Journal of Computational Science & Engineering 60 (2023) 1-10

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

Quantitative structure activity relationship study on IC50 of novel evodiamine derivatives against gastric cancer by machine learning

Zixuan Yang[1*], Guoliang Li[2]

[1]Qingdao No.58 Senior High School, Qingdao, 266071, China

[2]School of Computer Science and Technology, Qingdao University, Qingdao, 266071, China

Abstract: **A quantitative structure activity relationship (QSAR) study was performed based on a data set of 42 novel evodiamine derivatives that can against stomach cancer. 551 chemical descriptors were calculated from molecular structure of the compounds. The QSAR models were built using the heuristic method (HM), support vector machine (SVM) and gene expression programming (GEP). Model established by SVM is the best one among all these three models generated, with R^2=0.9542, for training set and R^2=0.9316 for test set. The leave-one-out (LOO) method was chosen as the cross-validation method, and the result of LOO is R^2=0.8924. The predicted results of SVM model are consistent with the experimental results. This study reveals 4 key descriptors of novel evodiamine derivatives and will help to screen out novel and effective drugs in the future.**

Keywords: QSAR, Support vector machine, Heuristic method, Gene expression programming, Evodiamine derivatives

Received 3 March 2023, Revised 7 March 2023, Accepted 28 March 2023

*Corresponding Author: Zixuan Yang, hedgehog_qd@163.com

1. Introduction

Stomach cancer, also called the gastric cancer, is a kind of malignant tumor originating from the gastric mucosal epithelium, and it is mainly related to helicobacter pylori, dietary, and heredity. The stomach cancer is the 5th most common cancer worldwide, and also the third leading cause of death worldwide [1]. According to the World Health Organization (WHO) latest estimates, the stomach cancer has affected 1,089,103 people around the world in 2020, and 768,793 died because of it among those people. Those numbers are still increasing through years. Among all the continents, Asia is by far the highest incidence among all the regions, which has 819,944 people under the disease, which also took 75.3% of all patients worldwide in 2021[2, 3].

At present, an increasing number of modern treatments have been developed to deal with stomach cancer, such as chemotherapeutic drugs which were derived from natural products and analogues. However, Research and development process of such drugs has indicated that drugs based on nature products are hard to develop without modification, which is caused by insufficient activity, high structural complexity, significant toxicity and so on [4, 5].

Natural product evodiamine (EVO) was firstly isolated from a Chinese medicine Wu-Chu-Yu, which was used to deal with problems related to stomach and spleen in the history [6]. The previous study has proved that it can help to lower the blood pressure and can be used to deal with inflammatory and tumor [7, 8]. The evodiamine derivatives also exhibits excellent broad-spectrum anti-tumor activity during the experiment [4]. Because of its simple structure and outstanding effect in treating gastric cancer, many evodiamine derivatives have been investigated for gastric cancer inhibitory activity on IC_{50} values of SGC-7901 cell lines. However, the traditional methods measuring IC_{50} value of evodiamine derivatives always needs great manpower and material resources, so the quantitative structure-activity relationship (QSAR) method is used to predict their properties accurately and effectively.

Quantitative structure-activity relationship (QSAR) method is based on the theory that the properties of chemical compounds are expressed by many measured physicochemical properties, which can be described as descriptors. This method can mainly help to predict the properties of new compounds quickly and precisely by establishing models based on several precisely selected molecular descriptors rather than chemical experiment, which can save a lot of human and time resources. The QSAR is normally developed to predict the IC_{50} value of drugs because the physicochemical descriptors of drugs are assumed to influence with IC_{50}.

In this study, 3 models which by heuristic method (HM), gene expression programming (GEP), support vector machine (SVM) were established to predict IC_{50} values of evodiamine derivatives. Firstly, the software CODESSA was used to calculate descriptors from the data set, and then the data set were separated into the training set (33 chemical compounds) which was used to train the models and the testing set (9 chemical compounds) which was used to evaluate the models. Finally, 4 kinds of descriptors were chosen to train the model with those three methods. The LOO method was used as the cross-validation method to test the stability of the models by those three methods. The result shows that the performance and stability of model developed by SVM is better than these of models built by HM and GEP.

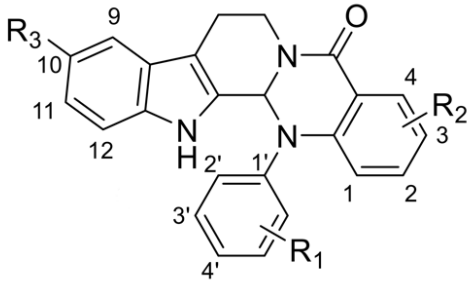
2023@ The American Computational Science Society. All rights reserved.

2. Material and method

2.1. Data set

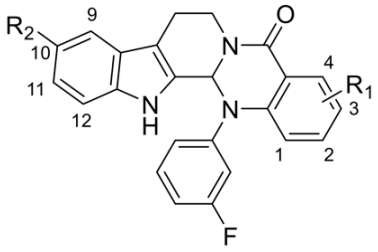
A complete list of compounds name and their half maximal inhibitory concentration (IC_{50}) are shown in Table 1. All those compounds were collected from two cancer research papers focusing on evodiamine derivatives [4, 5]. The data for these 42 compounds were randomly split into the training and test sets. The training set has 33 compounds, and it is used to establish the models, while the test set has 9 compounds, and it is used to evaluate the capability of models.

Table 1a. Measured IC50 value of compounds 1-13 *The compounds of training set



Compound	R_1	R_2	R_3	Measured IC_{50}	Predicted IC_{50}		
					HM	GEP	SVM
1	2'-F	H	H	8.90	6.37	8.08	8.80
2	3'-F	H	H	1.67	2.93	2.05	1.77
3	3'-Cl	H	H	2.33	3.92	0.44	2.23
4	3'-Br	H	H	6.41	5.38	5.07	6.31
5	3'-Me	H	H	5.04	3.74	4.37	4.68
6	3'-F	3-OMe	H	0.86	2.45	1.40	0.96
7	3'-F	3-Cl	H	1.36	1.69	0.80	1.46
8	3'-F	3-Br	H	3.86	3.68	2.60	3.76
9	3'-F	4-Me	H	1.94	3.37	4.22	2.04
10	3'-F	H	10-Me	3.61	3.70	3.46	3.71
11	3'-F	H	10-Cl	0.85	4.28	0.31	0.95
12	3'-F	3-Me	10-OMe	2.01	2.23	1.72	1.91
13	3'-F	3-Me	10-Cl	1.86	2.47	1.20	1.96

Table 1b. Measured IC50 value of compounds 18-42 *The compounds of training set

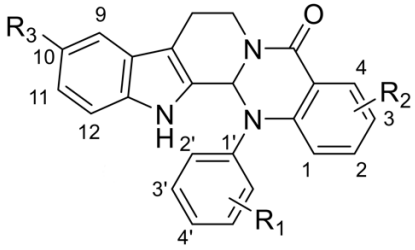


Compound	R_1	R_2	Measured IC_{50}	Predicted IC_{50}		
				HM	GEP	SVM
14	1-F	H	2.36	2.73	3.56	2.46
15	1-Me	H	1.55	2.55	1.66	1.45
16	2-F	H	6.62	3.61	5.09	6.52
17	3-CF_3	H	3.32	0.35	-8.91	0.93
18	3-F	H	0.89	0.60	1.82	0.99
19	3-Br	H	3.86	2.79	3.44	3.76

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

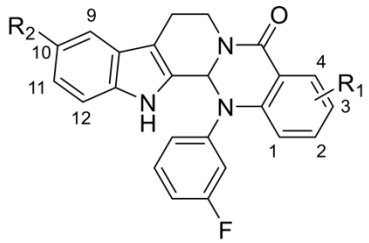
20	3-OMe	H	0.86	1.39	1.01	0.96
21	3-Me	H	0.69	0.37	1.31	0.84
22	4-Me	H	1.94	2.83	2.50	1.84
23	H	10-OMe	5.60	4.24	4.37	5.50
24	H	10-Cl	0.85	3.81	1.73	0.95
25	H	10-Me	3.61	3.27	3.52	3.51
26	3-Me	10-OMe	2.01	1.50	1.56	1.91
27	3-Me	10-Cl	1.86	1.12	1.66	1.76
28	3-Me	10-Me	0.24	0.37	1.31	0.84
29	3-F	10-OMe	0.57	1.74	-0.61	0.67
30	3-F	10-Cl	0.89	1.12	1.40	0.99
31	3-F	10-Me	0.88	0.39	0.30	0.78
32	H	10-OH	0.28	1.23	-0.94	0.37
33	3-F	10-OH	0.13	-1.35	0.12	0.23

Table 1c. Measured IC50 values of compounds 34-37 *The compounds of test set



Compound	R_1	R_2	R_3	Measured IC_{50}	Predicted IC_{50}		
					HM	GEP	SVM
34	3'-CF_3	H	H	5.70	3.74	4.37	4.68
35	3'-F	3-Me	H	0.69	1.21	0.45	0.56
36	3'-F	H	10-OMe	5.60	3.85	4.03	4.97
37	3'-F	3-Me	10-Me	0.24	1.61	0.97	0.24

Table 1d. Measured IC50 values of compounds 38-42 *The compounds of test set



Compound	R_1	R_2	Measured IC_{50}	Predicted IC_{50}		
				HM	GEP	SVM
38	H	H	1.67	2.55	1.66	1.45
39	3-Cl	H	1.36	0.78	-0.49	1.42
40	4-F	H	1.86	1.94	3.07	1.24
41	3-F	10-OAc	0.27	0.34	1.12	0.96
42	3-F	10-OEt	2.21	2.46	2.42	2.29

2.2 Calculating Descriptors

Molecular descriptors were calculated from chemical compounds, which influence the prediction performance of QSAR model greatly. The steps of calculating the molecular descriptors are described as bellow (Fig. 1): Twice optimization was performed by using MM+ molecular mechanics force field and semi-empirical AM1 method in HyperChem software after 42 compounds were drawn in ChemDraw, which were obtained the lowest energy structure.

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

After that, the files were put into MOPAC software to get the MNO files, and then they were imported to CODESSA software [9] to help calculating descriptors, finally, four 4 classes of descriptors were calculated as follows: XY Shadow / XY Rectangle, Molecular weight, Relative negative charged SA [Zefirov's PC] and RPCS Relative positive charged SA [Zefirov's PC].

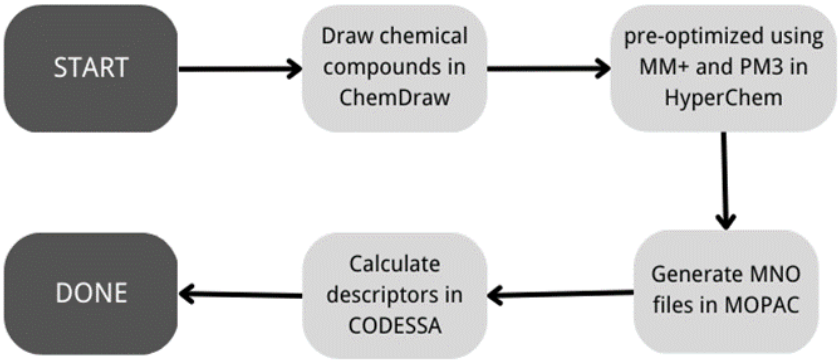


Figure 1. The whole progress of calculating chemical descriptors

2.3 Theory of Heuristic Method (HM)

After calculating all molecular descriptors, the HM in CODESSA was used to preselect descriptors and build several linear models. The high efficiency of calculation and no limits on the data set are the significant advantages of this method.

Before establishing HM models, descriptors should be preselected. Firstly, all the descriptors were checked to make sure all of them are universal for all chemical structures in this study, and the value of them are not constant. The descriptors which don't fit those rules above were discarded.

Secondly, the correlation matrix of descriptors is calculated and the nonlinear descriptors whose correlation coefficient is greater than 0.8 should be discarded.

After the preselection of descriptors, the number of candidate descriptors will greatly reduce and linear regression models with different number of parameters are subsequently developed.

In order to get the model with the best performance, statistical criteria such as correlation coefficient R^2 were also adopted to evaluate the models. The LOO method was used to do the cross-validation to check the stability of the models [10].

2.4 Theory of GEP

The GEP was developed by Ferreira in 1999 [11] which is the combination of genetic algorithm (GA) and genetic programming (GP). It is expressed with genes and has the advantages such as easy operation, and flexible, powerful ability to search the expression characteristics with a much higher efficiency than the traditional ways such as GA or GP algorithm [12].

The GEP includes chromosomes and the expression trees (ETs). The processes of gene encoding and translation is just like a one-to-one relationship between symbols of the chromosome and the functions or terminals they represent [13].

The steps for GEP are described below: At first, chromosomes will be randomly chosen, and act as the initial population. After that, the fitness of every individual in that chromosome will be evaluated. Then, the best ones will be selected for reproduction with modifications on the basis of their fitness to form new generation. And finally, the steps above will be repeated for new generation until the termination condition is met or the optimal solution is found [14].

The GEP is widely used in science and research because of its good performance. The software APS (Automatic Problem Solver) can easily establish the non-linear model based on the descriptors calculated by HM. However, there is a disadvantage of GEP that it is easy to fall into locally optimal solution which cannot have a good performance on the whole dataset.

2.5 Theory of Support Vector Machine (SVM)

The Support Vector Machine, which was developed by Vapnik [15] as a new supervised machine learning method, has been widely used because of its excellent performance in classification and regression. When SVM is used in regression, it is also called Support Vector Regression (SVR).

What a Support Vector Machine do is to make sure those support vectors are as far as they can apart from the hyperplane (Fig. 2). And it also needs to minimum the number of data points inside the area between two decision boundaries (blue and purple line), which makes the model has better robustness.

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

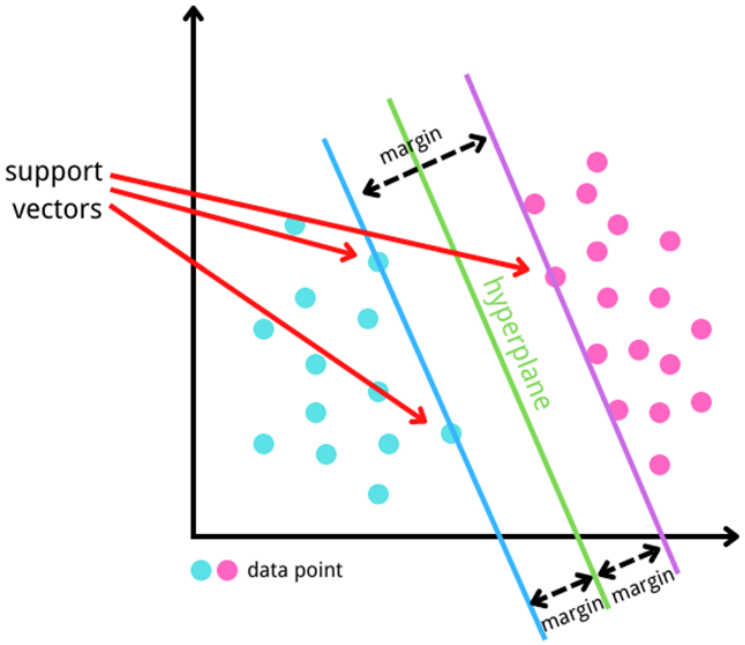


Figure 2. The progress of SVM

In SVR (Fig. 3), input data points are mapped into a higher dimension feature space through a non-linear transformation (kernel function) at first. Then, a linear regression is performed in the feature space. SVR is to find a hyperplane and maximize the number of data points included in the margin between the two decision boundaries (the blue line and the purple line in Fig. 3). If the hyperplane is written in the form like:

$$f(x) = \omega^T x + b$$

Then the decision boundaries (in Fig. 3) can be written as:

Purple line: $f(x) - \varepsilon$

Blue line: $f(x) + \varepsilon$

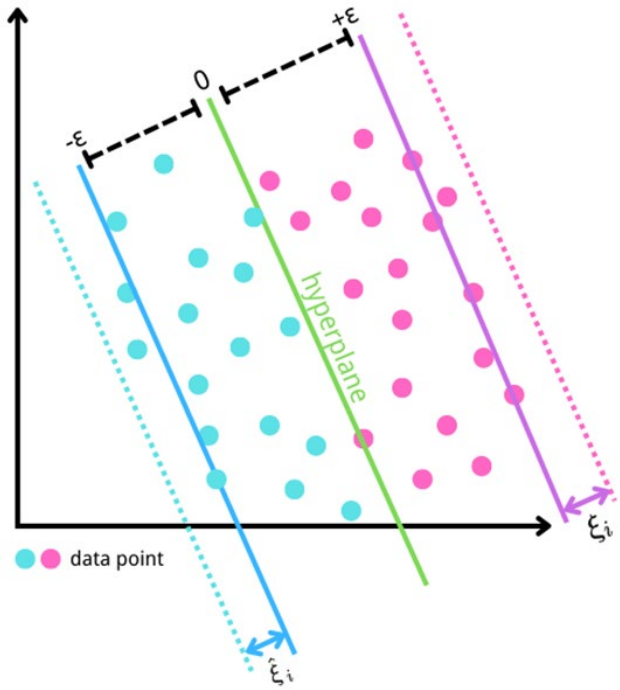


Figure 3. The work of SVR

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

As for the SVR problem can be written as:

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} \iota_\varepsilon\left(f(x_i) - y_i\right) (1)$$

Where C is a regularization constant which also called punish coefficient, and ι_ε is the ε-insensitive loss function:

$$\ell_\epsilon(z) = \begin{cases} 0, if \ |z| \le \epsilon \\ |z| - \epsilon, otherwise \end{cases} \quad (2)$$

After the equation (3) been added the slack variables (ξ_i and $\widehat{\xi_\iota}$) which are related to generalization ability, the formula (1) can be written as the equation (3):

$$\min_{\omega,b,\xi_i,\widehat{\xi_\iota}} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}(\xi_i + \widehat{\xi_\iota}) \quad (3)$$

The original idea of SVR is using a predefined function to map the input data points to high dimension feature space and then performs regression. In order to reduce the complexity of operations, kernel function written as the format $K(\cdot,\cdot)$ is adopted to realize inner product operation in the original space.

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \hat{\alpha}_i)K(x_i, x) + b \quad (4)$$

Inside the equation (4), N represents the number of support vectors, α_i and $\hat{\alpha}_i$ are Lagrange multipliers. $\mathrm{K}(x_i, \mathrm{x}) = \langle x_i, x_j \rangle$ is the kernel function which can be sigmoid kernel, polynomial kernel and linear kernel, radial basis kernel and so on. [16]

In this study, the Gaussian Radial Basis function (RBF) was chosen as the kernel function K.

3. Result and discussion

3.1. Results of HM

A total of 551 descriptors for all compounds were calculated by the CODESSA software. In order to find the set of descriptors that affect the IC_{50} value of evodiamine, linear models with 1 to 5 descriptors were established and the influence by number of descriptors are shown in Fig. 4. Fig. 4 also shows that there is no obvious increase of R^2 when the number of descriptors in the linear model ranges from 4 to 5. Therefore, 4 descriptors were selected, and their name and physical-chemical meaning and their coefficients are shown in Table 2. Table 3 illustrates the correlation matrix of the selected descriptors.

Table 2. The selected descriptors and their physical-chemical meaning, coefficient and T-test values

Descriptors	Physical-chemical meaning	Coefficient	T-test
S_{xy}	XY Shadow / XY Rectangle	65.042	5.8476
M	Molecular weight	0.051013	3.4451
RNCG	RNCS Relative negative charged SA [Zefirov's PC]	-0.25303	-2.8653
RPCG	RPCS Relative positive charged SA [Zefirov's PC]	-1.311	-2.3087

Table 3. Correlation matrix of the 4 descriptors

Descriptor	XY Shadow / XY Rectangle	Molecular weight	RNCS Relative negative charged SA [Zefirov's PC]	RPCS Relative positive charged SA [Zefirov's PC]
XY Shadow / XY Rectangle	1.00	-0.4484	0.2072	0.0325
Molecular weight	-0.4484	1.00	-0.1449	-0.1520
RNCS Relative negative charged SA [Zefirov's PC]	0.2072	-0.1449	1.00	-0.0121
RPCS Relative positive charged SA [Zefirov's PC]	0.0325	-0.1520	-0.0121	1.00

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

The R^2 and R^2cv values of the training set are shown in Fig. 4. It is clear that the R^2 and R^2cv are increased while the number of descriptors rises.

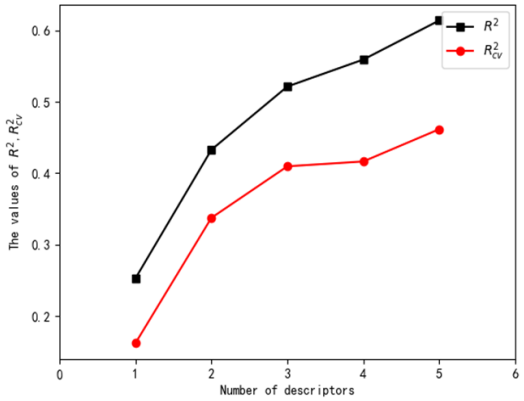


Figure 4. R^2 and R^2cv values affected by the number of descriptors

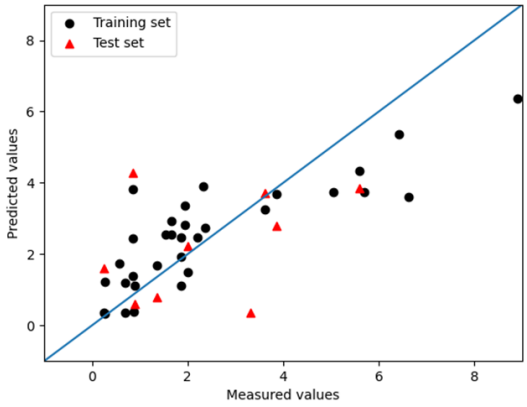


Figure 5. Plot of Measured and predicted IC_{50} by HM

The R^2 of the training set and testing set in the model are 0.559, 0.482. Furthermore, the LOO-CV of R^2 of the data set is 0.417.

3.2. Results of GEP

In order to compare the prediction result of the three models by GEP, HM and SVM, same descriptors selected by HM were used in constructing other models. The software automatic problem solver (APS) [17] was adopted to calculate and build the GEP model based on the selected descriptors in HM. The simplest function set is expressed as below equation (5):

$$IC50 = cos(((((x0 + x0) * sqrt(x1)) * sqrt(x0)) - x3)) + tan(sin(pow(10, fabs(sin(x1))))) + ((fabs(x0)/x3) * x0) + cos((cos(log(pow(10, x3))) - (sqrt(x1) * (x1/x0)))) + cos(tan(cos(((x2 - (x3 * x1)) * (x2 + x0))))) \quad (5)$$

In equation (5), x0, x1, x2, x3 represents XY Shadow, Molecular weight, RNCS Relative negative charged SA [Zefirov's PC] and RPCS Relative positive charged SA [Zefirov's PC].

The performance of GEP model is illustrated in Fig. 6. The R^2 of the training set and testing set in the model are 0.82 and 0.68. And the R^2 of LOO cross validation of the data set was 0.78.

3.3 Results of SVM

In this part, the SVM was used to predict the QSAR model based on the same dataset. As the part 2.5 shows, the result and performance of SVM is depends on coefficient C, kernel function k, and ϵ. In this model, a Gauss kernel is used as the kernel function k, which can greatly affect the performance of the model. It can be expressed as the following form: $k(x_i, x_j) = \exp\left(-\gamma|x_i - x_j|^2\right)$. The grid search was used to select the best value of C and γ. And the result was: $C = 626.81$ and $\gamma = 27.45$.

The IC_{50} it predicted in both training set and test set is shown below in Fig. 7. In this part, the best R^2 of the training set is 0.9542, while that of the test set is 0.9316. The LOO cross-validation of SVM is 0.8924.

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

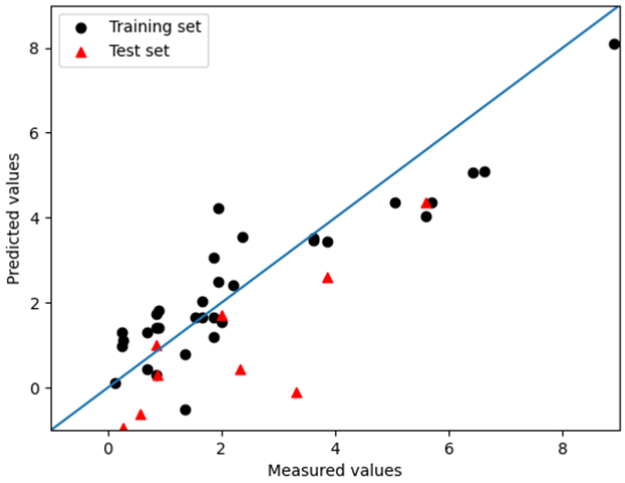


Figure 6. Plot of measured and predicted IC50 by GEP

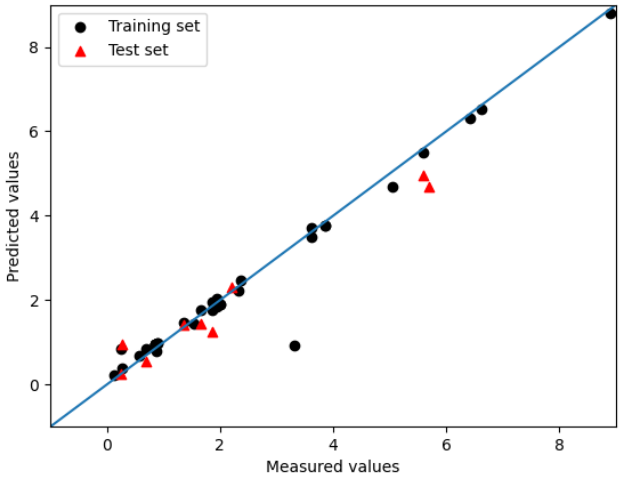


Figure 7: Plot of measured and predicted IC_{50} by SVM.

All predicted results of the three models is shown in Table 4.

Table 4. Comparison of models

Models	R^2 of training set	R^2 of test set
HM	0.559	0.482
GEP	0.82	0.68
SVM	0.9542	0.9316

Obviously, the prediction ability and stability of the non-linear model built by SVM is best among all the three models. The GEP method is easy to get into a local optimal solution, while the SVM is a convex optimization problem which can always get the global optimal solution. The SVM can also get a better result on complex situations than HM because it is a non-linear model construction method.

4. Discussion

One of the most important uses of QSAR models is providing the chemical insights of drug design, as predicting the IC_{50} value of molecules, and finding the factors which can affect the IC_{50}. In this study, four structural descriptors were used to generate the model.

The descriptor XY Shadow / XY Rectangle is belonging to the geometrical group, which shows the contour of the projection of the molecule on the plane defined by XY axes of the molecule. It has been proved that the shadow areas have a high correlation to molecular polarizability, molecular connectivity, molecular volume, and molar refraction [18]. It can be calculated through the equation (6).

$$S_k = \frac{1}{2}\oint_{(C)}(\gamma d\rho - \rho d\gamma) \quad (6)$$

In equation (6), C is the contour of the projection of the molecule on the plane defined by two principal axes of the molecule (k = XY), and γ is x, ρ is y.

And the descriptor Molecular Weight is belonging to the constitutional group, which shows the molecular weight of molecule.

Descriptor RNCG Relative negative charged SA (SAMNEG*RNCG) [Zefirov's PC] is belong to the CPSA group, which is an electrostatic descriptor that depends on the distribution of the charges on the molecule. The relative negative charge of the molecule and its surface area can influence activity [19]. It can be calculated through the equation (7).

$$RNCG = \frac{\delta_{max}}{\sum_A \delta_A} \qquad A \in \{\delta_A > 0\} \quad (7)$$

In equation (7), δ_{max} is the maximum atomic negative charge in the molecule, and δ_A is the negative atomic charge in the molecule.

And the descriptor RPCG Relative positive charged SA (SAMPOS*RPCG) [Zefirov's PC] is also in the group CPSA,

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

which reflects the total molecular surface area and properties of the function group and indicates interactions among polar molecules [20]. It can be calculated through the equation (8).

$$RPCG = \frac{\delta_{max}}{\sum_A \delta_A} \qquad A \in \{\delta_A > 0\} \quad (8)$$

In equation (8), δ_{max} is the maximum atomic positive charge in the molecule, and δ_A is the positive atomic charge in the molecule.

5. Conclusion

In this study, QSAR models were built based on totally 551 structural descriptors calculated from molecular structures. Three QSAR models were established by HM, GEP and SVM to predict the IC_{50} value of drugs for stomach cancer. The result shows that SVM is a very promising tool for further study in anti-stomach drug design aspect. Compared with GEP and HM, SVM has the best performance and efficiency on identifying and forecasting the relationship between structural descriptors and IC_{50} value of these compounds based on the same data set.

Conflict and interest

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

Acknowledgements

The authors are vary grateful to the APS softwareproviders.

References

[1] https://www.wcrf.org/cancer-trends/stomach-cancer-statistics/

[2] https://gco.iarc.fr/today/data/factsheets/cancers/7-Stomach-fact-sheet.pdf

[3] Katoh, Hiroto, and Shumpei Ishikawa. “Lifestyles, genetics, and future perspectives on gastric cancer in east Asian populations.” *Journal of human genetics* vol. 66,9 (2021): 887-899.

[4] Liang, Ziyi et al. “Design, synthesis and bioactivity evaluation of novel evodiamine derivatives with excellent potency against gastric cancer.” *European journal of medicinal chemistry* vol. 228 (2022): 113960.

[5] Liang, Ziyi et al. “Design, synthesis and bioactivity evaluation of favorable evodiamine derivative scaffold for developing cancer therapy.” *European journal of medicinal chemistry* vol. 239 (2022): 114530.

[6] Dong, Guoqiang et al. “New tricks for an old natural product: discovery of highly potent evodiamine derivatives as novel antitumor agents by systemic structure-activity relationship analysis and biological evaluations.” *Journal of medicinal chemistry* vol. 55,17 (2012): 7593-613.

[7] Hu, Xu et al. “Antiproliferative Effects of Alkaloid Evodiamine and Its Derivatives.” *International journal of molecular sciences* vol. 19,11 3403. 30 Oct. 2018.

[8] Gavaraskar, Kirti et al. “Therapeutic and cosmetic applications of Evodiamine and its derivatives--A patent review.” *Fitoterapia* vol. 106 (2015): 22-35.

[9] Katritzky, A.R., Ignatchenko, E., Barcock, R., Lobanov, V., Karelson, M., 1994a. Retention Times and Response Factors using a General QSPR Treatment. Anal. Chem. 66, 1799–1807.

[10] Katritzky, A R et al. “QSPR analysis of flash points.” *Journal of chemical information and computer sciences* vol. 41,6 (2001): 1521-30.

[11] Ferreira, Candida. "Gene expression programming: a new adaptive algorithm for solving problems." *arXiv preprint cs/0102027* (2001).

[12] Zheng Yi Research *on GEP algorithm and its decoding* method.2018 Guangzhou University,MA thesis.

[13] Ferreira, Candida. "Gene expression programming in problem solving." *Soft computing and industry: recent applications* (2002): 635-653.

[14] Yang, Huazhen, et al. "Studies on the IC50 of trisubstituted thiazoles as Cdc7 kinase inhibitors." *Letters in Drug Design & Discovery* 13.1 (2016): 33-42.

[15] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995).

[16] Zhihua, Z. (2016) “SVR,” in Machine Learning. Beijing: Tsinghua University Press, pp. 133–138.

[17] http://www.gepsoft.com/gepsoft

[18] Rohrbaugh, Robert H., and Peter C. Jurs. "Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships." *Analytica Chimica Acta* 199 (1987): 99-109.

[19] Gallorini, Marialucia et al. “Natural and Synthetic Xanthone Derivatives Counteract Oxidative Stress via Nrf2 Modulation in Inflamed Human Macrophages.” *International journal of molecular sciences* vol. 23,21 13319. 1 Nov. 2022.

[20] Yang, Dongmei et al. “Quantitative Structure Activity Relationship of Cinnamaldehyde Compounds against Wood-

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

Decaying Fungi." *Molecules (Basel, Switzerland)* vol. 21,11 1563. 17 Nov. 2016.

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