Deep Dti – (Prediction of Drug Target Interaction Using Deep Learning)

Tirumala Sree Vaishnavi[1], Y. Shashidhar Reddy[2], T. Sai Kiran Reddy[3], Dr Ruqsar Zaitoon[4]
Department of Information Technology, Vardhaman College of Engineering (Autonomous), Hyderabad, India

***Abstract*—The key to modern drug development is to uncover, identify and prepare drug molecular targets. However, it is challenging to widely use classic experimental procedures due to the impact of throughput, precision, and cost. Traditional experiments used to deduce these possible DTIs, or drug-target interactions. Thus, the development of efficient computational techniques to verify the drug-target interaction is crucial. Techniques: We built a deep learning-based model for DTIs prediction. Position Specific Scoring Matrix (PSSM) and Legendre Moment (LM) are used to extract the evolutionary properties of proteins, which are then linked to drug molecular substructure fingerprints to create feature vectors of drug-target couples.Then we utilized the Sparse principal Component Analysis (SPCA) to compress the characteristics of medicines and proteins into a consistent vector space.**

Finally, the deep long short-term memory (DeepLSTM) was designed to perform prediction. findings: A considerable improvement in DTIs prediction performance may be noticed on experimental findings, with AUC of 0.9951, 0.9705, 0.9951, 0.9206, respectively, on four classes relevant drug-target datasets. Additionally Preliminary trials demonstrate the significant benefit of the suggested characterisation approach on feature expressiveness and recognition. Additionally, we have demonstrated that the suggested approach can function effectively with tiny datasets.

I. INTRODUCTION

Background Drug targets are the core of drug research and development, and during the past few centuries, mankind have relied largely on hundreds of drug targets currently recognized for its ability to identify narcotics [1]. The number of approved medication targets still only represents a small portion of the human proteome, despite the fact that the number of pharmaceuticals that are known to interact with target proteins is growing. One of the most important aspects of drug screening and drug-directed synthesis is the identification of drug-target interactions, which is the initial stage in the creation of new medications. The structural space of therapeutic compounds and the genomic space of target proteins have been better understood thanks to high-throughput studies. Unfortunately, our knowledge of the link between the two areas is currently quite restricted because of the time-consuming and difficult experimental procedure [2, 3]. The fast growth of publicly accessible biological and chemical data has made it possible for researchers to revisit drug-target interactions (DTIs) and systematically learn and analyze heterogeneous new data using computational approaches.

The ChEMBL [4], DrugBank [5], and SuperTarget [6] are a few free databases that concentrate on the connections between medications and targets. These database contents form the gold standard datasets, which are necessary for the development of computer techniques to predict DTIs.

At present, the computational method for DTIs prediction may be categorized into three categories: the ligandbased approach, the docking approach and the feature learning approach. By comparing a drug or compound's chemical structure to that of active compounds of known targets, ligand-based approaches are frequently employed to estimate possible targets of action. Keiser et al. [3] suggested a method for inferring protein targets based on the chemical similarity of their ligands. Yamanishi and associates. [7–9] forecast unidentified drug-target interactions by using the structural similarity of chemicals and the amino acid sequence resemblance of proteins to a uniform space. Campillos et al. [6] predict the possible target proteins by comparing their phenotypic side effects. This form of ligand-based technique is simple

and successful in the event of strong chemical structural similarity, but it also limits the scope and accuracy of its use to a large extent. In order to determine prospective pharmacological targets of action, the docking method computes the form and electrical matching of medicines and probable targets in three-dimensional structures.

This approach ranks pharmacological targets by anticipating the interaction mechanism and affinity between a specific chemical and a target, thereby identifying potential pharmacological targets. Cheng and associates. [10] established a structure-based maximum affinity model. TarFisDock is a web service created by Li et al. [11] that uses docking techniques to find drug targets. Such approaches completely address the three-dimensional structural information of the target protein, however the molecular docking approach itself still has certain issues that have not yet been properly solved, such example protein flexibility, the precision of scoring functions, and solvent water molecules, which lead to reverse docking. The prediction accuracy of the approach is low. Another serious the issue with docking is that it can't be used with proteins with unclear 3D model. Thus far, proteins with known 3D structure are currently only a small part of all proteins. This drastically limits the promotion and popularization of this approach. Drug target relationships are treated as a two-class problem in a feature learning technique.

problem: interaction and non-interaction. Such techniques discover the possible patterns of a known compound-target pairs with machine learning algorithms, create prediction models through iterative optimization, and subsequently deduce possible DTIs. A methodical approach based on chemical, genomic, and pharmacological data was suggested by Yu et al. [12]. Faulon et al. Anticipated drug targets using the signature molecular descriptor. Even though these strategies have accelerated the finding of drug targets, there is still significant space for improvement.

In this study, we suggested a deep learning-based technique for locating unidentified DTIs. The proposed technique consists of three steps: (i) Representation for drug-target pairings. The drug molecules are encoded as fingerprint feature and the protein sequences features are getting by utilizing Legendre Moments (LMs) on Position Specific Scoring Matrix (PSSM) that provides evolutionary information about protein. (ii) Feature compression and fusion. The features dimension and information redundancy are reduced using the Sparse Principal Component Analysis (SPCA). (iii) Forecast.

The Deep Long Short-Term Memory (DeepLSTM) model is utilized for executing prediction tasks. The flow of our proposed model is depicted in The following. We apply the suggested approach to four significant DTIs datasets that include ion channels, enzymes, GPCRs and nuclear receptors. The findings are revealed to give the current state-of-the-art better performance. Algorithms for DTI prediction.

II. MATERIALS AND METHODS

Data collecting

We gathered data from the databases KEGG [14], DrugBank [5], and SuperTarget [6] regarding the interactions between pharmacological molecules and target proteins [14, 15].The data set is summarized in Table 1 based on the quantity of pharmacological molecules, target protein, and interactions.

This set of known DTIs are considered to be the gold standard for judging the performance of the proposed method. A network of pharmacological targets is created by connecting target proteins to medication molecules. All identified drug-target couples in the gold standard dataset are regarded as positive samples in order to extract positive datasets from the network.

The remaining drug-target pairs in the network are represented by the negative sample. Since the scale of the noninteraction pairs is substantially larger than that of the interaction pairs, the produced datasets are imbalanced. We randomly chose negative samples from the remaining drug-target pairs in the network until the number of negative samples matched the number of positive samples in order to address the bias brought on by unbalanced data sets.

Drug molecule characterization

The ability of substructure fingerprints in characterizing drug compounds has been verified in various research. Through the complete study of prior research results, PubChem fingerprint was employed to characterized each drug compounds. In this work, medications are encoded Boolean substructure vector expressing the existence or absence of relevant substructures in a molecule.

The PubChem database defines 881 chemical substructures in which each substructure is associated

to a certain location. Therefore, for a substructure present in the drug compound, the position corresponding to the substructure in the fingerprint vector is set to 1, otherwise, and the corresponding position is set to 0. As a result, an 881-dimensional vector was used to represent each medication [16].

Characterization of target proteins

Position specific scoring matrix**:** In order to identify proteins that are distantly related, the position specific scoring matrix (PSSM) was originally introduced.

Table 1 lists the chosen drug-target interaction data sets from the sources like DrugBank, SuperTarget, and KEGG databases.

Datasets	Interactions	Targets	Drugs
Enzyme	2926	664	445
Ion channel	1476	204	210
GPCR	635	95	223
Nuclear receptor	90	26	54

In recent years, PSSMs is widely employed in proteomics and genomics research, such as prediction of DNA or RNA binding sites and membrane protein types. This work uses PSSM to collect evolutionary information about amino acids and to encode proteins. The PSSM of protein A with N amino acids residue can be represented as where Ai → j is a score that represents chance of ith residue being mutated to j-th natural amino acid and N is the length of amino acids residue of sequence A, 20 means the 20 native amino acid types. Position Specific Iterated BLAST (PSI-BLAST) [17, 18] was used to obtain the PSSM for each protein sequence. The default parameters were selected, with the exception of three iterations [19, 20].

Moments of legendre

The size invariance, rotation invariance, and displacement invariance of the invariant moments make them a great global statistical feature that may be used to extract stability aspects. As a quick moment invariant feature extraction technique, Legendre moments (LMs) perform well in a variety of pattern recognition applications, including graphic analysis, target recognition, image processing, classification, and prediction. Here, we create a feature vector and further refine the evolutionary information in PSSM using the Legendre moment. LMs are continuous orthogonal moments, which can be utilized to represent objects with minimal information redundancy [21, 22]. As a consequence, utilizing LMs on PSSM of protein sequence, we have acquired 961 characteristics from each protein sequence by setting a, b = 30.

Feature compression and fusion

We got an 1842-dimensional drug target feature vector from each drug target pair by integrating drug substructure fingerprint characteristics (881-D) with protein LMs features (961-D). Sparse principal component analysis (SPCA) is used to integrate drug and target protein features into an organic whole, reduce feature dimension, and eliminate redundant information in order to minimize classifier calculation time, lower memory consumption, and eliminate noisy features from the original feature space. One evident flaw in classical principle component analysis (PCA) is that loadings are usually nonzero and each PC is a linear sum of all variables.As a result, combining two distinct feature types, like the medication and protein features generated here, frequently yields unexpected outcomes. The aforementioned issue is resolved by SPCA, an enhanced PCA that uses lasso (elastic net) to generate principal components with sparse loadings. Ultimately, we obtain a 400-dimensional refined feature vector to use as the classifier's input.

Building a DeepLSTM model

A unique recurrent neural network (RNN) architecture called LSTM outperforms conventional RNNs in terms of performance [25]. This section examines the use of LSTM architecture for drug target prediction.

The LSTM architecture replaces summation units with memory blocks, which is one of the main distinctions from regular RNN networks. memory blocks include gates (special multiplicative units), input, output, and forget gates, to regulate the flow of information, and self-connection memory cells to store the temporal state. The absence of memory cells makes it easier to comprehend how the gate unit operates. These gates lessen the effect of vanishing gradient issues on the prediction model by allowing the LSTM to store and retrieve data over extended periods of time. The input gate regulates the input activation flow that reaches the memory unit [26, 27]. The output gate controls the output flow of cell activation, which travels to other

areas of the network. To enable the LSTM network to process the continuous input stream, the forgetting gate is introduced to the cell as input via the unit's self-recursive link. Peephole connections, which enable gates to be modulated in accordance with the state values in the internal memory, can also be incorporated into the LSTM cell [28]. By stacking several LSTM layers, we created DeepLSTM [29, 30]. By distributing several levels over space, deep architecture can make better use of the parameters than simple three-tier architecture.

Avoid overfitting

Numerous parameters are frequently used to optimize neural networks. Nevertheless, these networks might have overfitting issues. By arbitrarily eliminating units from the neural network and their connections in the training train, dropout is employed to solve this problem. "dropout" refers to the process of separating a "sparse" network from the original network, which is made up of all the remaining units. Within in this work, we put the dropout rate at 0.5 in accordance with the prior study. There are thirty-five hidden layer units. May produce 235 distinct subnets while being trained. Within in the testing phase, a "mean network" strategy is adopted, which keeps all of the original network connection but cuts their efferent weights in half . compensate for the doubled number of active individuals [31, 32].

Settings for the experiment

Indicators of evaluation : In this work, we compute accuracy (ACu), true positive rate (TPR), specificity (SPC), positive predictive value (PPV), and Matthews's correlation coefficient (MCC) to assess our predictor's performance. The overall degree of prediction is shown using the ACu. The percentage of positive samples that were accurately predicted in the test findings is revealed by the TPR. The percentage of negative samples that were accurately predicted in the test findings is revealed by the SPC. The percentage of real positive samples among samples that were anticipated to be positive is shown by the PPV. A typical indicator of predicted performance for two classification issues is the MCC. The following is a definition of certain performance indicators:

$$ACu = (TN + TP) / (TN + FN + TP + FP).$$

$$TPR = TP / (FN + FP).$$

$$SPC = TN / (TN + FP).$$

$$PPV = TP / (TP + FP).$$

$$MCC = (TP * TN) + (FP*FN) / \sqrt{(TP+FP) * (TN + FN) *(TP + FN) * (TN + FP)}$$

In this case, FN, FP, TN, and TP stand for the number of false negatives, false positives, true negatives, and true positives, respectively. The area under the Receiver Operating Characteristic curve (AUC), which is used to gauge prediction quality, is computed [33–35].

Training models

We separated each of the four datasets into training, verification, and test sets. Test accounts the training set accounts for 10% of the total. Eight tenths of the remaining data are utilized as sets of validation. The training set is done to fit a use the validation set for the DeepLSTM prediction model to utilize the test set to confirm the model's performance after optimizing the DeepLSTM neural network weight. An additional advantage the purpose of the validation set is to avoid overfitting by early stopping: stop training the model when mistakes on validation dataset is trending upward instead of downward. This method lowers the model's training cost and prevents overfitting.

For the cell input, we employ hyperbolic tangent activation. units, cell output units, and logistic sigmoid for the units for input, output, and forget gates. The input to the LSTMs and RNNs is 40-dimensional features. The output layer employs softmax and is a fully connected network. Function to generate probability outcomes. To locate the optimal network architecture, we evaluate the effectiveness of DeepLSTM models with varying layer counts and units on the data used for validation. The quantity of concealed layers from 1 to 6 that were tested. Concerning the quantity, these were tested between 20 and 200 with stride s = 4. Lastly, the DeepLSTM model with four hidden 36 units and layers were found.Random values with a mean of 0 and a standard deviation of 0.1 were used to initialize the weights of the DeepLSTM. Using a dynamic learning rate with an initial value of 0.002, decay of 0.004, and momentum

of 0.5, we trained the model using mean squared error and the Nadam optimizer. The batch size was 64, and the time step was set at 1. Training was terminated either early if there was no new best error on the validation data or after a maximum of 500 iterations.

III. FINDINGS AND CONVERSATION (RESULTS)

Statistics of the prediction performance for the proposed Table 2 lists the models. Pay attention to enzyme data sets, our predictor has given satisfying result of 92.92% accuracy, along with of 99.31% sensitivity, of 86.57% specificity, With 88.04% accuracy, of 86.75% MCC and AUC of 0.9951. The other three data sets show the same positive outcomes. By applying our technique. The outcomes of our approach 91.97% accuracy on the ion channels dataset, in addition to 93.23% sensitivity, 90.87% specificity, 89.95% accuracy, 85.19% MCC, and 0.9705 AUC. The outcomes our technique achieved 91.80% accuracy, 83.71% sensitivity, and 100% specificity on the GPCRs dataset. 84.44% MCC, 0.9511 AUC, and 100% accuracy.Our approach yielded 91.11% accuracy, 95.24% sensitivity, 87.50% specificity, 86.96% precision, 83.76% MCC, and an AUC of 0.9206 on the nuclear receptor dataset. Our method's achievement of over 90% accuracy on nuclear receptor datasets with just 180 samples is really remarkable.

Table 2 ACu, TPR, SPC, PPV, accuracy, MCC, and AUC prediction performance for the four datasets

Model	Data Set	ACu (%)	TPR (%)	SPC(%)	PPV (%)	MCC (%)	AUC
DeepLSTM	enzymes	92.92	99.31	86.57	88.04	86.75	0.9951
	ion chan	91.97	93.23	90.87	89.95	85.19	0.9705
	GPCRs	91.80	83.71	100	100	84.44	0.9951
	nuclear rec	91.11	95.24	87.50	86.96	83.76	0.9206

This demonstrates unequivocally that our approach may deliver outstanding results when dealing with extremely little training samples. This is a significant benefit that will set it apart from other approaches. The following three factors are mostly responsible for the outstanding performance:1) our feature representation approach can efficiently extract the discriminative characteristics from medication molecular and target protein sequence;2) SPCA takes pleasure in benefits in a number of areas, including as high explained variance, computational efficiency, and the capacity to uncover significant variables that condense two different feature vectors into a unified feature space and extracts heterogeneous features; 3) The neural network's hierarchical structure allows it to transform the input data into new feature area that is better suited for finishing classification tasks.

Comparing this classifier model to others

Enzymes, ion channels, GPCRs, and nuclear receptor datasets were analyzed using two additional well-known classifiers (Multi-layer Perceptron and Support Vector Machines) in order to demonstrate the benefits of DeepLSTM. To be fair, all other settings are exactly the same, with the exception of

the various classifiers. We construct multi-layer perceptron (MLP) networks with the same number of neurons and hidden layers as the DeepLSTM network. The LIBSVM utility made the Support Vector Machine (SVM) accessible [36]. Grid search technology optimizes the parameters. Tables S1, S2, S3, and S4 in the Supplementary Material provide the 5-fold cross-validation results obtained by SVM. Table 3 displays the cross-validation average findings

for four datasets.Overall, the DeepLSTM produces the best prediction results, according to Table 2's summary.

The DeepLSTM obtained 92.92% accuracy in the enzymes data set, 91.97% in the ion channels data set, 91.80% in the GPCRs data set, and 91.1% in the nuclear receptors data set. and exceed SVM (89.88, 89.36, 85.43, 85.00%, respectively) and MLP (99.01, 87.58, 87.20, 88.89%, respectively). The DeepLSTM net produced AUCs of 0.9951 for the enzymes data set, 0.9705 for the ion channels data set, 0.9951 for the GPCRs data set, and 0.9206 for the nuclear receptors data set. Nevertheless, the MLP net achieves an average AUC of 0.9967 and 0.9972, respectively.

Table 3 ACu, TPR, SPC, PPV, accuracy, MCC, and AUC comparison using three classifiers on four datasets.

Datasets	Model	ACu (%)	TPR (%)	SPC (%)	PPV (%)	MCC (%)	AUC
Enzymes	MLP	90.01	100	80.06	83.31	81.67	0.9967
	SVM	89.88	92.31	87.53	88.12	81.77	0.9686
	DeepLSTM	92.92	99.31	86.57	88.04	86.75	0.9951
ion channels	MLP	87.58	100	75.22	80.07	77.61	0.9972
	SVM	89.36	85.95	92.74	92.23	80.93	0.9613
	DeepLSTM	91.97	93.23	90.87	89.95	85.19	0.9705
GPCRs	MLP	87.20	76.70	97.77	97.19	77.20	0.9853
	SVM	85.43	86.28	84.60	84.81	74.99	0.9230
	DeepLSTM	91.80	83.71	100	100	84.44	0.9951
nuclear rec	MLP	88.89	88.24	89.47	88.24	80.19	0.8421
	SVM	85.00	68.90	100	100	72.43	0.9910
	DeepLSTM	91.11	95.24	87.50	86.96	83.76	0.9206

0.9853 and 0.8421 over four datasets. The average AUC for the SVM is 0.9686, 0.9613, and 0.9230, respectively. and 0.9910 over four datasets. The suggested approach yields superior outcomes for five primary reasons. The first is that the deep neural network's hierarchical structure transforms the input data into a more complicated area that is more suited for finishing classification tasks. The second is that our DeepLSTM's design not only successfully prevents overfitting but also enables the rapid training of several neural networks, improving the network's performance. The third is that the LSTM's memory modules can store greater understanding, which contributes to more accurate choices made during the forecasting phase. The fourth is that the gradient disappearing issue is resolved using LSTM in the Back Propagation (BP) technique is useful for obtaining superior than MLP as a prediction model.The use comes in fifth of the validation set aids in the training of more adaptable models.

Compare with cutting-edge methods

In this part, we contrasted the AUC of our suggested approach with several cutting-edge techniques, such as DBSI [10], KBMF2K [37], and NetCBP [38].The paradigm for the four kinds of target proteins put forward by Wang et al. [39] and Yamanishi et al. [7–9]. Table 4 lists the outcomes of multiple approaches on four data sets. The AUC of the suggested technique is unquestionably better than the AUC of various alternative methods for the four datasets, as shown in Table 4, Our method's AUC value on the enzymes dataset is 16% greater than the average of various previous approaches. With regard to the nuclear receptors dataset, our method yields values that are 10% higher than the maximum and 21% lower than the lowest in a number of other ways. Our scheme clearly surpasses the other examined approaches, as evidenced by the clearly higher AUC. The fact that our approach can enhance the performance for drug-target interaction prediction is further supported by the

comparison results with other approaches. Actually, based on the findings displayed in Table 2, As we can see, the AUC values of the other two models—MLP-based and SVM-based—remain greater than those of a number of current methods. This demonstrates how our feature extraction approach can effectively capture the interaction data between drug targets and enhance the predictor's ability to forecast drug-target interactions.

IV. CONCLUSION

In this work, we have created a deep learning-based technique to use the sequences of chemicals and proteins to infer possible DTIs. We contrasted our procedure with a number of cutting-edge techniques in order to assess its effectiveness. The outcomes of the experiment demonstrated that this strategy performs noticeably better than others. We have shown preliminary evidence that DeepLSTM performs better on the DTIs task than conventional machine learning systems when compared to other classifiers. For the quantitative approach and characterization of drug-target.

Pairs, an intriguing plan was put out employing SPCA to combine protein evolutionary traits with PubChem fingerprints that were acquired by combining PSSM and LM Positive outcomes were noted when the three distinct classifiers are used in conjunction with the characterisation approach. These findings show that the suggested method has significant advantages in terms of feature expression and recognition. We have demonstrated that the suggested approach, which differs from the methodologies of its predecessors and operates in a unique manner, may function effectively with limited datasets. Additionally, we discovered that as dataset size increases, prediction quality keeps getting better. This highlights the usefulness of this model for training and applying very big datasets and implies that expanding the data size may result in more performance gains. Overall, the theoretical analysis and experimental findings provide compelling theoretical and empirical support for the effectiveness of the suggested approach for DTI prediction.

V. ABBREVIATIONS

ACC stands for accuracy; LMs for legendre moments; DTIs for drug-target interactions; and DeepLSTM for deep long short-term memory. Matthew's correlation coefficient, or MCC PPV stands for positive predictive value; MLP stands for multi-layer perceptron; PSSM stands for position-specific scoring matrix, RNN for recurrent neural network, and SPC for specificity. Sparse principal component analysis, or SPCA Support vector machines, or SVMs True positive rate, or TPR.

REFERENCES

[1] Knowles J, Gromo G. Target selection in drug discovery: a guide. 2003; Nat Rev Drug Discov. 2(1):63–9.

[2] Maria RD, Stassi G, and Marcucci F. Mesenchymal-epithelial transition: a novel target in the search for anticancer medications. 2016; Nat Rev Drug Discov. 15(5):311–25.

[3] Hufeisen SJ, Jensen NH, Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Kuijer MB, Matos RC, Tran TB. identifying novel molecular targets for well-known medications. Nature, 462 (7270), 2009, 175–81.

[4] Hersey A, Light Y, Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, McGlinchey S, Michalovich D, Allazikani B, and Hersey A. ChEMBL is a thorough bioactivity database for drug development. 2012; 40:1100–7; Nucleic Acids Res.

[5] Woolsey J, Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z. A comprehensive resource for in silico drug investigation and development is DrugBank. Nucleic Acids Res. 34:668–72, 2006.

[6] ther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ. Resources for investigating drug-target interactions include SuperTarget and Matador. Nucleic Acids Res. 36:919–22, 2007.

[7] Yamanishi Y, Bleakley K. Bipartite local models are used for supervised drug-target interaction prediction. Bioinformatics. 25(18):2397–403, 2009.

[8] Araki M, Gutteridge A, Honda W, Kanehisa M, Yamanishi Y. Drug-target interaction networks are predicted by integrating chemical and genomic spaces. Bioinformatics, 24(13), 2008:232–40

[9] Yamanishi Y, Goto S, Kotera M, and Kanehisa M. An integrated system that uses pharmacological, genetic, and chemical data to anticipate drug-target interactions. Bioinformatics, 26(12), 246-54 (2010).
[10] Zhou W, Huang J, Tang Y, Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G.using network-based inference to predict drug-target interactions and drug repositioning. 2012; 8(5):e1002503 PLoS Comput Biol.
[11] Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J.TarFisDock is a web server that uses a docking technique to find drug targets. Nucleic Acids Res. 2006; 34 (Web Server issue): 219–24.
[12] Enzyme at the genome scale predictions of drug-target interactions and metabolites using the signature molecular characterisation. Faulon JL, Misra M, Martin S, Sale K, and Sapra R. bioinformatics. 2008; 24(2):225–33.
[13] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Tokimatsu T, Okuda S, and Kawashima S. KEGG for tying genomes to life and environment. Nucleic Acids Res. 2008:36:480-4 (Database issue).
[14] Suzek TO, Jian Z, Wang J, Bryant SH, Wang Y, Xiao J. PubChem: a public information system for small molecule bioactivity analysis. Nucleic
[15] Acids Res. 2009;37 (Web server problem):623–33
[16] Weininger D, Weininger A, Weininger JL. SMILES. 2. SMILES notation generating algorithm. J Chem Inf Model. 29(2):97–101, 1989.
[17] Wang Y, Zhang J. PCVMZM: a probabilistic classification vector machines model combined with a Zernike moments descriptor can be used to predict protein–protein interactions from protein sequences. Int J Mol Sci., 2017.
[18] (5):1029–42. 18. Zhu L, Xia J, Wang B, Lei YK, and You ZH. Protein-protein prediction connections between principal component analysis and ensemble extreme learning machines using amino acid sequences. 2013; 14(S8):1–11.
[19] BMC Bioinform. You ZH, Wang YB, Li X, Jiang TH, Chen X, Zhou X, Wang L. using a deep sparse autoencoder deep neural network to predict protein-protein interactions from protein sequences. Mol BioSyst. 13(7):1336–45 (2017)
[20] Protein-protein interaction prediction by You ZH, Li L, Ji Z, Li M, and Guo S from amino acid sequences via a combination of extreme learning machines using the auto covariance descriptor. Memetic Computing, 2013. pages 80–5.
[21] Interaction detection by Wang YB, You ZH, Li LP, Huang YA, and Yi HC between proteins by extracting information using the Legendre moments descriptor PSSM contains discriminatory information. molecules. 2017; 22(8):1366– 79.