Regular Article

PHYSICAL CHEMISTRY RESEARCH
Published by the
Iranian Chemical Society
www.physchemres.org
info@physchemres.org

Phys. Chem. Res., Vol. 8, No. 2, 215-223, June 2020
DOI: 10.22036/pcr.2020.205099.1686

QSAR Study of Arylpyridone Oxime Based on the SVM and Elman Algorithms

W. Zhong-Yu*

School of Chemistry & Chemical Engineering, Xuzhou Institute of Technology, Xuzhou, China

(Received 13 October 2019, Accepted 28 January 2020)

Based on the topological chemistry theory, a new quantum chemistry method is used to calculate 91 electrical distance vector (M_i) of molecules in order to describe the chemical microenvironment. The multivariate stepwise regression method is used to screen important variables to obtain the best ternary equation. Correlation coefficients $R^2 = 0.887$ and $R^2_{CV} = 0.673$ were checked by F_{IT} and A_{IC} .The ternary variable is used as the input set, and the inhibition rate is used as the output set. The LS-SVM and Elman-ANN algorithms were used to fit them and compare with each other. The results show that R^2 values are 0.993 and 0.994, respectively. Their prediction ability is similar, while the stability of Elman is better, and the two-dimensional structure affecting the molecule is structural fragments such as =CH-, -CH<.

Keywords: LS-SVM, Elman-ANN, QSAR, Molecular electronegativity distance vector

INTRODUCTION

Wang Jia-yao *et al.* [1] measured the inhibition rate of target derivatives using Brassica napus as the experimental object. They showed that the novel arylpyridone oxime derivatives have obvious inhibitory effects on the sclerotinia sclerotiorum of rapeseed plants, inhibiting the growth of Sclerotinia sclerotiorum effectively. Wang Jia-yao *et al.* also showed that the antibacterial effect of arylpyridone oxime derivatives is better than that of the traditional antibacterial drug such as Azinc; the antibacterial activity of some derivatives can reach 100%. Therefore, it is necessary to study the derivatives and structural debris and their corresponding reaction mechanisms. Considering the low yield of organic substances and the high purity of the products, the cycle of chemical reaction and the cost are too high, So, it is impossible to carry out chemical analysis on the products one by one. Liu Shu-shen *et al.* [2] studied the possibility of QSAR method in predicting drug activity. Feng Chang-jun *et al.* [3-6] used QSAR method to study the drug molecules and achieved good prediction results. Wang chao *et al.* [7,8] used the MLR method to study the structure-activity relationship of a large number of different organic structures, however, the research using Elman and SVM algorithms has rarely been reported. Therefore, it is necessary to study QSAR of these compounds, in order to reduce the cost of the experimental measurements, calculating the topological index firstly, the molecular electrical distance vector is calculated by software, establishing QSAR model and using the neural network Elman algorithm and the support vector machine algorithm to fit and compare with each other.

*Corresponding author. E-mail: zhongyuwanxzit@163.com

RESEARCH OBJECT AND METHOD

Research Object Structure and Data

The inhibition rate of the target derivatives was measured using the nuclear pathogen of Brassica napus as an experimental tool. The basic structure of the novel arylpyridone oxime derivatives is shown in Fig. 1. The inhibitory activities (IC_{70}) of the substituents R_1, R_2, corresponding to Sclerotinia sclerotiorum, are shown in Table 1.

.

Theoretical Method of Calculation

Electrical distance vectors are topological indices used

Table 1. Substituents R1, R2 and Corresponding Inhibitory Activities of Sclerotinia Sclerotiorum

No.	R_1	R_2	M_{15}	M_{18}	M_{60}	IC_{70}
1	Ph	Ph	36.035	8.6452	0.56444	52.45
2	4-tBu-Ph	4-NO_2-Ph	46.6	8.3764	0.55673	11.34
3	CH_3	2-CH_3-Ph	22.504	5.8523	0.61808	100.00
4	3,4-2Cl-Ph	3-Py	40.27	23.815	0.70595	33.22
5	4-CH_3-Ph	4-Py	37.379	23.968	0.67066	25.76
6	3,4-2CH_3-Ph	4-Py	39.736	24.34	0.6755	10.97
7	2-CH_3-3-Cl-Ph	4-Py	40.024	23.727	0.67565	27.95
8	4-Br-Ph	4-Py	35.97	23.15	0.68207	25.76
9	4-Br-Ph	2-Cl-3-Py	40.38	24.163	0.67483	43.96
10	4-tBu-Ph	2-Cl-3-Py	38.343	18.949	0.73046	74.28
11	4-CH_3-Ph	2-Cl-3-Py	39.849	19.14	0.73078	60.94
12	3,4-2CH_3-Ph	2-Cl-3-Py	35.366	18.774	0.72612	68.60
13	4-CH_3-Ph	3-Cl-4-Py	39.131	18.484	0.73406	65.89
14	3,4-2CH_3-Ph	3-Cl-4-Py	38.821	20.222	0.69136	32.31
15	4-Br-Ph	3-Cl-4-Py	41.432	19.916	0.69638	53.20
16	4-Ph-Ph	3-Cl-4-Py	41.85	20.398	0.69556	50.77
17	3-NO_2-4-CH_3-Ph	3-Cl-4-Py	56.138	21.008	0.68385	17.28
18	4-tBu-Ph	3-Cl-4-Py	38.534	19.259	0.67241	36.46
19*	4-CH_3-Ph	4-Cl-3-Py	43.361	20.586	0.69565	27.55
20*	2,4-2CH_3-Ph	4-Cl-3-Py	37.737	20.289	0.72035	53.42
21*	4-Br-Ph	4-Cl-3-Py	37.684	19.581	0.72914	48.34
22*	4- (4'-Br-Ph) -Ph	4-Cl-3-Py	40.747	20.479	0.72468	52.39
23*	3,5-2Cl-Ph	4-Cl-3-Py	42.105	20.016	0.73038	45.10

Marked with * compound is randomly selected as a test set in a neural network.

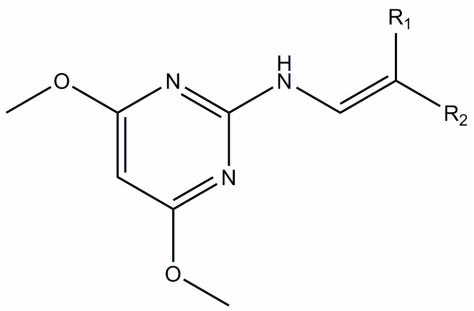

Fig. 1. Basic structure of arylpyridone oxime derivatives.

to characterize the molecular structure in different classes. In general, the topological index in the past cannot fully reflect the topological, geometric and electrical characteristics of molecules. Fully representing the topological index greatly increases the amount of calculation. In order to solve this problem, Liu Shu-Shen [9] proposed an electrical distance vector (M_i). The method can overcome the shortcomings of the general topological index described above. The specific electrical distance vector is calculated as follows:

Step1: Calculating the inherent properties, I_i, of non-hydrogen atoms,

$$I_i = \left(\frac{v}{4}\right)^{0.5} \left[\left(\frac{2}{n}\right)^2 d^n + 1\right] \Big/ d$$

where v is the number of valence electrons, I_i is the eigenvalue of the electrical state of the ith non-hydrogen atom (that is the intrinsic property of the atom), n is the number of main quantum.

Step2: Finding the solution of δ^v and δ:

$$\delta = \sigma - h$$
$$\delta^V = \sigma + \pi - h$$

where σ represents the number of electrons in which the atom participates in the σ orbital bond, π is the number of electrons in which the orbital of the atom participates, h is the number of hydrogen atoms connected to the ith atom, δ^v and δ are the atomic point valences of a the non-hydrogen atoms.

Since I_i ignores the influence of other atoms in the molecule, the perturbation effect of other atoms (ΔI_i) is considered:

$$\Delta I_i = \sum (I_i - I_j) \big/ (d_{ij})^2$$

where I_i, I_j are the intrinsic properties of non-hydrogen atoms i and j, d_{ij} is the number of atoms on the shortest path of atom i and atom j. I_i is added to ΔI_i to get the relative electrical properties, E_i, of the atom:

Step3: Calculating relative electrical properties E_i:

$$E_i = I_i + \Delta I_i$$

Step4: Calculating molecular electrical distance vector (M_i):

$$M_i = M_{mk} = \sum \left(E_i E_j\right) \big/ d_{ij}^{\ 2}$$

where the value of i is 1-91, and M_i is the molecular electrical distance vector.

For example, the number of valence electrons of -O- is 6, the number of non-hydrogen atoms connected to the non-hydrogen atom is 2, and the electric distance vector is 10. The complete methods of calculations and examples can be found in the literature [10-12].In this study, Matlab software was used to calculate the descriptors; see the results in Table 1.

Marked with * compound is randomly selected as a test set in a neural network.

MULTIPLE STEPWISE REGRESSION

Hypothesis y is the inhibition rate of the compound against Sclerotinia sclerotiorum, and its magnitude is

affected by m-1 non-random factors x_1, x_2, x_3, …x_{m-1} and the random factors. To establish a multiple linear regression equation, we have the following linear relationship between y and x_1, x_2, x_3, …x_{m-1}:

$$Y = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_m x_{i,m-1}, i = 1, 2, 3, ...n$$

For the established model, it must have excellent internal robustness and predictive ability. The internal robustness is usually the cross-validation coefficient R_{cv}^2; $R_{cv}^2 \geq 0.5$ indicates that the robustness of the model is good. At the same time, the variable expansion factor V_{IF} is used to judge the multiple correlations of the respective variables in the model. The V_{IF} definition is:

$$V_{IF} = \frac{1}{(1-\beta^2)}$$

where β^2 represents the determination coefficient between an independent variable and other variables. The correlation between the variables can be explained by the value of V_{IF}. The value of V_{IF} = 1 means that the independent variables are not related to each other. When V_{IF} is greater than 1 and less than 5, the correlation between each independent variable is small. When the value of V_{IF} is greater than 5, there is a significant multicollinearity of each independent variable. The A_{IC} and K_T functions are used to evaluate whether the quality of the model is accordance to the requirements. The definition formula is:

$$A_{IC} = R_{SS} \cdot \frac{(f+b)}{(f-b)^2}$$

$$K_T = R^2 \cdot \frac{(f-b-1)}{(f+b) \times (1-R^2)}$$

where R_{SS} is the variance sum, f is the number of compounds, and b is the number of variables. Theoretically, the A_{IC} value and the K_T value determine whether the model is stable or not. It is the key to the predictive model. If the A_{IC} value is smaller and the K_T value is larger, the more stable the model and the better the predictability will be.

In this paper, 91 electrical distance vectors are used as the independent variables and imported by the SPSS software, the inhibition rate is used as the dependent variable, and the multivariate stepwise regression is performed to reach the purpose of variable extraction. The stepwise regression results are shown in Table 2.

As can be seen from the Table 2, R^2 is the correlation coefficient, the variables of the fifth regression model are M_{15}, M_{18}, M_{60}, the value of $R^2 > 0.8$, the maximum value of R_{cv}^2 and F_{IT} are in the fifth model, and the A_{IC} is decreasing continuously, so we can get a three-element regression model:

$$IC_{70} = -71.773 - 1.648 M_{15} - 3.828 M_{18} + 367.988 M_{60}$$

$$N = 23,\ R^2 = 0.887,\ K_T = 39.331,\ F_{IT} = 5.736$$

These parameters can better reflect the relationship between the topological index and the inhibition rate, and the model has a good precision.

SVM AND ELMAN-ANN ALGORITHM PREDICTION

LS-SVM Model

We use the least squares support vector machine (LS-SVM) proposed by Suykens in order to solve the problem of classification and regression. This method uses a least square linear system as the loss function, instead of the traditional support vector machine using the quadratic programming method. The advantage of the LS-SVM method is that it can solve large-scale problems and simplify the operation, which can improve the learning speed and reduce the computational cost.

A kernel σ function is used to implement inner product operations on the transformed space; the decision function is:

$$f(x) = w \times \varphi(x) + b^* = \sum_{i=1}^{n} (a^* - a) y_i \sigma + b^*$$

where a is the Lagrangian multiplier, and σ is the kernel function. This paper uses the SVM toolbox in Matlab to establish model, normalizing the topology index firstly. The radial basis function (RBF) is selected as the kernel

Table 2. Compound Inhibition Rate Regression Results

Variables	R^2	V_{IF}	$R_{CV}{}^2$	A_{IC}	F_{IT}	K_T
M_{15}	0.479	1.000	0.279	363.9	0.804	15.623
M_{15}, M_{63}	0.712	1.084,1.084	0.465	308.6	1.977	19.758
M_{15}, M_{63}, M_{18}	0.821	1.206,1.090,1.143	0.520	317.4	3.351	22.916
M_{15}, M_{63}, M_{18}, M_{60}	0.888	2.387,6.299,5.852,10.698	0.641	273.6	5.286	27.714
M_{15}, M_{18}, M_{60}	0.887	1.182,2.087,1.851	0.673	216.1	5.736	39.331

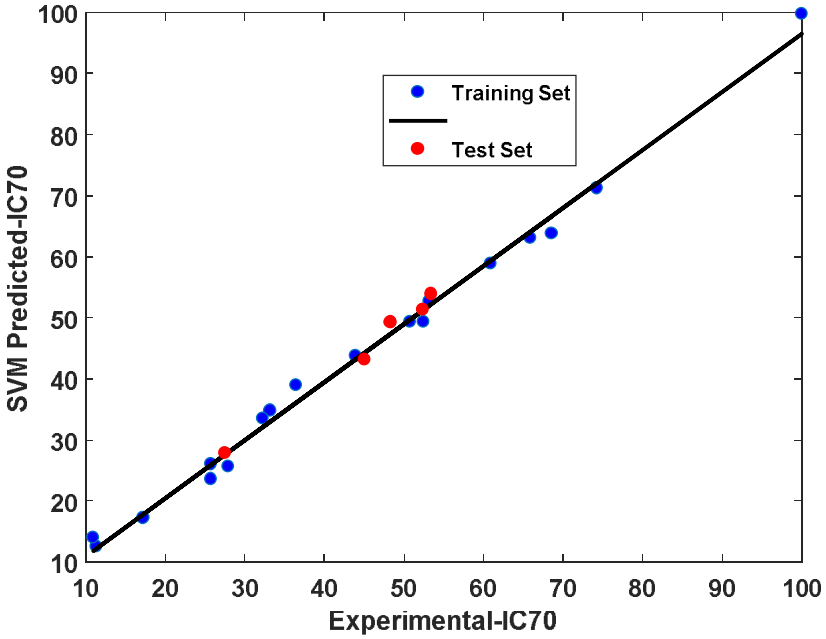

Fig. 2. Comparison of LS-SVM predictions and experimental values.

function, and the grid fast leave one method is used to search the optimal parameters of the model and get:

Optimal regularization parameter: $\gamma = 125.007128$

Optimal nuclear parameter: $\sigma^2 = 29.94732$

The three optimal variables obtained by multivariate stepwise regression are selected as the input set. LS-SVM is used here to predict training set and test set. The comparison between the predicted value and the experimental value is shown in Fig. 2. The numerical results are shown in Table 3.

The training set R^2 is 0.993, its RMSE is 1.374, the test set R^2 is 0.986, and its RMSE is 0.734. It can be seen that the LS-SVM algorithm has less prediction error and the accuracy is close to the experimental value, indicating that it has a good performance in prediction.

Neural Network Elman Algorithm

Unlike the general neural network, the Elman neural network adds a receiving layer based on the input layer,

Table 3. Comparison of the Normalized Values and Errors

No.	M'_{15}	M'_{18}	M'_{60}	IC'_{70}	LS-SVM			Elman-ANN		
					exp.	cal.	err.	exp.	cal.	err.
1	0.402	0.151	0.043	0.466	52.450	49.281	3.169	52.450	54.003	-1.553
2	0.716	0.137	0.000	0.004	11.340	12.482	-1.142	11.340	10.154	1.186
3	0.000	0.000	0.346	1.000	100.000	99.704	0.296	100.000	98.461	1.539
4	0.528	0.972	0.841	0.250	33.220	34.781	-1.561	33.220	32.521	0.699
5	0.442	0.980	0.642	0.166	25.760	25.997	-0.237	25.760	25.864	-0.104
6	0.512	1.000	0.670	0.000	10.970	14.001	-3.031	10.970	11.124	-0.154
7	0.521	0.967	0.671	0.191	27.950	25.627	2.323	27.950	25.962	1.988
8	0.400	0.936	0.707	0.166	25.760	23.564	2.196	25.760	23.068	2.692
9	0.531	0.990	0.666	0.371	43.960	43.708	0.252	43.960	44.076	-0.116
10	0.471	0.708	0.980	0.711	74.280	71.180	3.100	74.280	76.529	-2.249
11	0.516	0.719	0.982	0.561	60.940	58.827	2.113	60.940	60.967	-0.027
12	0.382	0.699	0.955	0.647	68.600	63.804	4.796	68.600	70.400	-1.800
13	0.494	0.683	1.000	0.617	65.890	63.012	2.878	65.890	65.135	0.755
14	0.485	0.777	0.759	0.240	32.310	33.449	-1.139	32.310	29.130	3.180
15	0.563	0.761	0.788	0.474	53.200	52.693	0.507	53.200	48.497	4.703
16	0.575	0.787	0.783	0.447	50.770	49.349	1.421	50.770	51.137	-0.367
17	1.000	0.820	0.717	0.071	17.280	17.189	0.091	17.280	16.705	0.575
18	0.477	0.725	0.652	0.286	36.460	38.960	-2.500	36.460	38.645	-2.185
19*	0.620	0.797	0.783	0.186	27.550	27.842	-0.292	27.550	29.579	-2.029
20*	0.453	0.781	0.923	0.477	53.420	53.867	-0.447	53.420	51.590	1.830
21*	0.451	0.743	0.972	0.420	48.340	49.252	-0.912	48.340	48.114	0.226
22*	0.542	0.791	0.947	0.465	52.390	51.279	1.111	52.390	52.534	-0.144
23*	0.583	0.766	0.979	0.383	45.100	43.092	2.008	45.100	47.165	-2.065

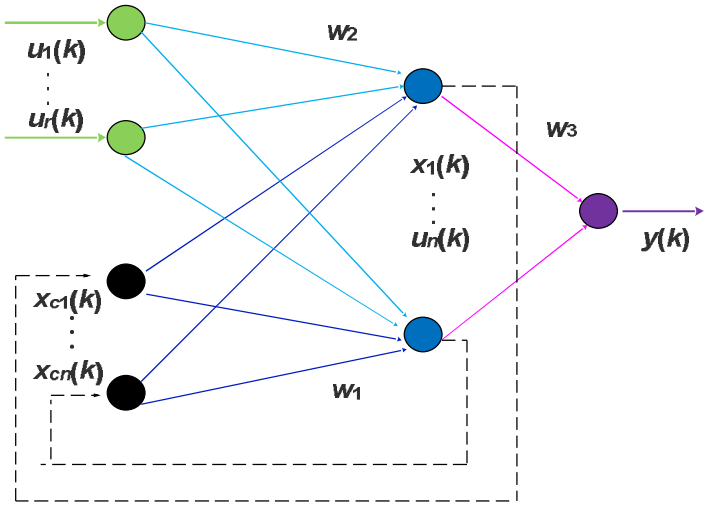

Fig. 3. Elman network structure.

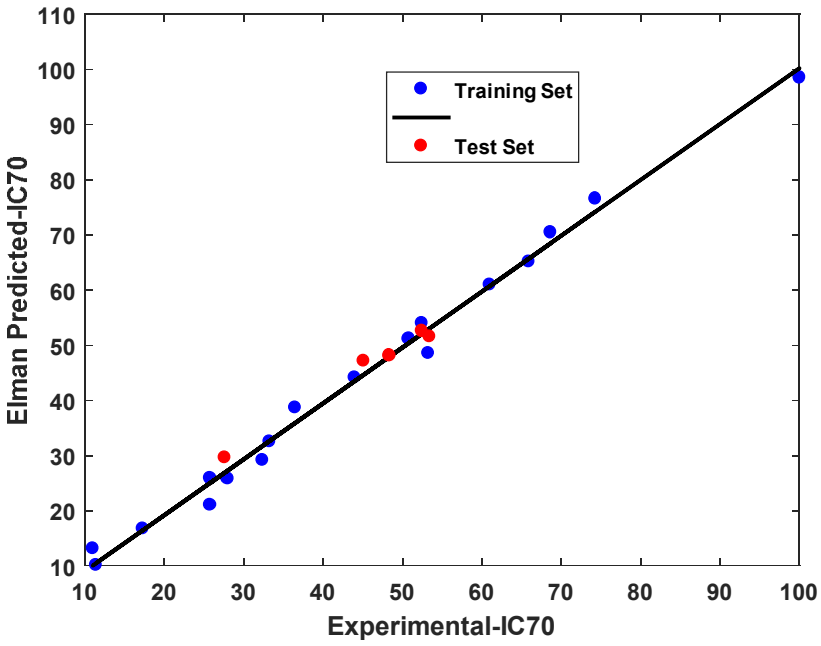

Fig. 4. Comparison of the predicted and experimental values of Elman-ANN.

hidden layer and the output layer. The receiving layer is used to feedback the hidden layer and used as the input layer in the next stage. It can store data dynamically and has a good characterization effect on historical data, especially suitable for nonlinear prediction. The network structure is shown in Fig. 3:

The mathematical model of the network is:

$$y(k) = g[w_3 x(k)]$$

$$x_c(k) = x(k-1)$$

$$x(k) = f[w_1 x_c(t) + w_2(u(k-1))]$$

where k is the number of iterations, u, x, x_c, y are the

r-dimensional input vector of the neural network, the *n*-dimensional output vector of the hidden layer, the output vector of the receiving layer and the output vector of the neural network, respectively. w_1 ,w_2 ,w_3 represent the matrix of connection weight values from the receiving layer to the hidden layer, the input layer to the hidden layer and the hidden layer to the output layer, respectively. $f(\cdot)$ is the activation function of the hidden layer neurons, $g(\cdot)$ is the activation function of the output layer neurons.

In order to avoid the phenomenon of saturation of neurons, this paper normalizes the three electrical distance vectors that are screened out. The calculation formula is:

$$M_i' = \frac{M_i - M_{\min}}{M_{\max} - M_{\min}}$$

where M_i is the *i*th original electrical distance vector, $M_{\min}$ is the minimum value of the original electrical distance vector, $M_{\max}$ is the maximum value of the original electrical distance vector, and M_i' represents the *i*th normalized value. In this paper, Matlab is used to normalize the electrical distance vector. The normalized values can be seen in Table 3.

For the number of neurons in the hidden layer, this paper uses Andrea's results, the number of neurons H in the hidden layer satisfies the relationship:

$$N = (I+1) \times H + (H+1) \times Q$$

where N is the sum of the weights in the neural network, I is the number of neurons in the input layer, H is the number of neurons in the hidden layer, Q is the number of neurons in the output layer, I is 3 and Q is 1. For N, there is $M/N > 1$, M is the number of samples, and the number of samples in this paper is 23, so only $N < 23$ can meet the requirements. In order to simplify the calculation, N of this paper takes 21 and $H = 4$. Then, Elman neural network was established as 3-4-1 neural network. In order to improve the computing speed of the computer, the hidden layer uses 'tansig' as the activation function, the output layer uses 'tansig' as the activation function.

The three variables obtained by stepwise regression are used as input sets. We use Elman neural network to train and the test set was used to verify. The comparison between the predicted values and the experimental values is shown in Fig. 4. The numerical results are shown in Table 3.

The training set R^2 is 0.994, its RMSE is 1.033, the test set R^2 is 0.985, and the RMSE is 0.535. Compared with LS-SVM, their prediction performance is similar, and the prediction ability is about the same.

DESCRIPTION OF THE QSAR RESULTS

(1) According to the results obtained by the stepwise regression equation, the magnitude of the influence inhibition rate is three electrical distance vectors: M_{15}, M_{18}, M_{60}. The main factors affecting the inhibition rate of compounds are electrical and topological environments. According to the Hall's research [13], the two-dimensional structure affecting the molecule is =CH- and -CH<.

(2) The multivariate stepwise regression equation used in this paper is far less fitting than SVM and Elman. In essence, multivariate stepwise regression is only a linear fitting between variables. Based on the results, the relationship between them tend to be more nonlinear, so the error of linear fitting is more significant.

(3) The prediction ability of SVM is close to Elman algorithm, while the stability of Elman algorithm is better.

REFERENCES

[1] Wang, J. Y.; Tao, H.; Jin, M., Design, synthesis and sterilization activity of aryl pyridine hydrazone compounds. *Chin. J. Org. Chem.* **2019**, *39*, 1044-1052, DOI: 10.6023/cjoc201810019.

[2] Liu, S. S.; Yin, C. S.; Wang, L. S., Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749-756, DOI: 10.1021/ci010245a.

[3] Feng, C. J.; Yang, W. H., Linear QSAR regression models for the prediction of bioconcentration factors of chloroanilines in fish by densityfunctional theory. *Chin. J. Struct. Chem.* **2014**, *33*, 830-834, DOI: 10.14102/j.cnki.0254-5861.2014.06.003

[4] Feng, C. J., Theoretical studies on quantitative structure-activity relationship and structural modification for 3-substituted sulfur-5-(2-hydroxy-

phenyl)-4H-1,2,4-triazole compounds. *Acta Chim. Sin*. **2012**, *70*, 512-518, DOI: 10.6023/A1107052

[5] Feng, C. J.; Yang, W. H.; Mu, L.L., Estimation and prediction of bioconcentration factors of nonionic organic chemicals in fish by electrotopological state indices and structural parameter. *Chin. J. Struct. Chem*. **2008**, *5*, 575-587, DOI: 10.14102/j.cnki.0254-5861.2008.05.013

[6] Feng, C. J.; Mu, L. L.; Yang, W. H.; Cai, K. Y., research on the bioconcentration factors of organic pollutants with topological indices and artificial neural network. *Acta Chim. Sin*. **2008**, *19*, 2093-2098, DOI: 10.6023/A1107052

[7] Wang, C.; Feng, C. J., QSAR studies on the Inhibitory activity of levofloxacin-thiadiazole HDACi conjugates to histone deacetylases. *Chin. J. Struct. Chem*. **2018**, *37*, 1679-1688, DOI: 10.14102/ j.cnki.0254-5861.2011-1827.

[8] Wang, C.; Feng, C. J., QSAR study of the action strength of DOM of phenyl-isopropyl-amine dopes using MLR and BP-ANN. *Chin. J. Struct. Chem*. **2017**, 10, 1720-1728, DOI: 10.14102/j.cnki.0254-5861.2011-1610.

[9] Liu, S. S.; Liu, Y.; Li, Z. L.; Cai, S. X., A novel molecular electronegativity-distance vector (MEDV). *Acta Chim. Sin*. **2000**, *11*, 1353-1357, DOI: 10.6023/A19060197.

[10] Zheng, Q. F.; Ju, Z.; Liu, S. S., Combined toxicity of dichlorvos and its metabolites to vibrio qinghaiensis sp.-Q67 and caenorhabditis elegans. *Acta Chim. Sin*. **2019**, *10*, 1008-1016, DOI: 10.6023/A19060197

[11] Liu, S. S.; Zhang, J.; Zhang, Y. H.; Qin, L. T., APTox: Assessment and prediction on toxicity of chemical mixtures. *Acta Chim. Sin*. **2012**, *14*, 1511-1517, DOI: 10.6023/A12050175

[12] Wang, M. C.; Liu, S. S.; Chen, F., Predicting the time-dependent Toxicities of Three triazine herbicide mixtures to V. qinghaiensis sp. Q67 using the extended concentration addition model. *Acta Chim. Sin*. **2014**, *1*, 56-60, DOI: 10.6023/A13101034

[13] Hall, L. H.; Mohney, B.; Kier, L. B., The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comp. Sci.* **1991**, 31, 76-82, DOI: 10.1021/ ci00001a012