

QSAR analysis of pyrimidine derivatives as VEGFR-2 receptor inhibitors to inhibit cancer using multiple linear regression and artificial neural network

Fariba Masoomi Sefiddashti, Saeid Asadpour, Hedayat Haddadi*, Shima Ghanavati Nasab

Department of Chemistry, Faculty of Sciences, Shahrekord University, Shahrekord, I.R. Iran.

Abstract

Background and purpose: In this study, the pharmacological activity of 33 compounds of furopyrimidine and thienopyrimidine as vascular endothelial growth factor receptor 2 (VEGFR-2) inhibitors to inhibit cancer was investigated. The most important angiogenesis inducer is VEGF endothelial growth factor, which exerts its activity by binding to two tyrosine kinase receptors called VEGFR-1 and VEGFR-2. Due to the critical role of VEGF in the pathological angiogenesis of this molecule, it is a valuable therapeutic target for anti-angiogenesis therapies.

Experimental approach: After calculating descriptors using SPSS software and stepwise selection method, 5 descriptors were used for modeling in multiple linear regression (MLR) and artificial neural network (ANN). The calibration series and the test series in this study included 26 and 7 combinations, respectively.

Findings/Results: The performance evaluation of models was determined by the R^2 , RMSE, and Q^2 statistic parameters. The R^2 values of MLR and ANN models were 0.889 and 0.998, respectively. Also, the value of RMSE in the ANN model was lower and its Q^2 value was higher than the MLR model.

Conclusion and implications: The results were evaluated by different statistical methods and it was concluded that the nonlinear neural network method is powerful to predict the pharmacological activity of similar compounds, and because of the complex and nonlinear relationships, the MLR was not capable of establishing a good model with high predictive power.

Keywords: Artificial neural network; Cancer; Multiple linear regression; Pyrimidine derivatives; QSAR.

INTRODUCTION

Cancer is one of the leading causes of worldwide mortality characterized by the loss of control of cell proliferation and almost most patients die without treatment (1,2). Angiogenesis is a physiological process in which new veins grow from existing veins and plays a key role in many pathological conditions such as tumor growth, metastasis, and so on. In adults, endothelial cells are silent in adolescence but are able to be activated in response to appropriate factors (3,4).

Angiogenesis plays a vital role in life, growth, and recovery, for example, in wound healing. However, the basis for the transformation of tumors from dormant to

malignant is due to this process. The most important angiogenesis inducer is the vascular endothelial growth factor (VEGF), which exerts its activity by binding to two tyrosine kinase receptors called VEGFR-1 and VEGFR-2 (4). Due to the critical role of VEGF in the pathological angiogenesis of this molecule, it is a valuable therapeutic target for anti-angiogenesis therapies (5).

Access this article online



Website: <http://rps.mui.ac.ir>

DOI: 10.4103/1735-5362.327506

*Corresponding author: H. Haddadi

Tel: +98-9126870183, Fax: +98-3814424419

Email: haddadi@sku.ac.ir

Pyrimidine is an aromatic heterocyclic organic compound similar to pyridine having nitrogen atoms at positions 1 and 3 in the ring (6). Pyrimidine derivatives have a wide range of pharmaceutical applications. There have been reports of pyrimidine derivatives as an antibacterial, analgesic, antiviral, anti-inflammatory, anti-HIV, antituberculosis, anticancer, anti-Parkinson, and antifungal as well as sleep medication in chemical sources. Among the reported medicinal properties of pyrimidines, anticancer activity is the most frequently reported (7,8).

The quantitative structure-activity relationship (QSAR) is a strategy of critical importance for chemistry and pharmacy, based on the idea that when the structural properties of a molecule change, the activity or property of the material changes, accordingly (9-11). QSAR models, mathematical equations related to the chemical structure of their biological activity, provide useful information for drug design and drug chemistry (12-15). These computational screening methods are a good alternative to the costly and laborious screening tests performed in laboratories. Therefore, there is a continuing effort among QSAR specialists to develop more efficient QSAR techniques to develop and discover more reliable approaches for pharmaceutical chemists in practice (16-18).

Following other papers in QSAR from our group members (19-21), the current study attempted to associate the pharmacological activity of some furopyrimidine and thienopyrimidine derivatives as VEGFR-2 inhibitors by using both MLR as an extension of linear regression and ANN as nonlinear methods which use several explanatory variables to predict the outcome of a response variable (22). A comparison of various linear and nonlinear modeling techniques in recent research has shown how different regression methods can affect the predictive power of QSAR models (23,24).

MATERIAL AND METHODS

Data sources

Two series of pyrimidine-based derivatives namely furo [2,3-d] pyrimidine and thieno [2,3-d] pyrimidine series, linked to either

biarylamide or biarylurea *via* an NH or ether linker were seen *in vitro* VEGFR-2 inhibitory activity. All inhibitors of VEGFR-2 and their biological activities (percent inhibition values) were taken from the Aziz's report (22). First, principal component analysis (PCA) was used to classify the molecules into calibration and test sets, then the data set is subdivided into a calibration set of 26 compounds and a test set of 7 compounds after PCA analysis for model evaluation. The chemical structures and the bioactivity values of all compounds are presented in Table 1.

Molecular model

All the 2D and 3D structures were drawn and built by ChemDraw and Chem3D software, respectively. Structures were optimized by MM2 algorithm in Chem3D. The theoretical molecular descriptors are derived from the chemical structure of the compounds. In order to calculate the theoretical descriptors, the molecular structures were constructed using ChemDraw Ultra version 15.0 and Chem3D Ultra version 15.0, then optimized using MM2 algorithm (25,26).

Molecular descriptors

A descriptor is the mathematics of a molecule that contains different sources of chemical information that is converted and encoded to counter chemicals, biological, and pharmaceutical problems. To develop 2D-QSAR models, different physicochemical descriptors are calculated for each of the compounds in the dataset using DRAGON software version 5.5- 2007 (27). Dragon is a program for calculating and producing a variety of molecular descriptors for different compounds and converts the information of molecules including bond energy, bond angle, bond type, molecular mass, electronic properties, and so on, into numeric form and stores them in descriptive format. These descriptors can be used to study and evaluate molecular structure-activity or structure-property relationships as well as to analyze the high-throughput similarity and screening molecule databases. In fact, the dragon is widely used in scientific studies as well as part of several QSAR collections.

Table 1. Structural formulae of compounds and their percent inhibition values

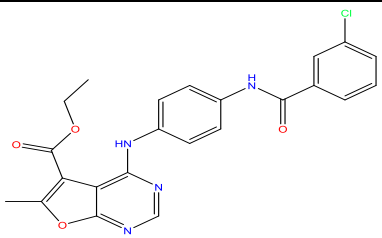
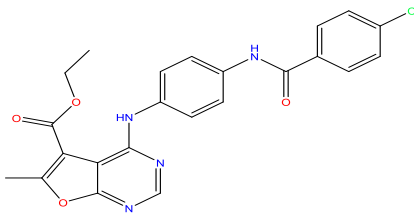
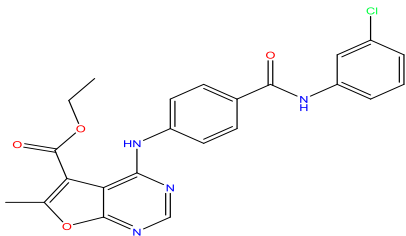
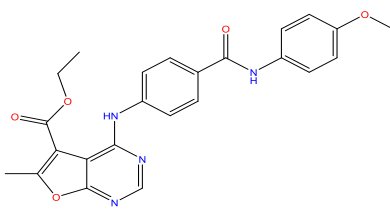
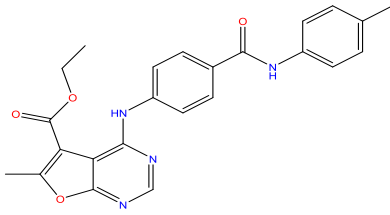
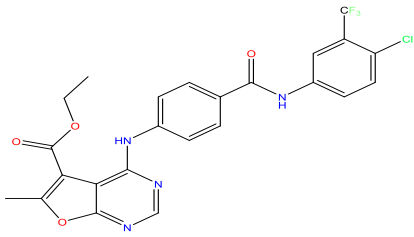
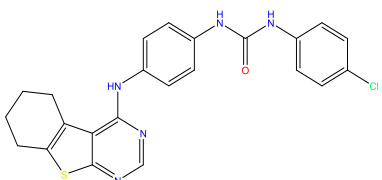
Order	Structure	% of inhibition
1		8
2		15
3		5
4		8
5		14
6		14
7		32

Table 1. (Continued)

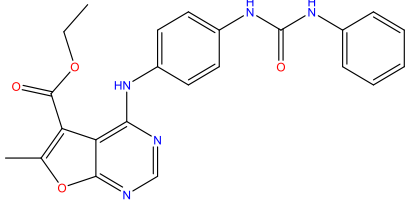
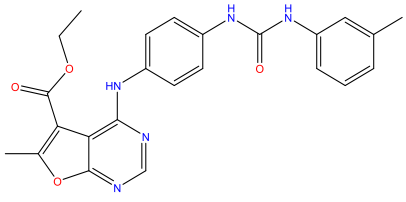
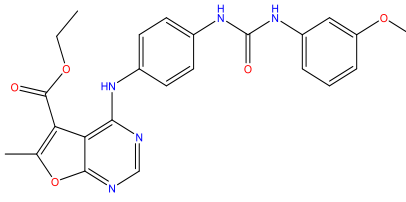
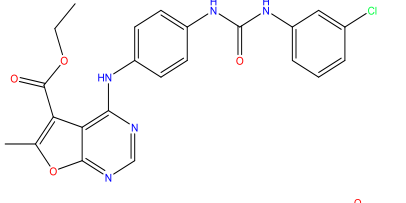
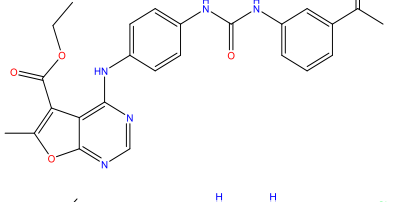
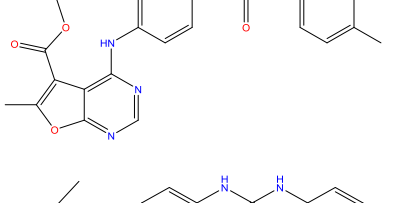
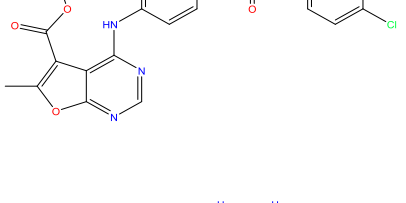
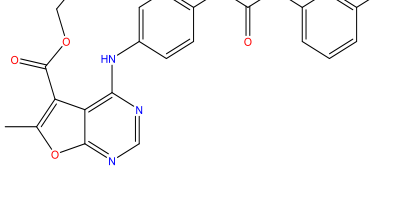
Order	Structure	% of inhibition
8		19
9		97
10		62
11		61
12		72
13		98
14		14
15		45

Table 1. (Continued)

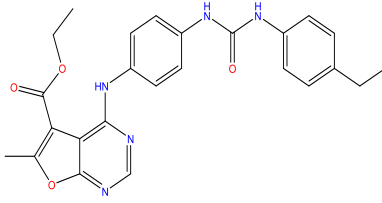
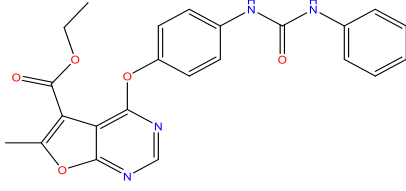
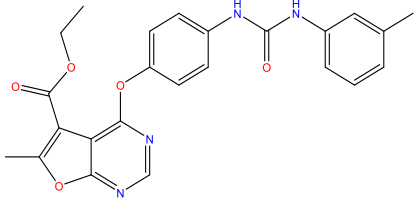
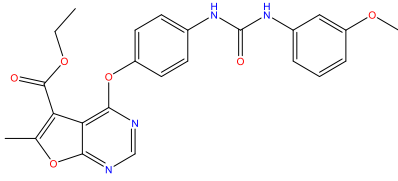
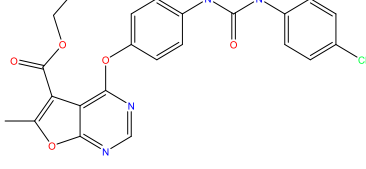
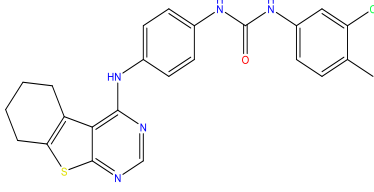
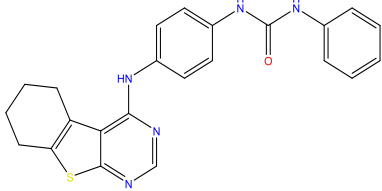
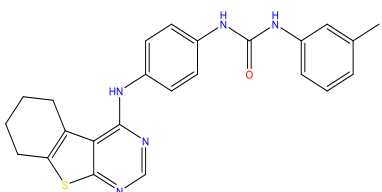
Order	Structure	% of inhibition
16		23
17		27
18		77
19		100
20		40
21		73
22		39
23		47

Table 1. (Continued)

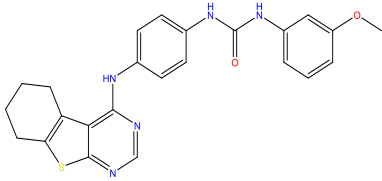
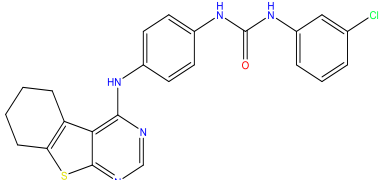
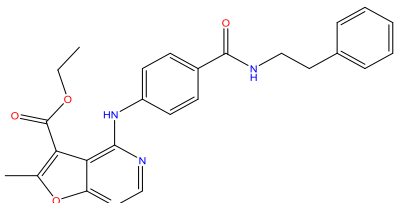
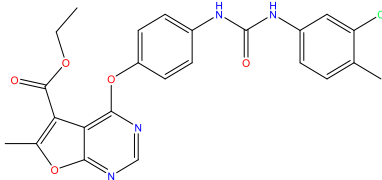
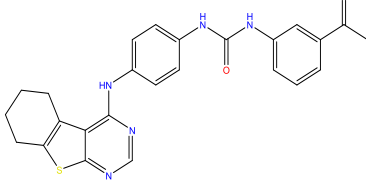
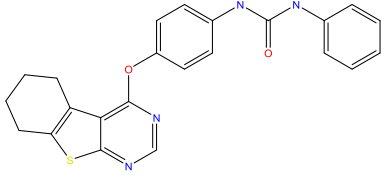
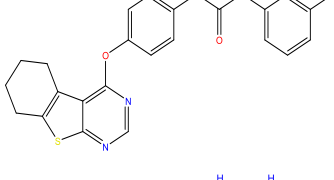
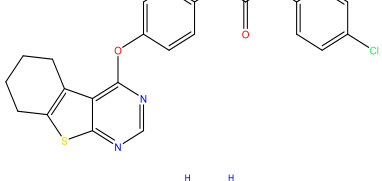
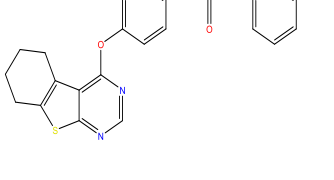
Order	Structure	% of inhibition
24		67
25		19
26		10
27		83
28		87
29		86
30		94
31		46
32		96

Table 1. (Continued)

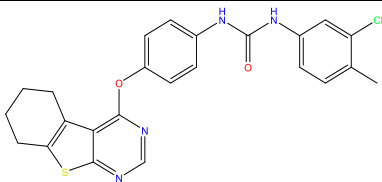
Order	Structure	% of inhibition
33		100

Table 2. R², RMSE, Q², adjusted R² values for models with the different number of descriptors.

Order	Adjusted R ²	Q ²	RMSE	R ²
1	0.527003	0.027606	22.33229	0.541784
2	0.704868	0.572353	17.35135	0.723314
3	0.778927	0.738329	14.74748	0.799653
4	0.827726	0.821183	12.83163	0.84926
5	0.869454	0.871045	10.97305	0.889852
6	0.884936	0.88934	10.18958	0.906511
7	0.909911	0.890616	10.14926	0.929618
8	0.926377	0.890344	10.2435	0.944782

R², Regression coefficient; RMSE, root-mean-square error.

Feature selection

Feature selection should be the first and most important step of model designing. Feature selection methods have been employed for selecting the best descriptors among the many descriptors containing low information for model construction or correlated with other descriptors without incurring much loss of information. In this study, three methods were used to reduce descriptors (28). It should be noted that the number of descriptors calculated could be reduced by some techniques.

Initially, among the pair of descriptors with a correlation coefficient above 0.95, one was eliminated by the Dragon software. Dragon reduced the number of 3224 calculated descriptors to 447. Then, descriptors that had constant or zero values that could not correlate the difference in structure to the difference in activity were removed. Then, the remaining descriptors were given to the software SPSS. The important descriptors are selected under a stepwise approach. In the stepwise strategy, a multiple-linear equation was built step by step. First, an initial model was determined, and then it was repeatedly changed by removing or adding a predictor variable based on stepping criteria for inclusion and exclusion. In each step, all variables were specified and evaluated to assign important descriptors. The SPSS presented a number of 8 proposed models by stepwise regression method.

RESULTS

Descriptor selection

First, the data set that consisted 33 compounds were divided into a calibration set of 26 compounds and a test set of 7 compounds with ratio 80% and 20%, respectively. Compounds number 2, 7, 9, 12, 19, 21, and 31 were selected as test sets and the remaining 33 compounds as a train set. In this study, the split of the data set was done with PCA. The equation must use the minimum number of descriptors to obtain the best fit and to achieve this, the stepwise regression method is used to find the best number of descriptors. Among the models given by the SPSS, after the sixth model, no considerable improvement in regression coefficient (R²) values were observed. For the appointed models, the values of the root-mean-square error (RMSE), (Q²), (R²), and R²_{adj} parameters are calculated as shown in Table 2.

The appropriate regression model is a model with the lowest number of descriptors to obtain the best fit and the number of compounds in the samples is best suited to be at least 5 times the number of descriptors and the descriptors should be orthogonal values (29,30). After analyzing the statistical parameters, according to the results shown in Table 2 and due to the changes in the slope of these parameters, model 5 with 5 descriptors was selected as the top

model, and modeling was performed with 5 descriptors. The characteristics of the descriptors used in this study are presented in Table 3, and their values are given in Table 4. The selected descriptors should be independent of each other because in their high dependence only the descriptor with a higher correlation

with the dependent variable is included in the model. A two-way correlation coefficient of descriptors was calculated by SPSS software and is presented in Table 5. The results showed that the behavior of the selected descriptors was independent and as can be seen, there is a little connection between the descriptors.

Table 3. Descriptors used in the 2D-QSAR study.

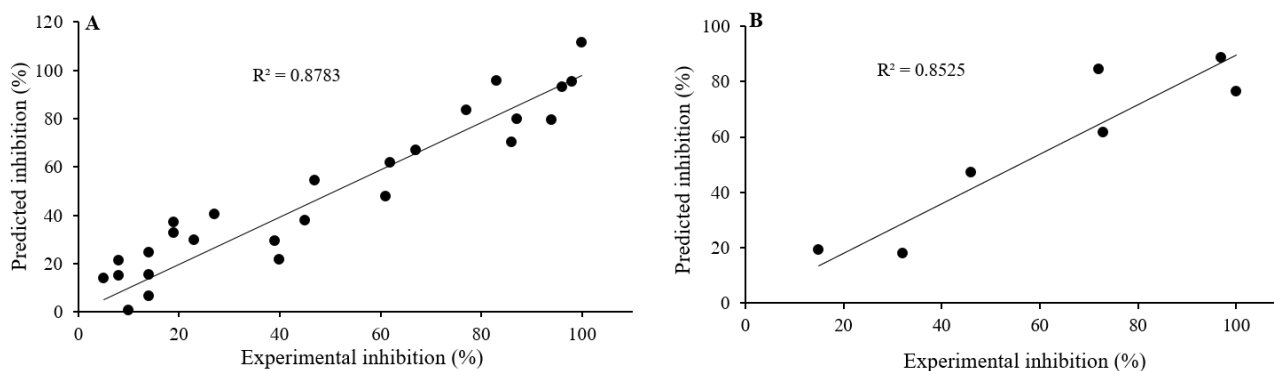
Descriptor types	Descriptor blocks type	Descriptor description
RDF035u	RDF descriptors	Radial distribution Function - 5.3 / unweighted
Mor24v	3DMoRSE	3D-MORSE-signal 24 / weighted by atomic van der volumes
EEig11r	3DMoRSE	Eigenvalue 11 from edge adj. matrix weighted by resonance integral
G2s	WHIM descriptors	2 nd component symmetry directional WHIM index / weighted by atomic electropological states
ATS3v	2Dauto correlation	Broto-Moreau autocorrelation of a topological structure-lag 3/ weighted by atomic van der Waals vol

Table 4. values of the obtained parameters of the studied derivatives of furopyrimidine and thienopyrimidine

Number	RDF035u	Mor24v	EEig11r	ATS3v	G2s
1	26.484	-0.41	2.044	3.739	0.167
2	27.26	-0.443	2.009	3.739	0.167
3	22.56	-0.276	2.14	3.739	0.167
4	23.356	-0.168	2.005	3.763	0.174
5	24.407	-0.304	2.005	3.739	0.167
6	24.234	-0.288	2.271	3.861	0.162
7	23.311	-0.237	2.011	3.746	0.165
8	26.9	-0.409	2.01	3.687	0.167
9	28.934	-0.395	2.179	3.736	0.182
10	27.84	-0.318	2.193	3.76	0.164
11	26.824	-0.369	2.164	3.736	0.165
12	26.841	-0.351	2.01	3.736	0.165
13	28.809	-0.377	2.329	3.805	0.18
14	29.849	-0.374	2.322	3.806	0.163
15	26.845	-0.416	2.162	3.754	0.165
16	28.53	-0.345	2.01	3.783	0.164
17	25.623	-0.279	2.007	3.664	0.167
18	28.411	-0.27	2.179	3.714	0.165
19	27.922	-0.246	2.193	3.739	0.164
20	25.499	-0.303	2.007	3.714	0.165
21	28.285	-0.21	2.328	3.785	0.164
22	26.108	-0.41	2.07	3.718	0.169
23	27.997	-0.402	2.179	3.766	0.168
24	27.069	-0.353	2.193	3.789	0.184
25	26.056	-0.364	2.164	3.766	0.168
26	26.003	-0.387	2.07	3.766	0.168
27	27.842	-0.396	2.333	3.833	0.167
28	28.994	-0.398	2.324	3.833	0.174
29	26.499	-0.143	2.068	3.696	0.169
30	28.306	-0.263	2.179	3.744	0.168
31	27.494	-0.042	2.193	3.768	0.167
32	26.049	-0.191	2.068	3.744	0.168
33	28.785	-0.19	2.333	3.813	0.176

Table 5. Correlation matrix between different obtained descriptors.

Descriptor types	RDF035u	Mor24v	EEig11r	ATS3v	G2s
RDF035u	1				
Mor24v	-0.2354	1			
EEig11r	0.562653	-0.02447	1		
ATS3v	0.255481	-0.08419	0.723526	1	
G2s	0.190981	-0.08694	0.209459	0.158042	1

**Fig. 1.** Predicted inhibition percent activities by multiple linear regression in comparison with experimental for (A) model and (B) test set.

Multiple Linear Regression (MLR)

MLR, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. MLR is modeling the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

The equation for MLR is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon \quad (1)$$

where, $i = n$ observations; y_i , dependent variable; x_i , explanatory variables; β_0 , y-intercept (constant term); β_p , slope coefficients for each explanatory variable; ϵ , the model error term (also known as the residuals). The multiple regression model is based on the following assumptions: a) there is a linear relationship between the dependent variables and the independent variables; b) the independent variables are not too highly correlated with each other; c) y_i observations are selected independently and randomly from the population; d) residuals should be normally distributed with a mean of 0 and variance σ .

At the center of MLR analysis is the task of fitting a single line through a scatter plot. More

specifically the multiple linear regression fits a line through a multi-dimensional space of data points. The simplest form has one dependent and two independent variables. The dependent variable may also be referred to as the outcome variable or regressing. The independent variables may also be referred to as the predictor variables or regressors.

After selecting the number of final descriptors to build the model, PCA analysis was used to classify the molecules into calibration and test sets. So, the data set is subdivided into a calibration set of 26 compounds to build the MLR model and a test set of 7 compounds for model evaluation.

We used Excel software and load the Analysis ToolPak add-in program. We used the regression in data analysis and by entering the data related to the calibration set (26 compounds), the MLR model was created. The results of which can also be seen in Table 6 and Fig. 1.

Then, equation (2) was the best MLR model that was selected by the regression method for furoprimidine and thienopyrimidine derivatives:

$$\text{Inhabitation percent} = 177.73 + 9.99 \text{ RDF035u} + 119.40 \text{ Mor24v} + 166.22 \text{ EEig11r} - 243.89 \text{ ATS3v} + 1196.81 \text{ G2s} \quad (2)$$

where, $N = 26$; $R^2 = 0.874$; $\text{RMSE} = 10.97$; $R^2_{\text{CV}} = 0.87$.

Table 6. Observed and calculated values of inhibition percent according to multiple linear regression method for the calibration and test sets.

Calibration set	Inhibition% (observed)	Inhibition% (predicted)	Residual	Relative error%
1	8	21.30	-13.30	-166.31
3	5	14.02	-9.02	-180.54
4	8	14.97	-6.97	-87.07
5	14	6.71	7.29	52.06
6	14	15.37	-1.37	-9.78
8	19	32.61	-13.61	-71.65
10	62	61.90	0.10	0.17
11	61	47.89	13.17	21.50
13	98	95.33	2.67	2.73
14	14	24.61	10.61	-75.75
15	45	37.76	7.24	16.09
16	23	29.55	-6.55	-28.48
17	27	40.48	-13.48	-49.92
18	77	83.43	-6.43	-8.35
20	40	21.79	18.21	45.54
22	39	29.38	9.62	24.66
23	47	54.44	-7.44	-15.83
24	67	66.88	0.12	0.18
25	19	37.08	-18.08	-95.14
26	10	0.65	-9.35	93.51
27	83	95.59	12.59	-15.17
28	87	79.83	7.17	8.24
29	86	70.21	15.78	18.37
30	94	79.49	14.51	15.43
32	96	93.04	-2.96	3.09
33	100	111.34	-11.34	-11.34
Test set				
2	15	19.31	4.31	-28.70
7	32	18.00	-14.00	93.51
9	97	88.72	8.28	8.54
12	72	84.33	12.33	-17.13
19	100	76.44	23.56	23.56
21	73	61.67	-11.33	15.52
31	46	47.07	-0.02	-2.32

The predicted values of the inhibition percentage of the calibration and test set datasets using this model were plotted against the experimental values and are shown in Fig. 1. The mentioned linear model was used to predict seven external test data that have never been used in the descriptor selection or model construction. The predicted values of the inhibition percent of the calibration set and the test set using the MLR equation are presented in Table 6.

Y-randomization test certifies the robustness of a QSAR model. The dependent parameter is shuffled randomly and a new QSAR model is developed applying the original independent parameter matrix. The new QSAR models (after several iterations) are expected to have low R² and Q² values. The results are shown in Table 7. The low R² and Q² values show that

the good results in our original model are not due to a chance correlation or structural dependency of the training set.

Artificial neural networks (ANN)

ANN is one of the main tools used in machine learning. As the “neural” part of their name suggests, they are brain-inspired systems that are intended to replicate the way that humans learn.

Table 7. R² and Q² values after several Y-randomization tests.

Iteration	R ²	Q ²
1	0.16	0.01
2	0.15	0.00
3	0.20	0.03
4	0.33	0.11
5	0.18	0.02
6	0.14	0.00
7	0.24	0.05

Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use (31-33). They are excellent tools for finding patterns that are far too complex or numerous for a human programmer to extract and teach the machine to recognize. In the network, it connects to each node of the connection layers and is influenced by the amount of weights affected by the units connected to it. During the random weight training and initial random crash, adjustments are made to find the minimum difference between the output value and the target value. After a sufficient number of training iterations, the ANN learns to recognize patterns in the data, so it can be used for predicting new input values (34,35).

The network used in this study consisted of three layers (an input layer, a hidden layer, and an output layer). The input nodes contain five parameters in the regression equation and one constant. The output neuron refers to the retention index. Before entering the neural

network, input data were stored at a ratio of 0 to 1. Inhibition percent values were also used with this rule. Sigmoid transfer functions were applied in all layers. The weights were adjusted through a backpropagation algorithm to correct the model behavior. This computer program is designed to generate the desired number of neurons in the hidden layer. In order to select the optimal model, different topological networks with different hidden units were performed. On the other hand, the values of learning factor, coefficient of movement, and core values of weight and bias were tested to find the best performance and fastest convergence. The predicted values of inhibition percent for training and test sets using the ANN model are presented in Table 8. The predicted values of the percentage of inhibition of the training data set are plotted using the ANN model against the experimental values and are shown in Fig. 2. Also, the residual plot of furopyrimidine and thienopyrimidine derivatives by ANN model is demonstrated in Fig. 3.

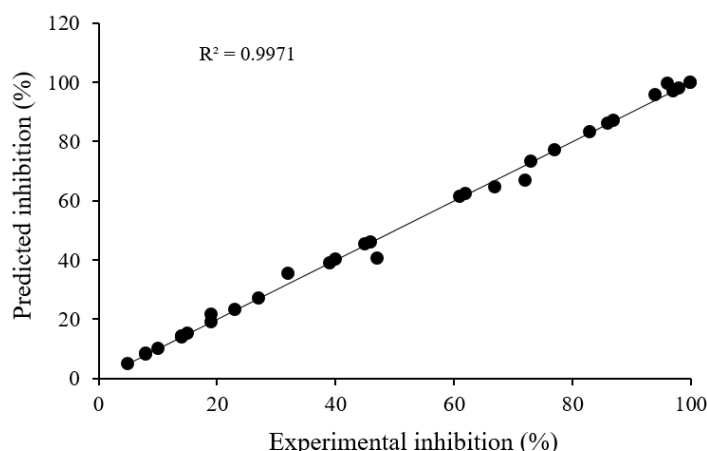


Fig. 2. Predicted inhibition percent activities by artificial neural network in comparison with experimental.

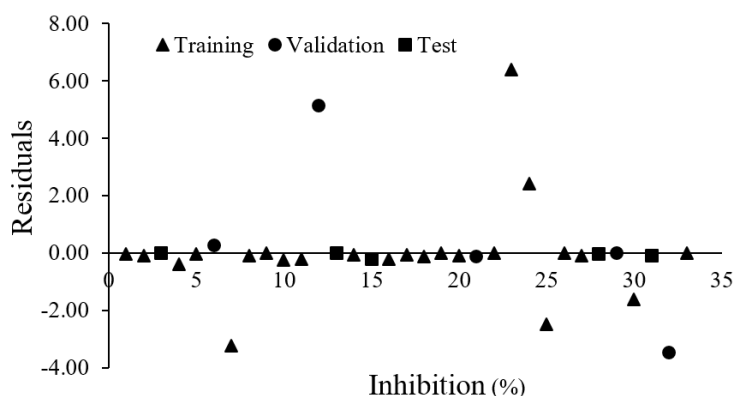


Fig. 3. Residual plot of furopyrimidine and thienopyrimidine derivatives by artificial neural network model.

Table 8. Observed values and calculated values of inhibition percent according to the artificial neural network method.

Training set	Inhibition% (observed)	Inhibition% (observed)	Residual	Relative error%
1	8	8.02	-0.02	0.25
2	15	15.08	-0.08	0.53
4	8	8.38	-0.38	4.75
5	14	14.03	-0.03	0.21
7	32	35.24	-3.24	10.13
8	19	19.10	-0.10	0.53
9	97	97.00	0.00	0.00
10	62	62.25	-0.25	0.40
11	61	61.2	-0.2	0.33
14	14	14.07	-0.07	0.50
16	23	23.21	-0.21	0.91
17	27	27.06	-0.06	0.22
18	77	77.13	-0.13	0.17
19	100	100.00	0.00	0.00
20	40	40.10	-0.10	0.25
22	39	39.01	-0.01	0.03
23	47	40.59	6.41	-13.64
24	67	64.58	2.42	-3.61
25	19	21.47	-2.47	13.00
26	10	10.01	-0.01	0.10
27	83	83.10	-0.10	0.12
30	94	95.62	-1.62	1.72
33	100	100.00	0.00	0.00
Validation set				
6	14	13.73	0.27	-1.93
12	72	66.86	5.14	-7.14
21	73	73.13	-0.13	0.18
29	86	86.01	-0.01	0.01
32	96	99.47	-3.47	3.61
Test set				
3	5	5.00	0.00	0.00
13	98	98.00	0.00	0.00
15	45	45.21	-0.21	0.47
28	87	87.02	-0.02	0.02
31	46	46.08	-0.08	0.17

N = 33; $R_{\text{train}} = 0.998$; $R_{\text{test}} = 0.999$; $R_{\text{validation}} = 0.999$; $R_{\text{all}} = 0.998$; $R^2_{\text{CV}} = 0.99998$; root-mean-square error = 1.78

Table 9. Performance comparison between models obtained by MLR and ANN.

Models	Calibration			Prediction		
	RMSE	R ²	Q ²	RMSE	R ²	Q ²
MLR	10.97	0.889	0.87	14.54	0.684	0.75
ANN	1.78	0.998	0.99	0.17	0.999	0.99

MLR, multiple linear regression; ANN, artificial neural network; RMSE, root-mean-square error; R², regression coefficient.

DISCUSSION

Modeling was performed with 5 descriptors named RDF035u, Mor24v, EEig11r, ATS3v, and G2s. The RDF035u descriptor belongs to the RDF family, Mor24v and EEig11r to the 3DMorSE family, the G2s to the WHIM family, and the ATS3v to the 2D-autocorrelation family. The 2D autocorrelation

descriptor is a subset of autocorrelation descriptors. This group of descriptors is molecular descriptors that are calculated based on the autocorrelation function (AC₁). In general, 2D autocorrelation descriptors express how an atomic property is distributed throughout the topology structure and can be calculated by summing the product of the terms containing the desired atomic property for the

final atoms in all paths of a given length. Among these descriptors, four descriptors including RDF035u, Mor24v, EEig11r, and G2s with positive coefficients and ATS3v with negative coefficient entered in the model. The positive coefficients of each descriptor indicate its direct effect on the activity and the negative coefficient indicates the inverse effect of the descriptor on the activity.

The ATS3v descriptor is a subset of the autocorrelation descriptors called Broto-Moreau, which is weighted by atomic and van der Waals volumes. This descriptor has a negative coefficient in the equation, meaning that increasing this descriptor reduces the inhibition activity of the VEGFR-2 receptor. 3DMoRSE descriptors (3D representation of molecule structure based on electron scattering) can be calculated from the equation used in electron diffraction studies, which allows the 3D representation of the molecule as fixed values. These descriptors are able to provide a link between the 3D structure of organic compounds and their physical, chemical, and

biological properties. Because these descriptors express the 3D arrangement of atoms without being related to the size of the molecule, they apply to a large number of molecules with large structural differences. The Mor24v descriptor in the equation, which is weighted by the atomic and van der Waals volumes, has a positive coefficient and has a direct effect on the inhibition index. WHIM descriptors of Cartesian coordinates of the 3D structure of a molecule are calculated using conformers with the least energy and include information about the size, shape, equation, and atomic distribution of the 3D structure of the molecule. The descriptor of G2s has a positive coefficient in the weighted equation with the electropathological state of Kier and Hall and has a direct effect on the inhibition percentage. RDF descriptors are based on measuring the atomic distance in the 3D representation of molecules, and in addition to the atomic distance, they provide other information about ring types, planar and non-planar systems, and types of atoms.

Table 10. Comparing values of inhibition percent experimental and predicted results using MLR and ANN methods.

Number	Inhibition (observed)	MLR (predicted)	Residual	Relative error (%)	ANN (predicted)	Residual	Relative error (%)
1	8	21.305	-13.305	-166.313	8.017	-0.017	-0.219
2	15	19.305	-4.305	-28.7	15.082	-0.082	-0.545
3	5	14.027	-9.027	-180.54	5.001	-0.001	-0.027
4	8	14.966	-6.966	-87.075	8.382	-0.382	-4.773
5	14	6.711	7.289	52.064	14.031	-0.031	-0.218
6	14	15.369	-1.369	-9.779	13.734	0.266	1.899
7	32	18.175	13.825	43.203	35.244	-3.244	-10.137
8	19	32.614	-13.614	-71.653	19.1	-0.1	-0.525
9	97	88.716	8.284	8.54	96.998	0.002	0.002
10	62	61.903	0.097	0.156	62.25	-0.25	-0.403
11	61	47.884	13.116	21.502	61.195	-0.195	-0.32
12	72	84.331	-12.331	-17.126	66.864	5.136	7.133
13	98	95.327	2.673	2.728	97.998	0.002	0.002
14	14	24.605	-10.605	-75.75	14.068	-0.068	-0.487
15	45	37.76	7.24	16.089	45.207	-0.207	-0.46
16	23	29.55	-6.55	-28.478	23.208	-0.208	-0.903
17	27	40.479	-13.479	-49.922	27.056	-0.056	-0.207
18	77	83.432	-6.432	-8.353	77.126	-0.126	-0.163
19	100	76.441	23.559	23.559	99.999	0.001	0.001
20	40	21.785	18.215	45.538	40.099	-0.099	-0.248
21	73	61.667	11.333	15.525	73.131	-0.131	-0.179
22	39	29.382	9.618	24.662	39.011	-0.011	-0.029
23	47	54.44	-7.44	-15.83	40.586	6.414	13.646
24	67	66.878	0.122	0.182	64.579	2.421	3.613
25	19	37.076	-18.076	-95.137	21.466	-2.466	-12.979
26	10	0.649	9.351	93.51	10.007	-0.007	-0.073
27	83	95.591	-12.591	-15.17	83.1	-0.1	-0.12
28	87	79.829	7.171	8.243	87.023	-0.023	-0.026
29	86	70.205	15.795	18.366	86.008	-0.008	-0.009
30	94	79.492	14.508	15.434	95.62	-1.62	-1.724
31	46	47.071	-1.071	-2.328	46.084	-0.084	-0.184
32	96	93.038	2.962	3.085	99.47	-3.47	-3.615
33	100	111.342	-11.342	-11.342	99.996	0.004	0.004

MLR, multiple linear regression; ANN, artificial neural network.

The RDF035u descriptor entered in the equation, which is not weighted with a specific property of the molecule, has a positive coefficient, and as it increases, the inhibition index increases.

The main performance parameters of the two models are shown in Table 9. As expected, according to the results shown in the table, all statistical parameters for the ANN model are better than the MLR model. Also, the results of the two models are compared in Table 9. The results of the analysis with two models indicate that the percentage of relative error obtained from the ANN model is much lower than the MLR model.

The ANN model containing a hidden layer with three nodes and a sigmoid transfer function could predict the activity of the VEGFR-2 inhibitory derivative with an absolute relative error of calibration and validation lower than 1% and that of prediction lower than 1%. Table 10 compares the predictions performances between models MLR and ANN.

CONCLUSION

QSAR analysis can greatly help us to comprehend the basic structural properties of the inhibitors required by its target, and thus to discover more promising chemical derivatives (36). The MM2 theory was used to optimize the 3D geometry of the molecules and DRAGON was used to calculate a diverse set of quantum chemical descriptors. As can be seen, the predicted values of the MLR method and the ANN technique are close to the experimental values, which demonstrates the ability to describe molecular topology in prediction. In the MLR method, a six-parameter equation containing a constant value and the coefficients of the 5 selected descriptors was obtained. The ANN model containing a hidden layer with three nodes and a sigmoid transfer function could predict the activity of the VEGFR-2 inhibitory derivative with an absolute relative error of calibration and validation lower than 1% and that of prediction lower than 1%. Comparing the results of MLR and ANN methods showed the superiority of the ANN method over MLR for predicting the activities.

Conflict of interest statement

All authors declared no conflict of interest in this study.

Authors' contribution

F. Masoomi Sefiddashti conducted the research, analyzed the data, and wrote the manuscript. Sh. Ghanavati Nasab conducted the research and participated in revising the manuscript. H. Haddadi and S. Asadpour supervised the project and revised the manuscript. The manuscript was reviewed by all authors.

REFERENCES

1. Wu HC, Chang DK, Huang CT. Targeted therapy for cancer. *J Cancer Mol.* 2006;2(2):57-66.
2. Asadpour S, Aramesh-Boroujeni Z, Jahani S. *In vitro* anticancer activity of parent and nano-encapsulated samarium(III) complex towards antimicrobial activity studies and FS-DNA/BSA binding affinity. *RSC Adv.* 2020;10(53):31979-31990. DOI: 10.1039/D0RA05280A.
3. Cimpean AM, Ribatti D, Raica M. A Historical Appraisal of Angiogenesis Assays Since Judah Folkman and Before. In: Gaetano S, editor. *Angiogenesis: Insights From a Systematic Overview.* New York: Nova Science; 2013. pp. 31-50.
4. Birbrair A, Zhang T, Wang ZM, Messi ML, Mintz A, Delbono O. Pericytes at the intersection between tissue regeneration and pathology. *Clin Sci (Lond).* 2015;128(2):81-93. DOI: 10.1042/CS20140278.
5. Niu G, Chen X. Vascular endothelial growth factor as an anti-angiogenic target for cancer therapy. *Curr Drug Targets.* 2010;11(8):1000-1017. DOI: 10.2174/138945010791591395.
6. Joule JA, Mills K. *Heterocyclic Chemistry at a Glance.* 2nd ed. John Wiley & Sons; 2012. pp. 48-61. DOI: 10.1002/9781118380208.
7. Amr AE-GE, Sabry NM, Abdulla MM. Synthesis, reactions, and anti-inflammatory activity of heterocyclic systems fused to a thiophene moiety using citrazinic acid as synthon. *Monatsh Chem.* 2007;138(7):699-707. DOI: 10.1007/s00706-007-0651-0.
8. Fujiwara N, Nakajima T, Ueda Y, Fujita H, Kawakami H. Novel piperidinympyrimidine derivatives as inhibitors of HIV-1 LTR activation. *Bioorg Med Chem.* 2008;16(22):9804-9816. DOI: 10.1016/j.bmc.2008.09.059.
9. Sabet R, Fassihi A, Moeinifard B. Preliminary MLR studies of antimicrobial activity of some 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives. *Res Pharm Sci.* 2007;2(2):103-112.
10. Masoomi Sefiddashti F, Haddadi H, Asadpour S, Ghanavati Nasab S. Prediction of IC₅₀ values of 2-benzyloxy benzamide derivatives using multiple

- linear regression and artificial neural network methods. *Iranian J Math Chem*. 2020;9(3):179-199. DOI: 10.22052/ijmc.2020.217837.1483.
11. Shahlaei M, Fassihi A, Saghaie L, Arkan E, Pourhossein A. A modeling study of aldehyde inhibitors of human cathepsin K using partial least squares method. *Res Pharm Sci*. 2011;6(2):71-80. PMID: 22224089.
 12. Ghanavati Nasab S, Semnani A, Marini F, Biancolillo A. Prediction of viscosity index and pour point in ester lubricants using quantitative structure-property relationship (QSPR). *Chemom Intell Lab Syst*. 2018;183:59-78. DOI: 10.1016/j.chemolab.2018.10.013.
 13. Shahlaei M, Saghaie L. Prediction of p38 map kinase inhibitory activity of 3,4-dihydropyrido [3,2-d] pyrimidone derivatives using an expert system based on principal component analysis and least square support vector machine. *Res Pharm Sci*. 2014;9(6):471-488. PMID: 26339262.
 14. Sadeghian-Rizi S, Sakhteman A, Hassanzadeh F. A quantitative structure-activity relationship (QSAR) study of some diaryl urea derivatives of B-RAF inhibitors. *Res Pharm Sci*. 2016;11(6):445-453. DOI: 10.4103/1735-5362.194869.
 15. Saghaie L, Shahlaei M, Fassihi A. Quantitative structure activities relationships of some 2-mercaptoimidazoles as CCR2 inhibitors using genetic algorithm-artificial neural networks. *Res Pharm Sci*. 2013;8(2):97-112. PMID: 24019819.
 16. Gramatica P, Papa E. QSAR modeling of bioconcentration factor by theoretical molecular descriptors. *QSAR Comb Sci*. 2003;22(3):374-385. DOI: 10.1002/qsar.200390027.
 17. Hansch C, Kurup A, Garg R, Gao H. Chembioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem Rev*. 2001;101(3):619-672. DOI: 10.1021/cr0000067.
 18. Peter SC, Dhanjal JK, Malik V, Radhakrishnan N, Jayakanthan M, Sundar D. Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. In: *Reference Module in Life Sciences*. Elsevier; 2018. pp. 661-676. DOI: 10.1016/B978-0-12-809633-8.20197-0.
 19. Norouzian MA, Asadpour S. Prediction of feed abrasive value by artificial neural networks and multiple linear regression. *Neural Comput Appl*. 2012;21(5):905-909. DOI: 10.1007/s00521-011-0579-5.
 20. Ghasemi J, Asadpour S, Abdolmaleki A. Prediction of gas chromatography/electron capture detector retention times of chlorinated pesticides, herbicides, and organohalides by multivariate chemometrics methods. *Anal Chim Acta*. 2007;588(2):200-206. DOI: 10.1016/j.aca.2007.02.027.
 21. Ghasemi J, Abdolmaleki A, Asadpour S, Shiri F. Prediction of solubility of nonionic solutes in anionic micelle (SDS) using a QSPR model. *QSAR Comb Sci*. 2008;27(3):338-346. DOI: 10.1002/QSAR.200730022.
 22. Aziz MA, Serya RAT, Lasheen DS, Abdel-Aziz AK, Esmat A, Mansour AM, et al. Discovery of potent VEGFR-2 inhibitors based on furopyrimidine and thienopyrimidine scaffolds as cancer targeting agents. *Sci Rep*. 2016;6:24460,1-20. DOI: 10.1038/srep24460.
 23. Talevi A, Goodarzi M, Ortiz EV, Duchowicz PR, Bellera CL, Pesce G, et al. Prediction of drug intestinal absorption by new linear and non-linear QSPR. *Eur J Med Chem*. 2011;46(1):218-228. DOI: 10.1016/j.ejmech.2010.11.005.
 24. Goodarzi M, Freitas MP, Jensen R. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3B inhibitory activities. *J Chem Inf Model*. 2009;49(4):824-832. DOI: 10.1021/ci9000103.
 25. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc*. 1985;107(13):3902-3209. DOI: 10.1021/ja00299a024.
 26. Young DC. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. John Wiley & Sons, Inc.; 2002. pp. 243-251. DOI: 10.1002/0471220655.
 27. Consonni V, Mauri A, Pavan M, Todeschini R. Dragon software: an easy approach to molecular descriptor calculations match. *Commun Math Comput Chem*. 2006;56:237-248.
 28. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep*. 2015;5:10312,1-13. DOI: 10.1038/srep10312.
 29. Zhou P, Mei H, Tian F, Wang J, Wu S, Li Z. A new two-dimensional approach to quantitative prediction for collision cross-section of more than 110 singly protonated peptides by a novel molecular electronegativity-interaction vector through quantitative structure-spectrometry relationship studies. *Front Chem China*. 2007;2(1):55-63. DOI: 10.1007/s11458-007-0012-x.
 30. Randic M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J Chem Inf Comput Sci*. 1991;31(2):311-320. DOI: 10.1021/ci00002a018.
 31. Gasteiger J, Zupan J. Neural networks in chemistry. *Angew Chemie Int Ed Engl*. 1993;32(4):503-527. DOI: 10.1002/anie.199305031.
 32. Salt DW, Yildiz N, Livingstone DJ, Tinsley CJ. The use of artificial neural networks in QSAR. *Pestic Sci*. 1992;36(2):161-170.
 33. Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure-activity

- relationship (QSAR) analysis. *J Med Chem.* 1990;33(9):2583-2590.
DOI: 10.1021/jm00171a037.
34. Guo W, Zhu P, Brodowsky H. The study for optimization of chromatographic condition by means of artificial neural networks. *Talanta.* 1997;44(11):1995-2001.
DOI: 10.1016/S0039-9140(96)02171-6.
35. Guo W, Lu Y, Zheng XM. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta.* 2000;51(3):479-488.
DOI: 10.1016/S0039-9140(99)00301-x.
36. Zhang X, Mao J, Li W, Koike K, Wang J. Improved 3D-QSAR prediction by multiple-conformational alignment: a case study on PTP1B inhibitors. *Comput Biol Chem.* 2019;107134.
DOI: 10.1016/j.compbiolchem.2019.107134.