[***molecules***](http://www.mdpi.com/journal/molecules)



# *Article*

**Prediction of Drug–Target Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures（**结合蛋白质序列和药物化学结构预测药物靶向相互作用网络**）**

# Fan-Rong Meng [[1]](#footnote-1), Zhu-Hong You 2,\*, Xing Chen 3,\*, Yong Zhou 1 and Ji-Yong An 1

1. School of Computer Science and Technology, China University of Mining（中国矿业大学） and Technology, Xuzhou 21116,

China; mengfr@cumt.edu.cn (F.-R.M.); yzhou@cumt.edu.cn (Y.Z.); ajy@cumt.edu.cn (J.-Y.A.)

1. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China
2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 21116, China

**\*** Correspondence: zhuhongyou@gmail.com (Z.-H.Y.); xingchen@amss.ac.cn (X.C.); Tel.: +86-181-6062-2862 (Z.-H.Y.)

Received: 27 May 2017; Accepted: 3 July 2017; Published: 5 July 2017

**Abstract:** Knowledge of drug–target interaction (DTI) plays an important role in discovering new drug candidates. Unfortunately, there are unavoidable shortcomings; including the time-consuming and expensive nature of the experimental method to predict DTI. Therefore, it motivates us to develop an effective computational method to predict DTI based on protein sequence. In the paper, we proposed a novel computational approach based on protein sequence, namely PDTPS (Predicting Drug Targets with Protein Sequence)（自己起名字） to predict DTI. The PDTPS method combines Bi-gram probabilities (BIGP), Position Specific Scoring Matrix (PSSM)( 位置特定计分矩阵), and Principal Component Analysis (PCA) with Relevance Vector Machine (RVM). In order to evaluate the prediction capacity of the PDTPS, the experiment was carried out on enzyme, ion channel, GPCR, and nuclear receptor datasets by using five-fold cross-validation tests. The proposed PDTPS method achieved average accuracy of 97.73%, 93.12%, 86.78%, and 87.78% on enzyme, ion channel, GPCR and nuclear receptor datasets, respectively. (这里写出了其预测功能是非常强大的。不过看起来是针对不同种类进行了分别的预测。)The experimental results showed that our method has good prediction performance. Furthermore, in order to further evaluate the prediction performance of the proposed PDTPS method, we compared it with the state-of-the-art support vector machine (SVM) classifier on enzyme and ion channel datasets, and other exiting methods on four datasets. The promising comparison results further demonstrate that the efficiency and robust of the proposed PDTPS method. This makes it a useful tool and suitable for predicting DTI, as well as other bioinformatics tasks.

**Keywords:** DTI; RVM; BIGP; PCA

develop fast and reliable computational methods for identifying drug–target interactions. Therefore, it is becoming more and more important to use computational approaches to detect DTI. （所以，计算手段很必要）The cost and time of experimental methods can be reduced and new potential drug–target interaction candidates can be found by using computational methods.（时间金钱省下来了，还可以发现新的交互关系。）

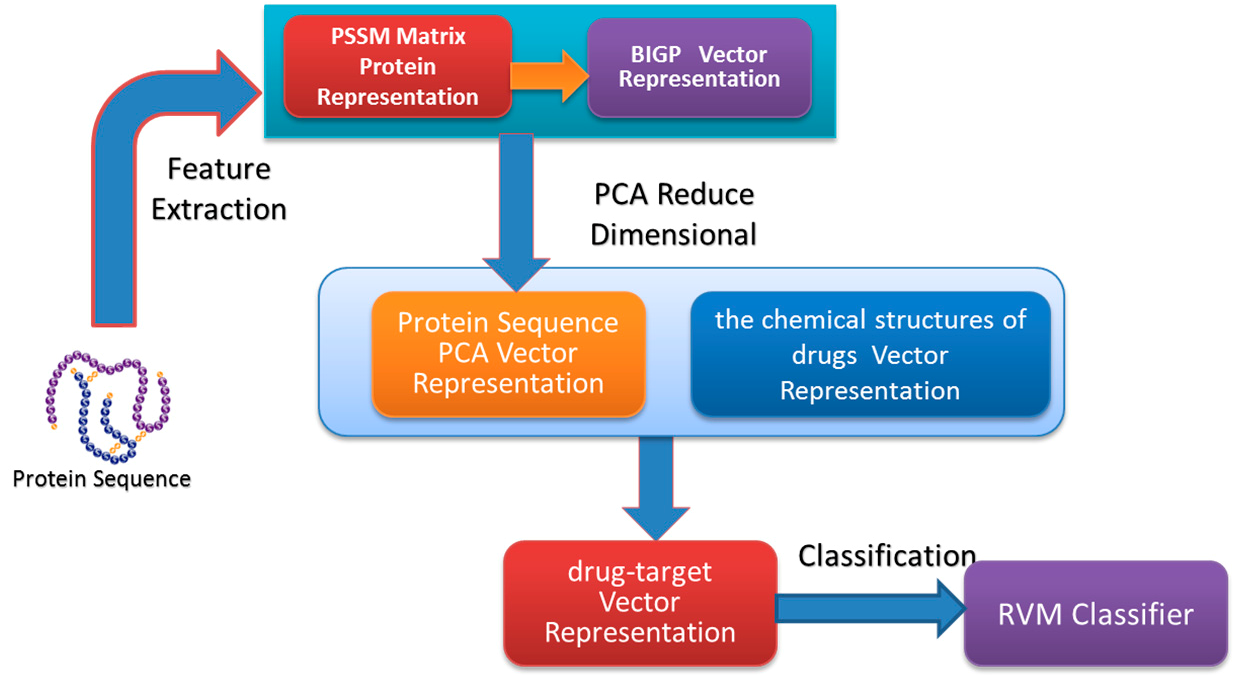
With the emergence of molecular medicine and the completion（完成） of the human genome project, the body of publicly-available knowledge of biology and chemistry is increasing rapidly.（讲现状，分子药物学和基因组学的发展，我们知道了很多知识） It makes the researchers restudy DTI questions by a systematic integration.（让我们重新学习药物靶点相关知识） A number of related databases that focus on drug–target relations have been constructed.（有了很多的数据库） We can freely obtain some of them from the public sector（n. 部门；扇形，扇区；象限仪；函数尺）, such as SuperTarget and Matador [6], Kyoto Encyclopedia of Genes and Genomes (KEGG) [7], DrugBank [8,9], Therapeutic Target Database (TTD) [10,11], etc. It is much useful for many researchers that a number of important experimental materials can be obtained from these databases to develop new computational approaches for identifying DTI on a genome-wide scale （数据多了，人们可以有更多的方法发展新的计算手段）[12,13].

All the time, in order to predict drug–target interactions, traditional computational methods are divided into the ligand-based virtual screening method and the docking approach.（介绍问题历史背景，基于配体的对接方法） The ligand-based virtual screening method compares the similarity of a given proteins represented based on chemical structure with a classic SAR framework,（这个是什么框架鸭？） which is used to predict DTI （基于配基的方法是，将给定化学结构的蛋白质与典型的sar架构进行比较，通过这个方法去预测DTI）[14]. However, there is an obvious shortcoming that the information of protein domains is not used for the method.（缺点在于不适用结构域信息） The docking simulation is a much useful molecular modeling method that can detect the positive interactions by using dynamic simulation when drug molecule and protein bound to each other（对接仿真是一种有用的技术，它可以测试分子和药物对接时的相互作用） [15–17]. However, the method has also a significant disadvantage that it can be only applied to proteins whose 3D structures are known.（但他只能预测已知3d结构的数据。） However, up to now, the proteins whose 3D structures are known comprise only a small part of all proteins.（知道的少） As a result, it is difficult to satisfy the experimental condition of the docking simulation method.（不能满足） Furthermore, the number of detected protein sequence data related to the known 3D structure data are increasing exponentially.（3D结构数据指数形式增长） Therefore, this promotes the need for developing new computational approaches based on protein sequence for detecting drug–target interactions.（所以需要新的方案去解决问题，扯犊子呢，数据多了就得换新方案？这个没关系鸭！）【这段介绍了历史2种方案的不足】

In recent years, a number of computational approaches have been proposed to predict drug–target interactions. For example, Yang et al. [18] developed a new computational method to detect multiple target optimal intervention（干预） solutions in a disease network.（最近几年有新的发展，包括多目标靶点干预机制的预测） The method attempts to identify effective points of intervention and the combination of interventions within a given disease network,which can best restore the disease network to a desired normal state. （这篇论文预期的是确定各种给定疾病的有效点，并恢复疾病网络到预期状态）Yan et al. [19] developed a representation of drug–target pairs based on drug chemical similarity and target sequence similarity and employed the random forest as classifier to build the prediction models.（基于药物相似度和蛋白质序列相似度，用随机森林的方案进行了预测。） By comparing the method and the state-of-the-art methods, it produces satisfying performance on the benchmark datasets.（比较后性能好） Kuang et al. [20] developed a novel method that proposed an eigenvalue（特征值） transformation technique and applied this technique to two representative algorithms for predicting DTI, the Regularized Least Squares classifier (RLS) and the semi-supervised link prediction classifier (SLP). （用了特征值？什么的，效果也好）The prediction results show that the method achieved better performance on drug–target interaction prediction. Bharadwaja et al. [21] proposed a new approach for identifying novel interactions for drugs and targets with no prior interaction information, which improved a machine learning method by integrating（v. 整合；积分；集成化） more correlated information of the drug compounds and extended it to a weighted profile method（其实这个方法我没看明白）. Peng et al. [22] proposed a prediction model name as NormMulInf which is a semi-supervised-based learning framework through collaborative filtering theory, employing labeled and unlabeled interaction information. Firstly, the method determines similarity principles, for example samples’ similarities and local correlations between samples’ labels by integrating biological information. Secondly, the similarity information can be integrated into the NormMulInf model, which solves the problem of augmented Lagrange multipliers. Wang et al. [23] proposed a new computational method, namely PDTD (Predicting Drug Targets with Domains), for identifying potential target proteins of new drugs based on derived interactions between drugs and protein domains.（这个是根据域结合来的） Zhang et al. [24] proposed a stacking-based ensemble learning method to boost performance of previous DTI prediction methods by using a state-of-the-art support vector machine (SVM) model as classifier to integrate the prediction results of previous methods. （Zhang等人[24]提出了一种基于累加的集成学习方法，利用最先进的支持向量机(SVM)模型作为分类器，综合以往方法的预测结果，提高了以往DTI预测方法的性能。）Although these methods have achieved good prediction accuracy, however, the proposed prediction model focuses on improving the prediction accuracy. Thus, there is still room to improve the prediction accuracy to identify DTI.（虽然很多人做了，但是还是可以做的）

In the paper, we proposed a novel computational approach based on protein sequence, namely PDTPS (Predicting Drug Targets with Protein Sequence), to predict drug–target interactions (DTI).（我们提出了一种基于序列的模式） The PDTPS method combines Bi-gram probabilities (BIGP), Position Specific Scoring Matrix (PSSM), and Principal Component Analysis （[自][数] 主成分分析）(PCA) with Relevance Vector Machine（关联向量机） (RVM). In order to evaluate the prediction capacity of the PDTPS, we carry out the experiment on enzyme, ion channel, GPCR, and nuclear receptor datasets by using five-fold cross-validation tests. （结合了4种方法，对1酶，2离子通道，3GPCR，4核受体）The proposed PDTPS method achieved average accuracy of 97.73%, 93.12%, 86.78%, and 87.78% on enzyme, ion channel, GPCR, and nuclear receptor datasets（非常高的结果） respectively. The experimental results showed that our method has good prediction performance. Furthermore, in order to further evaluate the prediction performance of the proposed PDTPS method, we compared it with the state-of-the-art（最高水平的） support vector machine (SVM) classifier on enzyme and ion channel datasets and other exiting methods on four datasets.（我们做的好，我们找了最高水平的支持向量机方法，用同样的数据运行结果好。）

The promising comparison results further demonstrate the efficiency and robustness of the proposed PDTPS method.（有鲁棒性和有效性） This makes it a useful tool and suitable for predicting DTI, as well as other bioinformatics tasks. （自己说这是个好工具）The flow chart（流程图） of the proposed prediction model is shown in Figure 1.



**Figure 1.** The flow chart of the proposed prediction model.

## 2. Results and Discussion

### 2.1. Performance of the Proposed Method

In order to verify the effectiveness of the proposed method, we carry out the experiment on enzyme, ion channel, GPCR, and nuclear receptor datasets through employing five-fold cross-validation tests respectively. For five-fold cross-validation, the whole dataset was divided into five parts; four parts of them were used as training samples, and one part of them was employed as testing samples.（为了验证该方法的有效性，我们分别采用5倍交叉验证试验，对酶、离子通道、GPCR和核受体数据集进行了实验。对于5倍交叉验证，将整个数据集分为5个部分;4个部分作为训练样本，1个部分作为测试样本。此外，实验中还需要对RVM分类器的几个参数进行优化。） In addition, there are several parameters that need be optimized for the RVM classifier in the experiment. Here, the ’ploy2’ function was selected as the kernel function, we also set up other parameters: width = 1, initapla = 1/N and beta = 0. Where width represents the width of ‘ploy2’ kernel function, N is the number of training samples, and beta represents classification. Tables 1–4 list the five-fold cross-validation tests prediction results by using the proposed approach on enzyme, ion channel, GPCR, and nuclear receptor datasets.

It can be observed from Tables 1–4 that the average Accuracy (Ac) and its standard deviation for enzymes, ion channels, GPCRs, and nuclear receptors is 97.73%, 93.12%, 86.77%, 87.78%, and 0.40%, 1.34%, 2.41%, and 3.17%, respectively. The corresponding average Sensitivity (Sn) and its standard deviation is 97.44%, 93.32%, 84.89%, 92.63%, and 1.04%, 1.54%, 4.04%, 11.53%, respectively. The corresponding average Precision (Pe) and its standard deviation is 98.01%, 92.96%, 87.91%, 85.19%, and 0.78%, 2.10%, 3.47%, 6.70%, respectively. At the same time, the average Matthews’s correlation coefficient (Mcc) and its standard deviation is 95.56%, 87.18%, 76.97%, 78.32%, and 0.76%, 2.28%, 3.64%, 4.72%, respectively. These experimental results indicated that the proposed method can obtain good prediction accuracy for predicting drug–target interactions.

The good prediction results of the proposed approach for drug–target interactions result from the correct choice of feature extraction method and classifier. Major improvements of the proposed feature extraction method can be divided into three following reasons: (1) Because PSSM not only describes the order information but also retains sufficient prior information, it can capture useful information from a given protein sequence; (2) The Bi-gram probabilities represented each protein PSSM and calculated the Bi-gram feature through employing the probability information PSSM contains. Because the Bi-gram features extracted from PSSMs can significantly reduce the sparsity level, this helps in improving the recognition performance; (3) For reducing the influence of noise for classifying and ensuring the integrity of feature information, we transformed the dimensions of each BIGP feature vector from 400 to 350 using Principal Component Analysis (PCA). Thus, it can be seen from these experimental results that the proposed BIGP method plays an essential role for improving prediction accuracy for predicting DTI.

**Table 1.** 5-fold cross validation results performed by proposed model on an enzyme dataset.

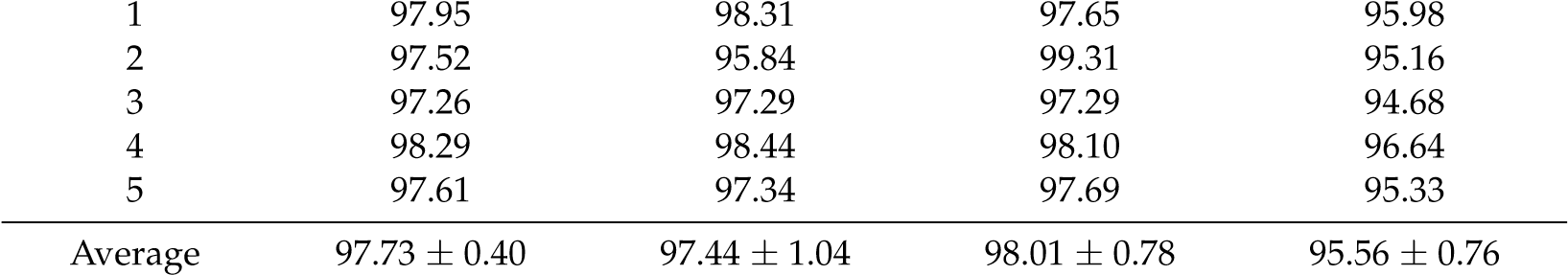
**Testing Set**

**Ac (%)**

**Sn (%)**

**Pe (%)**

**Mcc (%)**



**Table 2.** 5-fold cross validation results performed by proposed model on an ion channel dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing Set** | **Ac (%)** | **Sn (%)** | **Pe (%)** | **Mcc (%)** |
| 1 | 92.71 | 91.18 | 94.58 | 86.48 |
| 2 | 91.02 | 92.31 | 89.49 | 83.64 |
| 3 | 94.41 | 94.46 | 94.14 | 89.44 |
| 4 | 93.39 | 93.81 | 94.10 | 87.55 |
| 5 | 94.07 | 94.85 | 92.47 | 87.18 |
| Average | 93.12 ± 1.34 | 93.32 ± 1.54 | 92.96 ± 2.10 | 87.18 ± 2.28 |

**Table 3.** 5-fold cross validation results performed by proposed model on a GPCR dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing Set** | **Ac (%)** | **Sn (%)** | **Pe (%)** | **Mcc (%)** |
| 1 | 83.07 | 77.88 | 83.02 | 71.21 |
| 2 | 88.58 | 86.86 | 91.54 | 79.70 |
| 3 | 87.41 | 85.40 | 90.70 | 77.91 |
| 4 | 85.83 | 86.15 | 86.15 | 75.66 |
| 5 | 88.98 | 88.14 | 88.14 | 80.28 |
| Average | 86.77 ± 2.41 | 84.89 ± 4.04 | 87.91 ± 3.47 | 76.97 ± 3.64 |

**Table 4.** 5-fold cross validation results performed by proposed model on a nuclear receptor dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing Set** | **Ac (%)** | **Sn (%)** | **Pe (%)** | **Mcc (%)** |
| 1 | 83.33 | 73.68 | 93.33 | 71.79 |
| 2 | 88.89 | 100.0 | 80.00 | 80.00 |
| 3 | 91.67 | 100.0 | 86.96 | 84.05 |
| 4 | 86.11 | 100.0 | 76.19 | 75.59 |
| 5 | 88.89 | 89.47 | 89.47 | 80.19 |
| Average | 87.78 ± 3.17 | 92.63 ± 11.53 | 85.19 ± 6.70 | 78.32 ± 4.72 |

### 2.2. Comparison with the SVM-Based Method

The proposed method has achieved good prediction accuracy. In order to further evaluate the prediction performance of the RVM classifier, the comparison of prediction accuracy between the RVM classifier and the state-of-the-art support vector machine (SVM) classifier was carried out through employing the same feature extraction method on enzyme and ion channel datasets. We also adopted five-fold cross-validation tests to assess the prediction accuracy of the SVM classifier. The LIBSVM tool [25] of SVM was used to execute classification. In the experiment, we also optimized several parameters of the SVM classifier. We selected the radial basis function (RBF) as the kernel function, and the c and g parameters of the RBF kernel were set up (c = 0.5 and g = 0.6) by using a grid search method.

The comparison prediction results of RVM and SVM classifiers on enzyme and ion channel datasets are listed in Tables 5 and 6, respectively. At the same time, the comparison of ROC Curves between RVM and SVM classifiers are also shown in Figures 2 and 3 on enzyme and ion channel datasets, respectively. As displayed in Table 5, the RVM classifier obtained 97.73% average accuracy on the enzyme dataset, while 91.15% average accuracy was achieved by the SVM classifier. Similarly, it can be seen form Table 6 that 93.12% average accuracy was obtained by the RVM classifier and 87.77% average accuracy was achieved by the SVM classifier on the ion channel dataset. It can be observed from these results that the prediction accuracy obtained by the RVM classifier is significantly higher than that of the SVM classifier. In addition, as displayed in Figures 2 and 3, the ROC curves of the RVM classifier is also obviously better than that of the SVM classifier. The proposed method obtained good prediction results which may be attributable to two reasons: (1) because the RVM classifier greatly reduces the amount of calculation of the kernel function relative to the SVM classifier; which helps in improving the prediction performance; (2) the kernel functions required to meet the condition of Mercer is the obvious disadvantage of the SVM classifier; however, the RVM classifier overcame it and solved the problem. Thus, all of these experimental results indicate that the proposed prediction model might become a useful tool for predicting DTI, as well as performing other bioinformatics tasks.

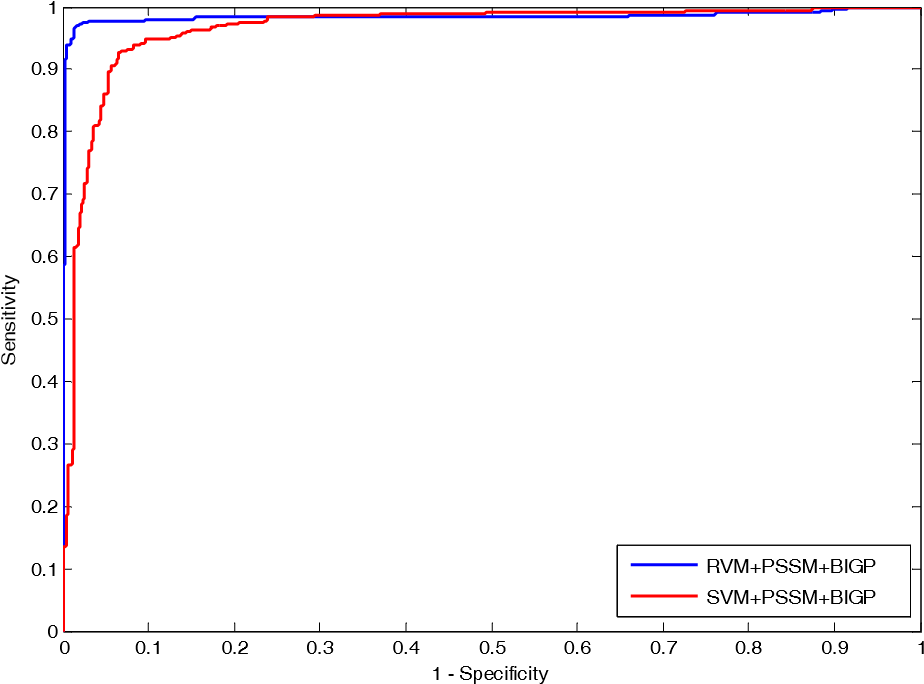
**Table 5.** 5-fold cross validation results performed by SVM and RVM classifiers on an enzyme dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing Set** | **Ac (%)** | **Sn (%)** | **Pe (%)** | **Mcc (%)** |
| RVM + PSSM + BIGP  1 | 97.95 | 98.31 | 97.65 | 95.98 |
| 2 | 97.52 | 95.84 | 99.31 | 95.16 |
| 3 | 97.26 | 97.29 | 97.29 | 94.68 |
| 4 | 98.29 | 98.44 | 98.10 | 96.64 |
| 5 | 97.61 | 97.34 | 97.69 | 95.33 |
| Average | 97.73 ± 0.40 | 97.44 ± 1.04 | 98.01 ± 0.78 | 95.56 ± 0.76 |
| SVM + PSSM + BIGP  1 | 90.94 | 90.56 | 91.48 | 83.52 |
| 2 | 89.49 | 91.18 | 88.67 | 81.15 |
| 3 | 90.60 | 93.06 | 88.85 | 82.93 |
| 4 | 92.48 | 94.11 | 90.95 | 86.08 |
| 5 | 92.24 | 93.97 | 90.29 | 85.67 |
| Average | 91.15 ± 1.23 | 92.57 ± 1.62 | 90.05 ± 1.25 | 83.87 ± 2.03 |

**Table 6.** 5-fold cross validation results performed by SVM and RVM classifier on an ion channel dataset.

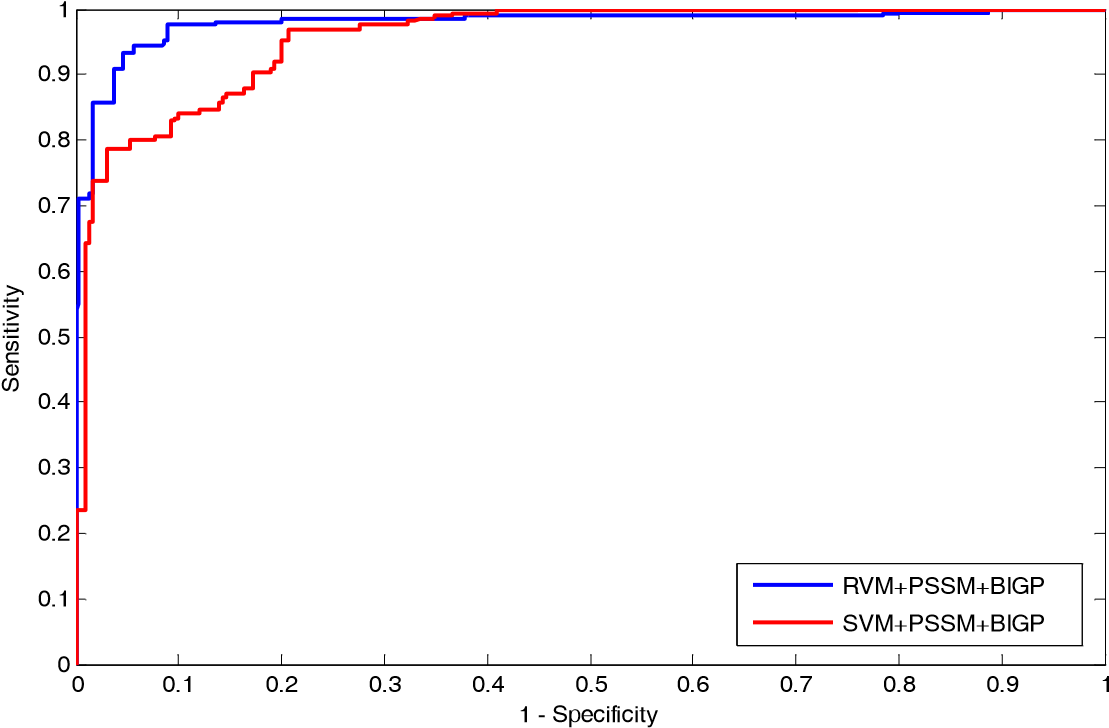
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing Set** | **Ac (%)** | **Sn (%)** | **Pe (%)** | **Mcc (%)** |
| RVM + PSSM + BIGP  1 | 92.71 | 91.18 | 94.58 | 86.48 |
| 2 | 91.02 | 92.31 | 89.49 | 83.64 |
| 3 | 94.41 | 94.46 | 94.14 | 89.44 |
| 4 | 93.39 | 93.81 | 94.10 | 87.55 |
| 5 | 94.07 | 94.85 | 92.47 | 87.18 |
| Average | 93.12 ± 1.34 | 93.32 ± 1.54 | 92.96 ± 2.10 | 87.18 ± 2.28 |
| SVM + PSSM+ BIGP  1 | 86.78 | 84.31 | 89.58 | 77.03 |
| 2 | 88.47 | 90.56 | 86.33 | 79.59 |
| 3 | 86.10 | 89.62 | 83.28 | 76.03 |
| 4 | 88.45 | 86.07 | 92.36 | 79.51 |
| 5 | 89.02 | 91.91 | 85.32 | 80.39 |
| Average | 87.77 ± 1.26 | 88.49 ± 3.18 | 87.37 ± 3.59 | 78.51 ± 1.87 |

Comparison of ROC Curves between RVM and SVM on enzyme dataset



**Figure 2.** Comparison of ROC curves performed between RVM and SVM on an enzyme dataset.

Comparison of ROC Curves between RVM and SVM on ion channel dataset



**Figure 3.** Comparison of ROC curves performed between RVM and SVM on an ion channel dataset.

### 2.3. Comparison with Other Methods

Up to now, a number of computational methods have been proposed for predicting drug target interactions. In our study, in order to further evaluate the prediction performance of the proposed method, we compared its prediction accuracy with four existing DTI predictors; DBSI [26], Yamanishi [27], KBMF2K [28], and NetCMP [29] on enzyme, ion channel, GPCR, and nuclear receptor datasets, respectively. These methods use the same strategy as the proposed method, however, they adopt different feature extraction methods and classifiers. Table 7 displays these comparison results. It can be observed from Table 7 that the prediction accuracy of the proposed approach is significantly higher than the other four methods on enzyme, ion channel, GPCR, and nuclear receptor datasets. The comparison results further demonstrated that the PDTPS can improve the prediction accuracy relative to current approaches. Due to using a good classifier and a novel feature extraction method, the proposed method achieved good prediction results. This makes the PDTPS a useful tool and suitable for predicting DTI.

**Table 7.** Comparison of predicting performance between our method and other methods on four Datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Our Method** | **DBSI [26]** | **Yamanishi [27]** | **KBMF2K [28]** | **NetCMP [29]** |
| Enzymes | 0.9773 | 0.8075 | 0.821 | 0.832 | 0.8251 |
| Icon Channels | 0.9312 | 0.8029 | 0.692 | 0.799 | 0.8034 |
| GPCRs | 0.8677 | 0.8022 | 0.811 | 0.857 | 0.8235 |
| Nuclear Receptors | 0.8778 | 0.7578 | 0.814 | 0.824 | 0.8394 |

## 3. Materials and Methods

### 3.1. Dataset

In this study, we carried out the experiment using the proposed method on four protein targets datasets: enzymes, ion channels, GPCRs, and nuclear receptors. These data can be freely obtained from the KEGG BRITE [7], BRENDA [30], SuperTarget [6], and Drug Bank [8] databases and were used as the gold-standard datasets by Yamanishi et al [27] The number of drugs known to target enzymes, ion channels, GPCRs, and nuclear receptors are 445, 210, 233, and 54, respectively.（药物数量是这些） The numbers of proteins known to be targeted by the drugs are 664, 204, 95, and 26 respectively.（蛋白质数量是这些） These drug–target pairs were carefully screened, 5127 pairs of them are known to interact with each other.（相互作用有这些） The numbers of known interactions involving enzymes, ion channels, GPCRs, and nuclear receptors are 2926, 1476, 635, and 90, respectively. （相互作用分别是这些）Then, all known interactions of the drug–target pairs were chosen as positive sample sets for four datasets in our experiment.（这些中作为正样本集）

A bipartite（双边的） graph is usually used to represent a drug–target interaction network, （双向图一般用于描绘药物靶点关系）whose nodes represent target proteins or drug molecules（端点代表药物或者蛋白质） and the edges describe the real drug–target interactions that have been already identified through experiments or other ways（边代表已经确定的连接）. It can be observed from bipartite graph that the number of the real drug–target interactions edges are small.（可以看出边少） Here, we take the enzyme dataset as an example; there are a total of 295,480 (445 × 664) connections in the corresponding bipartite and only 2926 edges of them are known drug–target interactions.（举例说明，对数多，有关系的少） Therefore, the possible number of negative samples (295,480 − 2926 = 29,2554) is significantly more than the number of positive samples (2926), which is a bias problem.（是一种偏见问题） In order to solve this problem, we randomly selected the negative samples as much as the positive sample.（随机选正负样本） As a result, there are 2926, 1476, 635, and 90 negative samples of enzymes, ion channels, GPCRs, and nuclear receptors datasets. （所以随机选取了相同数量的负样本）In other words, there are 5852, 2952, 1270, and 180 drug–target pairs of enzymes, ion channels, GPCRs, and nuclear receptors datasets in the experiment.（那负样本来源于是哪呢？）

### 3.2. Position Specific Scoring Matrix

Position Specific Scoring Matrix (PSSM) can be represented an M × 20 matrix *M* =

*Mij i* : 1 = 1 . . . *M*, *j*  , where M represents the length of a given protein sequence, 20 is the number of 20 amino acids, and *Mij* represents the score of the *jth* amino acid relative to the *ith* position for a query protein sequence [31]. The score *Mij* can be expressed as *Mij* =

20

∑ *p*(*i*, *k*) × *q*(*j*, *k*), where *p*(*i*, *k*) represents the appearing frequency of the *kth* amino acid at *k* = 1

position *i* of the probe, and *q*(*i*, *k*) is the value of Dayhoff’s mutation matrix between *jth* and *kth* amino acids. Thus, a high score represents a highly-conserved position; on the contrary, a low score represents a weakly-conserved position.

In the study, in order to create experimental datasets, we used Position Specific Iterated BLAST

(PSI-BLAST) [32] to construct PSSMs for each protein sequence. The e-value and number of iterations are set up as the default values in PSI-BLAST. For achieving highly and widely homologous sequences, an e-value of 0.001 and three iterations were selected. It is possible that features may be different if we use different parameters, however, in the work we concentrated on exploring general PSSM features for predicting DTI by employing mostly default settings. Thus, each PSSMs feature vector can be represented as M × 20 matrix by using PSI-BLAST, where M is the number of residues of a given protein sequence and the 20 columns are the number of 20 amino acids.

### 3.3. Bi-Gram Probabilities

The Bi-gram Probabilities (BIGP) have been used for protein fold recognition. （双谱概率(BIGP)被用于蛋白质折叠识别。） In the literature [33], it was described how to use a given protein’s original primary sequence or its consensus sequence for protein fold recognition. Instead, we employed the BIGP feature extraction method that the literature [34] proposed to represent a given protein sequence based on its PSSM (PSSM has been mentioned in the Section 3.2 of the paper). In detail, the bi-gram feature vector was computed through counting the bi-gram frequencies of occurrence in PSSM. It is assumed that *P* represents the PSSM of a protein sequence, which contains *L* rows and 20 columns, where *L* is the length of a given protein sequence and 20 columns represents a number of 20 amino acids. The PSSM element *Pij* can be interpreted as the relative probability of *jth* amino acid at the *ith* location of the primary protein

20

sequence, *Pij* can be expressed as *Pij* = ∑ *i* : 1 = 1 . . . *L*, *j* = 1 . . . 20. The frequency of occurrence

*j* = 1

of transition from *mth* amino acid to *nth* amino acid can be defined as follows:

*L*−1

*BIGPmn* = ∑ *Pi*,*mPi*+1,*n*1 ≤ *m* ≤ 20, 1 ≤ *n* ≤ 20 (1)

*i* = 1

Equation (1) gives 400 frequencies of occurrence *BIGPmn* for 400 bi-gram transitions, the matrix BIGP called the bi-gram occurrence matrix, the number of the 400 whose elements represent the bi-gram feature vector [34] are as follows:

*BF* = [*BGP*1,1, *BGP*1,2 . . . *BGP*1,20, *BGP*2,1, . . . *BGP*2,20, . . . . . . *BGP*20,1, . . . *BGP*20,20] (2)

These bi-gram features can also be expressed as follows:

*BF* = [*ϕ*1,, *ϕ*2, *ϕ*3, . . . *ϕu*,, . . . *ϕθ*] (3)

where *θ* = mn = 400 is the dimensionality of the feature vector BF, the *ϕ*u can be represented as follows:

 *BGP*1,*u* (1 ≤ *u* ≤ 20)

 *BGP*2,*u*−20 (21 ≤ *u* ≤ 40)

*ϕu* = . . . . . . (4)

 *BGP*20,*u*−380 (381 ≤ *u* ≤ 400)

Finally, each protein sequence was converted into a 400-dimensional vector by using BIGP method. In the paper, to reduce the influence of noise and improve the prediction accuracy, the dimensions of enzymes, ion channels, GPCRs, and nuclear receptors datasets were reduced from 400 to 350 by using Principal Component Analysis (PCA) method.

### 3.4. Relevance Vector Machine

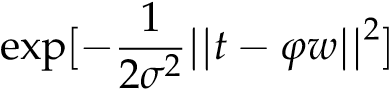
The related theory of the Relevance Vector Machine describes in details in the literature [35]. We assumed {*xn*, *tn*}*nN*= 1, *xn* ∈ *Rd* is the training set for binary classification question, where *tn* ∈ {0, 1} represents the training set label, *ti* is the testing set label, and *ti* = *yi* + *εi*, where

*N*

*yi* = *wTϕ*(*xi*) = ∑ *wjK xi*, *xj*) + *w*0 is the classification model; *εi* is the additional noise, with a

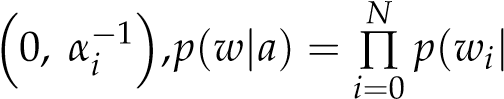
*j* = 1

mean value of zero and a variance of *σ*2, where *εi* ∼ *N*(0, *σ*2), *ti* ∼ *N*(*yi*, *σ*2). It is assumed that the training sets are independent and identically distributed; the vector *t* submits to as follows distribution:

*p**x*, *w*, *σ*2 ) = (2*πσ*2)−*N*/2  (5)

|  |  |
| --- | --- |
| where *ϕ* is defined as follows: |  |
|    1 *k*(*x*1, *x*1)··· *k*(*x*1, *xN*)  *ϕ* =  . . . . . . . . .      1 *k*(*xN*, *x*1) . . . *k*(*xN*, *xN*)  The training set label *t* is employed to detect the testing set label *t*∗, given by | (6) |
| Z *p*(*t*∗|*t*) = *p*(*t*∗|*w*, *σ*2)*p*(*w*, *σ*2|*t*)*dwdσ*2 | (7) |

Due to making the value of most components of the weight vector *w* zero and reducing the number of calculation of the kernel function, additional conditions are attached to the weight vector *w* Assuming that *wi* obeys a distribution with a mean value of zero and a variance of *αi*−1, the mean

*wi* ∼ *N**ai*) where *α* is a hyper-parameter vector of the prior distribution of the weight vector *w*.

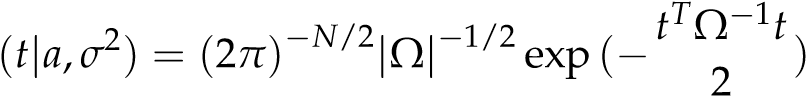
Z

|  |  |
| --- | --- |
| p(t∗|t) = p(t∗|w, a, σ2)p(w, a, σ2|t)dwdadσ2 | (8) |
| *p*(*t*∗|*w*, *a*, *σ*2) = *N*(*t*∗ |*y*(*x*∗; *w*), *σ*2) | (9) |

Because *p*(*w*, *a*, *σ*2|*t*) cannot be obtained by an integral, it must be resolved using a Bayesian formula, given as

|  |  |
| --- | --- |
| *p*(*w*, *a*, *σ*2|*t*) = *p*(*w*|*a*, *σ*2, *t*)*p*(*a*, *σ*2|*t*) | (10) |
| *p*(*w*|*a*, *σ*2, *t*) = *p*(*t*|*w*, *σ*2)*p*(*w*|*a*)/*p*(*t*|*a*, *σ*2) | (11) |

The integral of the product of *p w*, *a*,  and *p*(*w*|*a*) is as follows:

*p* (12)

|  |  |
| --- | --- |
| Ω = *σ*2*I* + *ϕA*−1*ϕT*, *A* = *diag*(*a*0,*a*1, . . . , *aN*) | (13) |
| *T*  *p*(*w*|*a*, *σ*2, *t*) = (2*π*)−(*N*+1)/2|Σ|−1/2 exp (−(*w* − *u*) (*w* − *u*)) 2 | (14) |
| Σ = (*σ*−2*ϕTϕ* + *A*)−1 | (15) |
| *u* = *σ*−2Σ*ϕTt* | (16) |

Because *p*(*a*, *σ*2|*t*)∝ *p*(*t*|*a*, *σ*2)*p*(*a*)*p*(*σ*2) and *p*(*a*, *σ*2|*t*) cannot be solved by means of integration, the solution is approximated using the maximum likelihood method, represented by

|  |  |
| --- | --- |
| (*aMP*, *σMP*2 ) = arg *maxp*(*t*|*a*, *σ*2)  *a*,*σ*2  The iterative process of *aMP* and *σMP*2 is given by: | (17) |

 (*σ new* ||*tµ*−*i ϕµ*||2 (18) 2) =







*a*

*new*

*i*

=

*γ*

*i*

2

*N*

−

∑

*N*

*i*

=

0

*µi*

 *γi* = 1 − *ai* ∑ *i*, *i*

Here ∑ *i*, *i* is *i*th element in the Σ diagonal and the initial value of α and *σ*2 can be decided via the approximation of *aMP* and *σMP*2 using Formula (15) continuously updated. After enough iterations, most of *ai* will be close to infinity, the corresponding parameters in *wi* will be zero, and other *ai* values will be close to finite. The resulting corresponding parameters *xi* of *ai* are now referred to as the

relevance vector.

### 3.5. Performance Evaluation

In the paper, we used the following evaluation criteria as a measure for evaluating the performance of the proposed classifier and feature extraction method in our experiment. There are Ac (Accuracy), Sn

(Sensitivity), Pe (precision), and Mcc (Matthews’s correlation coefficient). The definition is as follows:

Ac = *TP*+*FPTP*++*TNTN*+*FN*

Sn = *TPTP*+*TN*

Pe = *FPTP*+*TP* (19)

Mcc

) × (*TN*+*FN*)

=

(

*TP*

×

*TN*

)

−

(

*FP*

×

*FN*

)

√

(

*TP*

+

*FN*

)

×

(

*TN*

+

*FP*

)

×

(

*TP*

+

*FP*

where true positives (*TP*) represents the number of positive pairs that are predicted as interacting drug–target pairs, false positives (*FP*) is the count of negative pairs that are predicted as interacting drug–target pairs, true negatives (*TN*) is the total of negative pairs that are predicted as non-interacting drug–target pairs and false negatives (*FN*) represents the number of positive pairs that are predicted as non-interacting drug–target pairs. In addition, the Receiver Operating Curve (ROC) was established to evaluate the performance of the proposed approach in the experiment.

## 4. Conclusions

In the paper, we proposed a novel computational approach based on protein sequence, namely PDTPS (Predicting Drug Targets with Protein Sequence), to predict drug–target interactions (DTI).

The PDTPS method combines bi-gram probabilities (BIGP), Position Specific Scoring Matrix (PSSM), and Principal Component Analysis (PCA) with Relevance Vector Machine (RVM). In order to evaluate the prediction capacity of the PDTPS, we carried out the method on enzyme, ion channel, GPCR, and nuclear receptor datasets by using five-fold cross-validation tests. The proposed PDTPS method achieved average accuracy of 97.73%, 93.12%, 86.78%, and 87.78% on enzyme, ion channel, GPCR, and nuclear receptor datasets, respectively. The experimental results showed that our method has good prediction performance. Furthermore, in order to evaluate the prediction performance of the proposed PDTPS method, we compared it with the state-of-the-art support vector machine (SVM) classifier on enzyme and ion channel datasets and other existing methods on four datasets. The promising comparison results further demonstrate the efficiency and robustness of the proposed PDTPS method. This makes it a useful tool and suitable for predicting DTI, as well as performing other bioinformatics tasks. For future studies, more effective feature extraction approaches and machine learning algorithms can be developed for predicting DTI.

**Acknowledgments:** This work is supported in part by the National Science Foundation of China, under Grants

11631014, 61373086, 11301517, 61572506, in part by Guangdong Natural Science Foundation, under Grant 2014A030313555, and in part by the Shenzhen Scientific Research and Development Funding Program under grants JCYJ20140418095735569. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

**Author Contributions:** Fan-Rong Meng, Zhu-Hong You and Xing Chen conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; Yong Zhou and Ji-Yong An designed, performed and analyzed experiments and wrote the manuscript; all authors read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

# References

1. Wang, Y.C.; Yang, Z.X.; Wang, Y.; Deng, N.Y. Computationally Probing Drug-Protein Interactions via Support Vector Machine. *Lett. Drug Des. Discov.* **2010**, *7*, 370–378. [[CrossRef]](http://dx.doi.org/10.2174/157018010791163433)
2. Xia, Z.; Wu, L.Y.; Zhou, X.; Wong, S.T. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* **2010**, *4*, S6. [[CrossRef]](http://dx.doi.org/10.1186/1752-0509-4-S2-S6) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/20840733)
3. Landry, Y.; Gies, J.P. Drugs and their molecular targets: An updated overview. *Fundam. Clin. Pharmacol.* **2008**, *22*, 1–18. [[CrossRef]](http://dx.doi.org/10.1111/j.1472-8206.2007.00548.x) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/18251718)
4. Li, Q.; Lai, L. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinform.* **2007**, *8*, 1–11. [[CrossRef]](http://dx.doi.org/10.1186/1471-2105-8-353) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/17883836)
5. Overington, J.P.; Allazikani, B.; Hopkins, A.L. How many drug targets are there? *Nat. Rev. Drug Discov.* **2006**, *5*, 993–996. [[CrossRef]](http://dx.doi.org/10.1038/nrd2199) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/17139284)
6. Günther, S. SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **2008**, *36*, 919–922. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkm862) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/17942422)
7. Kanehisa, M.; Goto, S.; Hattori, M.; Aokikinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **2005**, *34*, 354–357. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkj102) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/16381885)
8. Wishart, D.S. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkm958) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/18048412)
9. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J.

DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, 668–672. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkj067) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/16381955)

1. Chen, X.; Ji, Z.L.; Chen, Y.Z. TTD: Therapeutic Target Database. *Nucleic Acids Rese.* **2002**, *30*, 412. [[CrossRef]](http://dx.doi.org/10.1093/nar/30.1.412)
2. Zhu, F.; Han, B.C.; Kumar, P.; Liu, X.H.; Ma, X.H.; Wei, X.N.; Huang, L.; Guo, Y.F.; Han, L.Y.; Zheng, C.J. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2010**, *38*, D787. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkp1014) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/19933260)
3. Kannadasan, R.; Saleembasha, M.S.; Emerson, I.A. A Frame Work for Learning Drug Designing through

Molecular Modelling Software Techniques and Biological Databases for Protein-Ligand Interactions. *Int. J.*

*Eng. Res. Afr.* **2016**, *27*, 111–118. [[CrossRef]](http://dx.doi.org/10.4028/www.scientific.net/JERA.27.111)

1. Rabelo, V.W.; Santos, T.F.; Terra, L.; Santana, M.V.; Castro, H.C.; Rodrigues, C.R.; Abreu, P.A. Targeting

CYP51 for drug design by the contributions of molecular modeling. *Fundam. Clin. Pharmacol.* **2016**, *31*, 37–53. [[CrossRef]](http://dx.doi.org/10.1111/fcp.12230) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/27487199)

1. Butina, D.; Segall, M.D.; Frankcombe, K. Predicting ADME properties in silico: Methods and models. *Drug Discov. Today（古董级论文也拿来说，这真是没谁了。。。）* **2002**, *7*, S83–S88. [[CrossRef]](http://dx.doi.org/10.1016/S1359-6446(02)02288-2)
2. Cheng, A.C.; Coleman, R.G.; Smyth, K.T.; Cao, Q.; Soulard, P.; Caffrey, D.R.; Salzberg, A.C.; Huang, E.S.

Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75. [[CrossRef]](http://dx.doi.org/10.1038/nbt1273) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/17211405)

1. Coleman, R.G.; Salzberg, A.C.; Cheng, A.C. Structure-based identification of small molecule binding sites using a free energy model. *J. Chem. Inf. Model.* **2006**, *46*, 2631. [[CrossRef]](http://dx.doi.org/10.1021/ci600229z) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/17125203)
2. Sousa, S.F.; Fernandes, P.A.; Ramos, M.J. Protein-ligand docking: Current status and future challenges. *Proteins Struct. Funct. Bioinform.* **2006**, *65*, 15–26. [[CrossRef]](http://dx.doi.org/10.1002/prot.21082) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/16862531)
3. Yang, K.; Bai, H.; Ouyang, Q.; Lai, L.; Tang, C. Finding multiple target optimal intervention in disease-related molecular network. *Mol. Syst. Biol.* **2008**, *4*, 228. [[CrossRef]](http://dx.doi.org/10.1038/msb.2008.60) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/18985027)
4. Niu, Y.Q. Supervised prediction of drug-target interactions by ensemble learning. *J. Chem. Pharm. Res.* **2014**, *6*, 1991–1999.
5. Kuang, Q.; Xu, X.; Li, R.; Dong, Y.; Li, Y.; Huang, Z.; Li, Y.; Li, M. An eigenvalue transformation technique for predicting drug-target interaction. *Sci. Rep.* **2015**, *5*, 13867. [[CrossRef]](http://dx.doi.org/10.1038/srep13867) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/26350590)
6. Bharadwaja, A. Similarity Based Learning Method for Drug taRget Interaction Prediction. M.Sc. Thesis, University of Windsor, Windsor, ON, Canada, 2014.
7. Peng, L.; Liao, B.; Zhu, W.; Li, K. Predicting Drug-Target Interactions with Multi-information Fusion. *IEEE J.*

*Biomed. Health Inform.* **2015**, *21*, 561–572. [[CrossRef]](http://dx.doi.org/10.1109/JBHI.2015.2513200) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/26731781)

1. Wang, Y.Y.; Nacher, J.C.; Zhao, X.M. Predicting drug targets based on protein domains. *Mol. Biosyst.* **2012**, *8*, 1528–1534. [[CrossRef]](http://dx.doi.org/10.1039/c2mb05450g) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/22402667)
2. Zhang, R. *An Ensemble Learning Approach for Improving Drug–Target Interactions Prediction*; Springer International Publishing: New York, NY, USA, 2015; pp. 433–442.
3. Chang, C.C.; Lin, C.J. *LIBSVM: A Library for Support Vector Machines*; ACM: New York, NY, USA, 2011; pp. 1–27.
4. Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of Drug-Target

Interactions and Drug Repositioning via Network-Based Inference. *PLos Comput. Biol.* **2012**, *8*, 357–372. [[CrossRef]](http://dx.doi.org/10.1371/journal.pcbi.1002503) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/22589709)

1. Yamanishi, Y.; Araki, M.A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232–i240. [[CrossRef]](http://dx.doi.org/10.1093/bioinformatics/btn162) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/18586719)
2. Gönen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **2012**, *28*, 2304–2310. [[CrossRef]](http://dx.doi.org/10.1093/bioinformatics/bts360) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/22730431)
3. Zong, W.; Huang, G.B.; Chen, Y. Weighted extreme learning machine for imbalance learning. *Neurocomputing* **2013**, *101*, 229–242. [[CrossRef]](http://dx.doi.org/10.1016/j.neucom.2012.08.010)
4. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. BRENDA, the enzyme database: Updates and major new developments. *Nucleic Acids Res.* **2004**, *32*, 431–433. [[CrossRef]](http://dx.doi.org/10.1093/nar/gkh081) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/14681450)
5. Gribskov, M.; Mclachlan, A.D.; Eisenberg, D. Profile analysis: Detection of distantly related proteins. *Proc. Nat. Acad. Sci. USA* **1987**, *84*, 4355–4358. [[CrossRef]](http://dx.doi.org/10.1073/pnas.84.13.4355) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/3474607)
6. Altschul, S.F.; Koonin, E.V. Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases. *Trends Biochem. Sci.* **1998**, *23*, 444–447. [[CrossRef]](http://dx.doi.org/10.1016/S0968-0004(98)01298-5)
7. Ghanty, P.; Pal, N.R. Prediction of Protein Folds: Extraction of New Features, Dimensionality Reduction, and Fusion of Heterogeneous Classifiers. *IEEE Trans. Nanobiosci.* **2009**, *8*, 100–110. [[CrossRef]](http://dx.doi.org/10.1109/TNB.2009.2016488) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/19278932)
8. Sharma, A.; Lyons, J.; Dehzangi, A.; Paliwal, K.K. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *Nanobiosci. IEEE Trans.* **2012**, *320*, 41–46. [[CrossRef]](http://dx.doi.org/10.1016/j.jtbi.2012.12.008) [[PubMed]](http://www.ncbi.nlm.nih.gov/pubmed/23246717)
9. Tipping, M.E. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.

**Sample Availability:** Samples of the compounds are available from the authors.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license [(http://creativecommons.org/licenses/by/4.0/)](http://creativecommons.org/licenses/by/4.0/./).

1. **. Introduction**

   The identification of drug–target interactions (DTI) has recently emerged as an area of intense research activity due to its important role in finding new proteins to target for drug development and discovering new drug candidates [1,2]（最近比较火热）. However, the target proteins of many drugs are not complete or even not known. （然而大家找不到）In the past years, much effort has been devoted to using experimental methods to identify drug–protein interactions（之前做了很多的努力）. But these experimental methods are both time-consuming and expensive（但是它们耗费时间和精力）. It often costs billions of dollars for developing a successful novel chemistry-based drug and takes nearly a decade for introducing the drug to market（花钱花时间多重新说一遍）. However, there are only few drug candidates that can be approved to reach the market by Food and Drug Administration (FDA) [3–5]（最后发现）. This is partially caused by the unacceptable toxicity（毒性） for those drug candidates with the satisfactory activity, due to the deficient（缺少） of the knowledge of drug–target interactions. Thus, it is necessary to

   *Molecules* **2017**, *22*, 1119; doi[:10.3390/molecules22071119](http://dx.doi.org/10.3390/molecules22071119) [www.mdpi.com/journal/molecules](http://www.mdpi.com/journal/molecules) [↑](#footnote-ref-1)