

# Prediction of Activities of Halogenated 2,4-Diphenyl Indeno [1,2b]pyridinol Derivatives Using QSAR Model Against Breast Cancer

Kaixuan Wang<sup>1</sup>, Hongzong Si<sup>2\*</sup>

<sup>1</sup>College of Computer Science & Technology of Qingdao University, Qingdao, 266071, China

<sup>2</sup>Laboratory of New Fibrous Materials and Modern Textile State Key Laboratory, Qingdao University, Qingdao, 266071, China

**Abstract:** Breast cancer is one of the most common cancers among women. It is the first killer of women, and the existing drugs are not enough to treat the disease. In this study, a quantitative structure-activity relationship model used for predicting the IC<sub>50</sub> value of those compounds was built by the chemical structures of a class of halogenated 2,4-diphenyl indeno[1,2b] pyridinol derivatives. All the 28 compounds were randomly split into a training set with 21 compounds as well as a test set with 7 compounds. The heuristic method in CODESSA program was utilized to optimize 28 compounds and establish linear models. Furthermore, four descriptors were selected and used to build a nonlinear model by gene expression programming method. The correlation coefficients  $R^2$ ,  $R^2_{cv}$ ,  $S^2$  in the heuristic method were 0.671, 0.520 and 0.079, while in gene expression programming, the  $R^2$  and  $S^2$  were 0.728, 0.055 in the training set and 0.688, 0.062 in the test set, respectively. Both the two methods had good prediction performance. In comparison, the gene expression programming method was more consistent with the experimental values. The nonlinear model was supposed to the design and development of targeted anti-breast cancer drugs.

**Keywords:** Halogenated 2,4-diphenyl indeno[1,2b]pyridinol derivatives; Quantitative structure-activity relationship; Heuristic method; Gene expression programming

Received 15 Jan 2021, Revised 20 May 2021, Accepted 22 June 2021

\* Corresponding Author: Hongzong Si

## 1. Introduction

Breast cancer is one of the most common cancers in the world. In 2018, there were 2.1 million new breast cancer patients worldwide, with an incidence rate of 11.6% [1]. And in recent years, the incidence of breast cancer in China is also increasing year by year. It ranks first among female malignant tumors [2]. Since this disease is malignant, the “cure rate” is not mentioned and the “survival rate” is discussed. The survival rate of patients with early breast cancer after treatment is generally high. Once the cancer cells metastasize, the survival rate of patients will be greatly reduced [3].

The therapeutic method and prognostic effect of breast cancer are affected by factors such as geographical difference, breast cancer type, staging and detection method. Previously, the disease was mainly treated with surgery or chemoradiotherapy in clinic, but the therapeutic effect was unsatisfactory [4]. Among the treatment, targeted therapy — targeted drugs are used to directly act on tumor cells that have been identified. Due to the characteristics of precision, low toxicity and high efficiency, targeted therapy has been widely used. However, targeted drugs still have some defects, especially resistance caused by molecular target mutations, long development cycle and high drug-developing costs [5]. Therefore, the task of developing more target drugs for the disease is still arduous.

It is an unrealistic job to conduct experimental evaluation of millions of drug-like compounds aiming

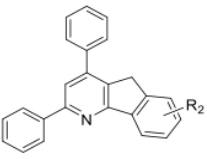
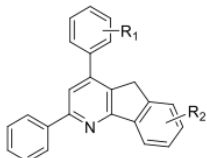
at thousands of targets by means of traditional drug-developing methods in a short period of time. By contrast, computer simulation through quantitative structure-activity relationship (QSAR) model is a feasible alternative for these tricky experimental screenings, which is environmental-friendly, cost-saving and time-saving [6]. By using the approach QSAR, we can choose an appropriate number of informative descriptors to describe compounds in order to build prediction models based on the heuristic method (HM) and gene expression programming (GEP). In 2013, Lee, Jana had successfully evaluated and predicted substrate properties and their affinity to breast cancer resistant protein on the basis of molecular structure description [7].

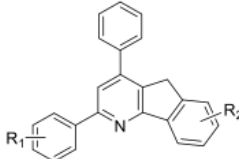
It has been reported that topoisomerase II $\alpha$ -targeted anticancer agents have a significant anti-breast cancer effect. Many papers had predicted anticancer drugs targeting DNA topoisomerase II $\alpha$ . These studies require extensive work to screen for new compounds. Some researchers have also applied the QSAR method to their studies on breast cancer drugs, but their models, descriptors, and molecular structure modification methods were completely different from ours. Researchers synthesized a series of hydroxy- and halogenated 2,4-diphenyl indeno[1,2-b]pyridinols and investigated the activities. This class of compounds was effective for resisting tumors [8]. The gene expression programming method can create an effective QSAR

1239

model to predict IC50 values.

**Table 1** Experimental and calculated log(IC50) of 28 compounds (HM and GEP)

Compound				Exp. log(IC50)	HM		GEP	
Structure	No.	R1	R2		Predict	Residue	Predict	Residue
	1	-	6'-OH	0.433	0.411	-0.022	0.590	0.157
	2*	-	7'-OH	0.582	0.384	-0.198	0.407	0.175
	3	-	8'-OH	0.543	0.472	-0.071	0.510	0.033
	4	-	9'-OH	0.146	0.330	0.184	0.338	0.192
	5	2-OH	6'-OH	-0.097	-0.419	-0.322	-0.465	0.368
	6	2-OH	7'-OH	0.000	-0.487	-0.487	-0.298	0.298
	7	2-OH	8'-OH	-0.824	-0.430	0.394	-0.314	0.510
	8*	2-OH	9'-OH	-0.770	-0.412	0.358	-0.292	0.477
	9	3-OH	6'-OH	-0.721	-0.395	0.326	-0.637	0.085
	10*	3-OH	7'-OH	-0.222	-0.162	0.059	-0.387	0.165
	11	3-OH	8'-OH	0.196	0.095	-0.101	0.029	0.166
	12	3-OH	9'-OH	-0.143	0.027	0.170	-0.179	0.037
	13	4-OH	6'-OH	0.107	-0.150	-0.258	0.197	0.090
	14	4-OH	7'-OH	0.395	0.332	-0.063	0.160	0.235
	15*	4-OH	8'-OH	0.140	0.562	0.422	0.510	0.370
	16	4-OH	9'-OH	0.616	0.338	-0.278	0.532	0.084

	17	2-OH	6'-OH	0.396	0.452	0.056	0.436	0.040
	18	2-OH	7'-OH	0.210	0.415	0.205	0.415	0.205
	19*	2-OH	8'-OH	0.272	0.391	0.119	0.368	0.096
	20	2-OH	9'-OH	0.983	0.391	-0.617	0.419	0.564
	21	3-OH	6'-OH	-0.284	-0.250	0.035	0.012	0.296
	22	3-OH	7'-OH	-0.569	-0.197	0.371	-0.500	0.069
	23	3-OH	8'-OH	-0.377	-0.255	0.122	-0.404	0.028
	24	3-OH	9'-OH	-0.367	-0.552	-0.185	-0.306	0.061
	25*	4-OH	6'-OH	-0.432	-0.470	-0.038	-0.409	0.022
	26	4-OH	7'-OH	-0.149	-0.231	-0.082	-0.355	0.206
27	4-OH	8'-OH	-0.244	-0.245	-0.001	-0.352	0.107	
28*	4-OH	9'-OH	-0.181	-0.280	-0.099	-0.254	0.073	

\*The compounds of test set

## 2. Materials and Methods

### 2.1. Data preparation

28 compounds and the IC50 values were provided by the literature [8]. We further took logarithm of these values we extracted. All the compounds' structures, their log (IC50) values and the predicted values are collected in Table 1. Moreover, all the data were randomly separated into two sets, a training set of 21 structures to construct the model and a test set of 7 structures to validate the prediction capability of the model.

### 2.2. Computation of the descriptors

The first step was to plot the 28 compounds in the program, ChemDraw. Then, these structures were imported into HyperChem software to do geometry optimization under the MM+ molecular mechanical force field and semi-empirical AM1 approaches [9]. After the step, we employed MOPAC6.0 program to get the .arc, .end and .mno files of each structure [10].

Ultimately, the exported file was submitted to the software CODESSA [11]. Some molecular descriptors were gained classified into 3 categories through computing, including electrostatic, topological and quantum mechanical classes. Descriptors, defined as numerical characteristics related to chemical structures, can reflect different characters of the molecular structure and would be used to set up an equation which was helpful to do the subsequent prediction [12].

### 2.3. Create the linear model by heuristic method

In order to select the descriptor that best reflects the correlation between the structure and biological activity of compounds, the optimal multiple linear regression method was used in HM method to screen the descriptors calculated by CODESSA. And the accuracy was guaranteed and tested by regression coefficient ( $R^2$ ), cross-validation regression coefficients ( $R_{cv}^2$ ) and standard deviation ( $S^2$ ). The details of the 4 descriptors are listed in Table 2.

#### 2.4. Create the nonlinear model by gene expression programming

GEP, a novel genetic algorithm, which combines the advantages of genetic programming and genetic algorithm, can solve complex problems with simple

coding [13]. Based on the gene expression law of biological inheritance, the GEP adopts the equal-length linear symbol as the genetic code, and the individual phenotype is an expression tree. After a good deal of operations, the algorithm can find an optimal solution.

**Table 2 Selected descriptors and statistical parameters**

Descriptor	Physical-chemical Meaning	Coefficient	T-test
Constant	-	-4.231E+02	-1.1893
TPCCMD	Tot point-charge comp. of the molecular dipole	7.507E-01	-5.451
RPCS	RPCS Relative positive charged SA (SAMPOS*RPCG) [Zefirov's PC]	4.901E-01	4.266
MBONA	Max bond order of a N atom	-1.206E+02	-2.664
MNRCNB	Max n-n repulsion for a C-N bond	3.451E+00	1.477

### Results

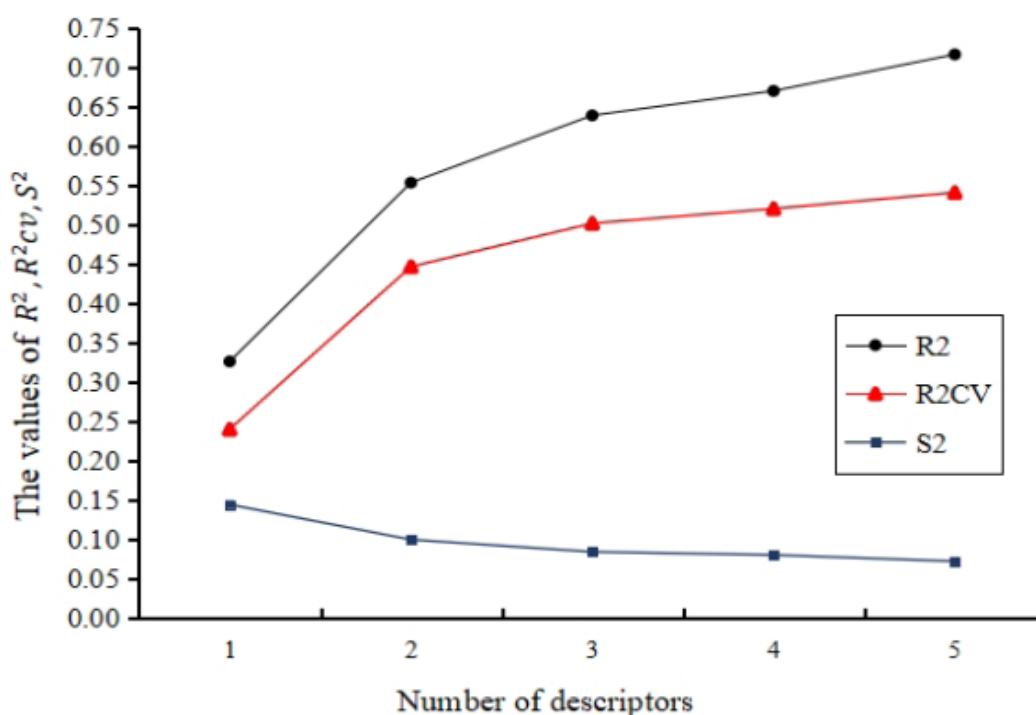
#### 3.1. Results of HM

In order to find the optimal number of descriptors describing the activity  $\log(\text{IC}_{50})$  value of compounds, we used heuristic method to deal with these structural parameters, and found the optimal equation with the number of descriptors from 1 to 5. When another descriptor was added, there was no obvious change in the statistical results, which indicated that the appropriate number of descriptors was reached. By calculating, we finally obtained 420 descriptors in the project CODESSA. In Figure 1, we could clearly get the information that the values of  $R^2$  and  $R^2_{cv}$  gradually

ascended as the number of descriptors increased, while the values of  $S^2$  decreased in the meanwhile. QSAR studies generally require a maximum of 1/5 of the sample size for the number of molecular descriptors [14]. Four descriptors with higher correlation were selected by HM method. We tested the correlation of every two descriptors, and as shown in the Table 3, the coefficient between any two descriptors is lower than 0.8, which ensured that these descriptors were independent of each other. And the linear model built by the 4 descriptors is demonstrated as follows (Figure 2):

$$\text{Log}(\text{IC}_{50}) = -423.09 - 0.751\text{TPCCMD} + 0.490\text{RPCS} - 120.64\text{MBONA} + 3.451\text{MNRCNB}$$

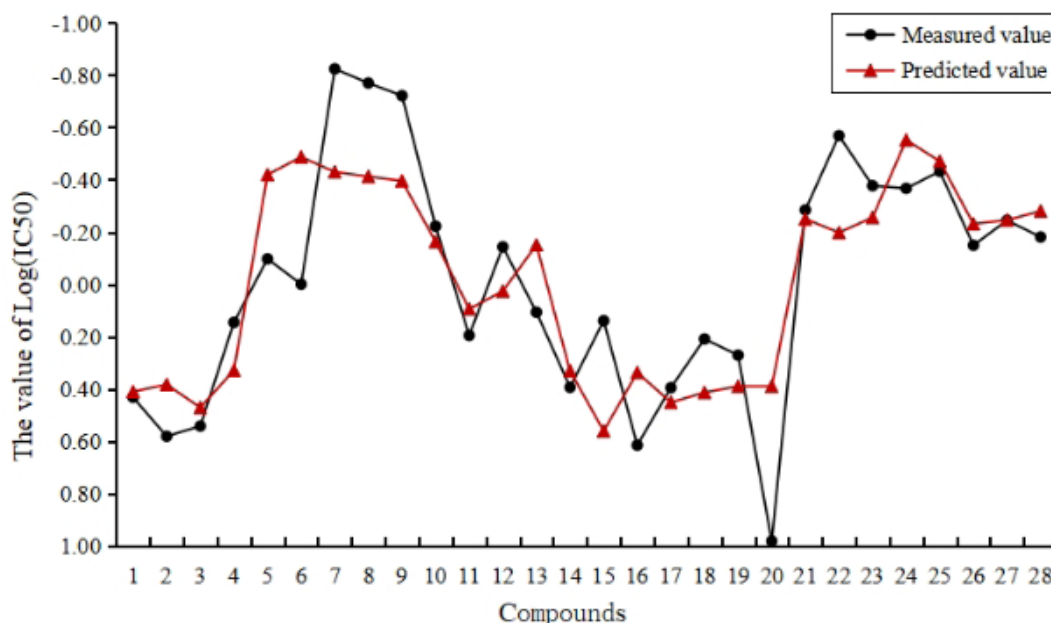
The correlation coefficients  $R^2$ ,  $R^2_{cv}$ ,  $S^2$  in the heuristic method were 0.671, 0.520 and 0.079.



**Figure 1. Influence of the number of descriptors on the  $R^2$ ,  $R^2_{cv}$  and  $S^2$ .**

**Table 3 Correlation matrix of the descriptors in the model**

	TPCCMD	RPCS	MBONA	MNRCNB
TPCCMD	1.000	0.341	0.151	0.364
RPCS		1.000	0.061	-0.100
MBONA			1.000	0.757
MNRCNB				1.000



**Figure 2. Plot of experimental and predicted log(IC50) values by heuristic method**

**3.2. Results of GEP**

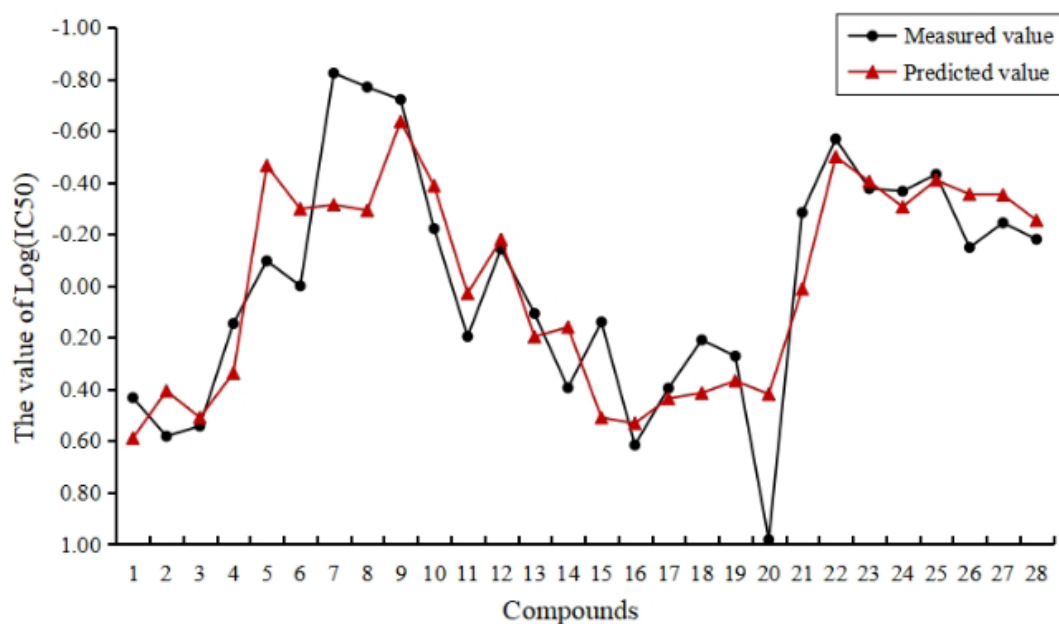
Training set and test set were put into the program automatic problem solver (APS) with four selected descriptors used as parameters [14]. The program modeled this function by training data, and after rounds of evolution, we finally established the nonlinear model of log(IC50) (Figure 3). Table 4 displays the relative coefficients. GEP algorithm calculates the expression

tree of four parameters. We transform the expression tree into a nonlinear mathematical equation, presents as:

$$\begin{aligned} \log(\text{IC}_{50}) = & 1 / (1 + \exp(-(\text{floor}((\text{RPCS} - \text{TPCCMD}) * \text{RPCS}) \\ & - (\text{floor}(\text{TPCCMD}) * \text{RPCS} * \text{TPCCMD})))) \\ & + 1 / (1 + \exp(-(\text{floor}(\text{RPCS} + \text{MNRCNB}) / \text{MNRCNB} + \\ & \text{TPCCMD} * \text{TPCCMD}))) \\ & + (\text{TPCCMD} / 2 \text{MNRCNB}) / ((\text{RPCS} + \text{MBONA}) \\ & \text{D} * (\text{MBONA} / \text{MNRCNB})) \end{aligned}$$

**Table 4. Parameters for the simple symbolic regression problem**

Parameter names	Values	Parameter names	Values
Number of chromosomes	100	Precision	0.01
Head size	8	1-Point recombination rate	0.3
Number of genes	5	2-Point recombination rate	0.3
Linking function	+	Gene recombination rate	0.1
Mutation rate	0.044	Gene transposition rate	0.1
Fitness function	MSE	IS transposition rate	0.1
Selection range	100	RIS transposition rate	0.1



**Figure 3. Predicted log(IC50) values versus experimental values by the gene expression programming.**

#### 4. Discussion

To evaluate the QSAR model, we commonly use the values of  $R^2$  and  $S^2$  [15].  $R^2$ ,  $S^2$  in training set are 0.728, 0.055, in test set are 0.688, 0.063 by GEP, respectively. At the same time,  $R^2$ ,  $S^2$  in linear model built by HM method are 0.6705 and 0.0793. The correlation coefficients of the training set and the test set of GEP were higher than that of HM, and the average errors the training set and the test set of GEP were lower than that of HM. What's more, comparing the curves of  $\log(\text{IC}_{50})$  fitted by the two algorithms (Figure 2, 3), we could clearly get the result that GEP method was better than HM method in fitting some experimental values. On the whole, the fitting ability of the latter is more prominent.

Based on the absolute values of the coefficients in the equation, the magnitude of the effect on the model is described as  $\text{MBONA} > \text{MNRCNB} > \text{TPCCMD} > \text{RPCS}$ . Next is a discussion about the impact of these four descriptors on the activities of all compounds.

$\text{IC}_{50}$  is that semi-inhibition rate, and represent the concentration of the drug when the inhibition rate is 50% [16]. Therefore, lower  $\text{IC}_{50}$  values imply that the derivatives have the high activity. The coefficient of MBONA is negative, others possess positive coefficients in the linear equation, thus MBONA has a positive impact for the activities of these compounds, while TPCCMD, RPCS and RPCS have negative effects on  $\text{IC}_{50}$ .

Bond order is the number of chemical bonds between a pair of atoms, indicating the stability of the bond. Bond order and bond length reflect the type and strength

of covalent bonds between atoms [17,18]. Bond order is inversely proportional to length: when the bond order increases, the bond length decreases. MBONA indicated the bond length of N atom would improve the activity of compounds in the experiment. MNRCNB is a descriptor related to the energy of quantum mechanics. The process driven by nuclear repulsion in the molecule is described by this energy, which may be related to the conformational (rotation, inversion) changes of the molecule or the atomic reactivity [19,20]. RPCS were defined as the product of the solvent accessible surface area of the largest positive atom and the relative positive charge (RPCG), characterizing charge redistribution [20,21]. The reactivity of molecules is related to the distribution of electrons in compounds, which has an influence on the activity of compounds. TPCCMD is a quantum chemical descriptor.  $\alpha$  polarizability, a molecular polarizability, reflects the volume of molecules and the interaction between drugs and targets. Polarity scale is closely related to hydrophobicity and electrophilicity. Hydrophobicity directly affects the binding of drugs to receptors [22]. Therefore, the descriptor TPCCMD also affects the value of  $\text{IC}_{50}$ .

These 4 descriptors indicate that the physical properties of the drug are the key factors of constructing the QSAR models and their prediction.

#### 5. Conclusion

We have built 2 different QSAR models in this study, and by comparison, GEP method has a better effect for the model. Furthermore, by analyzing the selected descriptor, arrived at a conclusion that increase MBONA

and decrease the value of TPCCMD, MNRCNB and RPCS is beneficial to improving the activity of the medicine. Selection of molecules by this model can help to avoid testing large numbers of compounds and greatly reduces the use of resources and time, also can develop greater potential in future drug design and modify work.

## References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, *A Cancer J for Clin*, 68 (2018) 394-424.
- [2] R.L. Siegel, K.D. Miller, A. Jemal, *Cancer J Clin*, 69 (2019) 7-34.
- [3] Y. Liang, H. Zhang, X. Song, *Semin Cancer Biol*, 60 (2020) 14-27.
- [4] H. Ye, X. Li, B. Qin, Z. Wang, *China Medical Herald*, 05 (2016) 118-121.
- [5] R.Y. Zhao, R.L. Zheng, *Guangdong Chemical Industry*, 14 (2020) 90-92+77.
- [6] O.A. Tarasova, A.F. Urusova, D.A. Filimonov, M.C. Nicklaus, A.V. Zakharov, V.V. Poroikov, *J Chem Inf Model*, 55 (2015) 1388-99.
- [7] Y. Lee, S. Jana, G. Acharya, C.H. Lee, *Chemistry Central Journal* 7, 23 (2013), <https://doi.org/10.1186/1752-153X-7-23>.
- [8] T.M. Kadayat, S. Park, A. Shrestha, H. Jo, S. Hwang, P. Katila, R. Shrestha, M.R. Nepal, K. Noh, S. Kim, W. Koh, K. Kim, Y. Jeon, T. Jeong, Y. Kwon, E. Lee, *Journal of Medicinal Chemistry*, 62 (2019) 8194-8234.
- [9] 4.0, Hypercube, 1994.
- [10] O.P.P Stewart, MOPAC 6.0, QCPE, No. 455, Quantum Chemistry Program Exchange; Indiana University: Bloomington, IN, 1989.
- [11] A.R. Katritzky, V.S. Lobanov, M. Karelson, R. Murugan, M.P. Grendze, J.E. Toomey, *Rev. Roum. Chim.*, 41 (1996) 851-868.
- [12] L.P. Chen, W.H. Chen, N. Shi, H. Yang, W. Xu, *Acta Physico-Chimica Sinica*, 12 (2012) 2790-2796.
- [13] C. Ferreira, New York:Springer-Verlag, 2002.
- [14] W. Ren, D.X. Kong, *Computers and Applied Chemistry*, 11 (2009) 1455-1458.
- [15] D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, *J Chem Inf Comput Sci*, 32 (1992) 306-316.
- [16] OECD (Organization for Economic Co-Operation and Development), OECD Environment Health and Safety Publications Series on Testing and Assessment No.69[R]. Paris, 2007.
- [17] <https://chem.libretexts.org/@go/page/1982>, 2020.
- [18] N.S. Zefirov, M.A. Kirpichenok, F.F. Izmailov, M.I. Trofimov, *Doklady Akademii Nauk SSSR*, 296 (1987) 883-887.
- [19] M.A. Kirpichenok, N.S. Zefirov, *Zhurnal Organicheskoi Khimii*, 23 (1987) 673-703.
- [20] D.T. Stanton, P.C. Jurs, *Anal. Chem.* 62 (1990) 2323.
- [21] D.T. Stanton, L.M. Egolf, P.C. Jurs, *J. Chem. Inf. Comput. Sci.* 32 (1992) 306.
- [22] H. Liu, Y. Hu, *Computational and mathematical methods in medicine vol*, 2014 (2014) 957154.