frontiers in Bioengineering and Biotechnology

METHODS
published: 03 April 2020
doi: 10.3389/fbioe.2020.00267

Dipeptide Frequency of Word Frequency and Graph Convolutional Networks for DTA Prediction

Xianfang Wang[1,2]*, *Yifeng Liu*[2], *Fan Lu*[2], *Hongfei Li*[2], *Peng Gao*[2] *and Dongqing Wei*[3]

[1] *School of Computer Science and Technology, Henan Institute of Technology, Xinxiang, China,* [2] *School of Computer and Information Engineering, Henan Normal University, Xinxiang, China,* [3] *School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China*

OPEN ACCESS

Edited by:
Wen Zhang,
Huazhong Agricultural University, China

Reviewed by:
Juexin Wang,
University of Missouri, United States
Pu-Feng Du,
Tianjin University, China
Liu Li,
Inner Mongolia University, China

Correspondence:
Xianfang Wang
2wangfang@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *30 January 2020*
Accepted: *13 March 2020*
Published: *03 April 2020*

Citation:
Wang X, Liu Y, Lu F, Li H, Gao P and Wei D (2020) Dipeptide Frequency of Word Frequency and Graph Convolutional Networks for DTA Prediction.
Front. Bioeng. Biotechnol. 8:267.
doi: 10.3389/fbioe.2020.00267

Deep learning is an effective method to capture drug-target binding affinity, but low accuracy is still an obstacle to be overcome. Thus, we propose a novel predictor for drug-target binding affinity based on dipeptide frequency of word frequency encoding and a hybrid graph convolutional network. Word frequency characteristics of natural language are used to improve the frequency characteristics of peptides to express target proteins. For each drug molecules, the five different features of drug atoms and the atomic bond relationships are expressed as graphs. The obtained protein features and graph structure are used as the input of convolution neural network and the input of graph convolution neural network, respectively. A prediction model is established to predict the drug affinity by calculating the hidden relationship. In the KIBA data set test experiment, the consistency coefficient of the model is 0.901, which is 0.01 higher than the existing model, and the MSE (mean square error) of the model is 0.126, which is 5% lower than the existing model. In Davis data set test experiment, the consistency coefficient of the model is 0.895, which is 0.006 higher than the existing model, and the MSE of the model is 0.220, which is 4% lower than the existing model. These results show that our proposed method can not only predict the affinity better than those existing models, but also outperform unitary deep learning approaches.

Keywords: drug-target binding affinity, dipeptide frequency of word frequency, graph convolutional network, variable importance measures, deep learning

INTRODUCTION

The discovery processes of the new drug are not only time consuming, but also cost expensively (Roses, 2008). It usually spends about $ 2.6 billion and 10–17 years on research and experimental processes (Yang et al., 2017). One of core method is to find novel targets for existing drugs (Santos et al., 2016) and overcome the current shortage capabilities of drug discovering (Chu et al., 2019). It not only reduces experimental cost, but also greatly shortens drug discovery time (Martin et al., 2018), by eliminating multiple experimental processes such as drug stability (Oprea and Mestres, 2012). How to discover novel target proteins between drugs and targets has become an important task for drug development. And successful identification of drug-target interactions (DTI) is a prerequisite in this task (Ezzat et al., 2019).

High-throughput screening (HTS) experiments are often used to identify the biological activity between drugs and targets, but this method has problems of expensive cost and consumable time

(Cohen, 2002). DTI prediction in silicon is one of the effective methods (Liu et al., 2012), and machine learning is a prevalent way (Yan et al., 2019). Support vector machine (SVM) (Keum and Nam, 2017) and random forest (RF) (Wang et al., 2018; Strobl et al., 2019) are often used as predictors in existing research (Olayan et al., 2018). Although these methods are effective, shallow learning models may simplify the relationship between drugs and targeted proteins (Nanni et al., 2020), which are limited by the size of the dataset (Keogh and Mueen, 2009). Deep learning methods have achieved remarkable results in many research areas, such as image processing (Zhou et al., 2020), natural language recognition (Rabovsky and McClelland, 2020), and bioinformatics (Khurana et al., 2018). Its main advantage is that hidden relationships are obtained by calculating of non-linear mapping relationships in original data.

DTI prediction is often considered as a binary classification problem in existing studies (Ban et al., 2019; Yan et al., 2019; Le et al., 2020), that whether or not is a correlation. However, the calculation methods ignore the degree information about DTI, which is the value of binding affinity. Binding affinity provides information about the strength of interactions between drug target (DT) pairs, usually expressed by measures such as dissociation constant (Kd), inhibition constant (Ki), or the half maximal inhibitory concentration (IC50) (Cer et al., 2009). Drug-target binding affinity (DTA) calculated by deep learning algorithms has important research significance.

DeepDTA is a predictive tool for Drug-target binding affinity (Ozturk et al., 2018), which is a Convolutional Neural Network (CNN) that using 1D coding and drug molecular to learn hidden relationships between features and predicting affinity. In order to obtain better model performance, WipeDTA (Öztürk et al., 2019) extracted four text-based information sources to represent proteins and drug structures on the basis of DeepDTA. GraphDTA is an effective prediction model (Nguyen and Venkatesh, 2019), its framework is graph convolutional network that the inputs are graph structure of drugs. OneHot encoding is used to represent protein sequences as input for convolutional neural network. However, these problems what lower expression ability of protein sequence and low prediction ability are caused by the loss of correlation of the OneHot encoding for each residue individually encoded.

In order to overcome the above problems, we propose a novel feature extraction method which is polypeptide frequency of word frequency based on natural language word frequency characteristics to enhance the ability of protein sequence expression. The network model is constructed by merging the graph convolutional network that calculates the graph structure of drugs and the convolutional neural network that calculates the hidden relationship of protein features. The results of output are combined as the input of two hidden layers for regression training and prediction of DTA.

DATA SETS AND FEATURE EXTRACTION

Data Sets

We use two datasets: KIBA dataset (Tang et al., 2014) and Davis dataset (Davis et al., 2011) (The data sets can obtain from **Supplementary Material**), as shown in **Table 1**. KIBA (Tang et al., 2014) was used as a benchmark dataset to evaluate the algorithm model. The Davis dataset (Davis et al., 2011) is lysed selectively using the kinase protein family and associated inhibitors for the dissofarence constant (K_d) value, including the affinity of 442 proteins and 68 drugs. We calculate ($_pK_d$) value (as shown in formula 1) regarding the Davis data set use literature processing method to show.

TABLE 1 | Units for magnetic properties number of data sets.

Data set	Number of proteins	Number of drugs	Number of correlations
Davis(pKd)	442	68	30,056
KIBA	229	2111	118,254

$$pKd = -\lg(\frac{Kd}{1e9}) \qquad (1)$$

It can be seen from **Table 1** that the number of true interrelationships in the KIBA dataset is about three times that of the statistical interrelationship. KIBA values are calculated based on combinations of different information sources such as IC50, K_i, and K_d. We used a filtered version of the KIBA data set, where each protein and ligand has ten interactions at least (He et al., 2017).

Drug Molecular Feature Extraction

The graphs of the drugs are constructed by using the GraphDTA (Nguyen and Venkatesh, 2019) method. It reflects interactions of internal atom for each SMILES compound. RDkit, open source chemical informatics package (G, 2013), is used to calculate the feature vectors of atom and adjacent atomic connection of drugs. The nodes of the graph represent the features of the drug's atoms, and the bonding bonds between the atoms are represented by the edges. The features vectors of the drug atomic are made up of five characteristics: atomic class, atomic rank, the total number of hydrogen atoms, implied value of atoms, and the existence or absence of aromatic groups. The atomic rank is the sum of the number of the bond between the current atom and neighboring atoms and the number of hydrogen atoms. The edge of graph represents the connection relation of adjacent atoms. The overall process is shown in **Figure 1**.

Protein Sequence Feature Extraction

Protein Sequence Representation

The first order-structure vectorization is a prerequisite for data analysis of protein sequences, formula 2 is used to discretize the primary structure of the protein.

$$S_n = R_1R_2R_3....R_i...R_L, \ (n \le K) \qquad (2)$$

where S_n is the nth protein data, R_i is the *ith* amino acid residue in the protein sequence, K is the number of protein sequences in the data set.

Polypeptide Frequency of Word Frequency

Term frequency-inverse document frequency (TF-IDF) algorithm plays an important role in Natural Language

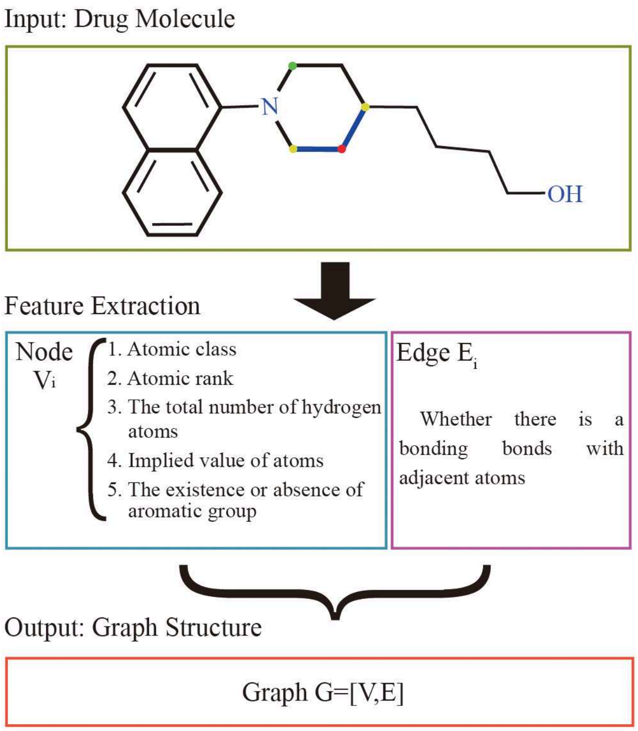

Input: Drug Molecule
N
OH
Feature Extraction
Node V_i
1. Atomic class
2. Atomic rank
3. The total number of hydrogen atoms
4. Implied value of atoms
5. The existence or absence of aromatic group
Edge E_i
Whether there is a bonding bonds with adjacent atoms
Output: Graph Structure
Graph G=[V,E]

FIGURE 1 | Drug molecular feature extraction process. The input of the extraction process of drugs is the drug molecular structure. Each atom is represented as a node by 5 different characteristics, and the bond between the atom and adjacent atoms is used as the edge set. The red atom has a binding bond with two yellow atoms, and no binding bond with the green atom. The set of nodes and the set of edges are made up of all the atoms together to form the graph structure representing the current drug molecule.

Processing (NLP) (Kaur and Jatinderkumar, 2019). TF-IDF is consisted of Term Frequency (TF) and Inverse Document Frequency (IDF). The algorithm of polypeptide frequency F (as shown in Formula 3) is similar to the calculation process of TF in bioinformatics.

$$F = (v_1, v_2, v_3, ..., v_{25^n})^T \tag{3}$$

Where, n is the number of 25 residues contained in the polypeptide, thus 25^n different polymers are formed by dehydration condensation, v_i represents the frequency of the *ith* feature of the polypeptide. The formula for v_i is as follows.

$$v_i = n_u / \sum_{u=1}^{25^n} n_u = n_u/(L-1) \tag{4}$$

where L represents the length of the protein sequence, n_u *r*epresents the occurrence times of uth dipeptide signature in the protein sequence.

IDF is the reversion document frequency to increase important weight of TF, as specified in formula 5.

$$IDF = \lg(\frac{N}{w_i}), (i = 1, 2, 3, 4, ..., 25^n) \tag{5}$$

where, in bioinformatics, N is the number of protein sequences in the data set, and w_i is the number of protein sequences which contain the *ith* polypeptide. From the formula, it can be known that the occurrence frequency of current words is inversely proportional to IDF, so TF-IDF algorithm will assign a lower feature for the high-frequency words. Which is not suitable for bioinformatics calculation. Therefore, we propose the polypeptide frequency of method word frequency, which can avoid this problem by only calculates the word frequency. As shown in formula 6:

$$WF = (wf_1, wf_2, wf_3, ..., wf_i, ..., wf_{25^n})^T \tag{6}$$

where, n is the number of residues that make up the polypeptide, and wf_i is the frequency of the *ith* polypeptide of word frequency, as shown in formula 7.

$$wf_i = \frac{w_i}{N} \times \frac{p_i}{L-1} \tag{7}$$

where, w_i is the number of protein sequences containing the *ith* peptide, N is the total number of proteins contained in the data set, p_i is the number of times that the *ith* peptide appears in the current protein, and L is the number of residues contained in the current protein.

Network Model Construction

A novel model that combining graph convolutional neural networks and convolutional neural networks are designed to regressively predict DTA. The multi-layers graph convolutional neural network is used to obtain the hidden relationships of drug graphs. The hidden relationships of the polypeptide frequency of word frequency are obtained through the convolutional neural network calculation. The output results of the two networks are combined as the input of fully connected layers. The complete process is shown in **Figure 2**.

Graph Convolutional Neural Network of Drug

We use the improved four types of graph convolutional neural networks by GraphDTA to discover potential relationships for the graph structure of drug features, which are GCN (Kipf and Welling, 2017), GAT (Veličković et al., 2018), GIN (Xu et al., 2019), GAT-GCN (Nguyen and Venkatesh, 2019). The linear connected layer that the inputs are results of graph convolutional neural networks maps to a 128-dimensional features vectors, which is consistent with the size of feature vectors for protein.

The GCN model is originally proposed by Kipf and Welling (2017) as a graph structure learner for semi-supervised classification. In order to meet the requirements of regression in our work, three graph convolutional units are made that include a GCN layer and a ReLU activation layer. The number of output channels is 78, 156, 312, respectively. And a fully connected layer of 1,024 neurons is created, the results are mapped to a 128-dimensional features vector in the output layer.

Graph Isomorphism Network (GIN) is an improved algorithm based on GCN. Injective aggregation updates the parameters and performs the feature vector mapping to obtain better model performance. The network model of the five-layer GIN layer is designed, and each GIN layer consists of two linear calculations with an output size of 32. The input and output layers are mapped into 128-dimensional features vectors.

Graph Attention Network (GAT) is different from the GCN model, the difference is that it calculates the corresponding hidden information for each node and introduces an attention mechanism when computing its neighboring nodes. The network model is designed using two GAT layers. In the first layer, the number of output channel is 78, and the number of attention nodes is 10. In the second layer, the number of output channel is 128, and the number of attention nodes is 1. The results are input to the output layer, which map to a 128-dimensional features vector.

Based on the GAT and GCN models, GAT- CCN integrates the advantages of the two models in series to obtain better model performance. The output channel of the GAT layer is 78, the number of attention nodes is 10. And the output channel of the GCN layer is 780. And a fully connected layer of 1,500 neurons is created, results are mapped to a 128-dimensional features vector in the output layer.

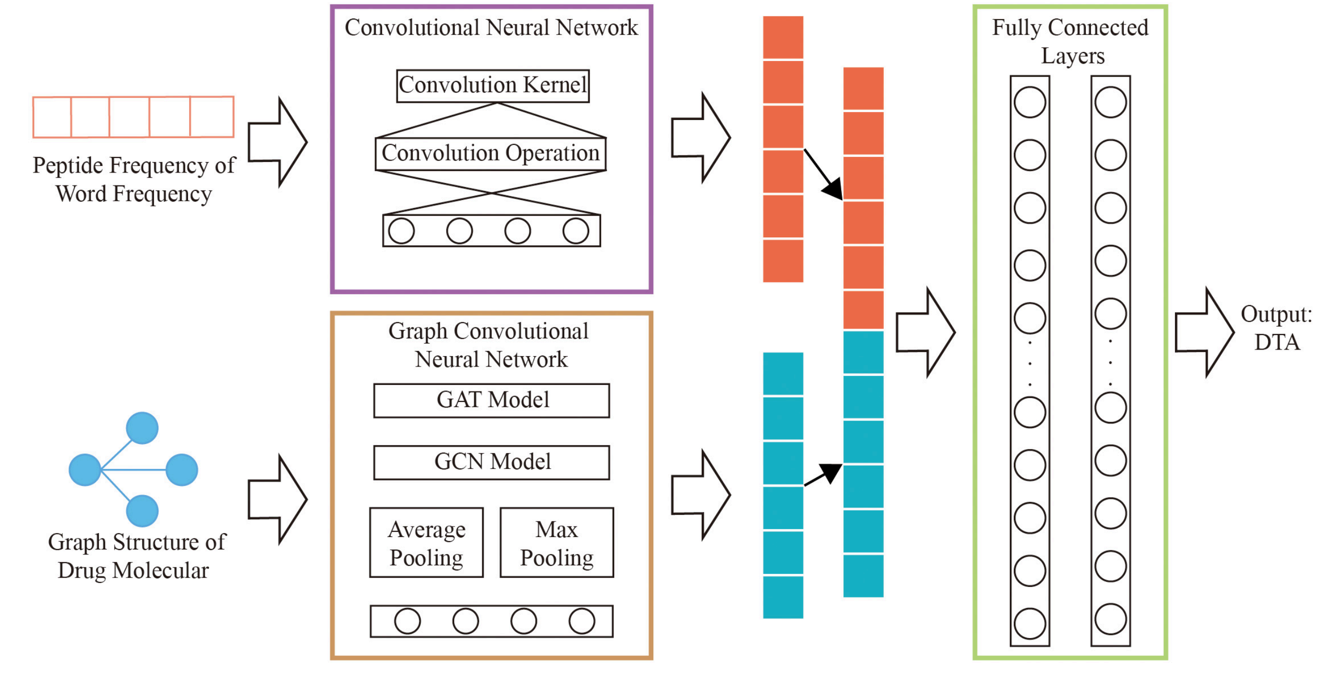

Peptide Frequency of Word Frequency
Convolutional Neural Network
Convolution Kernel
Convolution Operation
Graph Structure of Drug Molecular
Graph Convolutional Neural Network
GAT Model
GCN Model
Average Pooling
Max Pooling
Fully Connected Layers
Output: DTA

FIGURE 2 | Network structure diagram.

TABLE 2 | Comparative experimental results of word frequency feature of many different peptides.

Graph neural network model	Peptides	KIBA			Davis		
		MSE	CI	Pearson	MSE	CI	Pearson
GAT	1	0.758	0.372	0.323	0.740	0.649	0.436
	2	0.176	0.868	0.873	0.231	0.899	0.698
	3	0.187	0.858	0.823	0.244	0.861	0.659
GIN	1	0.427	0.696	0.569	0.472	0.802	0.634
	2	0.148	0.881	0.856	0.222	0.894	0.687
	3	0.151	0.871	0.851	0.239	0.882	0.685
GCN	1	0.803	0.431	0.341	0.834	0.408	0.337
	2	0.127	0.898	0.864	0.223	0.894	0.697
	3	0.151	0.873	0.846	0.247	0.887	0.691
GAT_GCN	1	0.624	0.798	0.698	0.743	0.644	0.434
	2	**0.126**	**0.901**	**0.893**	**0.220**	**0.899**	**0.701**
	3	0.191	0.852	0.839	0.224	0.896	0.693

The bold values are maximum.

Convolutional Neural Network of Protein

Convolutional neural network is used to obtain hidden relationships in vector of protein features. A 1D convolutional neural network is designed by analyzing the characteristic structure of protein word frequency and polypeptide frequency. The model contains a convolution kernel that the size is 32. The result of the convolution calculation is input to the fully connected layer for mapping to 256 neurons, keeping the size of the drug, and protein consistent.

We concatenate the feature vectors of proteins from convolutional neural networks and the feature vectors of drugs from graph convolutional neural networks. And they are input to two fully connected layers with 512 and 128 neutrons, respectively. And set the batch size to 512 and the learning rate to 0.00005.

RESULTS AND DISCUSSION

Performance Evaluation

In this work, the datasets are divided into two parts: training set and test set. That is, 80% of data instances are used for training, and 20% are for testing the models. The performances of our model are comprehensively compared by several experiment using evaluation metrics such as Concordance Index (CI), Mean Squared Error (MSE), as well as Pearson correlation coefficient. The evaluation indicators are consistent with WideDTA and GraphDTA. the performance of the predicted models of output continuous values is evaluated by CI, the formula is as follows.

$$CI = \frac{1}{Z} \sum_{\delta x > \delta y} h(bx - by) \tag{8}$$

where b_x is the prediction value for the larger affinity δ_x, b_y is the prediction value for the smaller affinity δ_y. Z is the normalization constant, *h(m)* is the step function, and as shown in the following formula:

$$h(m) = \begin{cases} 1, & if\ m > 0 \\ 0.5 & if\ m = 0 \\ 0 & if\ m < 0 \end{cases} \tag{9}$$

MSE is often used for the difference between the predicted value and the actual value vector, and it's an important index for evaluating regression models, the formula is as follows.

$$MSE = \frac{1}{n} \sum_{k=1}^{n} (b_k - \delta_k)^2 \tag{10}$$

where n is the number of data in the data set of KIBA or Davis, and other parameters have the same meaning as above.

Pearson correlation coefficient evaluates the difference of the affinity between the true value and the predicted value, the formula is as follows.

$$pearson = \frac{\mathrm{cov}(p, y)}{\sigma(p)\sigma(p)} \tag{11}$$

where *cov* indicates the co-variance, p is predicted values, y is original values, σ represents the standard deviation.

TABLE 3 | Comparison results of dipeptide features.

Features	KIBA			Davis		
	MSE	CI	Pearson	MSE	CI	Pearson
Dipeptide frequency of word frequency	**0.126**	**0.901**	**0.893**	**0.220**	**0.899**	**0.701**
Dipeptide frequency	0.148	0.882	0.857	0.239	0.881	0.690

The bold values are maximum.

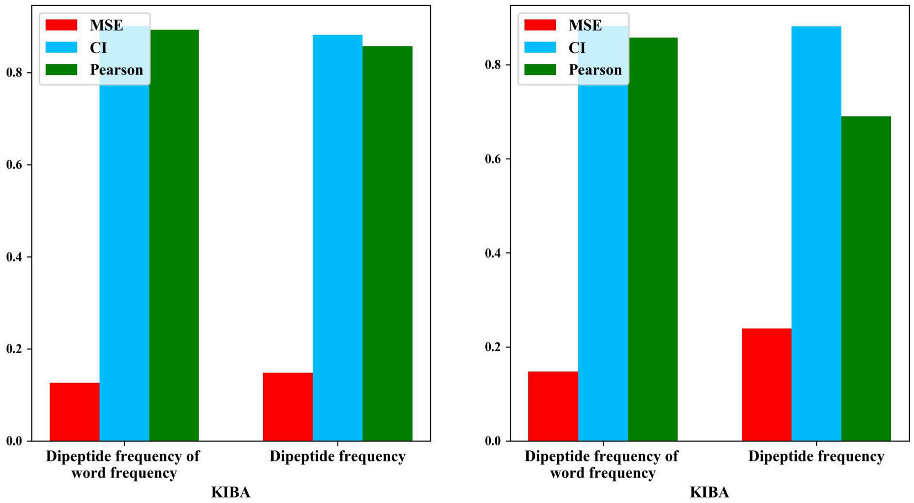


FIGURE 3 | Comparison results of dipeptide features.

Contrast Experiments and Analysis of Different Characteristics

In this study, we introduced polypeptide frequency of word frequency that was a novel way of protein feature extraction. The peptide frequency includes several methods. For every protein sequence, we calculated the word frequency characteristics and frequency of single peptide, dipeptides, as well as tripeptides. And different graph convolutional network models were designed to predict drug-target binding affinity. The results of comparative experiment are shown in **Table 2**.

When the protein sequence is represented by the word frequency dipeptide frequency and the GAT_GCN model, the model is the best predictor for 3 evaluation metrics yielding a CI of 0.901, a MSE of 0.126, and a Pearson of 0.893 in KIBA data set, and yielding a CI of 0.895, a MSE of 0.220 and a Pearson of 0.701 in Davis data set. When word frequency dipeptide frequency was used to represent protein sequences, compared with the second best GCN model, the CI and Pearson of GAT_GCN model in KIBA data set are increased by 0.03 and 0.029, respectively, and the MSE value decreases by 0.01. Compared with GAT and GIN models, the CI values of GAT_GCN model are 0.033 and 0.020 higher, the MSE values are reduced by 0.050 and 0.022, and Pearson values are increased by 0.020 and 0.037, respectively. In the Davis data set, the CI value of GAT_GCN model is same with GAT model as the next-highest model, the MSE value is reduced by 0.011, and Pearson is increased by 0.003. The CI value of the GAT_GCN model is 0.005 higher than the GCN and 0.002 higher than GIN. The MSE values are decreased by 0.003 and 0.005, and the Pearson values are increased by 0.004 and 0.006, respectively. So, the GAT_GCN model has the best performance in these four models.

When the GAT_GCN model is used as a graph calculator, compared with the word frequency single peptide frequency and the word frequency tripeptide frequency, the CI values of word

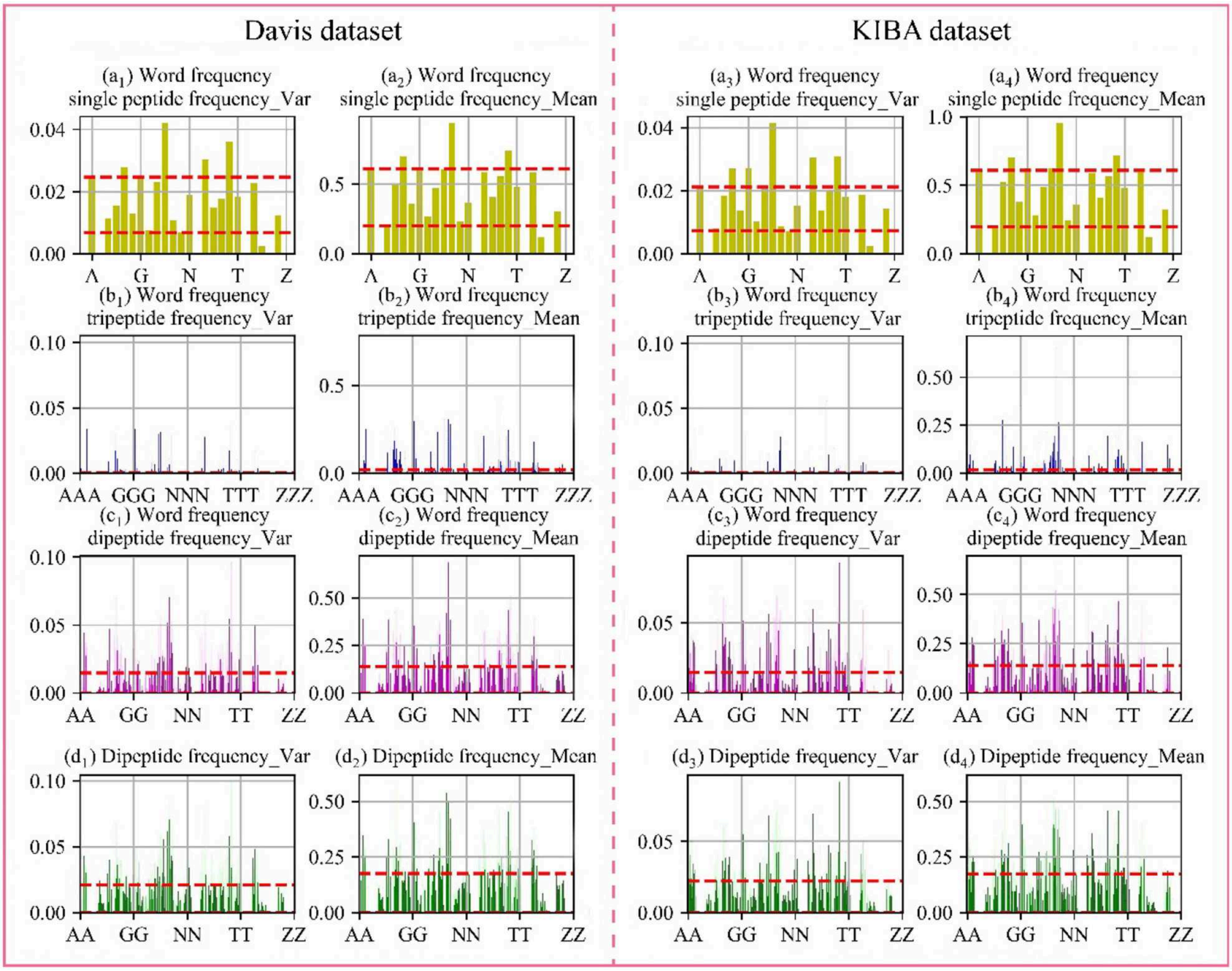


FIGURE 4 | Frequency chart of word frequency peptides. The X-axis is the peptide amino acid combination, and the Y-axis is the word frequency polypeptide frequency score. **(a)** Are the scores of the single peptide frequency of word frequency, **(b)** are the scores of the tripeptide frequency of word frequency, **(c)** are the scores of the dipeptide frequency of word frequency, and **(d)** are the scores of the dipeptide frequency. The red lines represent the upper and lower quartiles. The first and second columns are the Davis data set, and the first and second columns are the KIBA data set.

frequency dipeptide frequency in the KIBA dataset are higher by 0.103 and 0.049, the MSE values are reduced by 0.498 and 0.065, respectively. In the Davis data, CI values are 0.251 and 0.003 higher, the MSE values are decreased by 0.578 and 0.004, and Pearson values are increased by 0.195 and 0.054, respectively. The word frequency dipeptide frequency characteristics can also obtain the optimal index when combined with GIT, GAT, GCN models in the KIBA and Davis data sets, indicating that the word frequency dipeptide frequency characteristics have the best performance index compared to other characteristics.

Word Frequency Comparison Experiment

We also compared the differences in dipeptide frequencies with or without word frequency characteristics. The results are shown in **Table 3** and **Figure 3**.

After adding the word frequency characteristics based on the dipeptide frequency, the MSE decreased by 0.022 and the CI and Pearson increased by 0.009 and 0.033 in the KIBA data set, and MSE decreased by 0.019 and the CI and Pearson increased by 0.018 and 0.009 in Davis data set. This shows that the dipeptide frequency of word frequency is more conducive to the prediction of the classifier than the dipeptide frequency, and has better represented ability for protein sequences.

Analysis of Protein Features

Through the analysis of comparative experiments, we found that the model was obtained the best performance metrics when dipeptide frequency of word frequency be used to represent protein sequences. For every protein, we calculated the mean and variance in the Davis and KIBA datasets, respectively. The results are shown in **Figure 4**.

In the Davis dataset and the KIBA dataset, the distribution of score are basically same. The single peptide frequency of word frequency features scores are mainly concentrated between 0.20 and 0.61, and the variances are mainly concentrated between 0.007 and 0.220. Although there is a high features scores and large variance, the features have too high differences in the vectors of feature. And the number of features is only 25 dimensions, which contributes less to the spatially specific division of the model. Although the tripeptide frequency of word frequency features have a huge number of 15,625 dimensions, the scores are mainly distributed below 0.018, and the variances are mainly distributed below 0.003. The features have small differences between data, and there are a lot of features with value of 0. The scores of dipeptide frequency of word frequency characteristic mainly have a distribution range between 0 and 0.14, and the variances have a main distribution range between 0 and 0.0149, which has a good score and data difference.

Compared with the dipeptide frequency of word frequency, the score of dipeptide frequency are mainly distributed below 0.17, and the variances are mainly distributed between 0 and 0.021. Although it has a good score, the difference is high in vectors of feature, as same as the word frequency single peptide frequency. In order to discover the difference between the frequency characteristics of dipeptide and word frequency dipeptide, we draw a histogram of the frequency distribution of the two and perform overlapping processing, the results are shown in **Figure 5**.

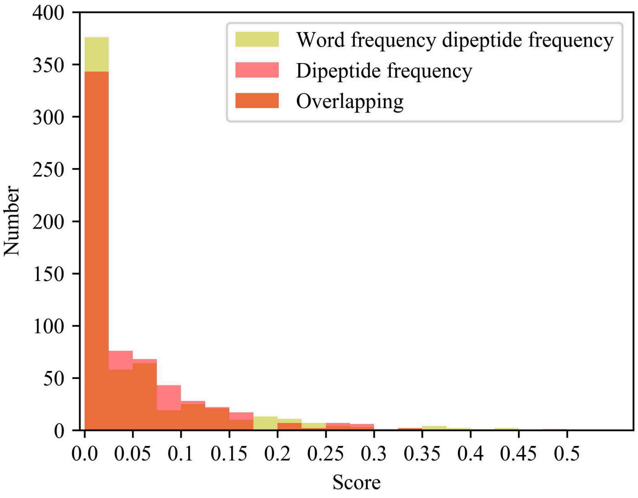
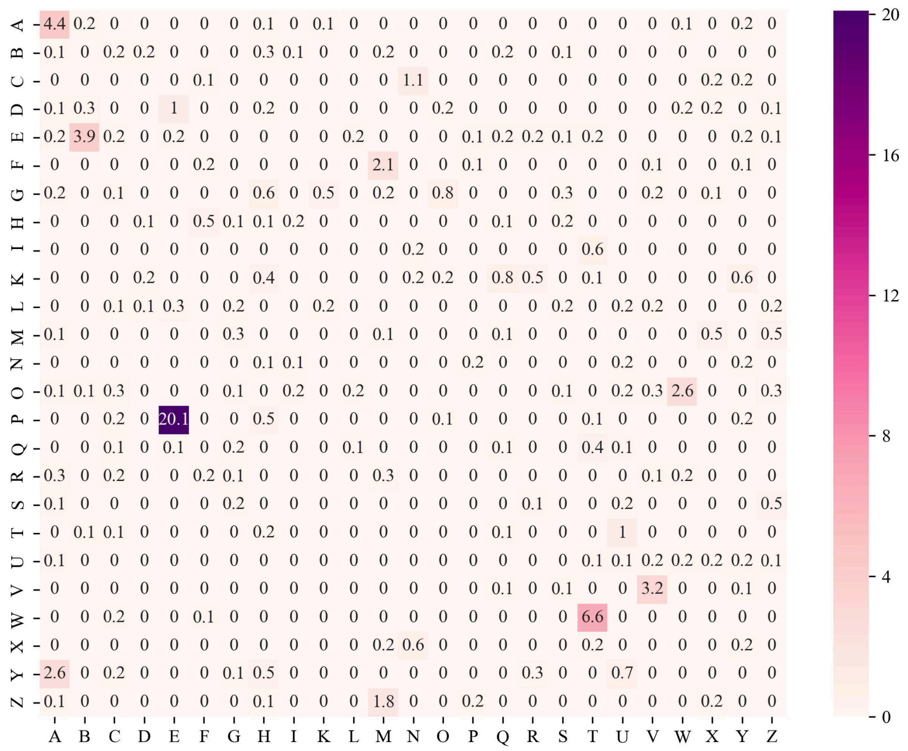


FIGURE 5 | Histogram of frequency distribution. Yellow represents the histogram of the frequency distribution of the word frequency dipeptide, red represents the histogram of the frequency distribution of the word frequency dipeptide, and orange represents the overlap between the two.

After adding the word frequency characteristics, the number of dipeptide frequency of word frequency features is less than that of the dipeptide frequency in the score intervals [0.025, 0.175] and [0.250, 0.300]. And the number of dipeptide frequency of word frequency features is more than that of dipeptide frequency features in the score intervals [0, 0.025] and [0.175, 0.250]. Dipeptide frequency features distribution is at [0, 0.35], and the dipeptide frequency of word frequency features distribution is at [0, 0.45], and the interval range is greater and more continuous. It shows that the frequency characteristics of words can play a role in reducing non-significant features and improving score difference.

Analysis of Variable Importance Measure

The protein dipeptide frequency of word frequency is composed of 625-dimensional features. The Variable Importance Measures (VIM) is used to analyze the contribution of each feature. In bioinformatics, Random Forest (RF) is a commonly used classification and regression model (Belgiu et al., 2016). And its unique advantage is to calculate VIM (Rawi et al., 2018), compared with other machine learning algorithms such as support vector machine (SVM). We used the RF model containing 10,000 decision trees to obtain the VIM score of features in the dipeptide frequency of word frequency, as shown in **Figure 6**. Features of non-zero VIM score have 199 dimensions, indicating that there's much noise in the vectors of features. The 27-dimensional features what a contribution >0.5% are listed in **Figure 7**. The top five dipeptide frequency of word frequency features are PE (20.1%), WT (6.6%), AA (4.4%), EB (3.9%), and VV (3.2%). This shows that PE (the combination of proline and glutamic acid) is significantly related to the affinity



	A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	4.4	0.2	0	0	0	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.2	0
B	0.1	0	0.2	0.2	0	0	0	0.3	0.1	0	0	0.2	0	0	0	0.2	0	0.1	0	0	0	0	0	0	0
C	0	0	0	0	0	0.1	0	0	0	0	0	0	1.1	0	0	0	0	0	0	0	0	0	0.2	0.2	0
D	0.1	0.3	0	0	1	0	0	0.2	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0.2	0.2	0	0.1
E	0.2	3.9	0.2	0	0.2	0	0	0	0	0	0.2	0	0	0	0.1	0.2	0.2	0.1	0.2	0	0	0	0	0.2	0.1
F	0	0	0	0	0	0.2	0	0	0	0	0	2.1	0	0	0.1	0	0	0	0	0	0.1	0	0	0.1	0
G	0.2	0	0.1	0	0	0	0	0.6	0	0.5	0	0.2	0	0.8	0	0	0	0.3	0	0	0.2	0	0.1	0	0
H	0	0	0	0.1	0	0.5	0.1	0.1	0.2	0	0	0	0	0	0	0.1	0	0.2	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0.6	0	0	0	0	0	0
K	0	0	0	0.2	0	0	0	0.4	0	0	0	0	0.2	0.2	0	0.8	0.5	0	0.1	0	0	0	0	0.6	0
L	0	0	0.1	0.1	0.3	0	0.2	0	0	0.2	0	0	0	0	0	0	0	0.2	0	0.2	0.2	0	0	0	0.2
M	0.1	0	0	0	0	0	0.3	0	0	0	0	0.1	0	0	0	0.1	0	0	0	0	0	0	0.5	0	0.5
N	0	0	0	0	0	0	0	0.1	0.1	0	0	0	0	0	0.2	0	0	0	0	0.2	0	0	0	0.2	0
O	0.1	0.1	0.3	0	0	0	0.1	0	0.2	0	0.2	0	0	0	0	0	0	0.1	0	0.2	0.3	2.6	0	0	0.3
P	0	0	0.2	0	20.1	0	0	0.5	0	0	0	0	0	0.1	0	0	0	0	0.1	0	0	0	0	0.2	0
Q	0	0	0.1	0	0.1	0	0.2	0	0	0	0.1	0	0	0	0	0.1	0	0	0.4	0.1	0	0	0	0	0
R	0.3	0	0.2	0	0	0.2	0.1	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0.1	0.2	0	0	0
S	0.1	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0.1	0	0	0.2	0	0	0	0	0.5
T	0	0.1	0.1	0	0	0	0	0.2	0	0	0	0	0	0	0	0.1	0	0	0	1	0	0	0	0	0
U	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.2	0.2	0.2	0.2	0.1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	3.2	0	0	0.1	0
W	0	0	0.2	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	6.6	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0.2	0.6	0	0	0	0	0	0.2	0	0	0	0	0.2	0
Y	2.6	0	0.2	0	0	0	0.1	0.5	0	0	0	0	0	0	0	0	0.3	0	0	0.7	0	0	0	0	0
Z	0.1	0	0	0	0	0	0	0.1	0	0	0	1.8	0	0	0.2	0	0	0	0	0	0	0	0.2	0	0

FIGURE 6 | Dipeptide frequency of word frequency VIM score. Its X-axis and Y-axis are 25 kinds of amino acids. Each point represents the importance score of the corresponding dipeptide frequency of word frequency characteristic variable. The color from white to purple represents the score from low to high.

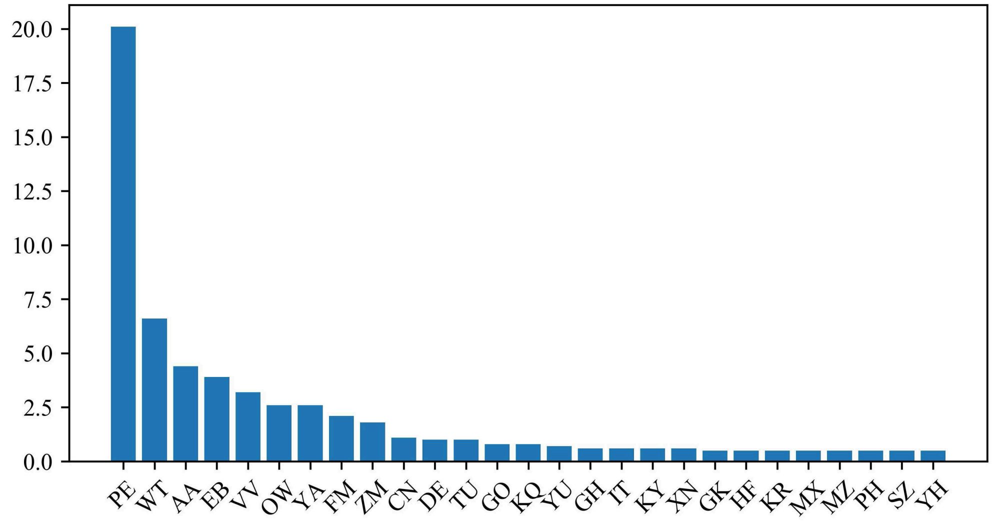


FIGURE 7 | Features ranking diagram with contribution >0.5%.

prediction, which is about three times of the second highest WT (the combination of tryptophan and threonine) and much larger than other combinations.

Comparison of Existing Models

The predictor of our work was compared with state-of-the-art methods what DeepDTA, WideDTA, and GraphDTA by using an

TABLE 4 | Algorithm comparison experiment results.

Features	KIBA		Davis	
	MSE	CI	MSE	CI
DeepDTA	0.194	0.863	0.261	0.878
WipeDTA	0.179	0.875	0.262	0.886
GraphDTA	0.139	0.891	0.229	0.893
This model	**0.126**	**0.901**	**0.220**	**0.899**

The bold values are maximum.

independent test set in Davis and KIBA. The results are shown in **Table 4**.

Our method outperformed state-of-the-art methods with two main quality metrics as CI and MSE in Davis and KIBA. Compared with the DeepDTA and WipeDTA models, our model reduced the MSE by 0.068 and 0.041, which increased the CI by 0.048 and 0.026, respectively in the KIBA dataset. And MSE decreased by 0.061 and 0.062, CI increased by 0.021 and 0.013, respectively, in the Davis data set. It shows that the graph neural network model with input as the graph structure of the drug can obtained better performance. Our method outperformed the GraphDTA model using the same graph convolutional neural network, the MSE decreased by 5% (0.007) and the CI increased by 0.01 in the KIBA data set, the MSE decreased by 4% (0.009) and the CI increased by 0.006 in the Davis data set. It shown that the dipeptide frequency of word frequency has better ability to express targeted proteins and can obtain better prediction models than 1D coding.

CONCLUSION

The DTA plays an important role in the discovery of new drugs. Dipeptide frequency of word frequency which is a novel feature extraction method is employed to represent protein sequences by natural language processing techniques. In addition, we use graphs to represent the drugs structure where the nodes is constructed by five different features and the edges represent atomic bond relationship. A network model is constructed, it is consisted of three parts: convolution neural network, graph convolution neural network, and fully connected layers. Convolutional neural network that input is dipeptide frequency of word frequency is to calculate hidden relationships of protein data. Graph Convolutional neural network is constructed to calculate hidden relationships for the graphs of drugs. The results of the two network models are mapped and combined to the fully connected layer predicting DTA. The results of peptide frequency comparison experiment showed that the dipeptide for the division of the spatial relationship was better than the monopeptide and tripeptide, so that the model performance can be obtained better. The results of the dipeptide frequency comparison experiment showed that adding word frequency characteristics for the dipeptide frequency can reduce the features difference. In comparison experiment of state-of-the-art model, our model has improved performance comparing with DeepDTA and WideDTA models, which indicating that the graphs can express the structure of drugs better. And experimental results show that our model has better performance than the GraphDTA model using graph convolutional neural network. In the KIBA dataset, MSE decreased by 5% (0.007) and CI increased by 0.01, and in the Davis dataset, MSE decreased by 4% (0.009) and CI increased by 0.006. It showed that the frequency characteristics of word frequency dipeptide could represent protein sequences better. Through the analysis of protein features, we observed that the vector have certain differences and intensity when the average score of the features is below 0.014 and the variance score is below 0.015, which are more conducive to the spatial division. In the analysis of variables importance, it was found that PE, WT, AA, EB, and VV had a high contribution to model prediction, among which PE (the combination of proline and glutamate) was highest by 20.1%. Besides drug discovery, the Dipeptide frequency of word frequency proposed in this work may also be applied in other field to represent protein sequence. Thus, it has the practical significance.

DATA AVAILABILITY STATEMENT

The datasets [Dives] for this study can be found in the [Comprehensive analysis of kinase inhibitor selectivity] [https://www.nature.com/articles/nbt.1990]. The datasets [KIBA] for this study can be found in the [Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis] [https://pubs.acs.org/doi/10.1021/ci400709d].

AUTHOR CONTRIBUTIONS

XW and YL designed the study and wrote the manuscript. FL translate manuscript. HL and PG analyzed data and drawn illustrations. DW provides theoretical guidance on Drug-Targets. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the grants from the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National Natural Science Foundation of China (Contract nos. 61802116, 61832019, 61503244), the State Key Lab of Microbial Metabolism and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00267/full#supplementary-material

REFERENCES

Ban, T., Ohue, M., and Akiyama, Y. (2019). NRLMFβ: β-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug-target interaction prediction. *Biochem. Biophys. Rep.* 18, 100615–100615. doi: 10.1016/j.bbrep.2019.01.008

Belgiu, M., Drăgut, L. J., and Sensing, R. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.* 114, 24–31. doi: 10.1016/j.isprsjprs.2016.01.011

Cer, R.Z., Mudunuri, U., Stephens, R., and Lebeda, F. J. (2009). IC50-to-K-i: a web-based tool for converting IC50 to K-i values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Res.* 37, W441–W445. doi: 10.1093/nar/gkp253

Chu, Y., Kaushik, A.C., Wang, X., Wang, W., Zhang, Y., Shan, X., et al. (2019). DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinform*. 2019:bbz152. doi: 10.1093/bib/bbz152

Cohen, P. (2002). Protein kinases - The major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* 1, 309–315. doi: 10.1038/nrd773

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051. doi: 10.1038/nbt.1990

Ezzat, A., Wu, M., Li, X.-L., and Kwoh, C.-K. (2019). Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinform.* 20, 1337–1357. doi: 10.1093/bib/bby002

G, L. (2013). *RDKit: Cheminformatics and Machine Learning Software*. Available online at: https://sourceforge.net/projects/rdkit/

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *J. Cheminform.* 9:24. doi: 10.1186/s13321-017-0209-z

Kaur, J., and Jatinderkumar, S. (2019). Designing punjabi poetry classifiers using machine learning and different textual features. *Int. Arab J. Inform. Tech.* 17, 38–44. doi: 10.34028/iajit/17/1/5

Keogh, E., and Mueen, A. (2009). Curse of dimensionality. *Ind. Eng. Chem.* 29, 48–53. doi: 10.1007/978-1-4899-7687-1_192

Keum, J., and Nam, H. (2017). SELF-BLM: prediction of drug-target interactions via self-training SVM. *PLoS ONE* 12:e0171839. doi: 10.1371/journal.pone.0171839

Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., Bensmail, H., and Mall, R., (2018). DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 34, 2605–2613. doi: 10.1093/bioinformatics/bty166

Kipf, T. N., and Welling, M. (2017). "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the International Conference on Learning Representations (ICLR). arXiv*:1609.02907.

Le, N. Q. K., Ho, Q. T., Yapp, E. K. Y., Ou, Y. Y., and Yeh, H. Y. (2020). DeepETC: a deep convolutional neural network architecture for investigating and classifying electron transport chain's complexes. *Neurocomputing* 375, 71–79. doi: 10.1016/j.neucom.2019.09.070

Liu, Z., Hong, F., Reagan, K., Xiaowei, X., Donna, M., William, S., et al., (2012). *In silico* drug repositioning - what we need to know. *Drug Discov. Today* 18, 110–115. doi: 10.1016/j.drudis.2012.08.005

Martin., E. M, Jane, N., and Louise, N. J. (2018). Protein kinase inhibitors: insights into drug design from structure. *Science* 303, 1800–1805. doi: 10.1126/science.1095920

Nanni, L., Lumini, A., Pasquali, F., and Brahnam, S. (2020). iProStruct2D: identifying protein structural classes by deep learning via 2D representations. *Exp. Systems Appl.* 142, 8. doi: 10.1016/j.eswa.2019.113019

Nguyen, T. A. L., and Venkatesh, S. H. (2019). GraphDTA: prediction of drug-target binding affinity using graph convolutional networks. *BioRxiv [preprint]*. doi: 10.1101/684662

Olayan, R. S., Ashoor, H., and Bajic, V. B. (2018). DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches (vol 34, pg 1164, 2018). *Bioinformatics* 34, 3779–3779. doi: 10.1093/bioinformatics/bty417

Oprea, T. I., and Mestres, J. (2012). Drug repurposing: far beyond new targets for old drugs. *Aaps J.* 14, 759–763. doi: 10.1208/s12248-012-9390-1

Ozturk, H., Ozgur, A., and Ozkirimli, E. (2018). DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* 34, 821–829. doi: 10.1093/bioinformatics/bty593

Öztürk, H., Ozkirimli, E., and Özgür, A. (2019). WideDTA: prediction of drug-target binding affinity. *Bioinformartics* 34, i821–i829.

Rabovsky, M., and McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Philos. Transac. R. Soc. Biol. Sci.* 375:20190313. doi: 10.1098/rstb.2019.0313

Rawi, R., Mall, R., Kunji, K., Shen, C.-H., Kwong, P. D., and Chuang, G.-Y. (2018). PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* 34, 1092–1098. doi: 10.1093/bioinformatics/btx662

Roses, A. D. (2008). Pharmacogenetics in drug discovery and development: a translational perspective. *Nat. Rev. Drug Discov.* 7, 807–817. doi: 10.1038/nrd2593

Santos, R., Oleg, U., Anna, G., Bento A., Ramesh, D., Cristian, B., et al. (2016). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov*. 16, 19–34. doi: 10.1038/nrd.2016.230

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2019). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinform.* 8, 25–20. doi: 10.1186/1471-2105-8-25

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inform. Model.* 54, 735–743. doi: 10.1021/ci400709d

Veličković, P., Cucurull, G., Casanova, A., Romero, A., and Bengio, Y. (2018). "Graph attention networks," in *Proceedings of the International Conference on Learning Representations (ICLR).*

Wang, L., You, Z.-H., Chen, X., Yan, X., Liu, G., and Zhang, W. (2018). RFDT: a rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. *Curr. Protein Peptide Sci.* 19, 445–454. doi: 10.2174/1389203718666161114111656

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). "How Powerful are Graph Neural Networks?," in *Proceedings of the International Conference on Learning Representations (ICLR).*

Yan, X. Y., Zhang, S. W., and He, C. R. (2019). Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods. *Comput. Biol. Chem.* 78, 460–467. doi: 10.1016/j.compbiolchem.2018.11.028

Yang, Q. H., Zhong, Y. N., Gillespie, C., Merritt, R., Bowman, B., George, M. G., et al. (2017). Assessing potential population impact of statin treatment for primary prevention of atherosclerotic cardiovascular diseases in the USA: population-based modelling study. *BMJ Open* 7:11. doi: 10.1136/bmjopen-2016-011684

Zhou, X., Sun, J., Tian, Y., Lu, B., Hang, Y. Y., and Chen, Q. S. (2020). Development of deep learning method for lead content prediction of lettuce leaf using hyperspectral images. *Int. J. Remote Sens.* 41, 2263–2276. doi: 10.1080/01431161.2019.1685721

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wang, Liu, Lu, Li, Gao and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.