Journal of Computational Science &Engineering 44 (2019) 1081-1093

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

QSAR Study of new synthetic oxazole-4-carboxamides as the inhibitor of GSK-3

Qingao Qin[1], Xingyuan Wang[1], Peijian Zhang[1], Hongzong Si[2*]

[1]School of Data Science and Software Engineering, Qingdao University, 266071, China

[2]Institute for Computational Science Engineering, Qingdao University, Qingdao, 266071, China

Abstract: A series of reported recently oxazole-4-carboxamides as the inhibitor of glycogen synthase kinase-3 (GSK-3) were firstly used to establish quantitative structural activity relationship models. Heuristic method was applied to select efficient molecular descriptors. At the same time, a linear model with correlation coefficient (R=0.93) and cross validation correlation coefficient (R_{CV}=0.90) was built. In addition, gene expression programing and back-propagation artificial neural network methods were carried out to establish nonlinear models. As a result, a five molecular descriptors nonlinear model with correlation coefficient (R) and mean square error (MSE) of 0.95 and 0.15 for the training set, 0.95 and 0.17 for the test set was developed by GEP. Moreover, a more robust nonlinear model with of 0.97 for the training set, 0.95 for the test set, 0.89 for the validation set and 0.96 for the whole data set was built by BPANN. As we can see, both GEP and BPANN are promising methods to explore nonlinear relationships for good statistical performance. Expectantly, these models can give assistance to discover and design new inhibitors of GSK-3 in the future.

Keywords: Neurodegenerative disease (ND); Quantitative structure activity relationship (QSAR); Gene expression programming (GEP); Back-error propagation artificial neural network (BPANN); Heuristic method (HM)

Received 3 October 2019, Revised 7 December 2019, Accepted 10 December 2019

*Corresponding Author: Hongzong Si, sihz03@126.com

1. Introduction

Glycogen synthase kinase-3 (GSK-3) is a gene family including GSK-3α and GSK-3β [1,2]. GSK-3α/β are widespread multifunctional kinases that phosphorylate many substrates on serine/ threonine residues [3]. In past years, GSK-3 has been indicated to play a critical role in many physiological processes. It has vital functions in many disorders and diseases, such as neurological disorders (Alzheimer's disease, amyotrophic lateral sclerosis) [4,5], cancer [6], metabolic disorders (heart disease, diabetes) [7,8], immune disorders [9]. Therefore, GSK-3 may be a therapeutic target for neurodegenerative disease, many cancers and other diseases [3]. Over several years, more and more GSK-3 inhibitors have been reported to treat neurodegenerative disease [10-14]. In fact, it is not sure that GSK-3 is a feasible target in neurodegenerative diseases in vivo for lack of suitable molecular tools. Thus, highly potent, selective, and brain penetrant inhibitors of GSK-3 are urgently needed [3]. However, the traditional method to synthesize new inhibitor molecules and test their biological activities has a long experimental period and huge human and financial expenditure. Thus, there is an urgent requirement for an efficient approach by computer to guide and help the study of novel drugs.

QSAR, a predictive tool, is used to establish computer-aided models which try to find a statistically significant correlation between a variety of molecular properties of molecule structures and their experimental biological activities. This method uses statistical techniques and optimizes new lead molecules to predict the biological activity of compounds [15,16]. As QSAR prediction can be rapidly performed for massive inventories of substances, it is very suitable for screening and discovering new drugs [17]. There are two kinds of methods (linear and nonlinear) for building QSAR models. Common linear methods include multiple linear regression (MLR), partial least squares (PLS) and principal component regression (PCR), heuristic method (HM), etc. We chose HM to select the optimal subset of molecular descriptors and establish linear models for its fast speed without limits on the number of descriptors. Nevertheless, the relationship between the structure and the activity is generally complex and nonlinear. Since the late 1980s, machine learning (ML) methods have been used to build nonlinear QSAR models such as artificial neural networks (ANN), support vector machines (SVM) and Random Forest (RF) for more predictive ability [18,19]. In this work, we used two ML methods, GEP and BPANN algorithms, to build nonlinear QSAR models.

GEP [20] not only has the search capability of genetic programming (GP), but also has the simplicity of genetic algorithm (GA), which makes automatic function discovery technology break through forward. Because of its excellent function mining abilities, it has been successfully applied on the prediction of cement strength and exhaust emissions [21,22].

ANN is a kind of information processing system by simulating the function of human brain and can perform massively parallel computations for data processing and knowledge representation [23,24]. ANN has many advantages in information processing, including nonlinearity, high parallelism, fault tolerance, adaptability, learning and generalization abilities, etc [25]. The BPANN, the most common multilayer neural network, is slightly influenced by data quality and

2019@ The American Computational Science Society. All rights reserved.

deviation problems for its particular back-propagation method. It is efficient for exploring nonlinear relationships [26]. Thus, the BPANN is a very popular model for artificial intelligence applications [27]. In the past years, BPANN has received widespread attention as an effective alternative method to predict the behavior of nonlinear systems [28,29]. It has been applied to predicting stand basal area [30], mechanical properties of steel strip [25], solubility of silymarin in supercritical carbon dioxide [31].

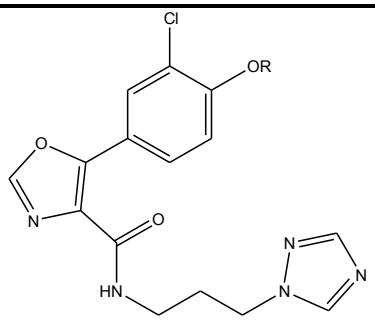
In this study, the predictive ability of the nonlinear model is better than the linear model. On the basis of same descriptors, nonlinear models were developed by GEP and BPANN. In addition, the nonlinear model by BPANN has better generalization ability and faster speed than the GEP model. So far, it is the first time that the QSAR method is applied to predicting the pIC50 of these oxazole-4-carboxamides as the inhibitor of GSK-3.

2. Experiment

2.1 Data set and its division

The 49 compounds were obtained from literature [32] and listed in Table 1, Table 2 and Table 3. The bioactivity of these compounds represented in IC_{50} (nm) was transformed into pIC_{50} ($= -log_{10}(IC_{50})$). In order to get more robust nonlinear models, the data set was divided into a training set of 39 compounds and a test set of 10 compounds at a approximate ratio of 4:1 in GEP. The test set was selected from the data set in multiples of 5. In BPANN, the 39 compounds in training set was further randomly divided into a new training set of 10 compounds and a validation set of 5 compounds. The training set was used to build models and the test set was used to evaluate the predictive capability of models.

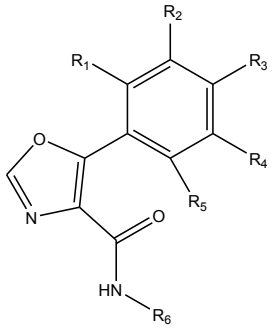
Table 1. The pIC_{50} values of compounds 1-5

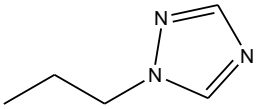


No.	R	pIC_{50}
1	CH_3	-3.010
2	H	-0.6532
3	CH_2CH_2F	-1.2625
4	$CH_2CH_2CH_2F$	-1.6730
5[T]	Bn	-2.4462

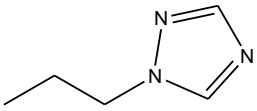
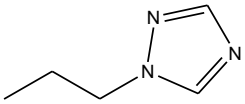
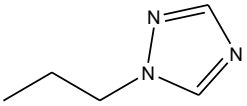
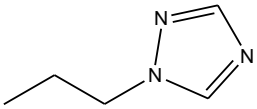
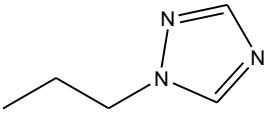
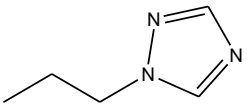
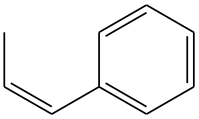
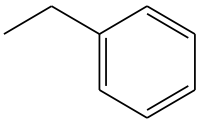
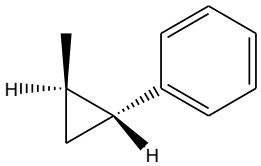
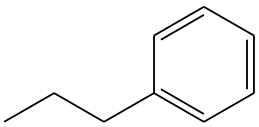
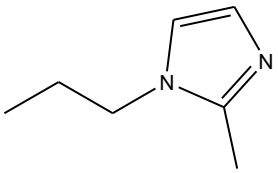
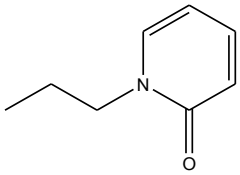
[T] is test set for BPANN and GEP.

Table 2. The pIC_{50} values of compounds 6-42

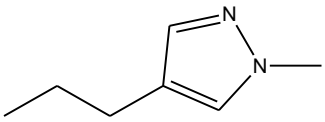
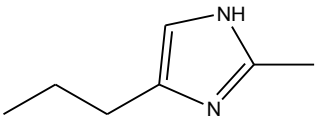
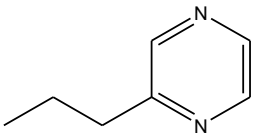
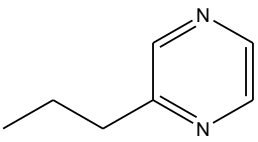
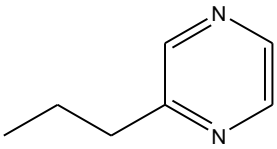
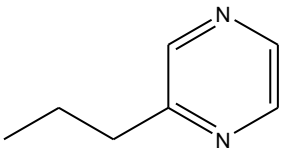
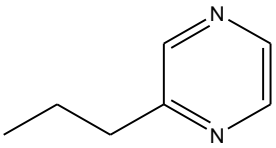
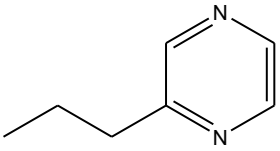
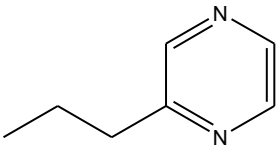
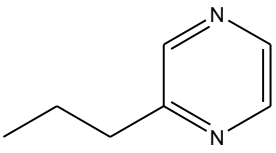
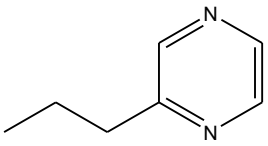


No.	R_1	R_2	R_3	R_4	R_5	R_6	pIC_{50}
6	H	F	F	H	H		-2.4053

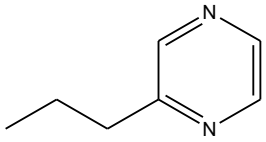
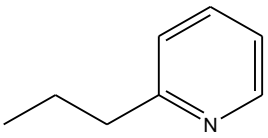
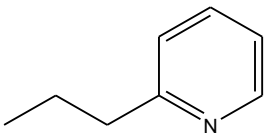
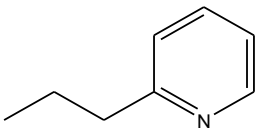
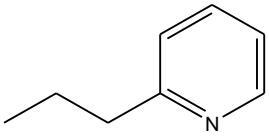
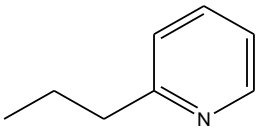
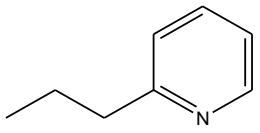
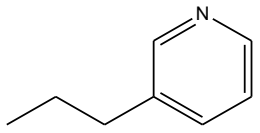
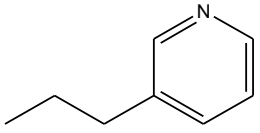
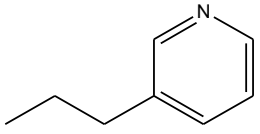
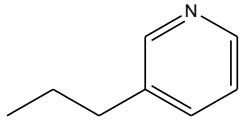
2019@ The American Computational Science Society. All rights reserved.

7	F	F	F	F	F		-2.6264
8	H	F	F	F	H		-2.6459
9	H	F	OCF_3	H	H		-2.5053
10^T	H	F	CF_3	H	H		-2.2572
11	H	F	OCH_3	F	H		-1.7716
12	H	F	OCH_3	H	H		-0.5051
13	H	F	OCH_3	H	H		-3.5835
14^V	H	F	OCH_3	H	H		-3.1629
15^T	H	F	OCH_3	H	H		-2.4019
16^V	H	F	OCH_3	H	H		-1.2279
17	H	F	OCH_3	H	H		-0.7076
18	H	F	OCH_3	H	H		-0.5563

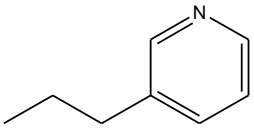
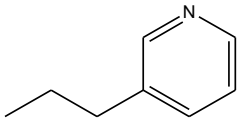
2019@ The American Computational Science Society. All rights reserved.

19	H	F	OCH_3	H	H		-1.0000
20^T	H	F	OCH_3	H	H		-1.2455
21	H	F	CF_3	H	H		-2.3886
22	H	F	F	F	H		-2.3976
23	H	F	F	F	H		-2.0927
24	H	F	OCH_3	F	H		-0.9731
25^T	H	F	OH	H	H		-0.6990
26	H	F	OCH_3	H	H		-0.3979
27	H	F	OCF_3	H	H		-3.6134
28	H	F	CN	H	H		-1.6031
29	H	CN	F	H	H		-1.7251

2019@ The American Computational Science Society. All rights reserved.

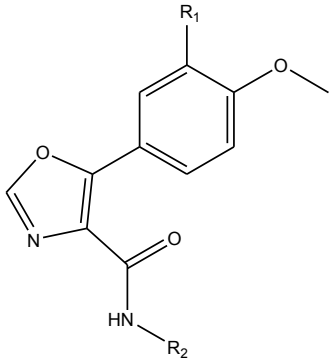
30^{T}	H	I	OCH_3	H	H		0.2218
31	H	F	CF_3	H	H		-2.9826
32	H	F	F	F	H		-2.3201
33^{V}	H	F	F	H	H		-2.2550
34	H	F	OCH_3	F	H		-1.0828
35^{T}	H	F	OCH_3	H	H		-0.1761
36	H	F	OH	H	H		-0.6232
37	H	F	CF_3	H	H		-1.7868
38	H	F	F	F	H		-2.6887
39	H	F	F	H	H		-1.7896
40^{T}	H	F	OCH_3	F	H		-0.8751

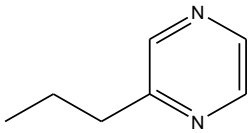
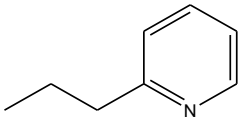
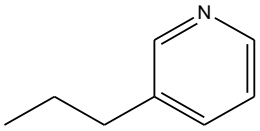
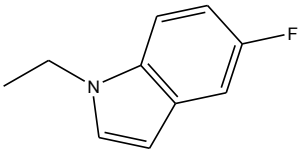
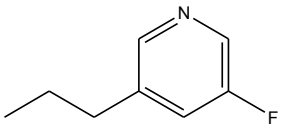
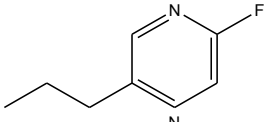
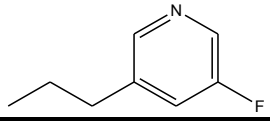
2019@ The American Computational Science Society. All rights reserved.

41[V]	H	F	OH	H	H		-0.3617
42	H	F	OCH_3	H	H		-0.3010

[V] is validation set for BPANN, [T] is test set for BPANN and GEP.

Table 3. The pIC_{50} values of compounds 43-49



No.	R_1	R_2	pIC_{50}
43	Cl		-0.1461
44	Cl		0.0458
45[T]	Cl		0.0458
46	Cl		-2.7298
47	Cl		0.3979
48[V]	Cl		0.0969
49[T]	I		1.5086

[V] is validation set for BPANN, [T] is test set for BPANN and GEP.

2019@ The American Computational Science Society. All rights reserved.

2.2 Calculation of the descriptors

The chemical structure of the compound was drawn in Chemdraw Ultra 8.0 software and roughly pre-optimized by Hyperchem [33] using MM+ molecular mechanics force field. To reach the lowest energy steady state of the compound, a more precise optimization was executed with semi-empirical AM1 method [34]. The molecular structure was optimized by using the Polak–Ribiere algorithm till the root mean square gradient became 0.01. The MOPAC and Hyperchem output files with pIC_{50} were used in CODESSA software to calculate 5 groups of descriptors including constitutional, electrostatic, topological, geometrical and quantum-chemical [35].

2.3 Linear models by Heuristic method (HM) [35]

HM which has been implemented in CODESSA was used to calculate molecular descriptors and build linear models. It has no software limits on the size of the data set and fast speed. The steps of HM are as follows: Firstly, all descriptors are checked to ensure that values are available for every structure and there is a variation in these values. In other words, descriptors whose values are not available or same for all structures in data set are removed. Secondly, the one-parameter models are built and descriptors are further screen by eliminating insignificant ones. Next, the selection of the best correlations is carried out and the number of descriptors is further decreased by removing highly relevant descriptors. The details of this step are: (a) all possible two-parameter models are tested by combining remaining descriptors. (b) 10 best pairs by the criteria in the two-parameter models are selected and processed further in next step. (c) For a correlation equation with n descriptors, a new descriptor which is lowly correlated with the previously selected n descriptors is used to obtain a correlation equation with n+1 descriptors. This step is repeated until the number of descriptors reaches the preset value. The criteria for evaluating model are the square of correlation coefficient (R^2), the F-test value (F), the t-test value (t), the square of standard error (s^3) and the square of cross-validation coefficient (R^2_{cv}). As a result, a linear model with 5 descriptors which were selected from 568 descriptors was developed by HM.

To evaluate the predictive ability of models, the Leave-One-Out Cross-Validation (LOOCV) technique is well implemented as follows: (1) For every experimental data point, the linear model is rebuilt by using the same descriptors for the data set without this point. (2) Then, the built model is used to predict the corresponding value of this data point. (3) Finally, the obtained array of predicted data points is linearly correlated with the array of experimental data points and the correlation coefficient of the last correlation is the cross-validation correlation coefficient.

However, it is worth noting that the factors influencing pIC_{50} of compounds are usually intricate. Thus, linear models are not enough to discover their correlation. Obviously, the nonlinear model is needed.

2.4 Nonlinear models by GEP

GEP [36], a new genetic algorithm based on genome and phenotype, is invented by Ferreira. It is essentially different from GA and GP in the essence of the individuals they using: in GA, individuals are fixed-length linear strings (chromosomes); in GP, individuals are nonlinear entities with different lengths and shapes (split tree). Then, in GEP, individuals are first encoded into a fixed-length linear string (genome or chromosome), and then expressed as nonlinear entities (simple graph representation or expression tree). Due to the combination of the advantages of GA and GP, GEP is 2-4 orders of magnitude more efficient than traditional GP methods for solving complex problems. Figure 1 shows the steps of GEP algorithm.

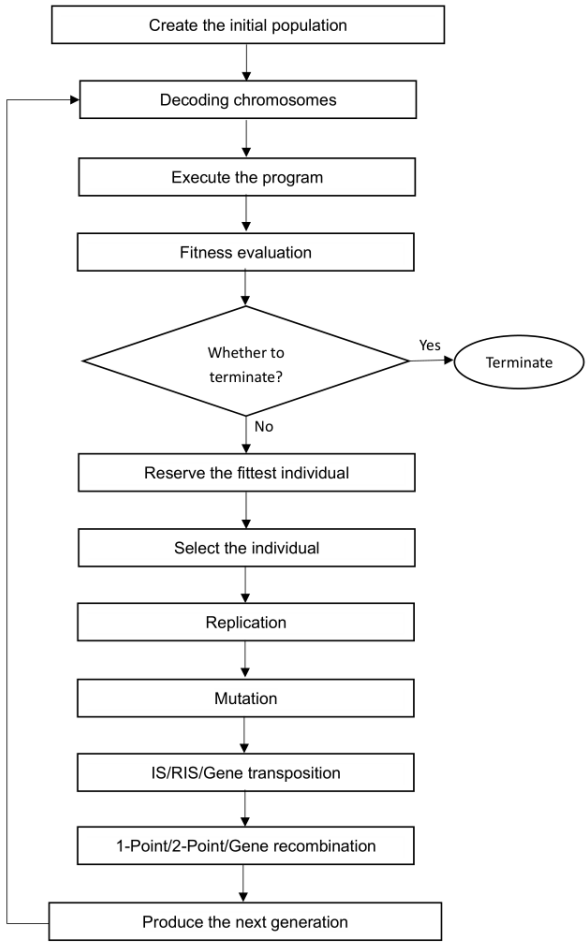


Figure 1. The flow chart of GEP

In the algorithm, the generation of an initial population refers to generating an initial population that consists of several individuals representing different solutions. The selection operator acts on the population. According to Darwin's the "survival of the fittest" principle, individuals with high fitness have a higher chance of survival. The genetic operator acts on the population and carries on genetic actions among the individuals of the

2019@ The American Computational Science Society. All rights reserved.

population to generate new offspring individuals. In GA, Genetic operators include 2 groups, recombination and mutation respectively. In GEP, in addition to single-point recombination, double-point recombination and single-point mutation those are similar to GA, genetic operators with unique actions and meanings such as insertion sequence (IS) and root insertion sequence (RIS). These genetic operators act on special genetic coding methods which form distinctive GEP [37]. The implementation techniques of GEP [38] mainly include the coding method, genetic operators, IS operations, recombination operators, and fitness function selection, etc. The brief introduction of the more important ones is shown below [37]:

1) The method of coding

The gene of GEP consists of head and tail. The head can contain both function and endpoint symbols while the tail can only contain endpoint symbols. For each problem, the following relationship between tail length t and head length h should be satisfied:

$$\boldsymbol{t = h(n-1)+1} \quad (1)$$

2) Fitness function

All evolutionary computation algorithms need to be evaluated by the problem-solving ability of the representative individual. In GEP, since the solution is usually a program, it is exactly an expression in many applications. The evaluation of the expression is to evaluate the consistency between the data calculated by the expression and the training data. There are two evaluation models proposed by Ferreira [36].

$$\boldsymbol{fitness = \sum_{i=1}^{n}\left(M - \left|\frac{y_i - \hat{y}_i}{y_i}\right|\right)} \quad (2)$$

$$\boldsymbol{fitness = \sum_{i=1}^{n}(M - |y_i - \hat{y}_i|)} \quad (3)$$

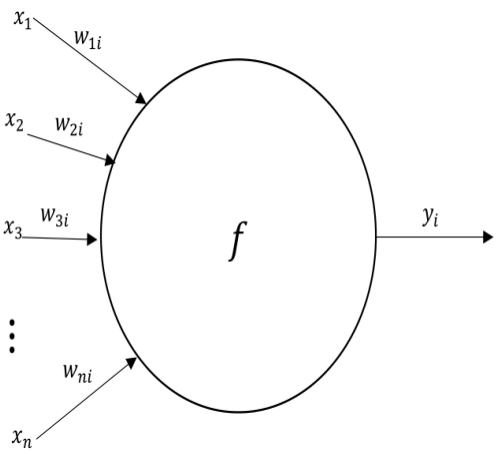


Figure 2. The Schematic of artificial neuron.

Indeed, the evaluation is made using relative or absolute errors. Both evaluation models have their own inherent disadvantages. Therefore, we selected mean squared error (MSE) as fitness function.

$$\boldsymbol{fitness = MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Let T be the training data set, containing n sets of data. T_i represents the input of i_{th} group data in the training set. $\boldsymbol{y_i}$ represents the experimental data corresponding to T_i, and $\hat{y}_i$ is the estimated value of $\boldsymbol{y_i}$ calculated from T_i using the formula. M is a constant. For more details about GEP, please refer to the literature [36]. In addition, we used software Automatic Problem Solver (APS) [39] which has implemented GEP algorithm to build nonlinear models.

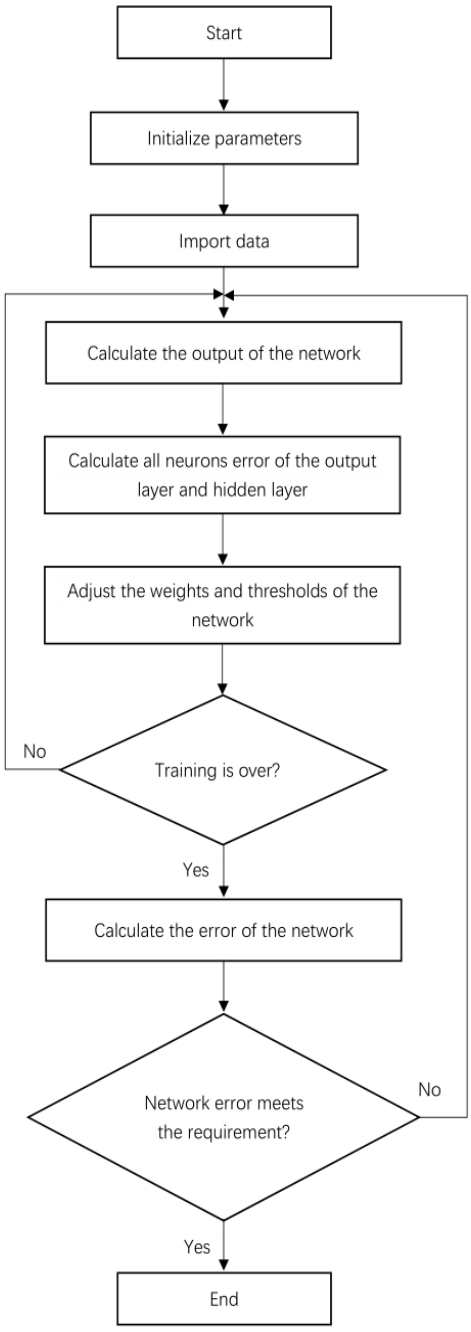


Figure 3. The Training flowchart of BPANN.

2019@ The American Computational Science Society. All rights reserved.

2.5 Nonlinear models by BPANN

BPANN, a supervised network, has been widely used for its excellent capability to train quickly on sparse data sets [31]. It is composed of adaptive simple information processing units called artificial neurons or nodes (Figure 2), which are closely connected with each other [25].

In general, it consists of three parts, namely the input layer, the output layer and the hidden layer which can be composed of multiple layers. However, some studies have indicated that there is rarely an advantage to using more than one hidden layer [26,40]. As a result, we trained several supervised feedforward neural networks with only one hidden layer containing several hidden nodes [30]. In fact, the process of BPANN is feedbacking the difference between the calculated output vector and the known training sample vector to the network and adjusting the strength of the connection (or weight coefficient) between layers to minimize the network error. The BPANN training flowchart is shown in Figure 3.

Please refer to literature [26] for more details about BPANN. There are various computer software packages for ANN analysis. We chose MATLAB [41] integrating many neural network algorithms for machine language program network. Three-layer BPANN (Figure 4) was decided to perform nonlinear regression [42].

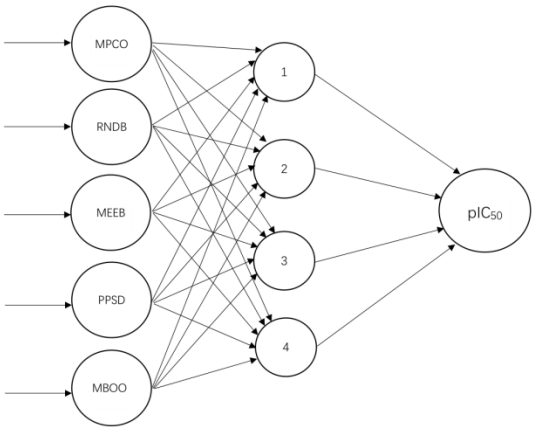


Figure 4. The structure of the 5-4-1 BPANN.

3. Results and discussion

3.1 Results of HM

565 molecular descriptors for all compounds were calculated by CODESSA software. To find the best linear model, 5 linear models with 1-5 descriptors respectively are built by HM. The R^2, R^2_{cv} and s^2 of these models are shown in Figure 5.

$$pIC_{50} = -3.6274 - 1.2602 \times 10^2 MPOC + 2.9923 \times 10 RNDB - 1.5720 MEEB + 2.7989 PPSD - 5.8205 \times 10^1 MBOO \quad (5)$$

It is easy to see that R^2 and R^2_{cv} increases when the number of descriptors increases. Instead, s^2 decreases when the number of descriptors increases. The QSAR models developed by using too many variables usually have poor generalization in statistical terms [43]. Therefore, the model with 5 descriptors is chosen as the best linear model (Figure 6) for the good statistical characteristics and low cardinality comparatively.

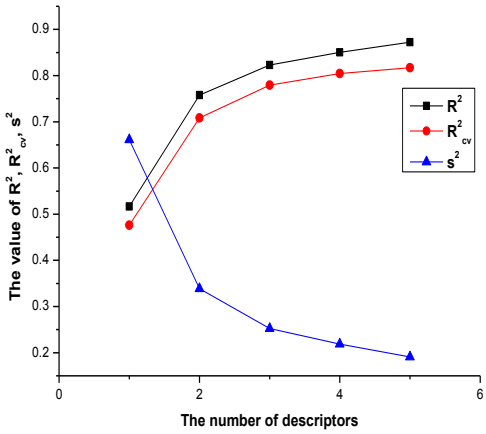


Figure 5. The influence of the number of descriptors on the correlation coefficient (R^2) the cross-validation correlation coefficient (R^2_{cv}) and the square of standard error (s^2) of linear models.

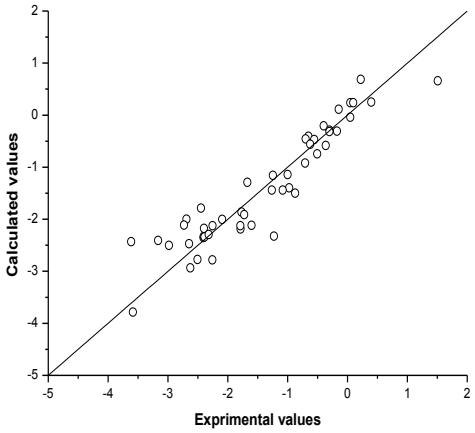


Figure 6. The plot of calculated pIC_{50} versus experimental values for all compounds by HM.

The involved descriptors and their physical-chemical meaning are shown in Table 4. Table 5 shows the correlation matrix of these descriptors. As we can see, the correlation coefficient of any two different descriptors is below 0.8 which indicates they are generally independent. The R, R_{cv}, F and s^2 of this model are 0.93, 0.90, 58.73 and 0.19. In general, the result is satisfactory but not good enough. Thus, we imported the corresponding descriptors with response variable (pIC_{50}) into APS and MATLAB to develop nonlinear models.

2019@ The American Computational Science Society. All rights reserved.

Table 4. The physical-chemical meaning, coefficient, DX and t-test values for the selected molecular descriptors of the best linear model by HM.

Descriptors	Physical-Chemical meaning	Coefficient	DX	t-test
Constant	Intercept	-3.6274	5.9207	-0.6127
MPOC	Min partial charge for a O atom [Zefirov's PC]	-126.02	9.1398	-13.788
RNDB	Relative number of double bonds	29.923	3.5546	8.4181
MEEB	Max exchange energy for a C-C bond	-1.572	0.2889	-5.4414
PPSD	Polarity parameter / square distance	2.7989	0.8963	3.1228
MBOO	Min (>0.1) bond order of a O atom	-58.205	21.323	-2.7297

Table 5. The correlation matrix of the descriptors in the best linear model

Descriptors	MPOC	RNDB	MEEB	PPSD	MBOO
MPOC	1	0.4614	-0.1716	0.0172	0.3075
RNDB		1	-0.2851	0.1145	0.1128
MEEB			1	-0.0245	-0.5429
PPSD				1	0.0298
MBOO					1

3.2 Results of GEP

To acquire a more robust model, the nonlinear models were developed by GEP in APS. We tried many function sets of GEP to build models. As a result, the optimal model with $R = 0.95, MSE = 0.15$ for the training set and $R = 0.95, MSE = 0.17$ for the test set was obtained. The mathematical formula of this model was shown as Eq (6).

$$pIC_{50} = ln(x_1) + \frac{x_3 + ln(x_1) - \frac{x_4}{x_4}}{\cos(x_3)} + \left(\tan\left(\frac{1}{\tan(\sin(x_4))}\right) + x_3 - x_0\right) + x_3 + tan\left(tan\left(\frac{1}{sin(x_1)\times(x_2 - x_2) + \tan(x_0)}\right)\right) + \tan(tan(\frac{1}{x_0})) + tan\left(\frac{1}{\tan(ln(x_2))}\right) + \left(\frac{\ln(x_4)}{\cos(\lg(x_1))}\%lg\left(\frac{1}{\sin(x_1)}\right)\right) + e^{\frac{x_3}{tan(cos(lg(\frac{1}{tan(x_3)})))}} + tan(\frac{1}{tan(cos(cos(\frac{1}{tan(ln(x_2))})))}) \quad (6)$$

x_0, x_1, x_2, x_3 and x_4 represent the descriptors of MPOC, PNDB, MEEB, PPSD and MBOO respectively (Table 2). The parameter setting of this model was listed in Table 6.

Table 6. The parameter setting for GEP

Parameter names	Values
Number of Chromosomes	100
Number of Genes	10
Function set	'+', '-', '*', '/', 'Mod', 'Sqrt', 'Exp', 'Ln', 'Log', 'Inv', 'Sin', 'Cos', 'Tan', 'Csc', 'Sec', 'Cot'
Gene head size	8
Linking function	+
Mutation rate	0.044
1-Point recombination rate	0.3
2-Point recombination rate	0.3
Gene recombination rate	0.1
IS transposition rate	0.1
RIS transposition rate	0.1
Gene transposition rate	0.1
Selection range	100
Precision	0.01

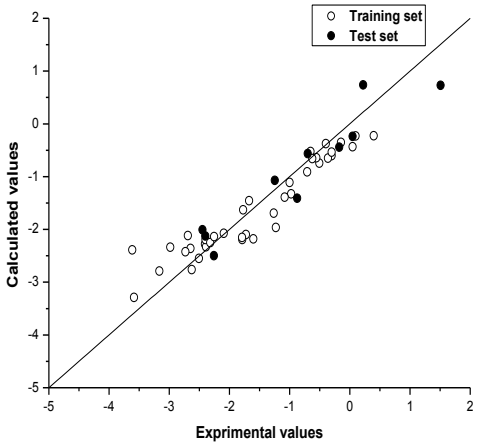


Figure 7. The plot of calculated pIC_{50} versus experimental values for the training set and test set by GEP.

2019@ The American Computational Science Society. All rights reserved.

The predictive result of GEP model was shown in Figure 7. As we can see, most of predicted values are well fitted with experimental values. From above, this model performed better than the linear model by HM.

3.3 Results of BPANN

To obtain models with more predictive and accurate capability, BPANN was trained according to Levenberg-Marquardt (LM) rule [44]. We imported 5 same descriptors from CODESSA as the input layer and the pIC_{50} of compounds as the output layer. To prevent larger numbers from overriding smaller ones and avoid premature saturation of hidden nodes which play a key role in learning process, the data was normalized within a uniform range ([0,1]) [30]. We combined the rules proposed by Andrea [45] and Tetko [46], the number of hidden layer nodes (H) was defined as:

$$\boldsymbol{M = (I + 1)H + (H + 1)O} \ (7)$$

where M is total weight of the network, I is the number of input layer nodes and O is the number of output layer nodes.

$$\boldsymbol{1.4 \le \rho\left(=\frac{N}{M}\right) < 2.2} \ (8)$$

where N is the number of samples.

In this study, $I = 5$, $O = 1$, so $3.0 < H <= 4.8$, the only permitted value of H is 4. Therefore, the structure of BPANN is 5-4-1 as shown in Figure 4. The activation function of network is Sigmoid function (Eq (9)) for hidden layer and simple linear function (Eq (10)) for output layer.

$$\boldsymbol{f(x) = \frac{1}{1 + e^{-x}}} \ (9)$$

$$\boldsymbol{f(x) = x} \ (10)$$

Training steps were set to a maximum of 1000 steps. Learning rate and training error were 0.01 and 1e-4, respectively. Moreover, the data set was randomly divided into three parts for training, validation and testing at ratio 70:10:20 (Table 1,2,3). In the end, a robust nonlinear model (Figure 8) with $R_{training} = 0.97$, $R_{test} = 0.95, R_{cv} = 0.89, R_{all} = 0.96$ was established. The weight and threshold of the BPANN was shown in Table 7.

Table 7. The weight and threshold of the BPANN (IW: the weight between input layer and hidden layer, LW: the weight between hidden layer and output layer)

	Weight					Thresold
IW	-0.0661	1.2645	1.4563	-0.5306	-1.1420	1.8097
	0.5981	-0.1638	1.9594	-1.3045	0.1353	-0.0779
	1.3317	0.2154	-0.4065	0.0843	0.7454	0.9211
	0.7299	1.0855	1.2818	0.0225	-0.0449	2.0189
LW	0.3498	-1.1632	-0.8095	0.1848		-0.4655

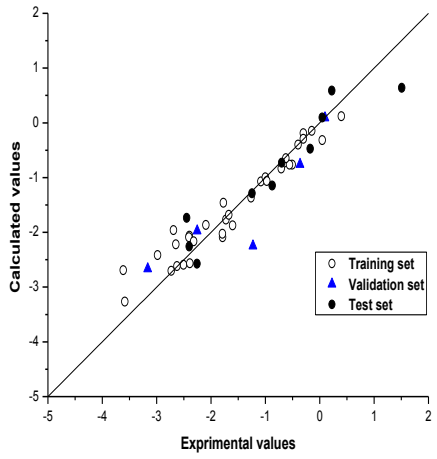


Figure 8. The plot of calculated pIC50 versus experimental values for the training set, validation set and test set by BPANN.

3.4 Discussion of the selected descriptors

To obtain some insight into factors influencing the pIC_{50}, more detailed discussion is carried out below. The 5 selected descriptors are sorted by the absolute value of t-test ($|t|$) which indicates the significance of the descriptor: MPCO > RNDB > MEEB > PPSD > MBOO (Table 4). As we can see, MPCO, RNDB and MEEB are the most significant descriptors.

(a) RNDB, a constitutional descriptor, is relative number of double bonds [47]. Due to its the max positive coefficient, the pIC_{50} has the biggest increase when RNDB increases.

(b) MPCO and PPSD are electrostatic descriptors. MPCO represents min partial charge for a O atom in the molecule. PPSD is a polarity parameter which indicates the polarity of the molecule. It is calculated as follows [48]:

$$\boldsymbol{P = \frac{(Q_{max} - Q_{min})}{R_{mm}^2}} \ (11)$$

Here, Q_{max} is the most positive atomic partial charge in the molecule, Q_{min} is the most negative atomic partial charge in the molecule and R_{mm} is distance between the

2019@ The American Computational Science Society. All rights reserved.

most positive and the most negative atomic partial charges in the molecule. Moreover, the empirical partial charges in the molecule are calculated by using Zefirov's method [49,50].

$$Q_i = f(\chi_i) \quad (12)$$

$$\chi_i = (\prod_{k=1}^{n} \chi_k)^{1/(n+1)} \quad (13)$$

Where, χ_i is electronegativities of isolated atoms and n is the number of atoms in the first coordination sphere of a given atom. Since the coefficient of PPSD is positive, the increasing of the value of PPSD leads to an increase in the pIC_{50}. However, MPCO has the opposite influence on pIC_{50} for its negative coefficient. In addition, whether the largest $|t|$ or the smallest negative coefficient indicate that MPCO plays a key role in the pIC_{50}. It's further explained that the electric property of molecule has a critical influence on inhibiting GSK-3.

(c) Both MEEB and MBOO are quantum chemical descriptors. MEEB is calculated as follows:

$$E_{exc}(AB) = \sum_{\mu,\nu \epsilon A} \sum_{\lambda,\sigma \epsilon B} P_{\mu\nu} P_{\lambda\sigma} \langle \mu\nu|\lambda\sigma\rangle \quad (14)$$

Where, A is given atomic species, B is another atomic species, $P_{\mu\nu}$, $P_{\lambda\sigma}$ are the density matrix elements over atomic basis $\langle \mu\nu\lambda\sigma\rangle$ and $\langle \mu\nu|\lambda\sigma\rangle$ are the electron repulsion integrals on atomic basis $\{\mu\nu\lambda\sigma\}$ [51]. MBOO is a kind of bond order and calculated as:

$$P_{AB} = \sum_{i=1}^{occ} \sum_{\mu \epsilon A} \sum_{\nu \epsilon B} n_i c_{i\mu} c_{j\nu} \quad (15)$$

Here, A is given atomic species, B is another atomic species, occ is the number of molecular orbital (MO), n_i is the occupation number of the i_{th} MO and $c_{i\mu}$, $c_{j\nu}$ is MO coefficients for atomic orbitals m and n [52,53]. Both MEEB's and MBOO's coefficient is negative. Thus, when their values increase, the pIC_{50} and activity decreases.

From above, it can be concluded that the molecule with higher value of RNDB, PPSD and lower value of MPCO, MEEB, MBOO has higher activity and better inhibition for GSK-3. As Table 1,2,3 shows, most of molecules with Cl atoms have a higher pIC_{50} value. The only 2 molecules with I atom all have relatively high activity. Therefore, the molecule with similar structure and property may be an efficient inhibitor for GSK-3.

3.5. The BPANN model compare with GEP model

The comparison between BPANN model and GEP model is shown in Table 8. As we can see, the R and MSE of BPANN model are 0.97 and 0.0048 in training set, 0.95 and 0.0064 in test set. In the meantime, the R and MSE of GEP model are 0.95 and 0.1452 in training set, 0.95 and 0.1679 in test set. Obviously, the BPANN model has better statistical performance than the GEP model in both R and MSE. Therefore, the BPANN is a more promising nonlinear method.

Table 8. The R and MSE of models by BPANN and GEP

Method	Training test		Test set	
	R	MSE	R	MSE
BPANN	0.97	0.0048	0.95	0.0064
GEP	0.95	0.1452	0.95	0.1679

From above, the physical chemical property of the molecule plays a key role in influencing the pIC50. The 5 selected descriptors represent the most important physical chemical properties for predicting the pIC50 of oxazole-4-carboxamides as GSK-3 inhibitors.

4. Conclusion

The electric property of molecule is a key point to develop new GSK-3 inhibitors. Due to its global optimality of convergence and powerful prediction ability, GEP is a good method for nonlinear regression and function estimation. However, BPANN has better generalization ability for its capability to identify hidden patterns in data. Therefore, BPANN is a more promising method to discover the nonlinear relationship. Moreover, the mentioned models can provide some guidance for further design of new inhibitors of GSK-3 for its efficiency and accuracy.

References

[1] K. MacAulay, B. W. Doble, S. Patel, T. Hansotia, E. M. Sinclair, D. J. Drucker, A. Nagy, J. R. Woodgett. Cell metabolism, 6 (2007) 329-337.

[2] Kaidanovich-Beilin, T. V. Lipina, K. Takao, M. V. Eede, S. Hattori, C. Laliberté, M. Khan, K. Okamoto, J. W. Chambers, P. J. Fletcher, K. MacAulay, B. W. Doble, M. Henkelman, T. Miyakawa, J. Roder, J. R. Woodgett. Molecular brain, 2 (2009) 35.

[3] J. A. McCubrey, D. Rakus, A. Gizak, L. S. Steelman, S. L. Abrams, K. Lertpiriyapong, T. L. Fitzgerald, L. V. Yang, G. Montalto, M. Cervello, M. Libra, F. Nicoletti, A. Scalisi, F. Torino, C. Fenga, L. M. Neri, S. Marmiroli, L. Cocco, A. M. Martelli. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1863 (2016) 2942-2976.

[4] V. Palomo, D. I Perez, C. Gil, A. Martinez. Current medicinal chemistry, 18 (2011)

2019@ The American Computational Science Society. All rights reserved.

3028-3034.
[5] V. Forlenza, V. J. de Paula, R. Machado-Vieira, B. S. Diniz, W. F. Gattaz. Drugs & aging, 29 (2012) 335-342.
[6] T. Shimura. Journal of radiation research, 52 (2011) 539-544.
[7] H. Cheng, J. Woodgett, M. Maamari, T. Force. Journal of molecular and cellular cardiology, 51 (2011) 607-613.
[8] R. H. Belmaker. Current pharmaceutical design, 17 (2011) 2264-2277.
[9] R. S. Jope, C. J. Yuskaitis, E. Beurel. Neurochemical research, 32 (2007) 577-595.
[10] P. Cohen, M. Goedert. Nature reviews Drug discovery, 3 (2004) 479-487.
[11] Martinez, A. Castro, I. Dorronsoro, M. Alonso. Medicinal research reviews, 22 (2002) 373-384.
[12] H. Eldar-Finkelman, A. Martinez. Frontiers in molecular neuroscience, 4 (2011) 32.
[13] M. Alonso, A. Martinez. Current medicinal chemistry, 11 (2004) 755-763.
[14] Martinez, D. I. Perez. Journal of Alzheimer's Disease, 15 (2008) 181-191.
[15] M. M. Al Mogren, A. Zerroug, S. Belaidi, A. BenAmor, S. D. A. Al Harbi. Journal of King Saud University-Science, 2020.
[16] T. K. S. Ahamed, V. K. Rajan, K. Muraleedharan. Food Science and Human Wellness, 8 (2019) 53-62.
[17] K. K. Chinen, K. Klimenko, C. Taxvig, N. G. Nikolov, E. B. Wedebye. Computational Toxicology, 14 (2020) 100121.
[18] R. Burbidge, M. Trotter, B. Buxton, S. Holden. Computers & chemistry, 26 (2001) 5-14.
[19] D. T. Manallack, D. J. Livingstone. European Journal of Medicinal Chemistry, 34 (1999) 195-208.
[20] C. Ferreira. Springer, (2006).
[21] Baykasoğlu, T. Dereli, S. Tanış. Cement and concrete research, 34 (2004) 2083-2090.
[22] S. Bhowmik, A. Paul, R. Panua, S. K. Ghosh, D. Debroy. Fuel, 235 (2019) 317-325.
[23] Y. Hu, Y. Gu, H. Su, Z. Wang, J. Chu. Chinese Journal of Scientific Instrument, 3 (2003).
[24] A. Basheer, M. Hajmeer. Journal of microbiological methods, 43 (2000) 3-31.
[25] S. Jiang, L. Zhao, G. Wu. Journal of Wuhan University of Technology-Mater. Sci. Ed., 24 (2009) 791.
[26] D. E. Rumelhart. Parallel distributed processing, 1 (1986).
[27] D. Wu, Y. Ji. Journal of Agricultural Science, 7 (2015) 18.
[28] G. Scrinzi, L. Marzullo, D. Galvagni. Ecological Modelling, 206 (2007) 331-346.
[29] P. Dande, P. Samant. Tuberculosis, 108 (2018) 1-9.
[30] S. Che, X. Tan, C. Xiang, J. Sun, X. Hu, X. Zhang, A. Duan, J. Zhang. Journal of Forestry Research, 30 (2019) 1641-1649.
[31] G. Yang, Z. Li, Q. Shao, N. Feng. asian journal of pharmaceutical sciences, 12 (2017) 456-463.
[32] V. Bernard-Gauthier, A. V. Mossine, A. Knight, D. Patnaik, W. Zhao, C. Cheng, H. S. Krishnan, L. L. Xuan, P. S. Chindavong, S. A. Reis, J. M. Chen, X. Shao, J. Stauff, J. Arteaga, P. Sherman, N. Salem, D. Bonsall, B. Amaral, C. Varlow, L. Wells, L. Martarello, S. Patel, S. H. Liang, R. G. Kurumbail, S. J. Haggarty, P. J. H. Scott, N. Vasdev. Journal of medicinal chemistry, 62 (2019) 9600-9617.
[33] HyperChem. 4.0, Hypercube, 1994.
[34] P. P. Stewart. University of Indiana, Bloomington, (1989).
[35] R. Katritzky, V. S. Lobanov, M. Karelson. Gainesville, Florida, (1994).
[36] Ferreira. Complex System, 13 (2001) 87–129.
[37] Zuo. Sichuan University, (2004).
[38] Mitchell. Cambridge, Massachusetts. London, England, (1996).
[39] http://www.gepsoft.com/gepsoft.
[40] R. Lippmann. IEEE Assp magazine, 4 (1987) 4-22.
[41] https://www.mathworks.com/.
[42] H. Beale, M. T. Hagan, H. B. Demuth. The MathWorks, (2010).
[43] Ponzoni, V. Sebastián-Pérez, M. J. Martínez, C. Roca, C. D. C. Pérez, F. Cravero, G. E. Vazquez, J. A. Páez, M. F. Díaz, N. E. Campillo. Scientific reports, 9 (2019) 1-13.
[44] D. W. Marquardt. Journal of the society for Industrial and Applied Mathematics, 11 (1963) 431-441.
[45] T. A. Andrea, H. Kalayeh. Journal of Medicinal Chemistry, 34 (1991) 2824-2836.
[46] V. Tetko, D. J. Livingstone, A. I. Luik. Journal of chemical information and computer sciences, 35 (1995) 826-833.
[47] http://www.codessa-pro.com/descriptors/const.htm.
[48] Ośmiałowski, J. Halkiewicz, A. Radecki, R. Kaliszan. Journal of Chromatography A, 346 (1985) 53-60.
[49] S. Zefirov, M. A. Kirpichenok. DokL Akad. Nauk SSSR. 296 (1987) 883.
[50] M. A. Kirpichenok, N. S. Zefirov. Journal of Organic Chemistry of the USSR, 23 (1987) 607-623.
[51] E. Clementi. Computational aspects for large chemical systems. Springer Berlin, 19 (1980) 167.
[52] G. Csizmadia. Elsevier, (1976).
[53] B. Sannigrahi. Advances in quantum chemistry, 23 (1992) 301-351.

2019@ The American Computational Science Society. All rights reserved.