Journal of Computational Science & Engineering 64 (2023) 1010-1019

Available at http://www.asocse.org ISSN 1710-4068

Prediction of Inhibitory Activity of JNK3 in Alzheimer's Disease Based on Artificial Intelligence

Ziang Wu, Peijian Zhang*

College of Computer Science and Technology, Qingdao University, Qingdao 266071, China

Abstract: **The novel C-Jun N-terminal kinase 3(JNK3) inhibitor studied in this article is a promising therapeutic option for Alzheimer's disease, as it is believed to play a significant role in its treatment. The inhibitory effect of certain compounds can be accurately identified by measuring their IC_{50} values. However, traditional IC_{50} determination methods require substantial human and material resources. To better understand the Quantitative Structure-Activity Relationship (QSAR), we employed various machine learning techniques, including heuristic method(HM), random forest(RF), support vector machine(SVM), generalized regression neural network(GRNN) to establish a comprehensive QSAR model that better describes this relationship. Among them, the support vector machine model demonstrated the best performance, with R^2 and MSE values of 0.914, 0.944, 0.066, and 0.041 on the training and testing sets, respectively. The leave-one-out cross validation(LOO-CV) yielded Q^2 values of 0.92 and 0.905 for the training and testing sets, respectively. These results demonstrate that the support vector machine model can be used to predict the inhibitory activity of this novel inhibitor, providing valuable assistance in the design and synthesis of new JNK3 inhibitors based on 2-aryl-1-pyrimydyl-1,4,5,6-tetrahydrocyclopenta[d]imidazole-5-carboxamide derivatives. Furthermore, the five key descriptors revealed in this study will aid in the design and screening of novel drug molecules for Alzheimer's disease.**

Keywords: Alzheimer's disease, quantitative structure–activity relationship, Novel JNK3 inhibitors,random forest, support vector machine.

*Corresponding Author: Peijian Zhang, peijianzh@126.com

1. Introduction

Alzheimer's disease is a neurological disorder with irreversible symptoms[1]. Early manifestations may include mild memory loss, but as the condition progresses, patients gradually experience a decline in memory and cognitive abilities, eventually leading to a loss of daily functioning[2]. Research indicates that Alzheimer's disease may be linked to factors such as neuronal damage, resulting in cell death and impairment in the brain, leading to a decline in memory and cognitive function[1].

Previous studies have suggested that JNK3 is a significant contributor to neuronal apoptosis, making JNK3 inhibitors a promising treatment option for Alzheimer's disease[3]. In order to enhance the chemical diversity of JNK3 inhibitors, a novel class of JNK3 inhibitors, known as Novel 1,4,5,6-tetrahydrocyclopenta[d]imidazole-5-carboxamide-based JNK3 inhibitors[3], has been designed and evaluated for their inhibitory activity through IC_{50} measurements. The investigation of these new JNK3 inhibitors holds potential for offering new approaches in the treatment of Alzheimer's disease[3]. Therefore, researchers aim to develop safer and more effective drugs to alleviate the progression of the disease.

QSAR methods offer a rapid and accurate means of predicting the properties of synthesized compounds, while also saving significant manpower and resources[4, 5]. In contrast to traditional IC_{50} determination methods, QSAR provides an effective screening tool for identifying compounds with high biological activity.[6] By establishing QSAR models, we can predict the bioactivity of new compounds, thereby providing robust support for drug discovery and enabling quick and accurate evaluation of compound properties while conserving resources[7-9].

In this study, heuristic method, random forest, support vector machine, and generalized regression neural network were employed to establish QSAR models. Furthermore, RF demonstrated the reliability of descriptor selection and validated the effectiveness of applying the same descriptors to other models. Additionally, the support vector machine model exhibited excellent predictive performance, indicating its efficacy as a method for predicting the inhibitory activity of novel JNK3 inhibitors.In summary, this study provides valuable assistance in the design and synthesis of Novel 1,4,5,6-tetrahydrocyclopenta[d]imidazole-5-carboxamide-based JNK3 inhibitors, and the five key descriptors revealed in this study will aid in the design and screening of novel drug molecules for Alzheimer's disease.

2. Experimental

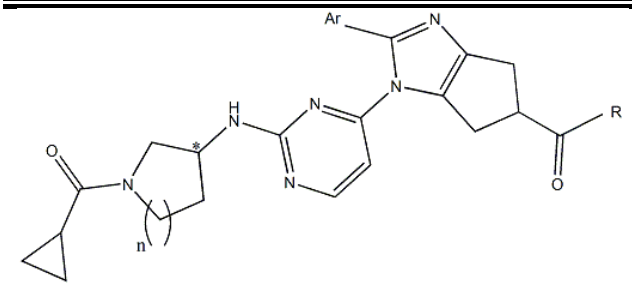
2.1. Data Preparation

This study initially collected the structures of drug molecules and their known activity data. To process the data, feature selection and extraction were conducted, such as molecular weight, atom types, and others. These tasks can be achieved using various feature selection and extraction algorithms. The compounds and their JNK3 inhibitory activity (IC_{50}) used in this study were obtained from the literature[3], totaling 52 compounds (Table 1). After randomly dividing these compounds into a training set and a test set in a 4:1 ratio, a model was established using the training set, which comprised 42 compounds, while the remaining 10 compounds constituted the test set, enabling the evaluation of the

2023@ The American Computational Science Society. All rights reserved.

model's performance.

Table 1. Measured and predicted lg (IC_{50}) of novel JNK3 inhibitors



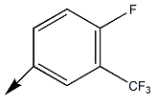
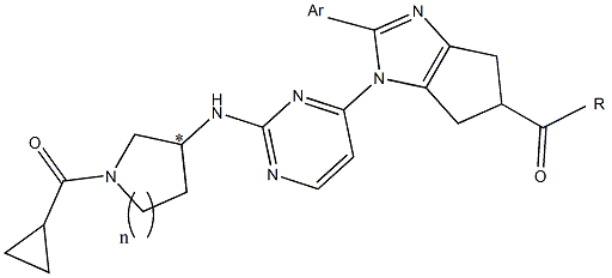
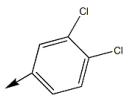
Compound	AR	*	n	R	Measured IC_{50}	Measured lg (IC_{50})	Predicted lg (IC_{50})			
							RF	HM	GRNN	SVM
16a		S	1	-OEt	453.0	2.656	2.551	2.743	2.611	2.556
20a*				-NH2	183.0	2.262	2.012	2.487	2.477	2.511
24a				-NHMe	601.0	2.779	2.352	2.404	2.734	2.679
17a		R	1	-OEt	63.1	1.8	1.888	2.5	1.842	1.9
21a				-NH2	300.0	2.477	2.253	2.673	2.474	2.377
25a				-NHMe	300.0	2.477	2.515	2.269	2.460	2.577
18a*		S	2	-OEt	52.8	1.723	1.772	1.529	1.794	1.542
22a				-NH2	6.39	0.806	1.015	1.373	0.987	0.906
26a				-NHMe	30.6	1.486	1.477	1.534	1.516	1.586
19a		R	2	-OEt	65.4	1.816	1.816	1.7	1.816	1.768
23a				-NH2	104.0	2.017	1.435	1.647	1.888	1.917
27a				-NHMe	700.0	2.845	1.795	1.339	1.818	1.745

Table 1. Continued

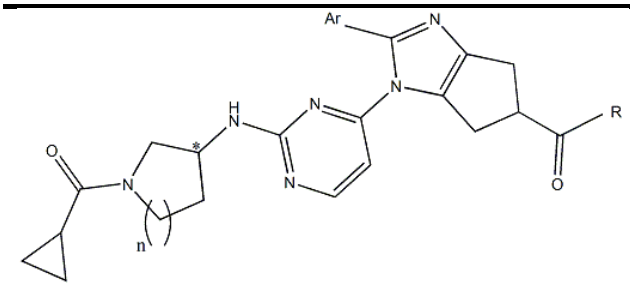


Compound	AR	*	n	R	Measured IC_{50}	Measured lg (IC_{50})	Predicted lg (IC_{50})			
							RF	HM	GRNN	SVM
16b*		S	1	-OEt	28.2	1.45	0.999	1.133	1.012	1.222
20b				-NH2	15.4	1.188	1.347	1.239	1.215	1.288
24b*				-NHMe	10.2	1.009	1.105	0.985	1.024	1.172

2023@ The American Computational Science Society. All rights reserved.

17b*	R	1	-OEt	4.33	0.636	0.75	0.954	0.669	0.489
21b			-NH2	16.4	1.215	1.164	1.024	1.203	1.115
25b			-NHMe	9.36	0.971	1.008	0.826	0.996	1.212
18b	S	2	-OEt	1.11	0.045	0.266	-0.016	0.020	0.063
22b*			-NH2	0.379	-0.421	0.357	-0.122	-0.09	-0.101
26b			-NHMe	1.20	0.079	0.302	-0.209	0.061	-0.021
19b*	R	2	-OEt	0.607	-0.217	0.226	0.168	-0.10	-0.114
23b			-NH2	0.730	-0.137	0.217	0.176	-0.03	-0.037
27b			-NHMe	4.83	0.684	1.239	1.356	1.704	1.775

Table 1. Continued



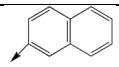
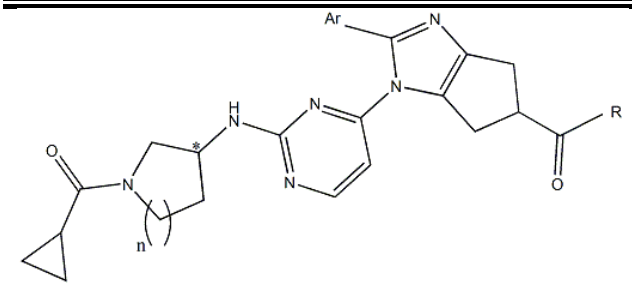
Compound	AR	*	n	R	Measured IC_{50}	Measured lg (IC_{50})	Predicted lg (IC_{50})			
							RF	HM	GRNN	SVM
16c		S	1	-OEt	23.4	1.369	1.272	1.089	1.232	1.406
20c				-NH2	14.8	1.17	1.244	1.289	1.197	1.27
17c		R	1	-OEt	6.82	0.843	0.862	0.935	0.897	0.839
21c				-NH2	24.9	1.396	1.118	1.008	1.372	1.296
18c		S	2	-OEt	0.716	-0.145	-0.089	-5.06E-03	-0.10	-0.045
22c*				-NH2	1.76	0.246	0.488	-0.041	0.036	0.074
26c				-NHMe	0.564	-0.248	-0.049	-0.067	-0.14	-0.148
19c		R	2	-OEt	0.779	-0.108	-0.072	0.191	-0.07	-0.008
23c				-NH2	1.72	0.236	0.259	0.199	0.157	0.136
27c				-NHMe	1.08	0.033	0.099	0.027	0.042	0.133

Table 1. Continued

2023@ The American Computational Science Society. All rights reserved.



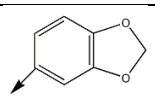
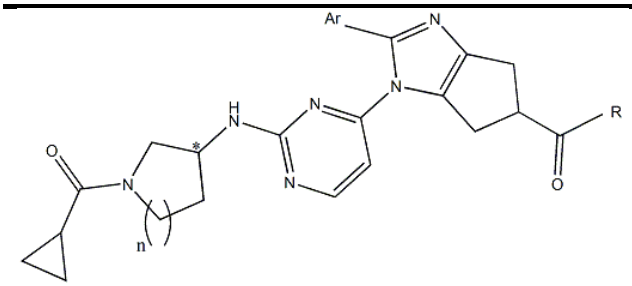
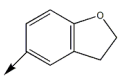
Compound	AR	*	n	R	Measured IC_{50}	Measured $lg(IC_{50})$	Predicted lg（IC_{50}）			
							RF	HM	GRNN	SVM
16d		S	1	-OEt	243.0	2.386	2.202	1.692	2.114	2.204
20d				-NH2	135.0	2.13	2.004	2.35	2.101	2.03
24d				-NHMe	180.0	2.255	2.092	2.02	2.183	2.155
17d		R	1	-OEt	49.9	1.698	1.919	1.612	1.905	1.798
21d*				-NH2	103.0	2.013	1.844	2.216	2.215	2.165
25d*				-NHMe	93.9	1.973	2.041	1.968	2.250	2.049
18d		S	2	-OEt	10.5	1.021	0.965	0.692	0.999	0.921
22d				-NH2	10.0	1	1.718	1.337	1.012	1.1
19d		R	2	-OEt	1.20	0.079	0.674	0.844	0.212	0.179
23d				-NH2	36.1	1.558	1.884	1.211	1.500	1.458

Table 1. Continued



Compound	AR	*	n	R	Measured IC_{50}	Measured $lg(IC_{50})$	Predicted lg（IC_{50}）			
							RF	HM	GRNN	SVM
16e		S	1	-OEt	158.0	2.199	1.835	1.941	2.041	2.099
20e				-NH2	87.3	1.941	1.923	2.051	1.903	1.841
17e		R	1	-OEt	21.5	1.332	1.509	1.795	1.478	1.432
21e				-NH2	63.0	1.799	1.817	1.954	1.811	1.85
18e		S	2	-OEt	31.4	1.497	1.497	1.09	1.477	1.397
22e				-NH2	9.06	0.957	1.142	1.039	0.975	1.057
19e		R	2	-OEt	10.4	1.017	1.06	1.08	1.031	0.917
23e				-NH2	8.92	0.95	1.046	1.006	0.948	0.85

*Test set.

2023@ The American Computational Science Society. All rights reserved.

2.2. Calculation of the Descriptors

The first step in calculating molecular descriptors is to use Chemdraw software to draw the molecular structure based on the chemical formula mentioned in the reference literature.

The second step involves using Hyperchem software to perform a series of optimizations on the molecular structure to obtain the lowest energy and most stable structure. During the process of optimizing the molecular structure, the molecule is initially pre-optimized using the MM+ molecular mechanics force field, followed by energy minimization using the semi-empirical (PM3) method[10].

The third step is to use MOPAC software to generate an MNO document[11], which serves as input for the CODESSA software. Finally, calculations are performed in the CODESSA software to obtain five categories of molecular descriptors[12]. Overall, the purpose of calculating molecular descriptors is to identify accurate descriptors, which are essential prerequisites for constructing QSAR models and play a significant role in the process.

2.3. Linear Model by HM

HM is a type of algorithm based on experience and rules, used to solve problems through search and optimization[13]. When selecting molecular descriptors, HM can help us quickly identify the optimal set from a large number of possible descriptors[14]. Therefore, we employ HM to choose the descriptors calculated by the CODESSA software.HM employs a preprocessing method to eliminate certain types of descriptors, including:

1. Descriptors that are not present in all compounds.
2. Descriptors that exhibit minimal variation across all structures.
3. Collinear descriptors, characterized by a correlation coefficient exceeding 0.8.[15]

In this study, the HM, RF, SVM, and GRNN models were evaluated using the R^2 and mean square error (MSE) tests. Additionally, leave-one-out cross-validation was employed to assess the robustness of all models, and the cross-validation correlation coefficient (Q^2) was used to describe the results.

2.4. Nonlinear Model by RF

Due to the influence of multiple complex factors, linear models cannot accurately predict the inhibitory activity (IC_{50} value) of new JNK3 inhibitors. Therefore, three nonlinear QSAR models, namely RF, SVM, and GRNN, were employed. Random Forest demonstrates superior performance in preventing overfitting and achieving higher accuracy. It can effectively handle complex samples, even in the presence of unhealthy samples such as noise and outliers, as well as high-dimensional and large-scale data, demonstrating its robustness[16]. Additionally, Random Forest can predict the biological activity of compounds and evaluate the importance of descriptors using the Gini coefficient[17].

In general，RF typically consists of two steps.

In the first step, feature selection is performed from all the descriptors obtained from CODESSA. In a random forest, each tree contributes to a subset of the most important descriptors in this subset[18]. The predictions from all the trees are then averaged to generate the most relevant descriptors. These descriptors are ranked in descending order based on their Gini coefficient, where a higher Gini coefficient indicates higher importance.

In the second step, the RF regression model is constructed. In the RF algorithm, three parameters need to be determined: the number of trees (n), the maximum depth of the trees (d), and the number of relevant features (f)[19].

This study involved the construction of two RF models. The first model utilized a large set of descriptors obtained from CODESSA, while the second model used descriptors selected by the HM algorithm. The purpose of the first model was to validate the descriptors chosen by HM, and the second model served as a comparison to other nonlinear methods. If the RF models showed comparable or superior prediction performance, the resulting Gini coefficients were used to explain descriptor importance. Different numbers of descriptors were selected based on Gini coefficients to build RF models. If there was significant overlap between the descriptors selected by the HM algorithm and those obtained from RF models, indicating comparable performance, it would confirm the validity of the HM-selected descriptors.

2.5. Nonlinear Model by SVM

In regression modeling using Support Vector Machines , a kernel function is utilized to map the input data into a higher-dimensional feature space[20]. This allows the construction of a linear model within this feature space to predict the target variable. The selection of an appropriate kernel function is crucial, as it should satisfy the properties of symmetry and positive definiteness[21]. In this study, the radial basis function (RBF) is chosen as the kernel function, represented by Equation 1.

$$K(x_i, x_j) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2) \quad (1)$$

Where γ is a hyperparameter of the kernel function, and $\left|x_i - x_j\right|$ represent the Euclidean distance between x_i and x_j.

Unlike traditional methods, the fundamental idea of Support Vector Machines is to endow the model with generalization capability rather than solely optimizing its performance on a given training set. This is achieved by minimizing the structural risk, which considers the upper bound of empirical risk balancing the capacity of the classifier (similar to its complexity) and the trade-off between minimizing training errors and reducing model complexity. SVM accomplishes this by defining the maximum margin hyperplane and extending the feature space, benefiting from the

2023@ The American Computational Science Society. All rights reserved.

definition of a kernel function. SVM demonstrates excellent performance in solving small-sample, nonlinear, and high-dimensional pattern recognition problems, and it has relatively low computational requirements and is amenable to parallel implementation, especially for large training datasets[22].

However, SVM still requires tuning of important parameters, such as the regularization parameter C and the parameters of the kernel function. The choice of these parameters can significantly impact the model's performance. For instance, the regularization parameter C controls the complexity of the model, where a large C value may lead to overfitting, while a small C value can result in underfitting[23]. The margin parameter ε is used to control the degree to which the entire training set satisfies the boundary conditions.

Finally, when using SVM, it is necessary to determine the penalty coefficient C, the ε-margin parameter, and the parameters of the radial basis kernel function. The objective function formula for SVM regression is shown in Equation 2.

$$\min \frac{1}{2}\left|\left|\omega\right|\right|^2 + C\sum_{i=1}^{N}\epsilon_i$$
$$\text{subject to} \quad y_i - \omega^T x_i - b \leq \epsilon_i, \epsilon_i \geq 0 \tag{2}$$

Where ω and b are model parameters, ϵ_i is the margin parameter, and C is the penalty coefficient.

2.6. Nonlinear Model by GRNN

The Generalized Regression Neural Network is widely used for various application prediction tasks due to its high generalization performance[24]. GRNN has the ability to rapidly process small datasets and performs exceptionally well in cases of low classification and fitting data accuracy[25]. Compared to other artificial neural networks, GRNN has several distinct advantages, making it a preferred choice as a prediction method for handling small data[26]. The structure of GRNN is illustrated in Figure 1:

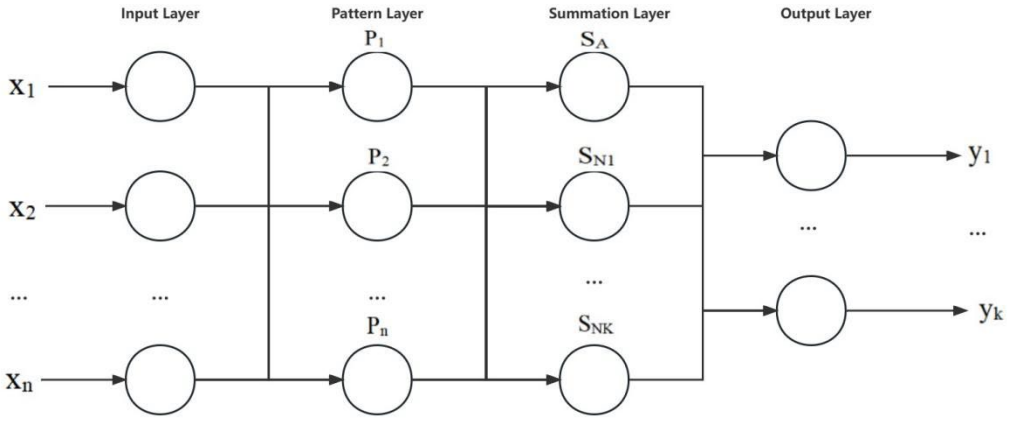


Figure 1. The structure of GRNN.

Input Layer: The input layer is the first layer of the GRNN model, and its role is to receive the dataset used for training or testing and pass it to the next layer.

Pattern Layer: The pattern layer is an unsupervised learning process used to extract latent patterns from the input data. The pattern layer uses input vectors to create a set of template vectors, with each template vector corresponding to a cluster in the input space. The number of template vectors is typically a small fraction of the dataset size.

Summation Layer: The computation layer is the core component of the GRNN. The computation layer assigns weights to each template vector by calculating the distance between the input vector and all template vectors. These weights are computed using a radial basis function (kernel function), typically a Gaussian kernel function. The computation layer generates an output vector by taking a weighted sum of all template vectors[27].

Output Layer: The output layer converts the output vector from the computation layer into the desired output format. It applies the necessary transformations or mappings to match the desired output representation. The output layer provides the final result of the prediction or inference process of the GRNN model..

3. Result and discussion

3.1. Results of HM

To select the most relevant descriptors for the novel 1,4,5,6-tetrahydrocyclopenta[d]imidazole-5-carboxamide-based

2023@ The American Computational Science Society. All rights reserved.

JNK3 inhibitors,a total of 498 descriptors were calculated in the CODESSA software. Based on this, the number of molecular descriptors in the linear model was gradually increased, and the corresponding R^2 and R^2_{cv} were recorded to determine the most relevant descriptors. As each additional molecular descriptor was included, there was no significant increase in R^2 and R^2_{cv}, indicating a saturation point for the number of descriptors. Eventually, five molecular descriptors were selected, and their corresponding R^2 and R^2_{cv} values were recorded to determine the most relevant descriptors. The effect of descriptor number on R^2 and R^2_{cv} is shown in Figure 2.

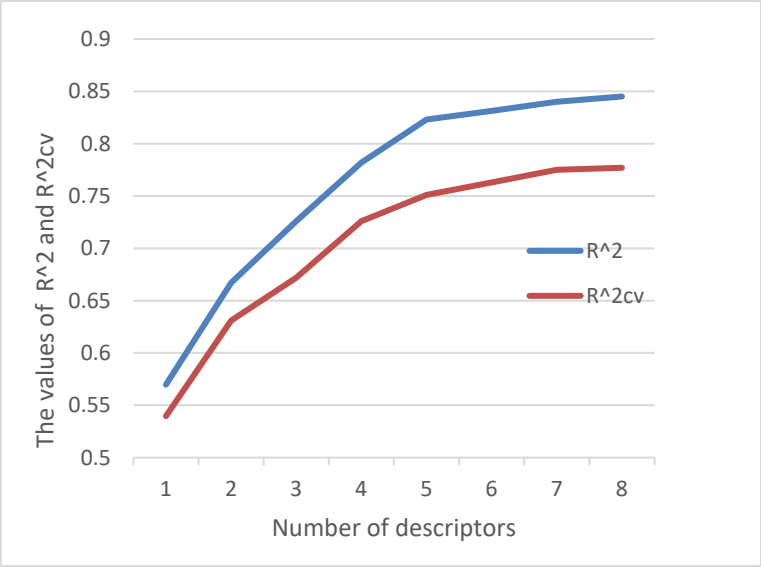


Figure 2. Influence of the Number of Descriptors on R^2 and R^2_{cv}

As shown in Fig. 2,with the number of descriptors increased from 1 to 5, there was a significant increase in the values of R^2 and R^2_{cv}. However, when the number of descriptors increased to 6, there was only a slight increase in R^2 and R^2_{cv}. Therefore, in this study, five molecular descriptors were chosen as the input features for the model.Physical-chemical meaning are shown in Table 2.Their correlation coefficients are shown in Table 3, which are significantly less than 0.8.

The training and testing set R^2 and MSE values in the model are 0.801, 0.939, 0.157, and 0.162, respectively. The HM modeling results are shown in Figure 3. The stability of the HM model was evaluated using Leave-One-Out cross-validation, resulting in Q^2 values of 0.771 and 0.785.

Table 2. The Descriptors Selected by HM and Their Physical-Chemical Meaning

Symbol	Physical-chemical meaning
KIO2	Kier&Hall index (order 2)
PPQQ	Polarity parameter (Qmax-Qmin)
XYS	XY Shadow / XY Rectangle
MOIB	Moment of inertia B
PPSD	Polarity parameter / square distance

Table 3. Correlation Matrix of the Five Descriptors

Descriptor	PPQQ	XYS	MOIB	PPSD	KIO2
PPQQ	1	0.059	-0.438	0.162	-0.134
XYS		1	-0.007	-0.267	0.153
MOIB			1	-0.137	-0.329
PPSD				1	0.285
KIO2					1

2023@ The American Computational Science Society. All rights reserved.

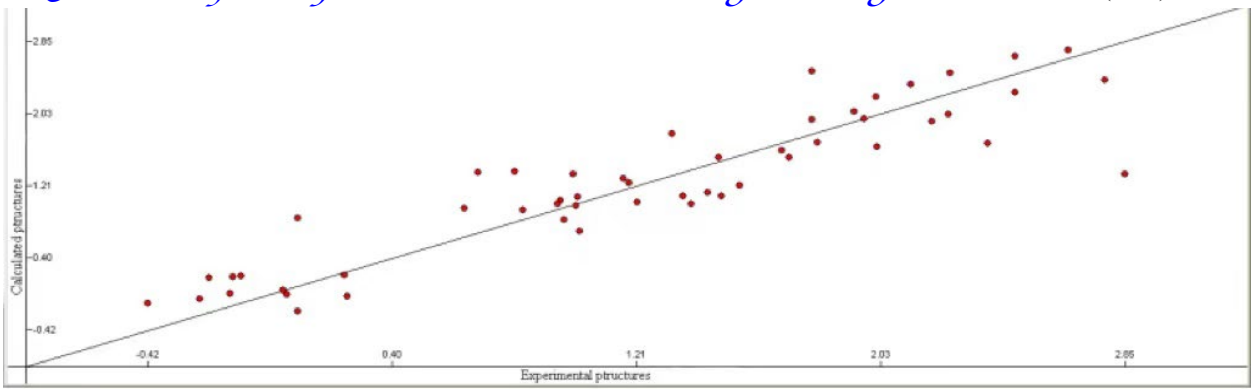


Figure 3. The Plot of Measured and Predicted lg (IC_{50}) by HM

3.2. Results of RF

After performing grid search optimization for the parameters, the optimal values are found to be n=10, f=5, while the value of d is set to the default in the Python library.The results of descriptor importance calculation based on Gini coefficients are shown in Figure 4. It can be observed that all 5 descriptors selected by HM are among the top 7 most important descriptors chosen by RF. This finding partly reveals the importance of these 5 descriptors and validates the effectiveness of HM. Therefore, the descriptors selected by HM (KIO2, PPQQ, XYS, MOIB, PPSD) are deemed highly significant, demonstrating that HM is an effective tool for descriptor selection.

Figure 5 displays the optimal prediction results of RF regression using the 5 descriptors determined by HM. The R^2 and MSE values for the training and test sets are 0.882 and 0.09, 0.836 and 0.139, respectively. During LOO-CV, the Q^2 values for the training and test sets are 0.889 and 0.763, respectively. Figure 5 represents the line plot for modeling using the random forest method.

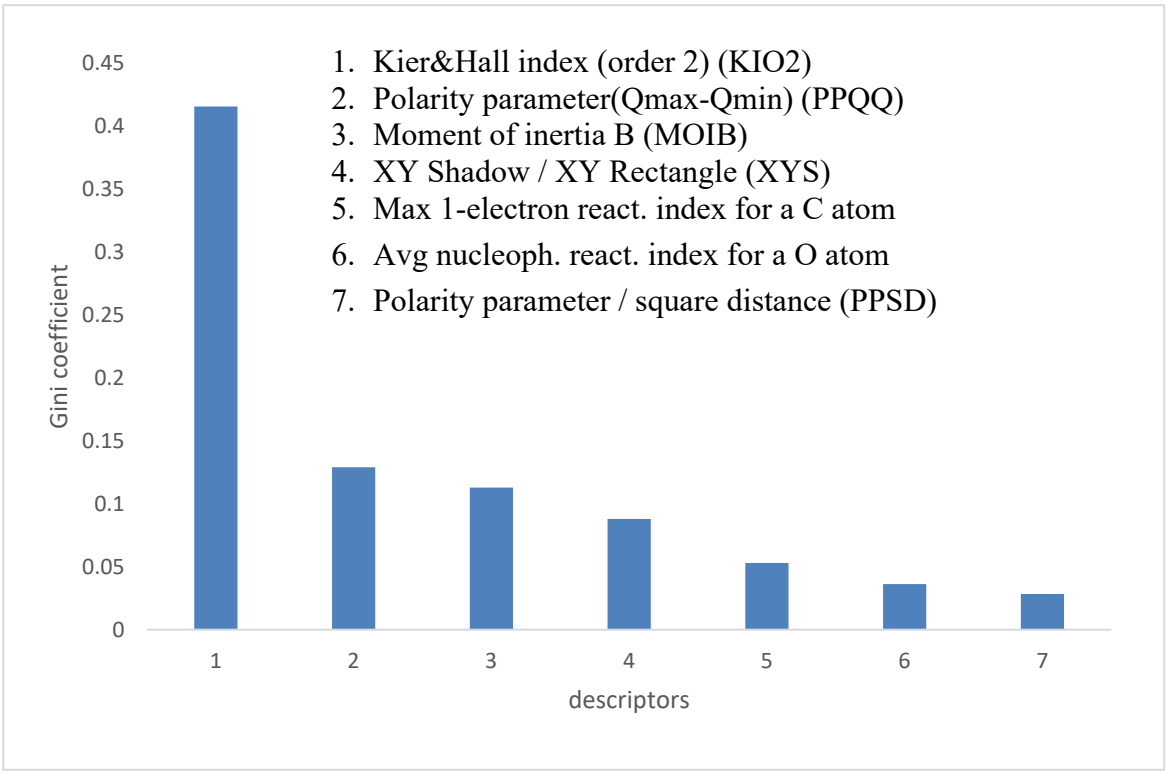


Figure 4.The Descriptors Selected by

2023@ The American Computational Science Society. All rights reserved.

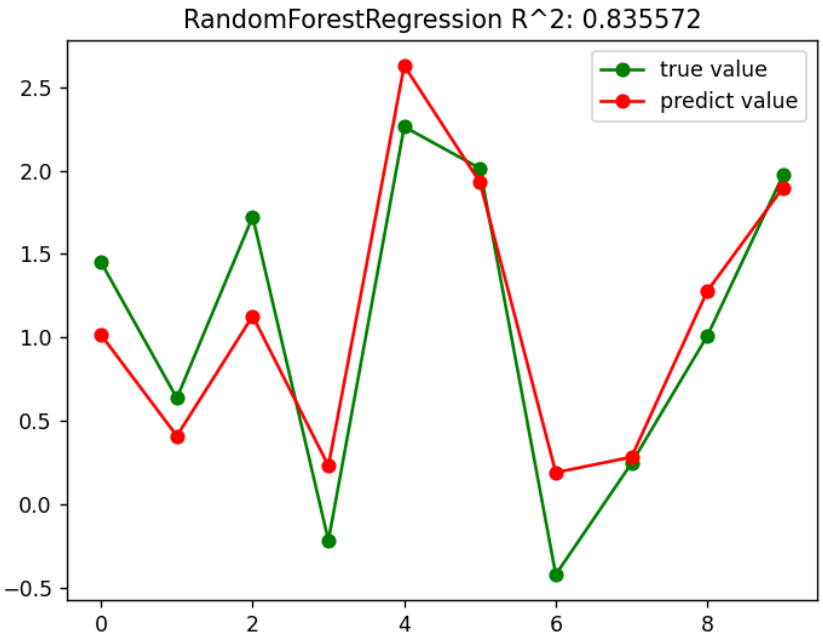


Figure 5. The line graph of the random forest modeling

3.3. Results of SVM

The study involved parameter tuning, and the optimal values for C, γ, and ε were determined to be 179.64, 5.43, and 0.2, respectively. In order to ensure comparability with the results of HM, RF, and GRNN, SVM regression modeling was performed using the 5 descriptors selected by HM. The optimal prediction results are shown in Figure 6, with R^2 and MSE values of 0.914, 0.944 for the training set and 0.066, 0.041 for the testing set, respectively. During LOO-CV, the Q^2 values for the training set and testing set were 0.92 and 0.905,respectively.

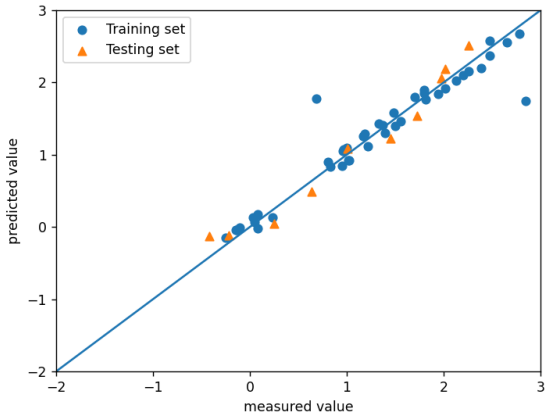


Figure 6.The Plot of Measured and Predicted lg (IC_{50}) by SVM

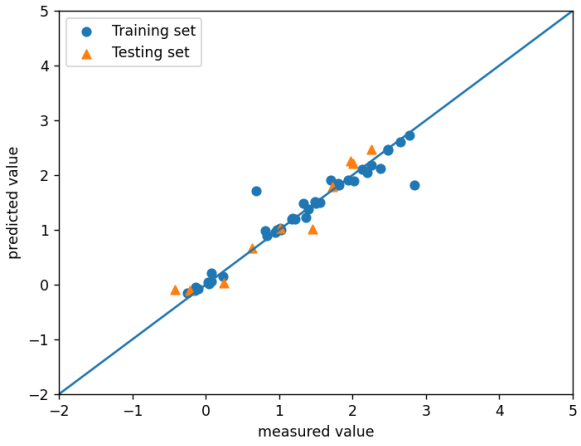


Figure 7. The Plot of Measured and Predicted lg (IC_{50}) by GRNN

3.4. Results of GRNN

GRNN utilized the values of five descriptors as input data and employed the Gaussian function as the transfer function in the pattern layer. Adjusting the value of the expansion parameter σ in the Gaussian function can affect the accuracy of the GRNN model. If the σ value is too small, it may lead to overfitting, while a too large value can result in underfitting. To find the optimal value, this study conducted tuning and eventually selected 0.2 as the optimal value. Experimental results demonstrated that when the σ value was between 0.1 and 0.3, the RMSE of the GRNN model for both training and

2023@ The American Computational Science Society. All rights reserved.

test sets reached the minimum. Ultimately, the GRNN model performed well on the training and test sets, achieving R^2 values of 0.913 and 0.943, and MSE values of 0.057 and 0.052, respectively. Furthermore, during LOO-CV, the Q^2 values for the training and test sets were 0.927 and 0.639, respectively. Overall, the predictive performance of the GRNN model was slightly lower than that of the SVM model. The optimal prediction results of GRNN are shown in Figure 7.

4. Comparison of different results

The prediction results of the four models and the LOO results are presented in Tables 4 and 5. In this study, five descriptors were selected for regression modeling. Based on the overall analysis, the following conclusions can be drawn:

1. Different models exhibit slight variations in their respective evaluation metrics. The SVM model performs well on both the training and test sets, and also shows the best performance in the cross-validation results. This indicates that the SVM model has good predictive ability, stability, and potentially better generalization capability.

2. The Generalized Regression Neural Network achieves prediction results comparable to SVM, but its model robustness is not as strong as SVM.

3. In both Random Forest and heuristic algorithms, the same important descriptors were selected, including KIO2, PPQQ, PPSD, XYS, and MOIB. This suggests that these descriptors are important for prediction, and the descriptors selected by the HM algorithm are effective. Particularly noteworthy is that KIO2 is consistently selected as the top node in the Random Forest models, indicating its highest Gini coefficient. Due to its low computational cost and good predictive performance, Random Forest plays a crucial role in machine learning studies on compound inhibition activity.

4. The constructed nonlinear models demonstrate better ability to describe complex real-world problems.

Therefore, when selecting models, it is necessary to consider different evaluation metrics and the specific characteristics of the problem in order to choose the most suitable model.

Among these five descriptors, KIO2 stands out as the most important one, consistently being selected as the top node in the RF model and exhibiting a significantly higher Gini coefficient compared to the other descriptors. KIO2 quantifies the shape and topological features of compounds by calculating the connectivity and distances between pairs of atoms within the molecule. It provides valuable insights into the interactions and spatial relationships among atoms, enabling the prediction of physicochemical properties and biological activity of molecules. The second notable descriptor, PPQQ, serves as a parameter for describing the polarity of compounds. It captures the polarity characteristics by calculating the difference between the maximum and minimum charges within the compound. PPQQ is widely used in molecular modeling, property prediction, and drug design, finding applications in various fields. Additionally, the MOIB descriptor allows us to gain an understanding of the rotational behavior of molecules in three-dimensional space, further enhancing our comprehension and prediction of their properties and behavior in chemical and physical processes. XYS, on the other hand, is employed to describe the shape and contour of molecules on a two-dimensional plane, providing valuable information about their structural characteristics. Lastly, PPSD is a descriptor that quantifies the polarity properties of chemical bonds in molecules and the strength of interactions between atoms. It helps in assessing the polarity characteristics and the nature of interatomic interactions within the molecule. In conclusion, these descriptors play a crucial role in characterizing and understanding the molecular properties of compounds, which will provide a research focus for the future design of novel JNK3 inhibitors.

Table 4. Comparison of Prediction Results of Different Methods

Method	R^2		MSE	
	Training set	Test set	Training set	Test set
SVM	0.914	0.944	0.066	0.041
RF	0.882	0.836	0.09	0.139
HM	0.801	0.939	0.157	0.162
GRNN	0.913	0.943	0.057	0.052

Table 5. Comparison of LOO-CV Results of Different Methods

Method	Q^2(Training set)	Q^2(Test set)
SVM	0.92	0.905
RF	0.889	0.763
HM	0.771	0.785
GRNN	0.927	0.639

5. Conclusion

2023@ The American Computational Science Society. All rights reserved.

In this study, a 5-parameter QSAR model was developed using methods such as SVM and GRNN to predict the compound diversity IC_{50} of 52 JNK3 inhibitors. The SVM model exhibited the strongest predictive ability and robustness. This study provides a new and effective method for predicting the inhibition activity IC_{50} of JNK3. Furthermore, this method can be applied to research on other compound properties and other properties. Additionally, the SVM model revealed 5 key descriptors that influence the activity of JNK3 inhibitors: Kier&Hall index (order 2), Polarity parameter (Qmax-Qmin), XY Shadow / XY Rectangle, Moment of inertia B, and Polarity parameter / square distance. These findings will guide future research in designing novel JNK3 inhibitors.

Conflict and interestComparison of different results

The authors confirm that this article content has no conflict of interest.

Acknowledgements

The authors are vary grateful to the APS softwareproviders.

References

1. Qin, P., et al., *Recent advances of small molecule JNK3 inhibitors for Alzheimer's disease.* Bioorganic Chemistry, 2022. **128**: p. 106090.
2. Bhatti, J.S., et al., *Stem cells in the treatment of Alzheimer's disease – Promises and pitfalls.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2023. **1869**(6): p. 166712.
3. Jun, J., et al., *Discovery of novel imidazole chemotypes as isoform-selective JNK3 inhibitors for the treatment of Alzheimer's disease.* European Journal of Medicinal Chemistry, 2023. **245**: p. 114894.
4. Si, H., et al., *Quantitative structure activity relationship study on EC50 of anti-HIV drugs.* Chemometrics and Intelligent Laboratory Systems, 2008. **90**(1): p. 15-24.
5. Si, Y., et al., *Novel quantitative structure-activity relationship model to predict activities of natural products against COVID-19.* Chem Biol Drug Des, 2021. **97**(4): p. 978-983.
6. Chen, Y., et al., *Network Pharmacology Integrated Molecular Docking to Reveal the Autism and Mechanism of Baohewan Heshiwei Wen Dan Tang.* Curr Pharm Des, 2022. **28**(39): p. 3231-3241.
7. Muratov, E., et al., *Existing and Developing Approaches for QSAR Analysis of Mixtures.* Molecular Informatics, 2012. **31**: p. 202-221.
8. Chen, Y., et al. *Design and Screening of New Lead Compounds for Autism Based on QSAR Model and Molecular Docking Studies*. Molecules, 2022. **27**, DOI: 10.3390/molecules27217285.
9. Si, Y., et al., *QSAR Model Study of 2,3,4,5-tetrahydro-1H-pyrido[4,3-b]indole of Cystic-brosis-transmembrane Conductance-regulator Gene Potentiators.* Letters in Drug Design & Discovery, 2021. **18**.
10. Coleman, W.F. and C.R. Arumainayagam, *HyperChem 5 (by Hypercube, Inc.).* Journal of Chemical Education, 1998. **75**(4): p. 416.
11. Früchtl, H.A., R.H. Nobes, and A. Bliznyuk, *Performance of mopac on parallel computers.* Journal of Molecular Structure: THEOCHEM, 2000. **506**(1): p. 87-97.
12. Stewart, J.J.P., *MOPAC: A semiempirical molecular orbital program.* Journal of Computer-Aided Molecular Design, 1990. **4**(1): p. 1-103.
13. Gawiejnowicz, S. and W. Gajda, *A note on exact algorithms and heuristics for generation of V- and Λ-shaped sequences.* Computers & Industrial Engineering, 2023. **178**: p. 109081.
14. Çelik, M.T. and S. Arslankaya, *Solution of the assembly line balancing problem using the rank positional weight method and Kilbridge and Wester heuristics method: An application in the cable industry.* Journal of Engineering Research, 2023: p. 100082.
15. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *QSPR: the correlation and quantitative prediction of chemical and physical properties from structure.* Chemical Society Reviews, 1995. **24**(4): p. 279-287.
16. Guo, W., et al., *Cost-effective assessment of in-service asphalt pavement condition based on Random Forests and regression analysis.* Construction and Building Materials, 2022. **330**: p. 127219.
17. Svetnik, V., et al., *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1947-1958.
18. Inthajak, K., et al., *A prediction approach for anti-HIV activity of HEPT compounds using random forest technique.* Monatshefte für Chemie - Chemical Monthly, 2017. **148**(10): p. 1697-1709.

2023@ The American Computational Science Society. All rights reserved.

19. Tyralis, H., G. Papacharalampous, and A. Langousis *A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources*. Water, 2019. **11**, DOI: 10.3390/w11050910.
20. Khemchandani, R., Jayadeva, and S. Chandra, *Regularized least squares fuzzy support vector regression for financial time series forecasting.* Expert Systems with Applications, 2009. **36**(1): p. 132-138.
21. Anyanwu, G.O., et al., *RBF-SVM kernel-based model for detecting DDoS attacks in SDN integrated vehicular network.* Ad Hoc Networks, 2023. **140**: p. 103026.
22. Fernandez, M., et al., *Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM).* Molecular Diversity, 2011. **15**(1): p. 269-289.
23. Waters, L., et al., *The use of a quantitative structure-activity relationship (QSAR) model to predict GABA-A receptor binding of newly emerging benzodiazepines.* Science & Justice, 2018. **58**(3): p. 219-225.
24. Bodyanskiy, Y.V., A.O. Deineko, and Y.V.J.R.E.C.S.C. Kutsenko, *SEQUENTIAL FUZZY CLUSTERING BASED ON NEURO-FUZZY APPROACH.* 2016(3).
25. Izonin, I., et al., *A GRNN-based Approach towards Prediction from Small Datasets in Medical Application.* Procedia Computer Science, 2021. **184**: p. 242-249.
26. Yakovyna, V. and N. Shakhovska, *Software failure time series prediction with RBF, GRNN, and LSTM neural networks.* Procedia Computer Science, 2022. **207**: p. 837-847.
27. Aengchuan, P., et al., *Prediction of the influence of castor oil–ethanol–diesel blends on single-cylinder diesel engine characteristics using generalized regression neural networks (GRNNs).* Energy Reports, 2022. **8**: p. 38-47.

2023@ The American Computational Science Society. All rights reserved.