Regular Article

ANALYTICAL AND BIOANALYTICAL CHEMISTRY RESEARCH

Published by the Iranian Chemical Society

Anal. Bioanal. Chem. Res., Vol. 9, No. 4, 331-339, September 2022.

Predicting the Anticonvulsant Activities of Phenylacetanilides Using Quantitative-structure-activity-relationship and Artificial Neural Network Methods

Javad Yousefi, S. Maryam Sajjadi* and Ahmad Bagheri

Faculty of Chemistry, Semnan University, Semnan, Iran

(Received 23 January 2022 Accepted 15 March 2022)

In this study, the anticonvulsant activity of phenylacetanilides compounds was predicted using QSAR and artificial neural network (ANN) models. Variety kinds of molecular descriptors were computed using Dragon for 30 monosubstituted phenylacetanilides. Then, seven out of 1600 descriptors were selected and used in ANN analysis. The complete set of 30 compounds was randomly divided into a training set of 80%, a test set of 10%, and a validation set of 10% compounds. Moreover, multiple linear regression (MLR) analysis was utilized to build a linear model by using the same descriptors and the results of this linear model were compared with the nonlinear ANN analysis. Correlation coefficient (R^2) and mean squared error (MSE) of the ANN and MLR models (for the whole dataset) were 0.85, 0.06816; and 0.6, 0.09792, respectively. The higher R^2 of ANN method revealed that the relationship between the descriptors and anticonvulsant activity of the compounds is non-linear.

Keywords: QSAR, Molecular descriptors, Artificial neural network, Anticonvulsant activity6

INTRODUCTION

Over the past years, QSAR/QSPR models have been widely recognized in a variety of fields such as physical, organic, analytical, drug design, chemistry, biochemistry, chemical and technology engineering, toxicology and biotechnology [1-3]. The success of the QSPR and QSAR methods can be explained with the insight expressed to determine the structure of chemical properties and biological activities as well as the possibility of estimating the properties of new chemical compounds without the need for synthesis and testing. Molecular descriptors are numerical indices that are related to the molecular structure, which encode some information about the molecular structure. Descriptors can include physical and chemical properties of molecules and theoretical indices computed by mathematical formulas or computational algorithms. There are various computer programs such as CoMFA [4], SOMFA [5], HINT [6], CoRSA [7], PRECLAV [8] and Dragon [9] used to find descriptors of a molecular structure. These programs introduce a large variety of descriptors for a molecule such as constitutional descriptors, topological descriptors, walk and path counts, 3D MoRSE descriptors and functional group counts [10-14].

In QSAR studies, models with fewer variables lead to reducing the complexity of the model, preventing overfitting/overtraining and diminishing computational time and improving the accuracy of predictions for new samples. Therefore, one important task is selecting a few relevant descriptors with minimum collinearity and maximum information about the compounds. In this regard, there are a lot of variable selection methods such as stepwise regression [15], multiple linear regression (MLR) [16], principal component analysis (PCA) [17], kernel stone, principal component regression (PCR) [18], particle least squares (PLS), genetic algorithm [19-21]. Apart from the importance of variable selection, the predictive ability of a QSAR model is influenced by the modeling techniques used to find the mathematical model between the descriptors and the biological activity. There are several linear and

*Corresponding author. E-mail: sajjadi@semnan.ac.ir

nonlinear models which are used in QSAR studies, such as MLR [22], PLS [23], nonlinear artificial neural networks [24,25], support vector machines [26], and multivariate adaptive regression splines [27].

Anticonvulsants are a diverse group of pharmacological agents which are used in the treatment of epileptic seizures. In spite of optimal use of anticonvulsant drugs, seizure control is failed for many people with seizures disorder and even the patients suffer the significant toxic side effects of the drugs. Therefore, it is required to develop new anticonvulsant drugs to improve the treatment of seizure disorder over the past years [28-30]. In this regard, QSAR techniques have been the most used approaches in the aided design of derivatives of known antiepileptic drugs or new compounds with this biological activity. They may assist to identify new compounds with improved activity and a better profile of side effects. Indeed, the combination of these techniques with traditional drug discovery methodologies increases greatly the chance of drug discovery in a sustainable and economical fashion.

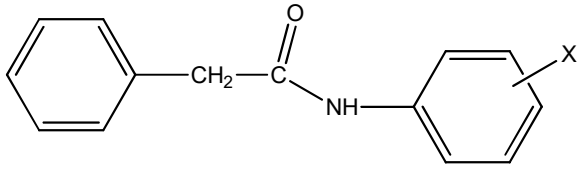
There are a lot of reports about the application of QSAR for predicting the anticonvulsant activity of some anticonvulsants [31,32]. In the present study, a nonlinear ANN algorithm is proposed to predict the anticonvulsant activity of 30 combinations of phenylacetanilides. The molecular descriptors were computed using Dragon software and the experimental anticonvulsant activity ($\log 1/ED_{50}$) of the molecules was obtained from the reported work by TARKO *et al*. [8]. The relationship between the structure and $\log 1/ED_{50}$ of the molecules was achieved using ANN and MLR analysis methods.

ANNs are generally nonlinear learning math systems and they need known input data set without any assumptions [33]. The ANN constructs a mapping of the input and output variables which can afterward be used to predict unknown output as a function of suitable inputs [34].

To our knowledge, there is no report on the study of the above anticonvulsant activity of compounds using ANN method. Therefore, we proposed an ANN approach to developing a QSPR model to improve R^2 and obtain a reliable prediction of the activity of the drug. Moreover, the system is analyzed by MLR method for comparison. Finally, it has been shown that the non-linear ANN model can make a satisfactory relationship between molecular descriptors and anticonvulsants activity.

MOLECULAR DATABASE AND METHOD

The database used as input by Dragon includes 30 monosubstituted phenylacetanilides with the common formula I, reported in Table 1. The anticonvulsant activity $\log 1/ED_{50}$ of these molecules has been taken from Ref. [8]. These activities were studied by the maximum electroshock possession method in mice.



I

MOLECULAR MODELING

In this paper, the structures of all the above compounds were examined. Each structure was drawn with Gauss View 5.0 and optimized with the semi-empirical method PM6 available in the Gaussian software. Molecular descriptors of the optimized structures were calculated by Dragon software. All statistical calculations were done with Matlab 7.0 software.

The number of descriptors was more than 3000 and the best set of them was selected by MATLAB programming. In the selection process, first, those descriptors that had a correlation of nearly 0.9 with the activities were chosen. Then a stepwise method was used to select the most significant descriptors among this set. In this step, seven descriptors out of 266 were chosen and represented in Table 2, which are used for further investigation.

The symbol and description of the selected descriptors are reported in Table 2. Of the seven selected descriptors, two of them belong to the 2D autocorrelation descriptors (MATS2m and MATS7m). In 2D autocorrelation descriptors, the molecule atoms show a set of discrete points in space, and the atomic property and function are investigated at those points. The symbol for each of the autocorrelation descriptors is represented by two indices d and w; where d stands for the lag and w stands for the

Table 1. Experimental Values for Phenylethanylidine Substitutes, Anticonvulsant Activity (8)

NO	X	log1/ED_{50} exp	NO	X	log1/ED_{50} exp
1	H	3.77	16	m-COMe	3.95
2	m-Me	3.75	17	m-OAc	3.48
3	m-Et	3.67	18	m-OEt	3.42
4	m-F	3.34	19	m-OSO2Me	3.77
5	m-Cl	3.4	20	p-Me	3.26
6	m-Br	3.32	21	p-F	3.49
7	m-I	2.64	22	p-OH	3.72
8	m-CF_3	2.84	23	p-OMe	3.78
9	m-OH	3.58	24	p-COMe	3.51
10	m-NH2	3.81	25	o-F	3.48
11	m-NHMe	4.03	26	o-OH	3.33
12	m-NHEt	3.91	27	o-NH2	3.40
13	m-OMe	3.22	28	o-OMe	3.43
14	m-CN	3.44	29	o-NO2	3.29
15	m-NO2	3.62	30	o-COMe	3.41

Table 2. The Selected Structural Descriptors for QSAR Analysis

ID	Name	Description	Block
1	MATS2m	Moran autocorrelation of lag2 weighted by mass	2D autocorrelations
2	MATS7m	Moran autocorrelation of lag7 weighted by mass	2D autocorrelations
3	RDF110m	Radial distribution function-110/weighted by mass	RDF descriptors
4	RDF030v	Radial distribution function-030/weighted by van der waals volume	RDF descriptors
5	RDF075v	Radial distribution function-075/weighted by van der waals volume	RDF descriptors
6	Mor16m	Signal 16/weighted by mass	3D-MoRSE descriptors
7	Mor22e	Signal 22/weighted by sanderson electronegativity	3D-MoRSE descriptors

weight. The weight can be p (polarizability), e (Sanderson electronegativity), m (relative atomic mass), and v (Vander Waals volume). MATS2m and MATS7m were mainly differentiated by atomic mass and Sanderson electronegativity weighted terms [35]. Figure 1 displays that MATS2m has a positive effect on anticonvulsant activities, which indicates that log1/ED_{50} is directly related to atomic Sanderson electronegativity; however, MATS7m shows a negative effect on log1/ED_{50}.

The third, fourth and fifth descriptors are RDF110m, RDF030v and RDF075v, which belong to the RDF descriptors. The RDF is independent of the atom number and the size of a molecule, it uniquely depends on the arrangement of the atoms, and it is invariant against the rotation and translation of the whole molecule. The RDF of an ensemble of n atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius R. The RDF110m is negatively correlated to log1/ED_{50}, while RDF030v and RDF075v have a positive effect on anticonvulsant activities.

Two other descriptors (Mor16m and Mor22e) belong to the 3D-MoRSE descriptors derived from infrared spectra

simulation using a generalized scattering function. These descriptors are proposed as signal 16 and 22⁄weighted and by atomic masses, which are assigned to the masses of the molecules.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks (ANNs) are computer programs that are defined by topology, computational characteristics of elements, and teaching rules which are widely used in QSAR/QSPR studies [36,37]. They include a series of layers where the information is transferred through the layers. The first layer is termed the input layer and each exterior variable is fed to each neuron. The last layer is the output layer whose neurons handle the output variables from the network. The layers of neurons between the input and output layers are called hidden layers, each of which may conduct the computations independently and may transfer the results yet to another layer. There are two main steps in the process of ANN: learning and validation. In learning or training, the weights are modified in answer to entrance information. Different learning methods such as BFGS [36,38,39], backpropagation [40], feed-forward [41], Elman backpropagation [42], feed-forward time-delay [43], Hopfield [44], and so on exist to measure the ability of models to generate the right answer.

ANN OPTIMIZATION

In ANN algorithm, the net input into the jth layer node (i[j]) is obtained from the sum of weighted outputs from the prior ith layer (o[i]) based on the following equation [45]:

$$\text{Net input to a node} = i[j] = \Sigma_i \{w[ji]\ o[i]\} \quad (1)$$

where, w[ji] is the weight factor. The weights of the links between the processing nodes are an important task during the learning process by the ANN model. The numerical values of the weight factors change based on the training data sets, in order to minimize the difference between the actual outputs and the model predicted outputs.

The weights are changed as follows during the ANN algorithm:

$$\Delta Wji + Wji \rightarrow Wji$$

$$\Delta Wji = \eta\ (t - o)\ xi \quad (2)$$

In this formula t is output **(**the objective function for the current sample), o is the output of perceptron and η is the learning factor. The role of the learning factor is to control the amount of weight change at each stage, which is usually a small value (*e.g.* 0.1) that decreases and would have less and less impact as the number of iterations increases.

In QSAR study of 30 monosubstituted phenylacetanilides, 7 selected structural descriptors of the molecules (Table 2) were used as input neurons in an ANN modeling and anticonvulsant activity was placed on the neuron in the output layer. The number of hidden layers and their neurons must be optimized to construct the ANN architectures. This model was developed in Matlab software using a 3-layer feed-forward back-propagation network.

To perform ANN algorithm, first, all inputs (descriptor values) and outputs were coded between -1 and +1 [46]. Next, the data is randomly divided into a three sets: training (80%), test (10%) and validation set (10%), including 24 data as the training set, 3 data as test set and 3 data as the validation set. The training set is used only for learning, (*i.e.*, to fit the weights of the network). The test set is applied to control the network parameters such as the number of hidden layers and their neurons and, or the number of training epochs. The validation set is employed only to define the extension efficiency of a trained neural network.

Table 3 demonstrates the main network characteristics in the MATLAB toolbox [47]. This includes network topology, training algorithm, and the number of data points of each of the three data sets (training, test and validation).

The performance of the ANN model was evaluated using mean square error (MSE), correlation coefficient (R^2), root mean square error (RMSE) as follows [48]:

$$MSE = \frac{\sum_i (y_{ANN,i} - y_{\exp,i})^2}{n} \quad (3)$$

$$RMSE = \left[\frac{1}{n \sum_{i=1}^{n} (\hat{y} - yi)^2} \right]_2 \quad (4)$$

Table 3. The Network Parameters in the MATLAB Toolbox

Topology	7 inputs, 1 output and 1 hidden layer with 3 neurons (7×3×1)
Data	Training set: 24 randomly selected data structures
	Test set: 3 randomly selected data structures
	Validation set: 3 randomly selected data structures
Beginning function	log-sigmoid
Training algorithm	Levenberge-Marquardt
Loss function conditions	Minimum MSE
Stopping conditions	The network stops in one of three way:
	Validation check > 10
	Minimum gradient < 10^{-7}
	Momentum speed > 10^{10}

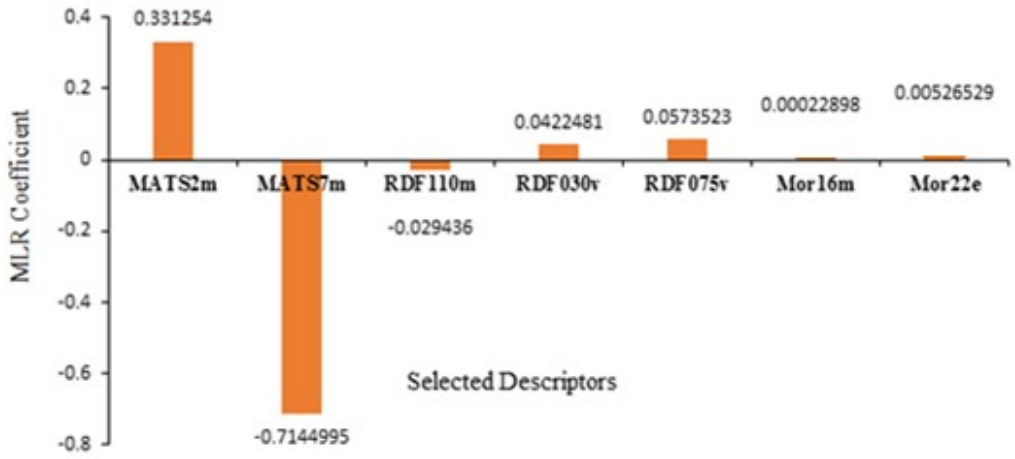


Fig. 1. Artificial neural network architecture with back-propagation algorithm.

$$R^2 = 1 - \frac{\sum_i (y_{ANN,i} - y_{\exp,i})^2}{\sum_i (y_{ANN,i} - y_m)^2} \quad (5)$$

$$R^2 adj = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right) \quad (6)$$

where $y_{ANN,i}$ and $y_{exp,i}$ are values of the predicted and experimental anticonvulsants activity of ith phenylacetanilides molecules, respectively, and y_m is the mean of y_{exp} in the above equations. Also *n* is the number of molecules in studied data set and *p* is the number of independent variables in generated model.

Minimizing the MSE of the test set is an important task for selecting the most suitable topology because it shows the ANN ability in the prediction the data which are not used during the training process.

The optimal number of neurons in hidden layers was 3 based on the minimum of the MSE of the test set (0.52918). Figure 1 illustrates the ANN architecture (7×3×1). Table 4 also shows the values of R^2, Adj, R-Square, MSE and RMSE for training, validation, and test sets using the ANN method.

Figure 2 illustrates a comparison between experimental and predicted results for fitting all data; moreover, the results for training, validation and test sets are reported in Table 5.

Table 4. Statistical Parameters of the ANN Model

Set of data	Residual sum of squares	R-Square	Adj.R-Square	MSE	RMSE
Total	0.70382	0.85403	0.84882	0.06816	0.04848
Training	0.45156	0.88175	0.87637	0.06523	0.14745
Test	0.04874	0.84641	0.69282	0.52918	0.51438
Validation	0.00027	0.99784	0.99567	0.03295	0.12835

Table 5. Statistical Parameters of the MLR Model

Set of data	Residual sum of squares	R-Square	Adj.R-Square	MSE	RMSE
Total	1.23499	0.60149	0.58555	0.09792	0.06136

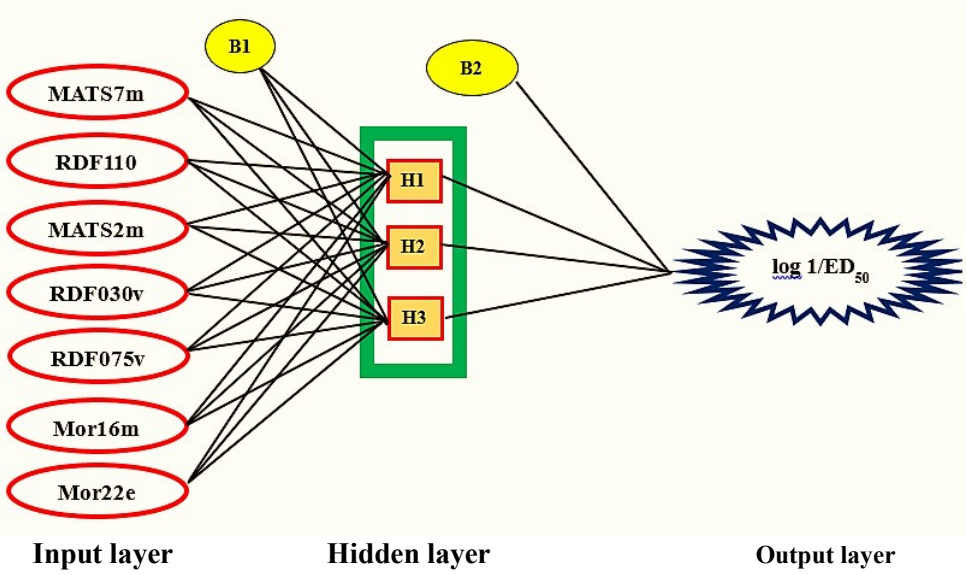


Fig. 2. The scatterplots of developed data (target) versus the ANN predicted model (output).

The correlation coefficient (R) between experimental and predicted results showed that ANN model is very efficient in detecting structure-activity correlation with predictive power.

For comparison, the MLR model also was applied to the data and the results were listed in Table 6 and shown in Fig. 3. The comparison results confirm that applying ANN method is necessary to analyze this data set because the obtained R^2 values by AAN are highly greater than those of MLR method. Therefore, ANN strategy is proposed to predit the anticonvulsants activities of the molecules in this study.

EFFECT OF INPUT VARIABLES

Weights are a series of numerical values located between the input descriptors and the hidden layer neurons, as well as between the neurons and the output layer. The

Table 6. Effective Weight Matrix for the ANN Model

Input descriptors							Hidden neurons	Hidden to out
MATS7m	RDF110	MATS2m	RDF030v	RDF075v	Mor16m	Mor22e		
0.1555	0.0379	0.8035	-0.1721	0.3571	1.1969	0.5381	H1	-2.3402
-0.5369	-1.9617	-1.5518	-0.0599	0.0019	-3.3851	-0.0241	H2	-1.7782
-0.1022	0.9827	-0.1324	0.1791	-1.0791	0.3175	-1.3012	H3	-1.4142
5.35	20.047	16.72	2.76	9.667	32.93	12.52	Relative importance (%)	

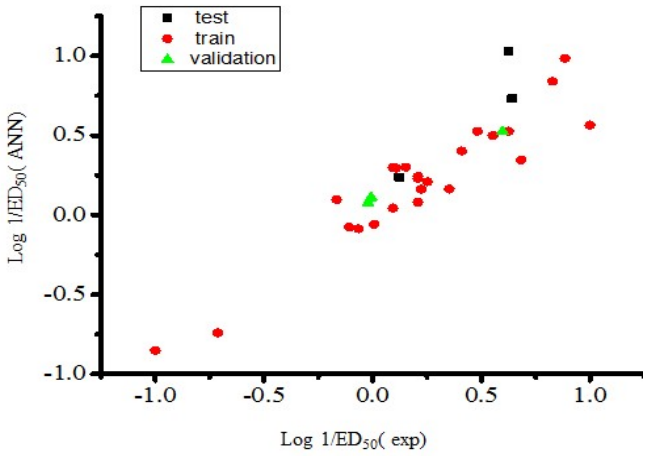


Fig. 3. The scatterplots of developed data (target) *versus* the MLR predicted model (output).

neural weight matrix can be used to evaluate the relative importance of each input variable on the output variable. The relative importance of each descriptor was determined by means of a numerical approach, the Garson method, as follows [1]:

$$Q_{md} = \frac{\sum_{n=1}^{h} \frac{|w_{mn} v_{nd}|}{\sum_{t=1}^{N} |w_{tn}|}}{\sum_{m=1}^{N} \sum_{n=1}^{h} \frac{|w_{mn} v_{nd}|}{\sum_{t=1}^{N} |w_{tn}|}}$$

where the weight between the m_{th} input and the n_{th} hidden unit is denoted by w_{tn} and v_{nd} is representative of the weight between the n_{th} hidden unit and the d_{th} output. In the final ANN model, the percentage of influence of the input variables on (log1/ED50) was calculated by incorporating input hidden and hidden-output connection weights. As the results reported in Table 5 show, the significance of the input descriptors increases in the following order:

Mor16m > RDF110 > MATS2m > Mor22e > RDF075v > MATS7m > RDF030v.

CONCLUSIONS

In summary, in the current study, a QSAR model was used to evaluate the relationship between the structure and the anticonvulsant activity of 30 monosubstituted phenylacetanilides. Dragon software was incorporated to calculate molecular descriptors where 7 descriptors were selected using an explained feature selection method and later were evaluated by ANN nonlinear model and MLR linear method. Correlation coefficient (R^2) and Mean squared error (MSE) of the ANN and MLR models (for the whole dataset) were 0.85, 0.06816 and 0.6, 0.09792, respectively. The results of this study support the efficiency of ANN in detecting the structure-activity correlation with high predictive power.

REFERENCES

[1] M. Fathollahi, H. Sajady, J. Therm. Anal. Calorim. 133 (2018) 1663.

[2] Y. Marrero-Ponce, J.E. Teran, E. Contreras-Torres, C.R. Garcia-Jacas, Y. Perez-Castillo, N. Cubillan, F. Peréz-Giménez, J.R. Valdés-Martini, J. Theor. Biol. 485 (2020) 110039.
[3] J.H. Shin, B.H. Lee, S.K. Lee, Bull Korean Chem Soc. 40 (2019) 819.
[4] A. Nasonov, Russ. J. Gen. Chem. 80 (2010) 2541.
[5] D.D. Robinson, P.J. Winn, P.D. Lyne, W.G. Richards, J. Med. Chem. 42(1999) 573.
[6] G.E. Kellogg, S.F. Semus, D.J. Abraham, J. Comput. Aided Mol. Des. 5 (1991) 545.
[7] O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, Analusis. 28 (2000) 637.
[8] L. Tarko, O. Ivanciuc, MATCH Commun. Math. Comput. Chem. 44 (2001) 201.
[9] F. Prado-Prado, X. García-Mera, M. Escobar, E. Sobarzo-Sánchez, M. Yañez, P. Riera-Fernandez, H. González-Díaz, Eur. J. Med. Chem. 46 (2011) 5838.
[10] A. Balaban, S AR QSAR Environ. Res. 8 (1998) 1.
[11] A.T. Balaban, From Chemical Topology to Three-dimensional Geometry, Springer Science & Business Media, New York, 2006.
[12] S.D. Bolboaca, L. Jantschi, M.V. Diudea, Curr. Comput. Aided. Drug. Des. 9 (2013) 195.
[13] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley, New York, 2000.
[14] R. Todeschini, V. Consonni (Eds.), Handbook of Molecular Descriptors, John Wiley & Sons, New York, 2008.
[15] A. Valadkhani, M. Asadollahi-Baboli, A. Mani-Varnosfaderani, J. Braz. Chem. Soc. 26 (2015) 619.
[16] B. Hemmateenejad, R. Miri, M. Tabarzad, M. Jafarpour, F. Zand, J. Mol. Struc.-Theochem. 684 (2004) 43.
[17] L.L. Hohrenk, M. Vosough, T.C. Schmidt, Anal. Chem. 91(2019) 9213.
[18] G. Ghasemi, A. Nemati Rashtehroodi, Anal. Bioanal. Chem. Res. 3 (2016) 253.
[19] M. Du, D. Zhang, Y. Hou, X. Zhao, Y. Li, Int. J. Environ. Res. Public Health. 16 (2019) 4156.
[20] S. Kim, K.H. Cho, Bull Korean Chem Soc. 40 (2019) 39.
[21] W. Ma, Y. Wang, D. Chu, H. Yan, J. Mol. Graph. Model. 92 (2019) 357.
[22] N. Bondarev, Russ. J. Gen. Chem. 89(2019) 1438.
[23] S.P. Bhujbal, S. Keretsu, S.J. Cho, Bull. Korean Chem. Soc. 40 (2019) 796.
[24] S.S. So, M. Karplus, J. Med. Chem. 39(1996) 5246.
[25] M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, J. Chem. Inf. Comput. Sci. 38 (1998) 726.
[26] H. Kavianpour, M. Vasighi, Amino Acids 49(2017) 261.
[27] T.S. Lee, I.F. Chen, Expert. Syst. Appl. 28(2005) 743.
[28] S. Malik, R.S. Bahare, S.A. Khan, Eur. J. Med. Chem. 67 (2013) 1.
[29] C. Salomé, E. Salomé-Grosjean, K.D. Park, P. Morieux, R. Swendiman, E. DeMarco, J.P. Stables, H. Kohn, J. Med. Chem. 53 (2010) 12885.
[30] S. Ulloora, A.V. Adhikari, R. Shabaraya, Chin. Chem. Lett. 24 (2013) 853.
[31] S. Bhutoria, N. Ghoshal, QSAR Comb. Sci. 27 (2008) 876.
[32] J.C.G. Martinez, P.R. Duchowicz, M.R Estrada, E.A.Castro, Match.-Commun. Math. Co. 67(2012) 745.
[33] L. Bertolaccini, P. Solli, A. Pardolesi, A. Pasini, J. Thorac. Dis. 9 (2017) 924.
[34] M.J. Willis, G.A. Montague, C. Di Massimo, M.T. Tham, A.J. Morris, Automatica. 28 (1992) 1181.
[35] M. Fernández, J. Caballero, A. Tundidor-Camba, Bioorganic Med. Chem. 14 (2006) 4137.
[36] V. Ramalingam, B. Palaniappan, N. Panchanatham, S. Palanivel, Expert. Syst. Appl. 31 (2006) 159.
[37] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley, New York, 1999.
[38] R. Fletcher, Comput. J. 13 (1970) 317.
[39] D. Goldfarb, Math. Comput. 24 (1970) 23.
[40] A.T.C. Goh, Artificial Intelligence in Engineering. 9 (1995) 143.
[41] A. Iqbal, S. Aftab, Int. J. Commun. Netw. Inf. Secur. 11 (2019) 19.
[42] S.S. Nidhyananthan, V. Shenbagalakshmi, Int. J. Speech Technol. 19 (2016) 577.
[43] S. Ahmed, J. Inf. Commun. Technol. 2 (2009) 1.
[44] I.R. Fermo, T.S. Cavali, L. Bonfim-Rocha, C.L. Srutkoske, F.C. Flores, C.M. Andrade , Food Bioprod. Process. 125 (2021)181.
[45] Y. Sun, Y. Peng, Y. Chen, A.J. Shukla, Adv. Drug

Deliv. Rev. 55 (2003) 1201.

[46] S. Aber, N. Daneshvar, S.M. Soroureddin, A. Chabok, K. Asadpour-Zeynali, Desalination. 211 (2007) 87.

[47] H. Demuth, M. Beale (Eds.), Neural Network Toolbox for Use with Matlab-User'S Guide Verion 3.0, The MathWorks, Inc. California, 1993.

[48] L. Saghaie, M. Shahlaei, A. Fassihi, Res. Pharm. Sci. 8 (2013) 97.