biology

MDPI

Article

MSPEDTI: Prediction of Drug–Target Interactions via Molecular Structure with Protein Evolutionary Information

Lei Wang [1,2,*], Leon Wong [1], Zhan-Heng Chen [3], Jing Hu [2], Xiao-Fei Sun [2], Yang Li [4] and Zhu-Hong You [1,5,*]

[1] Big Data and Intelligent Computing Research Center, Guangxi Academy of Sciences, Nanning 530007, China; lghuang@gxas.cn
[2] College of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China; hujing@uzz.edu.cn (J.H.); sxf@uzz.edu.cn (X.-F.S.)
[3] Computer Science and Technology, Tongji University, Shanghai 200092, China; chenzhanheng17@mails.ucas.ac.cn
[4] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China; 2021010123@mail.hfut.edu.cn
[5] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China
* Correspondence: leiwang@gxas.cn (L.W.); zhuhongyou@gmail.com (Z.-H.Y.); Tel.: +86-151-0632-2257 (L.W.); +86-173-9276-3836 (Z.-H.Y.)

Simple Summary: Drug discovery is the process of identifying potential new compounds through biological, chemical, and pharmacological means. Billions of dollars are spent each year on research aimed at discovering, designing, and developing new drugs for a wide range of diseases. However, the research and development of new drugs remain time-consuming and sometimes difficult to complete. With the development of new experimental techniques, huge amounts of data are generated at different stages of drug development. Biomedical research, especially in the field of drug discovery, is currently undergoing a major shift towards "big data" applications of artificial intelligence technologies. Therefore, a key challenge for future drug discovery research is the development of robust artificial-intelligence-based predictive tools for drug–target interactions (DTIs) that can study biomedical problems from multiple perspectives. In this study, a deep-learning-based prediction model for DTIs was designed by combining information on drug structure and protein evolution to provide theoretical support for drug research.

Abstract: The key to new drug discovery and development is first and foremost the search for molecular targets of drugs, thus advancing drug discovery and drug repositioning. However, traditional drug–target interactions (DTIs) is a costly, lengthy, high-risk, and low-success-rate system project. Therefore, more and more pharmaceutical companies are trying to use computational technologies to screen existing drug molecules and mine new drugs, leading to accelerating new drug development. In the current study, we designed a deep learning computational model MSPEDTI based on **M**olecular **S**tructure and **P**rotein **E**volutionary to predict the potential **DTI**s. The model first fuses protein evolutionary information and drug structure information, then a deep learning convolutional neural network (CNN) to mine its hidden features, and finally accurately predicts the associated DTIs by extreme learning machine (ELM). In cross-validation experiments, MSPEDTI achieved 94.19%, 90.95%, 87.95%, and 86.11% prediction accuracy in the gold-standard datasets enzymes, ion channels, G-protein-coupled receptors (GPCRs), and nuclear receptors, respectively. MSPEDTI showed its competitive ability in ablation experiments and comparison with previous excellent methods. Additionally, 7 of 10 potential DTIs predicted by MSPEDTI were substantiated by the classical database. These excellent outcomes demonstrate the ability of MSPEDTI to provide reliable drug candidate targets and strongly facilitate the development of drug repositioning and drug development.

Keywords: deep learning; drug–target interactions; extreme learning machine; convolutional neural network

Citation: Wang, L.; Wong, L.; Chen, Z.-H.; Hu, J.; Sun, X.-F.; Li, Y.; You, Z.-H. MSPEDTI: Prediction of Drug–Target Interactions via Molecular Structure with Protein Evolutionary Information. *Biology* **2022**, *11*, 740. https://doi.org/10.3390/biology11050740

Academic Editor: Armando Varela-Ramirez

Received: 22 April 2022
Accepted: 4 May 2022
Published: 13 May 2022

1. Introduction

Drug research is a global development problem. In the past few decades, the drug-targeted therapy strategy has achieved great success [1,2]. Finding specific drugs for targets is the focus of pharmaceutical research and development, which has made an indelible contribution to human health [3]. However, the rate of new drug development has been declining in recent years, and the cost of research and development has been rising [4]. The main reason for this is that the early screening of a large number of drug candidates in drug research still relies mainly on time-consuming and labor-intensive experimental methods, and the later discovery of unsatisfactory efficacy or toxic side effects of drugs leads to the failure of development. Therefore, efficient and high-throughput computational techniques in the early stages of drug research can play an important role in targeting and saving costs in early development [5–8].

With the rapid development of bioinformatics, many achievements have been achieved by using computational and simulation approaches to predict DTIs. Quantitative structure–activity relationship (QSAR) utilizes the physicochemical properties or structural parameters of the molecule to quantitatively study the interaction between small molecules and biological macromolecules by means of mathematics. Casañola-Marti et al. proposed a QSAR model for predicting anti-tyrosinase activity and demonstrated the effectiveness of the model in subsequent in vitro experiments, which greatly increased the rate of biochemical discovery of skin disease treatment [9]. Kar et al. proposed an approach to predict the carcinogenicity of drug compounds based on QSAR, which has been identified as a key factor in carcinogenicity by analyzing the contribution of molecular fragments to carcinogenicity [10]. Molecular docking (MD) is a computational simulation method for studying the optimal binding sites between drug molecules and target proteins by structural matching and energy matching and predicting their binding patterns and affinity [11]. Wallach et al. proposed a model to normalize docking scores through the virtually generated bait set that avoids the variability due to changes in physical properties when identifying active compounds in large screening libraries, thereby extending the applicability of the model [12].

Recently, computational methods for predicting DTIs based on protein target sequences have achieved excellent results and are favored by researchers for their use of reliable, high-quality characterization information enriched by raw data to ensure the accuracy of prediction results [13–18]. For instance, Lan et al. proposed a PUDT model combining protein target sequences and drug compound structures, which greatly improved the accuracy of DTI prediction using a weighted SVM classifier [19]. Cao et al. aimed to predict DTIs by using an extended structure–activity relationship method at the genome-scale level. In subsequent experiments, this approach gained good results [20].

In the present study, we combined protein sequence evolution with drug structure information to propose a deep learning MSPEDTI model to predict hidden DTIs. Concretely, MSPEDTI first fuses protein sequence information characterized by the Position-Specific Scoring Matrix (PSSM) and drug structure information characterized by molecular fingerprinting, and then automatically extracts them into continuous, low-dimensional, information-rich features using a deep learning CNN, thus avoiding the disadvantages of manual features such as tediousness, sparsity, and high dimensionality. Finally, the ELM classifier is used to accurately determine whether drug–target pairs are associated or not. In the gold-standard dataset, we evaluated MSPEDTI using the five-fold cross-validation (5CV) approach. Compared with other previous methods, MSPEDTI was able to learn valid biological characteristics for predicting DTIs and showed better performance. The robustness of MSPEDTI is also demonstrated by the experimental results of the case study, which can provide effective candidate targets for new drug research. The supporting data used in this study can be downloaded from https://github.com/look0012/MSPEDTI (accessed on 1 April 2022).

2. Materials and Methods

2.1. Gold-Standard Datasets

In the present study, we implemented the MSPEDTI model using the gold-standard datasets enzyme, GPCR, ion channel, and nuclear receptor, which were collated by Yamanishi et al. [21] from the BRENDA [22], KEGG [23,24], SuperTarget [25], and DrugBank [26] databases. After removing the redundant information, the numbers of DTI pairs contained in these datasets are 2926, 635, 1467, and 90, respectively. All of these pairs are constructed as positive datasets. Table 1 presents the statistical information for these gold-standard datasets.

Table 1. Statistical information for the four gold-standard datasets: the number of target proteins, drugs, and interaction pairs. Sparsity is the ratio of positive DTIs to all possible interactions.

Dataset	Target Proteins	Drugs	Interactions	Sparsity
Enzymes	664	445	2926	0.0099
Ion Channels	204	210	1467	0.0344
GPCRs	95	223	635	0.0299
Nuclear Receptors	26	54	90	0.0641

The corresponding negative dataset construction process is as follows: firstly, all drug–target interaction pairs are divided into drug and target components; secondly, these drug and target are recombined into DTI pairs, and the pairs of interactions are removed. Finally, these drug–target pairs are randomly selected to construct the negative dataset, which is the same size as the positive dataset.

2.2. Drug Structure Characterization

We employed molecular fingerprints in this study to characterize the drug structures for the purpose of numerical conversion. The design idea of fingerprints is to characterize the molecular structure using the form of a dictionary collection of molecular fragments, which converts a drug molecule into a binary vector of values by determining whether certain fragments, i.e., molecular substructures, are present in the molecule. It first divides the molecular structure to obtain the structural fragments, and then encodes the fragments of these molecular structures into numbers according to certain rules and corresponds to each bit of the binary string, thus combining them as a whole (binary string) as a characterization of the molecular structure.

At present, the commonly used molecular fingerprints are FP4 fingerprint, MACCS fingerprint, Estate fingerprint, and PubChem fingerprint, and their corresponding molecular structure fragment numbers of 307, 166, 79, and 801. In this experiment, molecular fingerprints from the PubChem database were selected to characterize the drug structure of DTIs. The drug molecule is decomposed into 881 substructures in this descriptor. Given a drug, encode its corresponding bit as 1 or 0 depending on whether its molecular substructure is present. The fingerprint is encoded in Base64 on the PubChem website and provides a text description of it in binary, available for download from https://pubchem.ncbi.nlm.nih.gov/ (accessed on 1 January 2018).

2.3. Target Protein Characterization

In the experiments, the Position-Specific Scoring Matrix (PSSM) was used to numerically characterize the target protein. The PSSM can effectively describe the evolutionary information of protein amino acids, and it is commonly used in protein secondary structure prediction [27], protein binding site prediction [28], disordered region prediction [29], and distantly related protein detection [30,31] domains. The PSSM is a matrix of

$H \times 20$, where H is the length of the protein, and 20 is the type of amino acid. The PSSM $Pssm = \{\Theta_{i,j} : i = 1 \cdots H \; and \; j = 1 \cdots 20\}$ can be expressed equationally as follows:

$$Pssm = \begin{bmatrix} \Theta_{1,1} & \Theta_{1,2} & \cdots & \Theta_{1,20} \\ \Theta_{2,1} & \Theta_{2,2} & \cdots & \Theta_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ \Theta_{H,1} & \Theta_{H,2} & \cdots & \Theta_{H,20} \end{bmatrix} \quad (1)$$

Here, the matrix element $\Theta_{i,j}$ indicates the probability that the i-th residue of the protein mutates to the *i-type* amino acid during the evolutionary process.

In the implementation, we utilized the Position-Specific Iterated BLAST (PSI-BLAST) [32] to calculate the PSSM by comparing it with the SwissProt database. We followed the previous study, setting the parameter iterations and e-value of the PSI-BLAST tool to 3 and 0.001 to obtain high homologous sequences in the experiment. The database and tool are available for download from http://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed on 18 March 2002).

2.4. Feature Extraction

In the MSPEDTI model, the convolution neural network (CNN) algorithm of deep learning is used to extract the hidden features of the protein. Deep learning can learn the intrinsic patterns and levels of representation of sample data, thus enabling machines to have the same analytical learning capabilities as humans. As one of the representative algorithms of deep learning, CNN is able to classify the input information in a translation-invariant manner by hierarchical structure, thus deeply mining the essential features of data. Therefore, we introduced it into MSPEDTI to greatly strengthen the model prediction capability.

CNN is a feedforward neural network with artificial neurons that respond to a portion of the surrounding units in the coverage area, including convolutional, pooling, sampling, fully connected, input, and output layers. With its special structure of local weight sharing, CNN has unique advantages in feature extraction, and its layout is closer to the actual biological neural network. CNN has unique superiority in feature extraction, with its special structure of local weight sharing, and its layout is closer to the actual biological neural network. Weight sharing reduces the complexity of the network, especially the feature that multidimensional input vectors can be directly input into the network, which avoids the complexity of data reconstruction in the process of feature extraction and classification. The structure diagram of CNN is shown in Figure 1. Assuming that C_i is the feature map of layer ith, its description can be:

$$C_i = g(C_{i-1} \cdot W_i + b_i) \quad (2)$$

Here, operator $\cdot$ indicates convolution operations, b_i indicates the offset vector, W_i indicates the weight matrix of the ith layer convolution kernel, and $g(x)$ indicates the activation function. The subsampling layer follows the convolutional layer and samples the feature map according to specific rules. Let C_i be the subsampling layer with the following sampling rules:

$$C_i = subsampling(C_{i-1}) \quad (3)$$

After multiple convolution and sampling, the features are classified by the fully connected layer to yield the data distribution Γ of the original input. Fundamentally, CNN can be regarded as a mathematical model that uses multilevel dimensional transformations to transform the original data C_0 into a new feature representation Γ.

$$\Gamma(i) = Map(P = p_i | C_0; \; (W, b)) \quad (4)$$

Here, Γ represents the feature representation, p_i indicates the ith label class, and C_0 represents the original data.

Minimizing the loss function $H(W, b)$ is the ultimate goal of CNN training. Therefore, CNNs are typically trained to solve the overfitting problem by controlling the fitting strength using the parameter θ and adjusting the loss function $L(W, b)$ by generalizing the norm.

$$L(W, b) = H(W, b) + \frac{\theta}{2} W^T W \tag{5}$$

CNNs normally update their network layer parameters (W, b) layer by layer by gradient descent in the training phase and control the backpropagation function to exploit the learning rate ε.

$$W_i = W_i - \varepsilon \frac{\partial E(W, b)}{\partial W_i} \tag{6}$$

$$b_i = b_i - \varepsilon \frac{\partial E(W, b)}{\partial b_i} \tag{7}$$

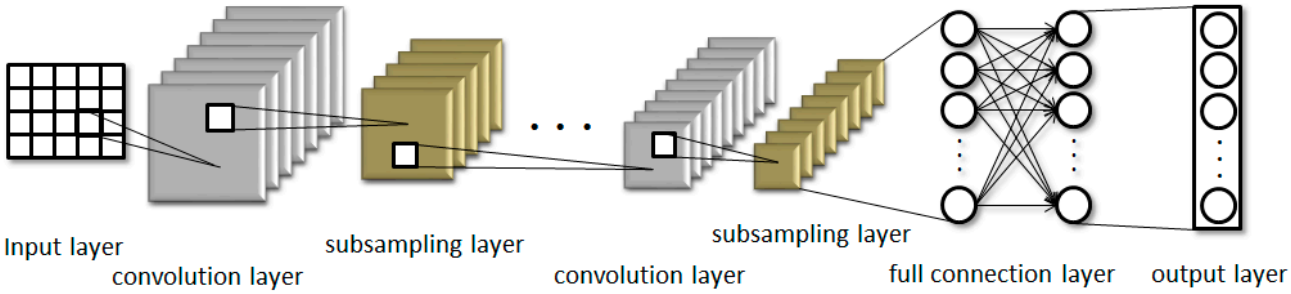

Figure 1. Schematic diagram of the structure of CNN.

2.5. Classification Prediction

The extreme learning machine (ELM) [33] is employed by MSPEDTI as a classifier to predict potentially associated DTIs. The ELM is a simple and effective single-hidden layer feedforward neural network learning algorithm that does not need to adjust the input weights of the network and the bias of the hidden elements during the execution and produces a unique optimal solution, so it has the advantages of fast learning and good generalization performance.

Given input samples (X_i, P_i) with L tagged, the ELM consisting of N neurons can be formulated as:

$$\sum_{i=1}^{N} V_i g(W_i \cdot X_j + b_i) = O_j, \quad j = 1, \ldots, L \tag{8}$$

where $X_i = [x_{i1}, x_{i2}, \ldots, x_{iL}]^T \in \mathbb{R}^L$, $P_i = [P_{i1}, P_{i2}, \ldots, P_{im}]^T \in \mathbb{R}^m$, $g(x)$ indicates the activation function, V_i indicates the output weight matrix, $W_i = [w_{i1}, w_{i2}, \ldots, w_{iL}]^T$ stands for the input weight matrix, $W_i \cdot X_j$ stands for the inner product of W_i and X_j, and b_i stands for the offset of the ith neurons.

To realize the minimization of the output error, i.e., the training goal of $\sum_{j=1}^{L} \|O_j - P_j\| = 0$, the ELM needs to optimize its hyperparameters.

$$\sum_{i=1}^{N} V_i g(W_i \cdot X_j + b_i) = P_j, \quad j = 1, \ldots, L \tag{9}$$

The equation can be simplified as follows:

$$SV = P \tag{10}$$

$$S = \begin{bmatrix} g(W_1 \cdot X_1 + b_1) & \cdots & g(W_N \cdot X_1 + b_N) \\ \vdots & \vdots & \vdots \\ g(W_1 \cdot X_L + b_1) & \cdots & g(W_N \cdot X_L + b_N) \end{bmatrix}_{L \times N} \quad V = \begin{bmatrix} V_1^T \\ \vdots \\ V_N^T \end{bmatrix}_{N \times m} \quad P = \begin{bmatrix} P_1^T \\ \vdots \\ P_L^T \end{bmatrix}_{L \times m} \tag{11}$$

Here, V means the output weight, P means the expected output, and S means the hidden layer neurons output. To gain optimal performance, we want the ELM to acquire $\hat{W}_i$, $\hat{b}_i$ and $\hat{V}_i$, that is:

$$\|S\left(\hat{W}_i, \hat{b}_i\right)\hat{V}_i - P\| = \min_{W,b,V} \|S(W_i, b_i)V_i - P\| \quad i = 1, 2, \cdots, N \tag{12}$$

This equates to minimizing the loss function

$$E = \sum_{j=1}^{L} (\sum_{i=1}^{N} V_i g(W_i \cdot X_j + b_i) - P_j)^2 \tag{13}$$

By the principle of the ELM algorithm, when the input weight W_i and the offset b_i of the hidden layer are ascertained, the ELM is able to uniquely obtain its output matrix. Therefore, the training problem of the ELM is transformed into the problem of solving the linear equation $SV = P$ with a minimal and unique interpretation.

3. Results

3.1. Evaluation Indicators

We measured the performance of MSPEDTI in the present study using the evaluation indicators calculated by the five-fold cross-validation method (5CV). The 5CV approach first splits the whole dataset D into five subsets $D_1, \ldots, D_5$, which are roughly equal in size and do not intersect with each other. When testing subset D_i, the remaining subsets $D - D_i$ are fed into the classifier as the training set. Loop this operation until all subsets have been tested. The performance of MSPEDTI was evaluated by the average results and deviations of the five experiments. There are several evaluation indicators calculated through 5CV, which are described by the following equations.

$$Accu. = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$Sen. = \frac{TP}{TP + FN} \tag{15}$$

$$Spec. = \frac{TN}{TN + FP} \tag{16}$$

$$Prec. = \frac{TP}{TP + FP} \tag{17}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{18}$$

where TP means true positive, TN means true negative, FP means false positive, and FN means false negative. Additionally, we plotted the operating characteristic curve (ROC) generated by 5CV and calculated its area under the curve (AUC) [34,35].

ROC is an essential metric for assessing the comprehensive performance of the model, which visualizes the variation between specificity and sensitivity and is displayed graphically. It computes a set of specificities and sensitivities by setting multiple different thresholds for successive variables, and then plots curves by using 1-specificity as abscissa and sensitivity as ordinate.

3.2. Assessment of Performance

Gold-standard dataset enzymes, ion channels, GPCRs, and nuclear receptors were used to measure the capabilities of MSPEDTI in the experiment. The detailed outcomes of 5CV obtained by MSPEDTI on these datasets are listed in Tables 2–5, respectively. From these tables, it is possible to observe that MSPEDTI accomplished satisfactory prediction accuracy, with values of 94.19%, 90.95%, 87.95%, and 86.11%, and their standard deviations were 0.41%, 1.10%, 1.51%, and 4.39%, respectively. In the enzyme dataset, the accuracy of all five MSPEDTI experiments was higher than 93.85%, with the highest result reaching 94.87%, and their standard deviations values were 94.87%, 94.27%, 93.85%, 94.02%, and 93.94%, respectively. MSPEDTI achieved good results of 88.51%, 81.95%, 76.41%, and 72.46% on MCC, which was used to measure classification performance, and its standard deviations were 0.89%, 2.24%, 2.88%, and 8.97%, respectively. On the comprehensive performance assessment index AUC, MSPEDTI gained 94.37%, 90.88%, 88.02%, and 86.63%, with standard deviations of 0.59%, 0.97%, 2.88%, and 4.77%, respectively. Additionally, MSPEDTI also yielded more satisfactory outcomes in terms of sensitivity and precision. The ROC curves produced by MSPEDTI for 5CV on the four gold-standard datasets are shown in Figures 2–5.

Table 2. MSPEDTI outcomes for 5CV on enzyme dataset.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	94.87	91.23	98.75	90.04	95.12
2	94.27	93.14	95.26	88.57	94.77
3	93.85	89.80	97.78	87.99	94.32
4	94.02	93.07	94.71	88.04	93.98
5	93.94	92.33	95.15	87.91	93.68
Average	**94.19 ± 0.41**	**91.91 ± 1.41**	**96.33 ± 1.81**	**88.51 ± 0.89**	**94.37 ± 0.59**

Table 3. MSPEDTI outcomes for 5CV on ion channel dataset.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	90.17	88.44	91.55	80.38	89.99
2	89.83	90.70	89.51	79.65	90.14
3	92.20	90.26	94.56	84.50	91.66
4	90.51	91.86	89.44	81.05	90.46
5	92.06	90.27	93.73	84.18	92.15
Average	**90.95 ± 1.10**	**90.31 ± 1.23**	**91.76 ± 2.36**	**81.95 ± 2.24**	**90.88 ± 0.97**

Table 4. MSPEDTI outcomes for 5CV on GPCR dataset.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	86.61	92.68	82.01	73.89	85.37
2	89.76	95.74	87.10	79.53	91.90
3	88.98	95.58	82.44	78.82	88.46
4	88.19	92.86	84.78	76.74	89.39
5	86.22	93.94	82.12	73.07	85.00
Average	**87.95 ± 1.51**	**94.16 ± 1.45**	**83.69 ± 2.22**	**76.41 ± 2.88**	**88.02 ± 2.88**

Table 5. MSPEDTI outcomes for 5CV on nuclear receptor dataset.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	91.67	86.96	100.00	84.05	94.98
2	80.56	85.71	70.59	61.51	84.74
3	88.89	85.00	94.44	78.26	85.63
4	83.33	83.33	83.33	66.67	83.02
5	86.11	86.67	81.25	71.81	84.76
Average	**86.11 ± 4.39**	**85.53 ± 1.45**	**85.92 ± 11.56**	**72.46 ± 8.97**	**86.63 ± 4.77**

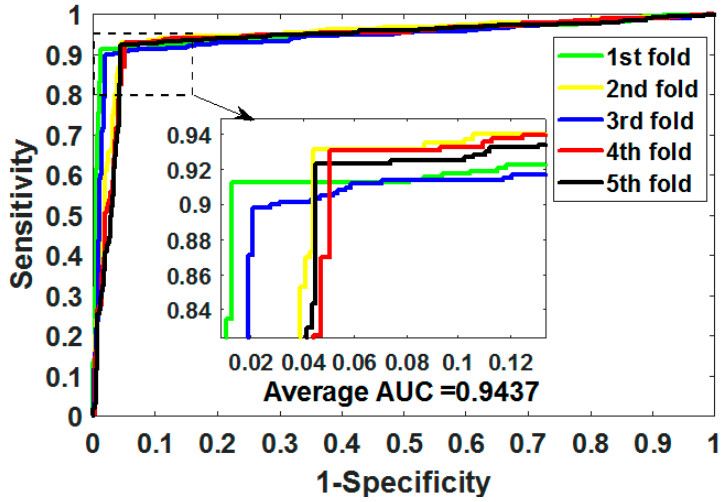

Figure 2. ROC of 5CV mapped by MSPEDTI on enzyme dataset.

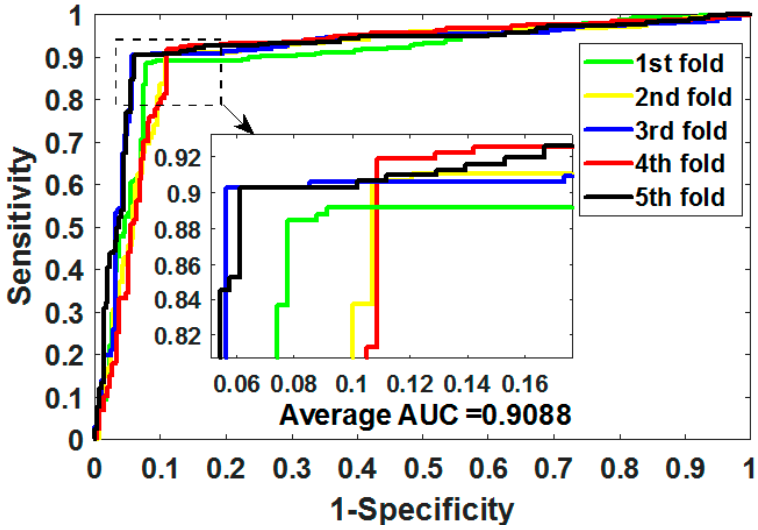

Figure 3. ROC of 5CV mapped by MSPEDTI on ion channel dataset.

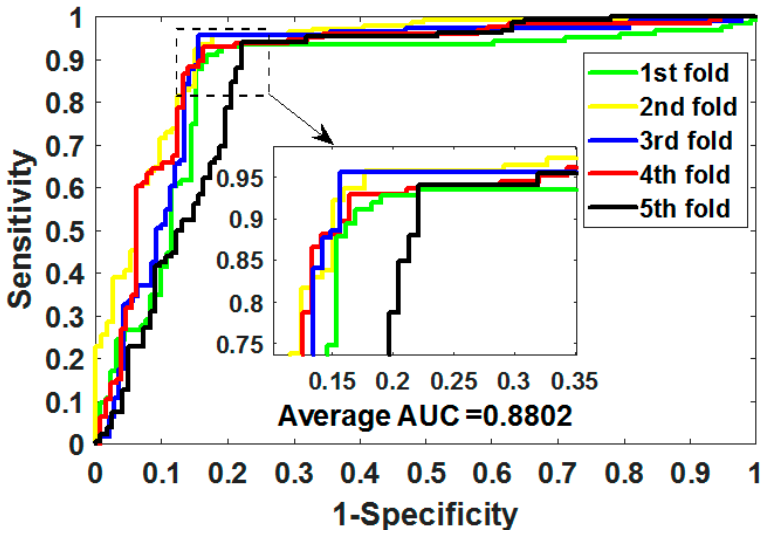

Figure 4. ROC of 5CV mapped by MSPEDTI on GPCR dataset.

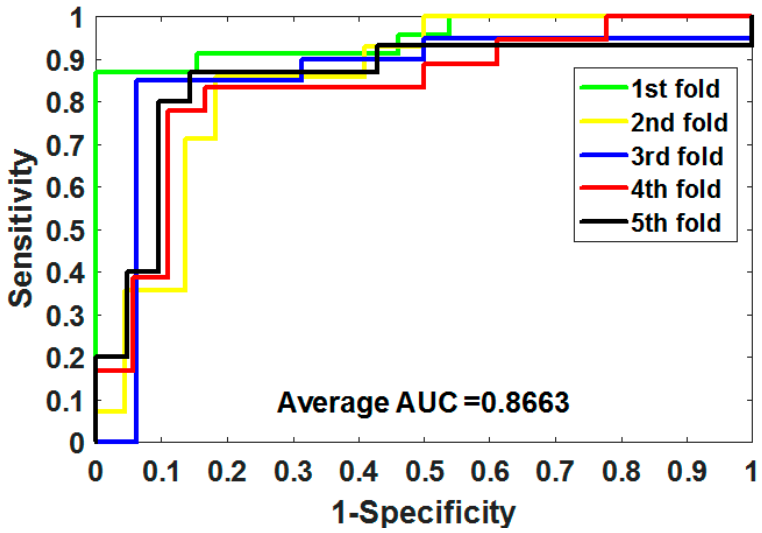

Figure 5. ROC of 5CV mapped by MSPEDTI on nuclear receptor dataset.

3.3. Comparison of Different Descriptor Model

To estimate the impact of feature descriptors on MSPEDTI performance, we compared it with the two-dimensional principal component analysis (2DPCA) descriptor model. 2DPCA is an advanced version of the principal component analysis algorithm [36], which does not need to convert raw data into one-dimensional vectors, which is equivalent to removing the correlation of the row vector or column vector of the matrix. So, it can directly calculate the covariance training sample matrix and has the advantage of calculating the feature vectors quickly.

To validate the representation capability of the features extracted by CNN, we compared it with the 2DPCA descriptor on the ion channel dataset. In the interest of fairness, the other modules in MSPEDTI were kept unchanged, and only the feature extraction module was replaced. The 5CV results produced by the two descriptor models on the ion channel dataset are shown in Table 6, in which it can be observed that the MSPEDTI-generated results are higher than the 2DPCA descriptor model. The experimental outcomes of the contrast indicated that the CNN algorithm extracts the features better than the 2DPCA algorithm in our model. Figure 6 shows the ROC curve plotted on the ion channel by utilizing the 2DPCA descriptor method.

Table 6. Comparison results of the 2DPCA descriptor model and MSPEDTI on ion channel.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	84.75	84.90	84.90	69.49	86.41
2	82.03	82.31	80.00	64.02	81.24
3	82.37	82.84	82.84	64.72	83.35
4	80.68	84.23	78.93	61.47	81.22
5	82.77	82.00	83.67	65.56	83.12
Average	**82.52 ± 1.47**	**83.26 ± 1.25**	**82.07 ± 2.52**	**65.05 ± 2.91**	**83.07 ± 2.12**
MSPEDTI	**90.95 ± 1.10**	**90.31 ± 1.23**	**91.76 ± 2.36**	**81.95 ± 2.24**	**90.88 ± 0.97**

3.4. Comparison with Different Classifier Model

To validate whether the classifier helps to improve the performance of MSPEDTI, we compared it with the SVM classifier model in the same dataset. The learning strategy of SVM is to maximize the sample interval, thus converting it to the solution of the convex quadratic programming problem [37,38]. Similar to the ablation experiments for the descriptor model, in the comparisons of the classifier models, we only replaced the ELM classifier with the SVM classifier and left the other modules unchanged.

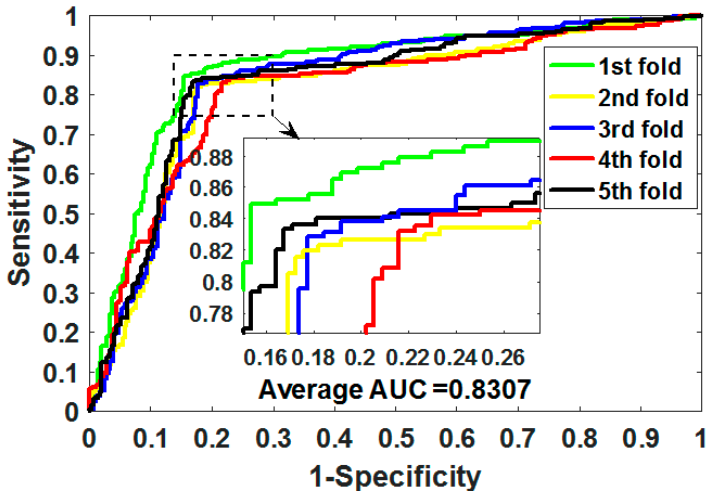

Figure 6. ROC curves plotted by the 2DPCA descriptor model on ion channel.

Table 7 presents the 5CV experimental outcomes of the MSPEDTI and SVM classifier model on the ion channel dataset. It is possible to observe from the table that the SVM classifier model performs well, and the accuracy, AUC, MCC, precision, and sensitivity are 86.48%, 86.64%, 73.05%, 83.86%, and 89.05%, respectively. However, compared with the ELM classifier, there are still some gaps, and the values of the above evaluation criteria are lower by 4.47%, 1.26%, 7.90%, 8.90%, and 4.24% respectively. These results indicate that the ELM classifier is indeed helpful to improve the prediction performance of MSPEDTI. Figure 7 shows the ROC curve plotted on the ion channel through utilizing the SVM classifier model.

Table 7. Comparison outcomes of SVM model and MSPEDTI on ion channel.

Test Set	Accu. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	85.76	90.14	81.42	71.81	85.08
2	85.93	89.04	82.70	71.94	87.90
3	85.76	87.34	84.04	71.46	84.80
4	86.61	89.49	83.73	73.34	87.10
5	88.34	89.26	87.41	76.70	88.33
Average	**86.48 ± 1.10**	**89.05 ± 1.04**	**83.86 ± 2.24**	**73.05 ± 2.16**	**86.64 ± 1.62**
MSPEDTI	**90.95 ± 1.10**	**90.31 ± 1.23**	**91.76 ± 2.36**	**81.95 ± 2.24**	**90.88 ± 0.97**

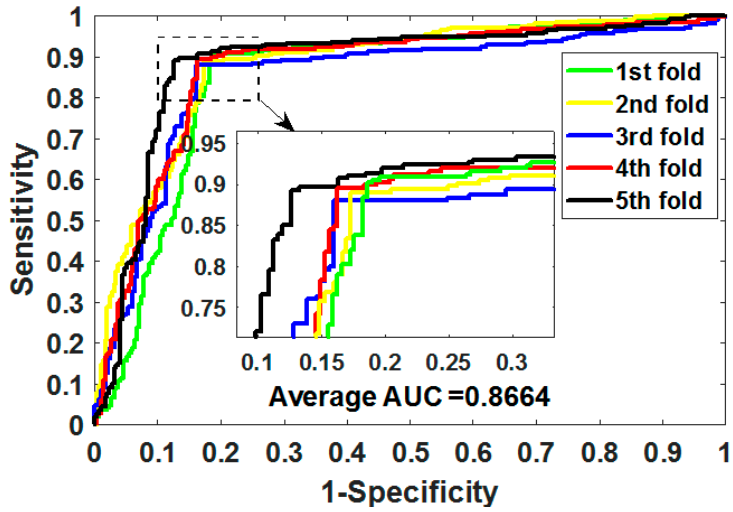

Figure 7. ROC curves plotted by the SVM classifier model on ion channel.

3.5. Comparison with Previous Approaches

We compared MSPEDTI with previous methods in the gold-standard dataset to assess its ability to predict DTIs in a more intuitive way. Here, we picked the metric AUC, which best reflects the overall comprehensive capability of the model as the evaluation criterion. The AUC values resulting from these previous methods, including Yamanishi [4], DBSI [39], KBMF2K [40], Temerinac-Ott [41], NLCS [42], WNN-GIP [43], SIMCOMP [42], and NetCBP [44], are aggregated in Table 8. It can be observed from the table that MSPEDTI yielded optimal results in all four gold-standard datasets over the previous method. This suggests that the strategy of combining the CNN algorithm with the ELM classifier used by MSPEDTI can greatly enhance the ability to predict DTIs.

Table 8. Comparison of AUC with previous methods in the gold-standard dataset.

Method	Enzymes	Ion Channels	GPCRs	Nuclear Receptors
SIMCOMP	86.30	77.60	86.70	85.60
NLCS	83.70	75.30	85.30	81.50
Temerinac-Ott	83.20	79.90	85.70	82.40
Yamanishi	82.10	69.20	81.10	81.40
KBMF2K	83.20	79.90	85.70	82.40
WNN-GIP	86.10	77.50	87.20	83.90
DBSI	80.75	80.29	80.22	75.78
NetCBP	82.51	80.34	82.35	83.94
MSPEDTI	**94.37**	**90.88**	**88.02**	**86.63**

3.6. Case Studies

To further verify MSPEDTI's ability in predicting new pairs, we trained it using all available data and predicted the unknown DTIs with the trained model. We searched the SuperTarget database [25] for the 10 highest-ranked DTI pairs of predicted associations. SuperTarget is a publicly available classic database that stores information about DTIs, and it currently collects 332,828 DTIs. Table 9 lists the top ten DTIs with the highest predictive score, from which we can see that seven potential DTIs were validated in the SuperTarget database. These outcomes indicated that MSPEDTI has outstanding capabilities in predicting new DTIs. Notably, while the rest of the three DTI interactions were not found in the current database, there is also the possibility of interaction between them.

Table 9. Top 10 DTI pairs predicted by MSPEDTI.

Drug ID	Drug Name	Taregt Protein ID	Target Protein Name	Validation Source
D00951	Medroxyprogesteroneacetate	hsa2099	ESR1_HUMAN	SuperTarget
D00542	Bromochlorotrifluoroethane	hsa1571	CP2E1_HUMAN	SuperTarget
D03365	Transdermal Nicotine	hsa1137	ACHA4_HUMAN	SuperTarget
D00049	Nikotinsaeure	hsa 8843	G109B_HUMAN	SuperTarget
D00160	Epsilcapramine	hsa7298	TYSY_HUMAN	unconfirmed
D00771	Chlorzoxazone	hsa1374	CPT1A_HUMAN	unconfirmed
D00139	Xanthotoxine	hsa1543	CP1A1_HUMAN	SuperTarget
D00964	Letrozole	hsa1215	CMA1_HUMAN	unconfirmed
D00585	Mifepristone	hsa2099	ESR1_HUMAN	SuperTarget
D00437	Nifedipine Monohydrochloride	hsa1559	CP2C9_HUMAN	SuperTarget

4. Discussion

Accurate identification of the target protein of the drug can improve the efficacy of the drug and reduce side effects, thereby improving people's health. In the current study, we presented a model MSPEDTI to predict DTI on the basis of protein evolution and molecular structures. The model takes full advantage of the protein evolutionary

information and drug molecular information and uses a deep learning algorithm to mine the deep association between them. The experimental outcomes in the four gold-standard datasets revealed that the MSPEDTI model has outstanding performance.

However, there are still some shortcomings in our method: firstly, the number of DTIs known at present is still relatively small, and the model cannot be trained adequately; secondly, the parameters of the deep learning algorithm used in the model need to be further optimized to avoid overfitting in some cases; finally, how to integrate more biological information into the model is still worth further study.

5. Conclusions

In the present work, we designed a deep learning model MSPEDTI for predicting DTI on the basis of drug structure and protein evolution information. The model deeply excavates hidden features in protein evolutionary information by CNN, combines them with drug molecular fingerprint features, and uses ELM to efficiently predict potential DITs. The model on the gold-standard datasets enzymes, GPCRs, ion channels, and nuclear receptors, attained better 5CV results. To evaluate whether the modules used by MSPEDTI contribute to boost model performance, we implemented ablation experiments and compared them with other descriptor and classifier models. Furthermore, 7 of the 10 DTIs predicted by MSPEDTI were substantiated in authoritative databases. The exceptional results as mentioned above indicate that MSPEDTI has outstanding ability to predict DTIs and can provide reliable candidate targets for drug research. In the next step of our research, we will try to optimize the deep learning feature extraction method to mine more useful information from the raw data.

Author Contributions: Conceptualization, L.W. (Lei Wang), L.W. (Leon Wong) and Z.-H.C.; methodology, J.H., X.-F.S. and Y.L.; writing—original draft preparation, L.W. (Lei Wang), writing—review and editing, L.W. (Leon Wong) and Z.-H.Y.; investigation, L.W. (Lei Wang); funding acquisition, L.W. (Lei Wang) and Z.-H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China, under Grants 62172355 and 61702444, in part by the Tianshan Youth—Excellent Youth, under Grant 2019Q029, in part by the West Light Foundation of the Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008, and in part by the Qingtan Scholar Talent Project of Zaozhuang University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank all anonymous reviewers for their constructive advice.

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

References

1. Mamoshina, P.; Volosnikova, M.; Ozerov, I.V.; Putin, E.; Skibina, E.; Cortese, F.; Zhavoronkov, A. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* **2018**, *9*, 242. [CrossRef] [PubMed]
2. Xuan, P.; Sun, C.; Zhang, T.; Ye, Y.; Shen, T.; Dong, Y. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Front. Genet.* **2019**, *10*, 459. [CrossRef] [PubMed]
3. Landry, Y.; Gies, J.-P. Drugs and their molecular targets: An updated overview. *Fundam. Clin. Pharmacol.* **2008**, *22*, 1–18. [CrossRef] [PubMed]
4. Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **2010**, *26*, i246–i254. [CrossRef] [PubMed]
5. Wang, L.; You, Z.H.; Chen, X.; Li, J.Q.; Yan, X.; Zhang, W.; Huang, Y.A. An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences. *Oncotarget* **2017**, *8*, 5149. [CrossRef] [PubMed]
6. Zhu, S.; Bing, J.; Min, X.; Lin, C.; Zeng, X. Prediction of drug–gene interaction by using Metapath2vec. *Front. Genet.* **2018**, *9*, 248. [CrossRef]

7. Wang, L.; You, Z.-H.; Zhou, X.; Yan, X.; Li, H.-Y.; Huang, Y.-A. NMFCDA: Combining randomization-based neural network with non-negative matrix factorization for predicting CircRNA-disease association. *Appl. Soft Comput.* **2021**, *110*, 107629. [CrossRef]
8. Wang, L.; Yan, X.; You, Z.-H.; Zhou, X.; Li, H.-Y.; Huang, Y.-A. SGANRDA: Semi-supervised generative adversarial networks for predicting circRNA–disease associations. *Brief. Bioinform.* **2021**, *22*, bbab028. [CrossRef]
9. Casañola-Martin, G.M.; Marrero-Ponce, Y.; Khan, M.T.H.; Khan, S.B.; Torrens, F.; Pérez-Jiménez, F.; Rescigno, A.; Abad, C. Bond-Based 2D Quadratic Fingerprints in QSAR Studies: Virtual and In vitro Tyrosinase Inhibitory Activity Elucidation. *Chem. Biol. Drug Des.* **2010**, *76*, 538–545. [CrossRef]
10. Kar, S.; Roy, K. Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. *Indian J. Biochem. Biophys.* **2011**, *48*, 111–122.
11. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489. [CrossRef] [PubMed]
12. Wallach, I.; Jaitly, N.; Nguyen, K.; Schapira, M.; Lilien, R. Normalizing molecular docking rankings using virtually generated decoys. *J. Chem. Inf. Modeling* **2011**, *51*, 1817. [CrossRef] [PubMed]
13. Wang, L.; You, Z.H.; Chen, X.; Yan, X.; Liu, G.; Zhang, W. RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions Using Drug Structure and Protein Sequence Information. *Curr. Protein Pept. Sci.* **2018**, *19*, 445–454. [CrossRef] [PubMed]
14. Zhao, L.; Wang, J.; Pang, L.; Liu, Y.; Zhang, J. GANsDTA: Predicting Drug-Target Binding Affinity Using GANs. *Front. Genet.* **2019**, *10*, 1243. [CrossRef] [PubMed]
15. Yang, X.; Kui, L.; Tang, M.; Li, D.; Wei, K.; Chen, W.; Miao, J.; Dong, Y. High-throughput transcriptome profiling in drug and biomarker discovery. *Front. Genet.* **2020**, *11*, 19. [CrossRef]
16. Wang, L.; You, Z.-H.; Huang, D.-S.; Li, J.-Q. MGRCDA: Metagraph Recommendation Method for Predicting CircRNA-Disease Association. In *IEEE Transactions on Cybernetics*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–9.
17. Wang, L.; You, Z.-H.; Li, J.-Q.; Huang, Y.-A. IMS-CDA: Prediction of CircRNA-Disease Associations From the Integration of Multisource Similarity Information With Deep Stacked Autoencoder Model. In *IEEE Transactions on Cybernetics*; IEEE: Piscataway, NJ, USA, 2020.
18. Li, H.-Y.; You, Z.-H.; Wang, L.; Yan, X.; Li, Z.-W. DF-MDA: An effective diffusion-based computational model for predicting miRNA-disease association. *Mol. Ther.* **2021**, *29*, 1501–1511. [CrossRef]
19. Lan, W.; Wang, J.; Li, M.; Wu, F.-X.; Pan, Y. Predicting drug-target interaction based on sequence and structure information. *IFAC-PapersOnLine* **2015**, *48*, 12–16. [CrossRef]
20. Cao, D.-S.; Liu, S.; Xu, Q.-S.; Lu, H.-M.; Huang, J.-H.; Hu, Q.-N.; Liang, Y.-Z. Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Anal. Chim. Acta* **2012**, *752*, 1–10. [CrossRef]
21. Yamanishi, Y.; Araki, M.; Gutteridge, 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]
22. 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*, D431–D433. [CrossRef]
23. Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357. [CrossRef] [PubMed]
24. Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2009**, *38* (Suppl. 1), D355–D360. [CrossRef] [PubMed]
25. Gunther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; et al. SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **2008**, *36*, D919–D922. [CrossRef] [PubMed]
26. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906. [CrossRef] [PubMed]
27. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195–202. [CrossRef] [PubMed]
28. Chen, X.-W.; Jeong, J.C. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **2009**, *25*, 585–591. [CrossRef] [PubMed]
29. Jones, D.T.; Ward, J.J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins Struct. Funct. Bioinform.* **2003**, *53*, 573–578. [CrossRef]
30. Gao, Z.G.; Wang, L.; Xia, S.X.; You, Z.H.; Yan, X.; Zhou, Y. Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM. *Biomed Res. Int.* **2016**, *2016*, 8. [CrossRef]
31. Wang, L.; You, Z.-H.; Xia, S.-X.; Chen, X.; Yan, X.; Zhou, Y.; Liu, F. An improved efficient rotation forest algorithm to predict the interactions among proteins. *Soft Comput.* **2017**, *22*, 3373–3381. [CrossRef]
32. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
33. Huang, G.B.; Wang, D.H.; Lan, Y. Extreme learning machines: A survey. *Int. J. Mach. Learn. Cybern.* **2011**, *2*, 107–122. [CrossRef]
34. Wang, L.; You, Z.-H.; Yan, X.; Xia, S.-X.; Liu, F.; Li, L.-P.; Zhang, W.; Zhou, Y. Using Two-dimensional Principal Component Analysis and Rotation Forest for Prediction of Protein-Protein Interactions. *Sci. Rep.* **2018**, *8*, 12874. [CrossRef] [PubMed]

35. Ghadermarzi, S.; Li, X.; Li, M.; Kurgan, L. Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins. *Front. Genet.* **2019**, *10*, 1075. [CrossRef] [PubMed]
36. Yang, J.; Zhang, D.; Frangi, A.F.; Yang, J.Y. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 131–137. [CrossRef]
37. Cao, D.-S.; Liang, Y.-Z.; Xu, Q.-S.; Hu, Q.-N.; Zhang, L.-X.; Fu, G.-H. Exploring nonlinear relationships in chemical data using kernel-based methods. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 106–115. [CrossRef]
38. Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z.; Chen, X.; Li, H.-D. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J. Chemom.* **2010**, *24*, 584–595. [CrossRef]
39. 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*, e1002503. [CrossRef]
40. Gonen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **2012**, *28*, 2304–2310. [CrossRef]
41. Temerinac-Ott, M.; Naik, A.W.; Murphy, R.F. Deciding when to stop: Efficient experimentation to learn to predict drug-target interactions. *BMC Bioinform.* **2015**, *16*, 1–10. [CrossRef]
42. Öztürk, H.; Ozkirimli, E.; Özgür, A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC Bioinform.* **2016**, *17*, 1–11. [CrossRef]
43. Van, L.T.; Marchiori, E. Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. *PLoS ONE* **2013**, *8*, e66952.
44. Chen, H.; Zhang, Z. A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks. *PLoS ONE* **2013**, *8*, e62975. [CrossRef] [PubMed]