local structure of the DTIs. Subsequently, drug-drug and target-target similarity networks are constructed using the resource allocation algorithm. The collaborative filtering algorithm was used to score the topological information of known drug targets. Finally, a linear combination of the drug-target similarity model and target-drug similarity model is innovatively proposed to obtain the final prediction results. More detailed information on similarity-based drug-target interactions can be seen in the review article published by Wang and Kurgan^[12].

1.3 The Network-Based Model

Accurate prediction of drug-target interactions (DTI) through biological data can reduce the time and economic cost of drug development, and DTI prediction methods based on similarity networks are receiving increasing attention. Currently, many studies have focused on the prediction of DTI. However, these methods do not consider the features of drugs and targets in multiple networks, nor consider how they can be extracted and combined. And a recent network called a heterogeneous bio information network can describe drugs and targets from multiple perspectives, and biological network-based models can obtain more comprehensive biological information^[13,14]. Figure 2 illustrates an example of a protein interaction network, specifically for the protein clorf95.

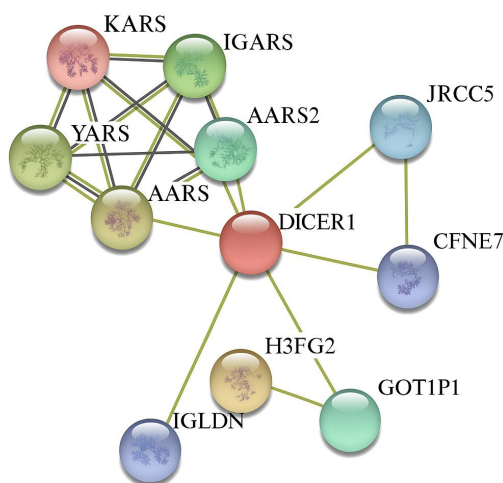


Figure 2 Example diagram of the protein interaction network of clorf95 from the STRING database

An and Yu^[15] introduced a network embedding framework, NEDTP, for predicting DTIs in multiple networks. NEDTP constructs a node similarity network from 15 heterogeneous information networks. The network is composed of 15 heterogeneous networks, Including nine drug-related networks: 1) clinically reported drug-drug interactions, 2) Drug-disease association, 3) Drug-side-effect association, 4) Chemical similarity, 5) Treatment similarity derived from the

anatomical therapeutic chemical classification system, 6) Drug-drug similarity derived from the target sequence, 7) Gene ontology (GO) biological processes, 8) GO cell components and 9) GO molecular function; And six protein-correlation networks: 1) proteinCprotein interactions, 2) The protein-disease association, 3) The protein sequence similarity, 4) The GO biological process, 5) GO cell components and 6) GO molecular functions. Next, a random walk is applied to extract the topological information for each node in the network, which is learned as a low-dimensional vector. Finally, a gradient-boosting decision tree model is constructed to complete the classification task of drug-protein pairs.

From the above introduction we can know, although molecular docking has extensive applications in biology, there are still two serious problems to be solved, first of all, we need to know the target's 3D structure to calculate the binding affinity of drug and candidate target, but many targets' 3D structure, especially G protease has been not available. Second, the simulation is time-consuming and requires substantial computational resources. On the other hand, the similarity in machine learning is more efficient than molecular docking, it can allow large-scale prediction, to check for further experimental screening of more promising targets, so compared to molecular docking, the model based on similarity and network-based model with the use of machine learning method, faster than molecular docking to occupy the DTI market. We all know that DTI is the most important step in feature extraction, only feature extraction is detailed enough and targeted, the DTI prediction will be more accurate, the mainstream deep learning method on the market now provides a variety of effective methods to extract features and can effectively with the DTI prediction training, such as convolutional neural network, graph neural network, recurrent neural network, and so on. From the current research situation, deep learning will be a rising spring shoot in the field of DTI and is likely to replace machine learning to quickly occupy the market of DTI and even more fields.

Recently appeared an end-to-end deep learning method to predict drug-protein interactions, the study by Monteiro et al.^[16], proposed a deep learning architecture model, model using convolutional neural network (CNN), from the protein sequence (amino acid sequence) and drug SMILES (simplified molecular input line input system) string get the one-dimensional representation. These representations can be interpreted as features expressing local dependence or patterns, and can then be used in a fully connected neural network (FCNN), as a binary classifier. Since the model uses the grid search method to find the best parameters, the training time takes a long time, reaching more than a month.

Wu et al.^[17] proposed a model of drug-protein interaction called BridgeDPI, which is also an end-to-end deep-learning model and predicts DPIs by introducing hypernodes. Hypernodes build bridges between proteins/drugs that capture information about PPA and DDA and then obtain predictions of DTI.

So, based on the above conditions, we proposed a deep learning model CBSG_DPI to predict drug-protein interaction, the model to extract the linear features of drugs and protein, as well as the CNN extraction features, and input to the GCN finally obtained drug-protein interaction

prediction results, the model abandoned the traditional grid search method to find the best parameters, thus greatly improve the training speed, and the accuracy is improved.

2 Method

2.1 Model Architecture

Based on the previous research results, we propose a deep learning-based CBSG_DPI model, whose architecture is shown in Figure 3. This model incorporates two primary inputs: The amino acid sequence of the protein and the drug SMILES string. For processing the protein's amino acid sequence, the model initially employs a three-layer feedforward network for handling K-mer features, followed by the integration of features extracted via the CT method and Bert. Additionally, a convolutional neural network (CNN) is utilized for extracting sequence features of proteins. These two sets of features are then amalgamated to obtain the final protein embedding. Regarding the drug SMILES, the approach begins with a two-layer feedforward network to process molecular fingerprint features, complemented by integrating drug features extracted through the SMILES2Vec method. In parallel with the protein embedding process, the CNN is also applied to extract sequence features of the drug. Subsequently, these two sets of features are combined to formulate the final drug embedding. The model then introduces supernodes to create a bridge between proteins and drugs, employing a graph neural network (GNN) to learn the associations between proteins/drugs. Finally, we obtain elemental-level products of protein and drug export after GNN and then use linear layers activated by sigmoid to predict interactions.

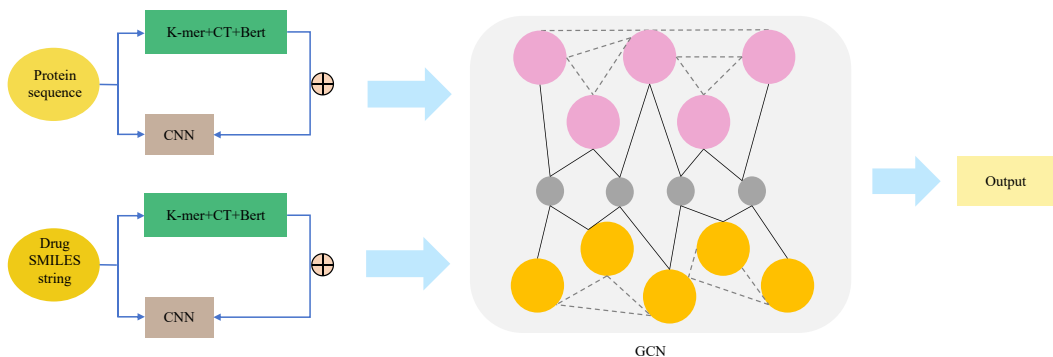


Figure 3 Model architecture

2.2 The Feature Extraction of the Proteins

2.2.1 Protein Features Were Extracted by the CT Method

Given that PPI information is integral to our GCN model, a significant computational challenge lies in identifying an effective method to accurately encapsulate the crucial information of PPIs from sequences. To address this problem, inspired by Shen et al.^[18], we used the CT method to extract sequence features of the protein, which considers the properties of an amino acid and its neighboring amino acids and treats any three consecutive amino acids as one unit.

Thus, it can be distinguished by the class of amino acids, that is, composed of three amino acids belonging to the same category, which can be processed identically because they can be considered to play similar roles in processing PPI. By calculating the frequency of each type of amino acid triplet, the PPI information from protein sequences is projected into a homogeneous vector space. The underlying principle of the CT method is illustrated in Figure 4.

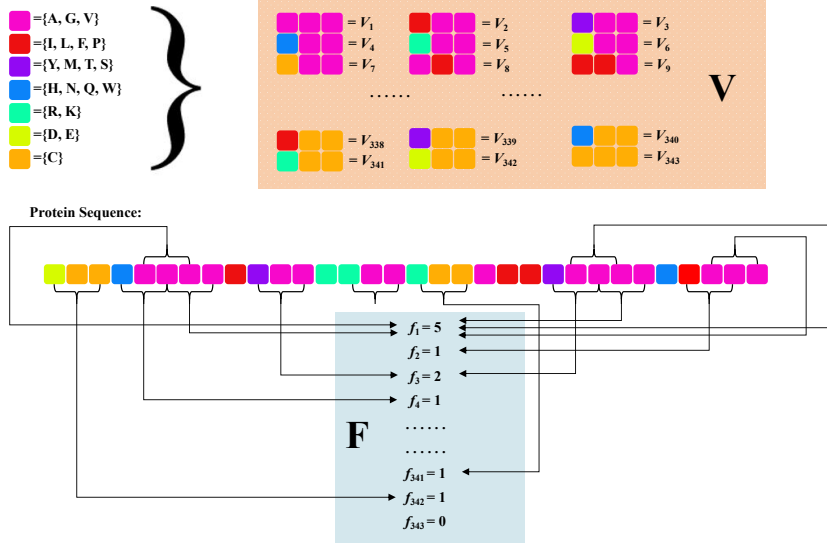


Figure 4 Computational principle of extracting protein features by the CT method

As detailed above, the initial step in the CT method involves classifying all amino acids into seven distinct classes. This classification is based on the volume of their dipoles and side chains. Amino acids within the same class may undergo synonymous mutations due to their similar characteristics. Consequently, each amino acid in a protein sequence can be substituted with its respective class label, thereby transforming the protein sequence into a string of integers. Subsequently, the method employs a sliding window of size 3 to traverse the entire sequence, step by step, calculating the occurrence frequency of each possible triplet combination. Since we treat any three consecutive amino acids as one unit (triplet) and these amino acids are divided into seven classes, there are $7 * 7 * 7$ different combinations such that the size of the feature vectors obtained by the CT method is 343. Information about the protein sequence can be projected into the homogeneous vector space by calculating the frequency of each triplet type.

2.2.2 Bert Method to Extract Features of Proteins

Bert, originally developed for natural language processing^[19], has increasingly found applications in the field of bioinformatics^[20,21], so to more comprehensively obtain characteristic information of proteins, inspired by Notin et al.^[21], we use the modified Bert method pro-Bert to extract sequence features of proteins. Similar to the Bert model, instead of sentences, the amino acid sequence of a protein, each amino acid can be a word in the sentence. We then pre-train the model using the BFD100 and UniRef100 databases^[22,23] and adjust the hyperparameters as

outlined in Table 1, resulting in a fine-tuned Pro-Bert model specifically optimized for protein sequences.

Table 1 Hyperparameters modified by pro-Bert compared to the Bert model

Hyperparameter	ProBert	
Dataset	BFD100	
Number of Layers	30	
Hidden Layers Size	1,024	
Hidden Layers Intermediate Size	4,096	
Number of Headers	16	
Positional Encoding Limits	40 K	
Dropout	0.0	
Target Length	512/2,048	
Memory Length	-	
Masking Probability	15%	
Local Batch Size	32/6	30/5
Global Batch Size	32,768/6,144	15,360/2,560
Optimizer	Lamb	
Learning Rate	0.002	
Weight Decay	0.01	
Training Steps	800 K/200 K	300 K/100 K
Warm-up Steps	140 K/20 K	40 K/0 K
Mixed Precision	None	
Number of Parameters	420 M	
System	TPU Pod	
Number of Nodes	128	64
Number of GPUs/TPUs	1,024	512

2.2.3 The k-mer Features of the Proteins

For proteins, k-mer is a classical and efficient method for protein feature extraction. The k-mer signature can describe the type and number of amino acid functional groups, which is very important in the prediction of drug-protein interactions. Therefore, we added the k-mer feature as a protein feature, inspired by Wu et al.^[17].

We set k to 1, 2, 3, and thus we get $20^1 + 20^2 + 20^3 = 8420$ k-mer features.

$$x^i = (x_1^i, x_2^i, \dots, x_{8420}^i). \quad (1)$$

In this framework, $x^i_1 \dots 20x^i_{21 \dots 420}$ and $x^i_{421 \dots 8420}$ represents the 1-mer, 2-mer, and 3-mer features of the protein i , respectively. To mitigate the influence of protein length on the analysis,

each section is horizontally normalized, as delineated in Equation (2).

$$\text{normalized}(x_{1...20}^i) = \frac{x_{1...20}^i - \text{mean}(x_{1...20}^i)}{\text{std}(x_{1...20}^i)}. \quad (2)$$

In the model, $x_{1...20}^i$ denotes the 1-mer features of protein i . The $\text{mean}(\cdot)$ and standard deviation (std) are computed for these features. For the calculated mean and standard deviation. For the 2mer section $x_{21...420}^i$ And the 3-mer section, $x_{421...8420}^i$ also to do the same thing.

2.3 Characteristic Extraction of Drugs

2.3.1 Drug Features Were Extracted by the SMILES2Vec Method

SMILES, an acronym for Simplified Molecular Input Line Entry System, represents a method for encoding the structure of chemical compounds in a linear, textual format suitable for computer processing. Distinct from traditional chemical formulas, SMILES employs a unique set of writing rules to compactly encode structural information of chemicals. For example, letters represent atoms, and in some cases, what type of atoms remain. For instance, different letters signify different atoms, with ‘c’ and ‘C’ denoting aromatic carbon and aliphatic carbon, respectively. Special characters like “=” indicate the type of bond (the connection between atoms). Lops are represented by packaged numbers and side chains are indicated in parentheses. Thus, when fully trained, the chemist can read SMILES and infer the structure of the chemical. From this structural information, the more complex properties can be predicted. The format of SMILES, comprising ASCII strings, bears resemblance to sentences in natural language processing (NLP). Influenced by the Word2Vec model^[24] in NLP, Goh et al.^[25] proposed a feature extraction method SMILES2Vec. While this method was first used in the drug-drug interaction model of Han et al.^[26], demonstrated promising results. Therefore, under the influence of these, we also added SMILES2Vec extracted features into the model to extract drug features. The fundamental principle of feature extraction using SMILES2Vec is illustrated in Figure 5.

As depicted in the aforementioned figure, the SMILES sequence is first segmented into a series of word packets. Each unique word in these packets corresponds to a vector element. Following this, the SMILES structure is converted into a one-hot encoded matrix. This process involves assessing each character in the SMILES string; characters present in the word packet are assigned a value of 1, while absent characters are designated as 0. Consequently, this step results in the generation of a one-hot matrix whose dimensions correlate to the word packet vector. Finally, to facilitate feature extraction, dimensionality reduction is performed on the one-hot matrix using Principal Component Analysis (PCA).

2.3.2 The FP Characteristics of the Drug

In drug discovery and virtual screening, molecular fingerprinting (FP) is recognized as an effective technique. Molecular FP involves encoding the topological information of a molecular structure into a vector, which can accurately represent a drug molecule. Following the approach suggested by Wu et al.^[17], we utilized MorganFC^[27] to characterize molecules. This method effectively encodes a drug as a vector with a dimensionality of 1024. $F^j = (b_1, b_2, \dots, b_{1024})$ In

SMILES [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

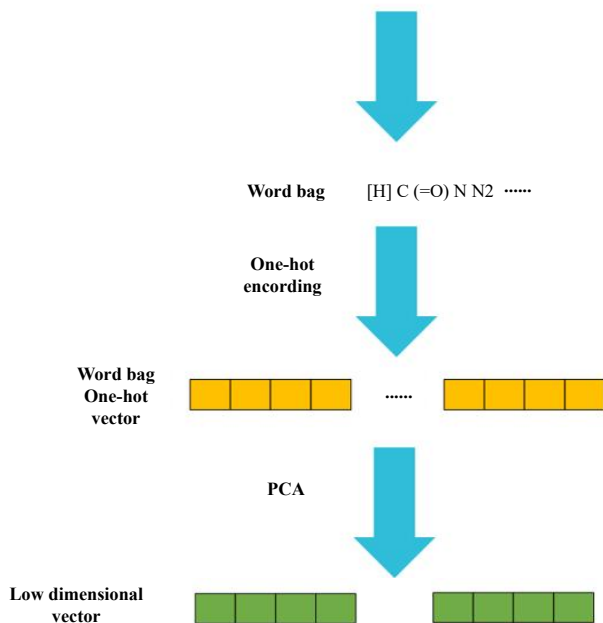


Figure 5 Principle for extracting drug features using SMILES2Vec

this context, F^j represents the molecular fingerprint (FP) characteristic of the drug j , $b_{1...11024}$ is the binary value in the molecular fingerprint.

2.4 Fusion of the Drug and the Protein Feature

In the process of combining the resultant protein vector A^i and drug vector F^j , direct construction of a relationship between them is not feasible due to the difference in their dimensions and their existence in distinct vector spaces. To address this, we employ several layers of neural networks to map both A^i and F^j into a common vector space. Additionally, to extract the sequence features of proteins and drugs effectively, we utilize a Convolutional Neural Network (CNN) with maximum pooling. This approach allows for the separate extraction and subsequent fusion of these sequence features.

$$u^i = f_p(A^i) + p^i, \quad (3)$$

$$v^j = f_d(f^j) + d^j. \quad (4)$$

In our model, $f_p(\cdot)$ And $f_d(\cdot)$ represent the nonlinear transformation layer of the protein and the drug, respectively, p^i and d^i sequence features were extracted for the CNNs. Finally, with u^i as an embedding of protein i , using the v^j as an embedding of the drug j .

2.5 The Data Set

In this study, the model's training and predictive analyses were conducted using the datasets

utilized in the End-to-End model by Monteiro et al.^[16]. These datasets encompass the Drug-Target Interaction (DTI) dataset, along with specific datasets for proteins and drugs. Detailed descriptions of these datasets are provided below.

2.5.1 DTI Data Set

The DTI dataset, originating from the work of Coelho et al. human Coelho et al.^[28] is utilized as a benchmark in this study. Positive samples were sourced from DrugBank^[29] and the dataset of Yamanishi et al.^[30], excluding samples associated with specific classes of protein targets and those with uncensored status. On the other hand, negative samples were obtained from the BioLip^[31] and BindingDB^[32] datasets, employing a bioactivity threshold of $10\mu M$ to discern weak binding interactions. Then a positive and negative ratio of 1.5, resulted in 7206 positive and 10912 negative samples in the training set of DTI pairs and 3530 positive and 5297 negative samples in the test set. In addition, they used only the positive samples of Yamanishi et al.^[30] and DrugBank^[29] for training and testing, respectively. The final drugs with sequence similarity greater than 85% used 16011 DTI pairs for the training set and 7926 DTI pairs for the test set. The outcomes of this dataset utilization are summarized in Table 2.

Table 2 The DTI data used for the End model

	Positive	Negative	Total
Train	5,839	10,172	16,011
Testing	3,012	4,914	7,926

2.5.2 Protein Data Set

For the End models, protein sequences were sourced exclusively from the UniProt^[33]. Given that proteins consist of a unique amino acid sequence varying in length, each amino acid in these sequences is treated as an individual feature. Therefore, it is necessary to define a threshold based on its length to ensure that each protein has the same number and type (order) characteristics. So this paper used a 95% information threshold to obtain a protein sequence maximum length of 1205. Sequences exceeding this length were excluded, while those shorter than the threshold were augmented to meet the defined standard.

2.5.3 Drug Dataset

For the End model, the SMILES dataset for drugs incorporates IDs from various databases including the Protein Data Bank (PDB)^[34], KEGG^[35], ZINC^[36], and DrugBank^[29]. To extract the SMILES strings, these IDs first need to be converted to PubChem^[37] compound is to extract the SMILES string. The PyPDB^[38], BioServices^[39], and PubChemPy^[40] from the Python package were used for conversion and extraction. Similar to the protein sequence, a threshold based on its length was used to obtain a maximum length of 90. Table 3 presents the number of unique proteins, drugs, and targets in the processed training and test datasets.

Table 3 Number of unique proteins, drugs, and targets in the training and test dataset after processing

	Unique		Number of Targets	
	Targets	Drugs	1	> 1
Training	1,790	9,583	8,026	1,557
Testing	1,068	5,718	4,884	834

2.6 Evaluation Index

For the performance comparison, we used the following evaluation measures:

- 1) Accuracy: To predict the accuracy of the classification

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (5)$$

- 2) F1-score: The harmonic mean between precision and recall.

$$F1\text{-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (6)$$

- 3) The area under the receiver operating characteristic curve (AUROC): Measuring the trade-off between the TP rate and FP rate.

2.7 Parameter Selection

Grid search, also known as exhaustive search, is a method of parameter tuning that involves searching through all candidate parameters in a stepwise manner. Through the loop, try every possibility, from all the parameters in the validation set accuracy parameters, the best parameter is the final result, akin to finding the maximum value in the array. Since the grid search takes a large amount of time to determine the best parameters, we used the Adam optimizer^[41] with a learning rate of 0.001 to adjust the parameters during training. To mitigate overfitting, an L2 regularization was added to the loss function. For each step, a batch of protein-drug pairs was randomly selected to run the gradient descent algorithm, with each batch size set to 512. For other hyperparameters settings such as the number of layers, neurons, and dropout rate, values are selected by many experiments based on the performance of the validation set. The maximum number of epochs was set to 100, and the model achieving the best Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) on the validation set was saved. This final model was then evaluated on the test set by accuracy indicators (ACC) and AUC.

3 Result

3.1 Model Comparison

We used the above-mentioned evaluation metric to train the dataset used by the End model into our model CBSG-DPI and compared the training results with the BridgeDPI and End-to-End models to obtain the results shown in Table 4.

Table 4 Results of the CBSG_DPI against the BridgeDPI and End-to-End models

Datasets/Evaluation metrics	CBSG_DPI	BridgeDPI	End-to-End
ACC	0.926	0.914	0.923
AUC	0.976	0.973	0.966
F1	0.895	0.875	0.895

The above results demonstrate that our proposed model CBSG_DPI outperforms BridgeDPI, primarily due to the integration of theCT and Bert methods for extracting protein features. And added the SMILES2Vec method to extract the characteristics of the drug, the result is even better; at the same time, our model is also slightly improved compared to End-to-End, and since the End-to-End model uses the traditional method of grid search to find the best parameters, resulting in a training duration of over a month. In contrast, our model eschews this traditional approach, leading to a significant increase in running speed, surpassing hundreds of times that of the End-to-End model. Furthermore, the drug-protein interactions predicted by CBSG_DPI were also compared with those predicted by BridgeDPI, as illustrated in Figures 6 to 7.

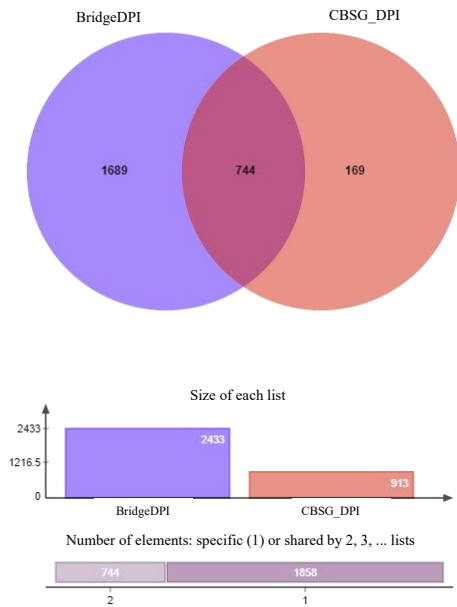
**Figure 6** Number and intersection of true positives predicted by BridgeDPI and CBSG_DPI

Figure 6 reveals that BridgeDPI predicts 2,433 true positives, the number of true positives predicted by CBSG_DPI is 913. Among these, 744 true positives are common to both models, while only the number of true positives predicted by BridgeDPI is 1,689. Only CBSG_DPI predicted 169 true positives.

From Figure 7, it's observed that BridgeDPI predicts a total of 6,348 positives, in contrast

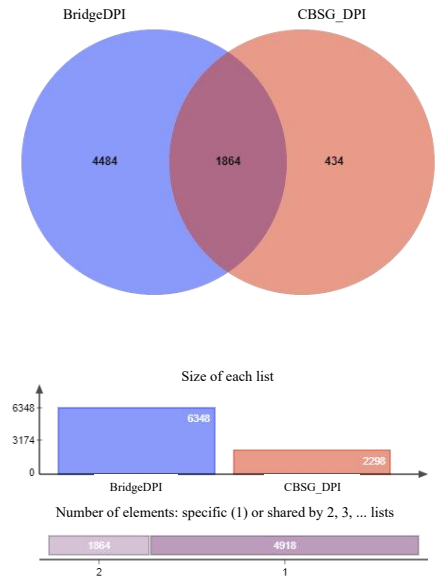


Figure 7 Number and intersection of all positives predicted by BridgeDPI and CBSG_DPI

to 2,298 by CBSG_DPI. Both models predict 1,864 positives, while only 4,484 positives were predicted by BridgeDPI. Only 434 were predicted to be positive by CBSG_DPI.

Therefore, according to the Venn figure above, we calculated the positive rate of the two models and drew the bar chart as shown in Figure 8.

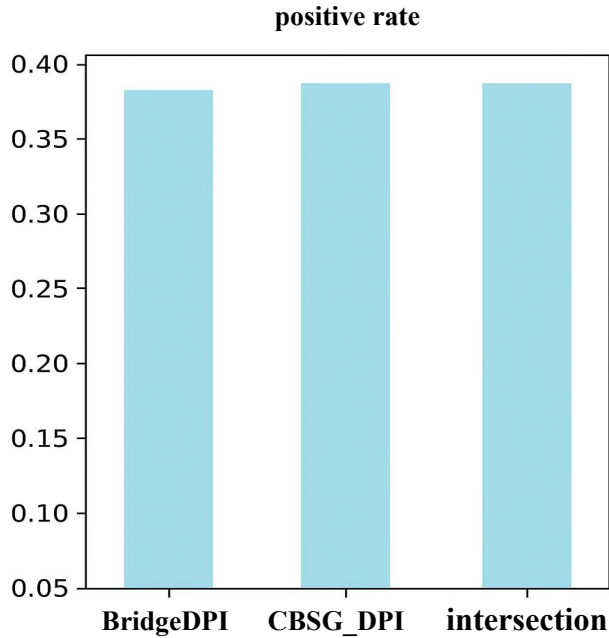


Figure 8 BridgeDPI, CBSG_DPI and their intersections positive rate

Subsequently, as depicted in Figure 8, we computed the positive rates of both models and represented them in a bar chart. The positive rate of CBSG_DPI is better than that of BridgeDPI. Additionally, the positive rate of our CBSG_DPI model more closely aligns with the rate observed at the intersection of the two models. Overall, these results suggest that the CBSG_DPI model demonstrates greater accuracy compared to BridgeDPI.

3.2 Ablation Experiments

Since the innovation point of our model is the addition of CT and Bert methods to extract protein features and SMILES2Vec methods for the extraction of drugs^[42], we need to verify whether adding these methods is useful, so we conducted the following ablation experiments^[43]. First, we removed one of the methods of extracting the features of proteins, and then compared them with our model, and obtained the results shown in Table 5.

Table 5 Test results of one method to extract protein features

Datasets/Evaluation metrics	CBSG_DPI	CBSG_DPI-Bert	CBSG_DPI-CT	CBSG_DPI-k-mer
ACC	0.926	0.888	0.918	0.917
AUC	0.976	0.964	0.974	0.974
F1	0.895	0.829	0.881	0.881

As indicated by the results, employing any single method independently did not yield optimal outcomes. The highest accuracy was achieved only when all these methods were integrated into our CBSG_DPI model. Applying a similar approach, we also conducted the same test for the method of extracting drug features, and the test results are presented in Table 6.

Table 6 Test results of one method

Datasets/Evaluation metrics	CBSG_DPI	CBSG_DPI-smile2Vec	CBSG_DPI-FP feature
ACC	0.926	0.915	0.903
AUC	0.976	0.973	0.968
F1	0.895	0.877	0.856

As can be seen from the above results, like proteins, the best results cannot be achieved when using one method alone, and the best accuracy can only be achieved by using both methods, our CBSG_DPI model.

4 Conclusion

This study introduces CBSG_DPI, an innovative model for predicting drug-protein interactions, which distinctively integrates CT, Bert, and SMILES2Vec methods for advanced feature extraction. This pioneering composite approach, unprecedented in prior research, significantly enhances the model’s proficiency in processing complex biological information, thereby achieving marked improvements in both accuracy and efficiency. In comparative analysis with existing models like End-to-End and BridgeDPI, CBSG_DPI exhibits superior predictive accuracy. By

employing grid search methods for optimal parameter tuning, the model demonstrates heightened efficiency and accuracy, especially in handling intricate biological datasets. Moreover, diverging from traditional parameter identification methods, CBSG.DPI adopts a more efficacious parameter tuning strategy, enhancing both operational efficiency and practical applicability in real-world scenarios.

The integration of these innovative techniques and the substantial improvement in performance underscore the pivotal role of CBSG.DPI in accelerating the drug discovery and design process. The model's focus on accurately predicting drug-protein interactions is instrumental in facilitating more efficient, cost-effective drug development. Such advancements hold significant implications for personalized medicine and optimized patient treatment strategies, potentially transforming patient outcomes and the overall efficacy of the healthcare system. Additionally, the incorporation of cutting-edge technologies like artificial intelligence and machine learning in drug discovery heralds a paradigm shift in the pharmaceutical industry, paving the way for novel approaches in drug development. Embracing these advancements, the progression in AI and machine learning is set to continually reshape the landscape of drug discovery and medical research, promoting more precise, targeted treatment methodologies and unlocking new avenues in medical breakthroughs with a broad societal impact.

References

- [1] Molina D M, Jafari R, Ignatushchenko M, et al. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science*, 2013, 341(6141): 84–87.
- [2] Roux K J, Kim D I, Burke B, et al. BioID: A screen for protein-protein interactions. *Current Protocols in Protein Science*, 2018, 91(1): 19.23.1–19.23.15.
- [3] Syal K, Wang W, Shan X, et al. Plasmonic imaging of protein interactions with single bacterial cells. *Biosensors and Bioelectronics*, 2015, 63: 131–137.
- [4] Schiebel J, Radeva N, Köster H, et al. One question, multiple answers: Biochemical and biophysical screening methods retrieve deviating fragment hit lists. *ChemMedChem*, 2015, 10(9): 1511–1521.
- [5] Seidel S A I, Dijkman P M, Lea W A, et al. Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods*, 2013, 59(3): 301–315.
- [6] Luo H, Mattes W, Mendrick D L, et al. Molecular docking for identification of potential targets for drug repurposing. *Current Topics in Medicinal Chemistry*, 2016, 16(30): 3636–3645.
- [7] Ton A T, Gentile F, Hsing M, et al. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Molecular Informatics*, 2020, 39(8): 2000028.
- [8] Wu Z, Li W, Liu G, et al. Network-based methods for prediction of drug-target interactions. *Front Pharmacol*, 2018, 9: 1134.
- [9] Xuan P, Zhang Y, Cui H, et al. Integrating multi-scale neighboring topologies and cross-modal similarities for drug-protein interaction prediction. *Briefings in Bioinformatics*, 2021, 22(5): Bbab119.
- [10] Lin X, Xu S, Liu X, et al. Detecting drug-target interactions with feature similarity fusion and molecular graphs. *Biology*, 2022, 11(7): 967.

- [11] Wang W, Wang Y, Zhang Y, et al. PPDTS: Predicting potential drug-target interactions based on network similarity. *IET Systems Biology*, 2022, 16(1): 18–27.
- [12] Wang C, Kurgan L. Survey of similarity-based prediction of drug-protein interactions. *Current Medicinal Chemistry*, 2020, 27(35): 5856–5886.
- [13] Davis M I, Hunt J P, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 2011, 29(11): 1046–1051.
- [14] Tang J, Szwarzajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 2014, 54(3): 735–743.
- [15] An Q, Yu L. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Briefings in Bioinformatics*, 2021, 22(6): Bbab275.
- [16] Monteiro N R C, Ribeiro B, Arrais J P. Drug-target interaction prediction: End-to-end deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, 18(6): 2364–2374.
- [17] Wu Y, Gao M, Zeng M, et al. BridgeDPI: A novel graph neural network for predicting drug-protein interactions. *Bioinformatics*, 2022, 38(9): 2571–2578.
- [18] Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 2007, 104(11): 4337–4341.
- [19] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020, 36(4): 1234–1240.
- [20] Wang S, Guo Y, Wang Y, et al. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, 429–436.
- [21] Notin P, Kollasch A W, Ritter D, et al. ProteinGym: Large-scale benchmarks for protein design and fitness prediction. *BioRxiv*, 2023, 07: 570727.
- [22] Wishart D S, Feunang Y D, Guo A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 2018, 46(D1): 1074–1082.
- [23] Yang H, Qin C, Li Y H, et al. Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, 2016, 44(D1): 1069–1074.
- [24] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. *arXiv preprint*, 2013, arXiv: 1301.3781.
- [25] Goh G B, Hodas N O, Siegel C, et al. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint*, 2017, arXiv: 1712.02034.
- [26] Han X, Xie R, Li X, et al. Smilegnn: Drug-drug interaction prediction based on the smiles and graph neural network. *Life*, 2022, 12(2): 319.
- [27] Cereto-Massagu A, Ojeda M J, Valls C, et al. Molecular fingerprint similarity search in virtual screening. *Methods*, 2015, 71: 58–63.
- [28] Coelho E D, Arrais J P, Oliveira J L. Computational discovery of putative leads for drug repositioning through drug-target interaction prediction. *PLoS Computational Biology*, 2016, 12(11): E1005219.

- [29] Wishart D S, Knox C, Guo A C, et al. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 2008, 36(suppl_1): 901–906.
- [30] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 2008, 24(13): 232–240.
- [31] Yang J, Roy A, Zhang Y. BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, 2012, 41(D1): 1096–1103.
- [32] Liu T, Lin Y, Wen X, et al. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 2007, 35(suppl_1): 198–201.
- [33] UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 2017, 45(D1): 158–169.
- [34] Berman H M, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Research*, 2000, 28(1): 235–242.
- [35] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 2000, 28(1): 27–30.
- [36] Irwin J J, Shoichet B K. ZINC-a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 2005, 45(1): 177–182.
- [37] Kim S, Thiessen P A, Bolton E E, et al. PubChem substance and compound databases. *Nucleic Acids Research*, 2016, 44(D1): 1202–1213.
- [38] Gilpin W. PyPDB: A Python API for the protein data bank. *Bioinformatics*, 2016, 32(1): 159–160.
- [39] Cokelaer T, Pultz D, Harder L M, et al. BioServices: A common Python package to access biological Web Services programmatically. *Bioinformatics*, 2013, 29(24): 3241–3242.
- [40] Swain M. PubChemPy documentation. 2014.
- [41] Kingma D P, Ba J. Adam: A method for stochastic optimization. *arXiv preprint*, 2014, arXiv: 1412.6980.
- [42] Di L, Gu Y, Qian G, et al. A prediction framework for turning period structures in COVID-19 epidemic and its application to practical emergency risk management. *Journal of Systems Science and Information*, 2022, 10(4): 309–337.
- [43] Devaraj S, Madian N, Subramaniam G, et al. Improved integrated deep model for Pap-smear cell analysis. *Journal of Systems Science and Information*, 2024, 12(1), 113–124.