Molecular Diversity (2021) 25:1811–1825
https://doi.org/10.1007/s11030-020-10164-5

ORIGINAL ARTICLE

A simple and robust model to predict the inhibitory activity of α-glucosidase inhibitors through combined QSAR modeling and molecular docking techniques

Elaheh Izadpanah[1] · Siavash Riahi[2] · Zeinab Abbasi-Radmoghaddam[2] · Sajjad Gharaghani[3] · Mohammad Mohammadi-Khanaposhtanai[4]

Received: 26 March 2020 / Accepted: 27 November 2020 / Published online: 9 February 2021

Abstract

Quantitative structure–activity relationships (QSAR) and molecular docking studies have been performed on a series of 35 α-glucosidase inhibitory derivatives. The QSAR models have been developed by genetic algorithm-multiple linear regression (GA-MLR) and least squares-support vector machine (LS-SVM) methods to correlate the conformational descriptors to the inhibitory activity. The obtained models with 5 descriptors were validated and illustrated to be statistically significant. They had desirable prediction based on squared correlation coefficient (R^2), cross-validated correlation coefficient (Q^2), root-mean-squares error (RMSE) and Fisher (F) parameters ($R^2 = 0.951$, $Q^2 = 0.931$, RMSE = 0.121, and $F = 114.629$ for GA-MLR model, and $R^2 = 0.989$, $Q^2 = 0.987$, RMSE = 0.056 and $F = 543.754$ for LS-SVM model). The crucial descriptor named DELS was explored to have the highest correlation with the inhibitory activity and thus has been chosen to build a simple model. The QSAR model developed with this mono-descriptor showed appropriate results of the predicted model using LS-SVM method ($R^2 = 0.888$, $Q^2 = 0.872$, RMSE = 0.185 and $F = 221.459$). Also, molecular docking which focuses on the interaction between ligands and α-glucosidase in the protein active site considered different binding positions to find the best binding mode. It helped the QSAR study to propose more comprehensive details of the compounds structures and was used to design more active compounds. The most active designed compound had a high inhibitory activity of 9.22 that can be proposed for the treatment of diabetes type 2.

✉ Siavash Riahi
riahi@ut.ac.ir

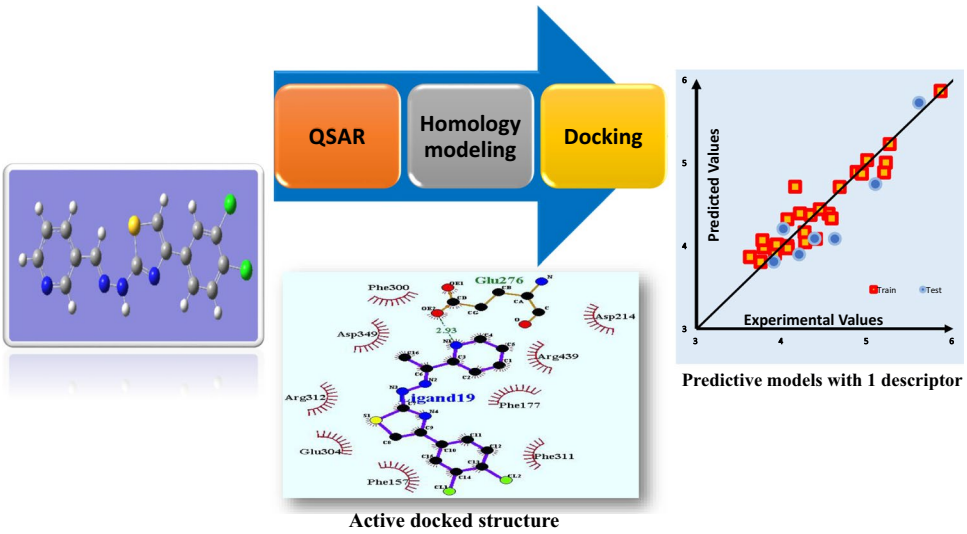
[1] College of Engineering, Faculty of Caspian, University of Tehran, Tehran, Iran

[2] Institute of Petroleum Engineering, School of Chemical Engineering, College of Engineering, University of Tehran, P.O Box: 113654563, Tehran, Iran

[3] Laboratory of Bioinformatics and Drug Design (LBD), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

[4] College of Engineering, Faculty of Fouman, University of Tehran, Tehran, Iran

Graphic abstract



Keywords QSAR · Inhibitory activity · Docking · α-Glucosidase · Diabetes · Medicinal chemistry

Introduction

Diabetes is known as the seventh reason for death worldwide [1], and about 438 million people will suffer from this disease by 2030 [2]. In the diabetes category, diabetes type 2 is the most common illness which includes about 80–90% of diabetic cases [3]. There are two important factors with a major influence on diabetic problems. The first one is the insulin hormone which is released by the pancreas and converts glucose to the required energy of the cells, and the second one is the enzyme α-glucosidase that breaks long-chain carbohydrates into small ones such as glucose and fructose. In the body of a diabetic patient, enough insulin is not released or it does not work properly, while α-glucosidase continues its activity. This leads to the aggregation of glucose in the blood (hyperglycemia) which can hurt different organs of the body, especially the nerves and blood vessels [4]. Therefore, inhibiting the catalytic activity of α-glucosidase is considered as a solution to control the amount of glucose in the blood particularly in individuals with diabetes mellitus type 2 [5, 6]. Some different glycosidic inhibitors of α-glucosidase have been used such as miglitol [7], voglibose and acarbose [8]; although these inhibitors are effective, they have some side effects such as flatulence, diarrhea and abdominal discomfort and have to be used in combination of other medications to increase efficiency [9]. Thus, great attention is paid to discover or design novel and efficient inhibitors. Among various admixtures, heterocyclic compounds are notable options. They can be used to synthesize and produce new drugs due to their synergy with most of the molecular targets [10]. Pyridine as a heterocyclic compound and primitive section of many natural compounds have received much attention to be used in a new generation of drugs [11].

Pharmacists believe that the chemical properties of any segment in a drug depend on its structure. So, structural knowledge is required to anticipate pharmaceutical function. In order to save time and investment in the process of designing effective medicines, more useful methods than trial and error are required and QSAR serves as a beneficial computer tool for this purpose. This method makes a rational relationship between the structure of compounds and their properties and finally predicts the biological activities of the compounds to be prepared. QSAR is a mathematical relationship between biological activity and chemical properties of compounds in form of $Y_i = F_i\,(X_1, X_2, \ldots, X_n)$ where Y_i is the dependent variable (IC_{50}) and X_i refers to a molecular descriptor as the independent variable [12–16].

Molecular docking as a complementary tool for QSAR modeling is an advantageous method to calculate the descriptors that contain significant structural information of the compounds. The main role of this method is to seek different orientations of ligands in protein active sites as the receptor. In this way, molecular docking generates a series of complexes and predicts the best orientation for ligand binding [17].

A survey on recent works about the inhibition of α-glucosidase reveals that different compounds have been studied to inhibit its activity using a computer analytical tool as QSAR [1, 4, 7, 18, 19]. But there is not any investigation

on arylated hydrazinyl thiazole derivatives that possess inhibition properties against this enzyme. Hence, in this study, the required effective concentration for causing 50% inhibition (IC_{50}) of α-glucosidase for 35 arylated hydrazinyl thiazole-based pyridine derivatives has been predicted by the QSAR models. In this process, two different modeling methods, namely MLR and LS-SVM, were used to predict the inhibitory activity. On the other hand, molecular docking has been used to interpret the binding interactions of the compounds and calculate all available descriptors. These compounds were considered as ligands, and molecular docking describes the different binding positions of the ligands in the active sites of the target protein (α-glucosidase). This work was especially aimed to build the easiest model for the descriptor. This goal needs to follow the QSAR and molecular docking procedures. Then validation of the calculated models is required. After all of these efforts to find the most simple model, the statistical results of the predictive model will be compared with previous works in this scope. Moreover, some new compounds have been designed by combining QSAR and molecular docking results with improved inhibitory activities.

Materials and methods

Data set

The data set is obtained from the research of Ali et al. [11] shown in Table 1. It consists of 39 arylated hydrazinyl thiazole-based pyridine derivatives which were synthesized by two-level reaction patterns. These heterocyclic compounds consist of favorable α-glucosidase inhibitory activity. On the other hand, the new synthetic compounds have constructive likeliness as for example pyridine ring, thiazole ring and hydrazine moiety. The remarkable point is the existence of the same amidine moiety as in the antidiabetic agent "metformin."

In this work, 4 compounds containing NO_2 (ionic compounds) have been removed and the remained 35 derivatives were studied. The IC_{50} values varied in a range of 1.4–168 μM. They were converted to their equivalent pIC_{50} ($-\log IC_{50}$) values. Figure 1 and Table 1 exhibit the chemical structures and experimental inhibitory activity values of these compounds.

Geometry optimization of compounds

Three-dimensional (3D) structures of the compounds were pre-optimized based on minimum energy molecular geometries by the HyperChem package (Ver. 7.0) [20]. The RM1 optimization method was used for the initial optimization of compounds. The HyperChem output files were introduced to Gaussian software [21], and optimization of compounds was performed based on a semi-empirical (PM6) method with a frequency cycle to find the lowest energy level in every compound (the most stable state of a compound).

Molecular descriptors calculation and selection

QSAR modeling needs some favorite descriptors to describe the relationship between the chemical structure and activity of the molecules. There is various software with a different theoretical basis for this purpose. Here, Dragon software (Ver 7.0) [22] has been used to calculate descriptors. It contains about 4485 descriptors which are divided into several categories including topological and geometrical, ring descriptors, 2D autocorrelation, GETAWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors, physical properties which include zero-, one-, two- and three-dimensional descriptors. In the first step, about 2100 descriptors were calculated for QSAR analysis. If two descriptors have a correlation coefficient higher than 0.9, one of them has to be excluded [23]. Also, all the duplicate and zero descriptors are useless descriptors that have to be removed. So, the number of remained descriptors is reduced to about 978 descriptors. To build the final QSAR model, these descriptors should be chosen proportional to the data set [24, 25].

Model construction and evaluation parameters

QSAR models were developed using the genetic algorithm (GA) technique. GA provided the final practical descriptors of the model, and GA-MLR as a linear method and LS-SVM as a nonlinear method were applied to construct the QSAR models. To evaluate the model, the data set was divided into two subsets: a test set and a train set. The model is built based on the train set, and its efficiency is analyzed based on its performance on the test set. The *y*-scrambling method was used to choose the test set. In this method, all compounds sorted with descending data and about 20% of the data was chosen as the test set (7–10 compounds).

Assessment of the model performance was performed via the leave-one-out (LOO) cross-validation method. This is the most popular method to evaluate a QSAR model. In this method, there is a sample set of n members. Each member would be set aside in turn, and the modeling would be applied to the other $n-1$ remaining members. This process will be continued until all members are put aside once. Every time, the R^2 parameter is evaluated and the values closer to unity lead to less error for activity prediction [26]. The applicability domain and some other important parameters such as *RMSE* and F have to be studied for a more thorough validation as an inevitable step of QSAR modeling.

Table 1 Chemical structures, experimental and predicted inhibitory activity (pIC_{50}) values (μM)

Compound	SMILES	pIC_{50}	$MLR._{pred}$*	$SVM._{pred}$**
1	C(=S)(N)N/N=C/c1cccnc1	3.92	3.87	3.96
2	c1(c2nc(C/N=C/c3cccnc3)sc2)ccccc1	3.75	3.94	3.87
3	c1c(c2nc(N/N=C/c3cccnc3)sc2)ccc(c1)c1ccccc1	3.62	3.56	3.72
4	NCc1ccc(c2nc(N/N=C/c3cccnc3)sc2)cc1	4.21	4.4	4.39
5	c1ccncc1/C=N/NC1=NC[C](C=S1)c1cccc(c1)Br	3.94	3.97	3.98
6	c1ccncc1/C=N/NC(=S)/N=C(\Br)/c1ccc(cc1)Br	4.01	3.98	4.05
7	c1ccncc1/C=N/Nc1nc(cs1)c1ccc(cc1)Cl	4.54	4.45	4.43
8	c1ccncc1/C=N/Nc1nc(cs1)c1cc(ccc1)Cl	4.33	4.35	4.32
9	c1ccncc1/C=N/Nc1nc(cs1)c1cc(c(cc1)Cl)Cl	5.21	5.22	5.24
10	c1ccncc1/C=N/Nc1nc(cs1)c1c(cc(cc1)Cl)Cl	4.15	4.26	4.21
12	c1ccncc1/C=N/Nc1nc(cs1)c1cc(ccc1)O	4.87	4.88	4.94
13	c1cc(ncc1)/C(=N/NC(=S)N)/C	3.91	4.24	4.01
14	C1C[C@H](NCC1)/C(=N/Nc1nc(cs1)c1ccccc1)/C	4.61	4.47	4.46
15	C1C[C@H](NCC1)/C(=N/Nc1nc(cs1)c1ccc(cc1)c1ccccc1)/C	4.06	3.98	4.09
16	C1C[C@H](NCC1)/C(=N/Nc1nc(cs1)c1cccc(c1)Br)/C	4.26	4.51	4.47
17	C1C[C@H](NCC1)/C(=N/Nc1nc(cs1)c1ccc(cc1)Br)/C	4.38	4.37	4.3
18	c1cc(ncc1)/C(=N/Nc1nc(cs1)c1cc(ccc1)Cl)/C	5.25	5.22	5.13
19	c1cc(ncc1)/C(=N/Nc1nc(cs1)c1cc(c(cc1)Cl)Cl)/C	5.85	5.8	5.63
21	c1cc(ncc1)/C(=N/Nc1nc(cs1)c1cccc(c1)O)/C	5.19	5.35	5.21
22	c1c(cncc1)/C(=N/NC(=S)N)/C	4.06	3.89	3.96
23	c1c(cncc1)/C(=N/Nc1nc(cs1)c1ccccc1)/C	4.44	4.23	4.3
24	c1c(cncc1)/C(=N\Nc1nc(cs1)c1ccc(cc1)c1ccccc1)/C	3.78	3.9	3.96
25	c1c(cncc1)/C(=N\Nc1nc(cs1)c1cccc(c1)Br)/C	4.39	4.4	4.33
26	c1c(cncc1)/C(=N/Nc1nc(cs1)c1ccc(cc1)Br)/C	4.27	4.3	4.35
27	c1c(cncc1)/C(=N\Nc1nc(cs1)c1cc(ccc1)Cl)/C	4.67	4.82	4.76
28	c1c(cncc1)/C(=N/Nc1nc(cs1)c1ccc(c(c1)Cl)Cl)/C	5.6	5.65	5.59
30	c1c(cncc1)/C(=N\Nc1nc(cs1)c1cccc(c1)O)/C	5.09	5.15	5.04
31	c1cnccc1/C(=N/NC(=S)N)/C	3.93	3.95	3.96
32	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1ccccc1)C	3.9	4.01	4.03
33	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1ccc(cc1)c1ccccc1)C	3.77	3.7	3.87
34	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1cc(ccc1)Br)C	4.19	4.11	4.09
35	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1ccc(cc1)Br)C	4.07	4.16	4.12
36	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1cc(ccc1)Cl)C	4.58	4.52	4.5
37	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1ccc(c(c1)Cl)Cl)C	4.93	4.82	4.88
39	c1c(ccnc1)[C@@H](NNc1nc(cs1)c1cc(ccc1)O)C	4.99	4.91	4.92

*$MLR._{pred}$ = Predicted pIC_{50} values by GA-MLR method

**$SVM._{pred}$ = Predicted pIC_{50} values by LS-SVM method

As stated earlier, to verify a QSAR model, usually, the LOO cross-validation procedure is applied. The outcome represented by the cross-validated correlation coefficient (R^2), which is calculated based on the below formula:

$$R^2 = 1 - \frac{\sum \left(y_i - \hat{y}_i\right)^2}{\sum \sum (-\bar{y})^2} \quad (1)$$

Here y_i,$\hat{y}_i$ and $\bar{y}$ are the actual, estimated and averaged (over the entire data set) activities, respectively [25]. According to the literature, a good model should pass the following conditions [19, 27]:

$$Q^2 > 0.5 \quad (2)$$

$$R^2 > 0.6 \quad (3)$$

$$\left(R^2 - R_0^2\right)/R^2 < 0.1 \text{ or } \left(R^2 - R_0^{2\prime}\right)/R^2 < 0.1 \quad (4)$$

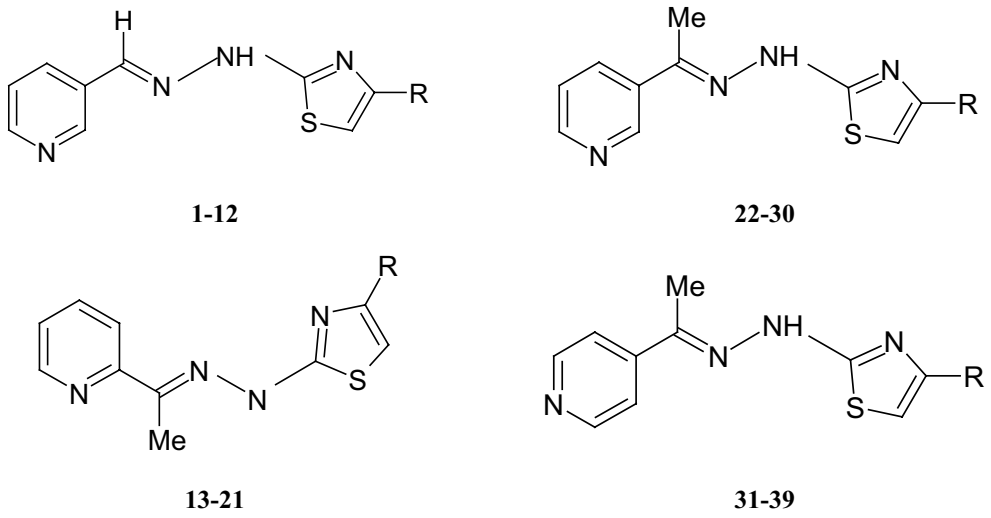


Fig. 1 The basic structure of different pyridine derivatives under study based on hydrazinyl

$$0.85 \le k \le 1.15 \tag{5}$$

Q^2 coefficient of leave-one-out cross-validation, R^2 squared correlation coefficient, k slope of the regression line through the origin, R_0^2 regression of the anticipated activities opposed to observed activities.

Other important statistical parameters that are required to have a perfect comparison between different models are defined as: S standard error of estimation, F Fischer ratio.

And RMSE values calculated as follows [27]:

$$\mathrm{RMSE} = \sqrt{\frac{\sum \left(y_i - y_0\right)^2}{n_s}} \tag{6}$$

where y_i the experimental value of the activity, y_0 the predicted value of inhibitory activity using the model, n_s the number of molecules in the data set, lower values of "S" and "RMSE" together with a higher measure of "F" means that the model can forecast the biological activity with lower error, and it can reveal the high prediction potential of the QSAR models.

Applicability domain

The applicability domain is a theoretical space in which the predictions of QSAR are reliable. There are different approaches to determine the applicability domain, but here the most common method, i.e., the William plot is used. It involves the calculation of the standardized residuals versus leverage amounts. Calculation of the leverage (h_i) for each compound and its threshold is defined in Eqs. (7) and (8), respectively. Compounds with leverage more than warning leverage (h^*) usually had a great influence on the model. A point in the right side of h^* with a residual more than 3 or less than -3 is known as the over-fitted point.

$$h_i = x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x_i \tag{7}$$

$$h^* = \frac{3(k+1)}{n} \tag{8}$$

In Eq. (7), x_i is the descriptor vector of the query molecule and X is the $k \times n$ matrix containing the k descriptor values for the train molecules (n members).

In Eq. (8), k is the number of descriptors in the selected model, and n is defined as the number of objects in the train set [28].

Molecular docking

Molecular docking is an accurate approach to predict the binding affinity and orientation of ligands to the target molecules which is enzyme α-glucosidase in our study. Since the 3D structure of the protein was not available in the protein data bank, the homology modeling method is applied as an alternative solution. This method predicts the structure of an unknown protein based on the structure of similar proteins from the same family [29]. In this study, the homology modeling was used with the template 3A47 [30].

Then molecular docking was run by AutoDock 4.2 software [31]. For all docking parameters, standard values were used. A two-dimensional schematic representations of the docking results including binding sites and interactions of inhibitor with ligands were proposed using LIGPLOT [32].

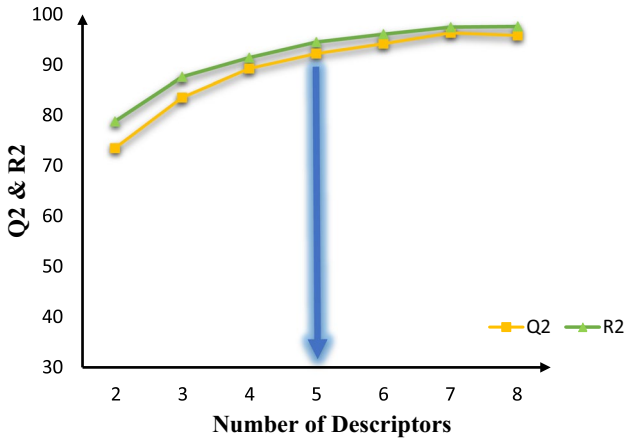


Fig. 2 Breaking point plot of the model to find the best number of descriptors to build the final QSAR model

QSAR and molecular docking could be applied for designing new inhibitors. According to the basic structures, i.e., arylated hydrazinyl thiazole-based pyridine scaffold, new inhibitors have been designed to reduce the inhibitory level. The results of QSAR and molecular docking of the main compounds were carefully investigated to detect the most effective basic structures. Then, the best structures were modified by replacing some of their branches with various useful components. So, some new basic structures were produced (about 126).

Results and discussion

MLR and LS-SVM models

To find a statistically rational QSAR model, the number of independent variables has to be determined through a reliable approach so that in this study the final number of model descriptors was set by the "breaking point" method. This method is based on the sloping trend of statistical parameters versus the number of descriptors. Figure 2 shows that the slope of the breaking point diagram starts to drop off from the fifth descriptor. As far as the smallest suitable number of descriptors is concerned, the breaking point is the optimum number of descriptors [19], which is 5 in this case.

GA was used to select the most effective descriptors in a huge space of different features. The selected descriptors were then assessed to be incorporated in the final model. Consecutively, the final models were built on the 5 selected descriptors as presented in Table 2.

The linear function including the selected variables was obtained using GA-MLR method as below:

$$\begin{aligned} pIC_{50} = &-2.127 - 0.007(D/Dtr05) + 0.248(DELS) \\ &+ 2.586(GATS4s) + 11.277(G1p) - 1.605(H4m) \end{aligned} \quad (9)$$

The equation indicates descriptors G1p and GATS4s have the highest coefficients in the model, and they have a direct relationship with pIC_{50}. To better illustrate the influence of these variables, their correlation with each other and pIC_{50} was calculated and the results are collected in Table 3. It has been demonstrated that descriptor DELS provides the highest correlation with the inhibitory activity which makes it a

Table 2 Molecular descriptors of the best model with 5 descriptors

Molecular descriptor	Type	Description
D/Dtr05	Ring	Distance/detour ring index of order 5
DELS	Topological indices	2D descriptor from molecular graphs and independent of conformations
GATS4s	2D autocorrelations	Geary autocorrelation of lag 4 weighted by I-state
G1p	WHIM	Geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes
H4m	GETAWAY	H Autocorrelation of lag 4/weighted by atomic masses

Table 3 The correlation between 5 descriptors of the best-structured model and the inhibitory activity (pIC_{50})

	pIC50	D/Dtr05	DELS	GATS4s	G1p	H4m
pIC50	1					
D/Dtr05	0.305	1				
DELS	0.796	0.445	1			
GATS4s	0.223	0.626	0.162	1		
G1p	−0.218	−0.676	−0.344	−0.446	1	
H4m	−0.074	0.180	0.355	0.395	−0.004	1

Table 4 The statistical results of GA-MLR and LS-SVM models with 5 descriptors

	Overall				Train set				Test set	
	R^2	Q^2	RMSE	F	RMSE	F	Q^2	R^2	RMSE	R^2
GA-MLR	0.951	0.931	0.121	114.629	0.100	98.075	0.946	0.963	0.179	0.919
LS-SVM	0.989	0.987	0.056	543.754	0.950	99.456	0.953	0.966	0.144	0.947

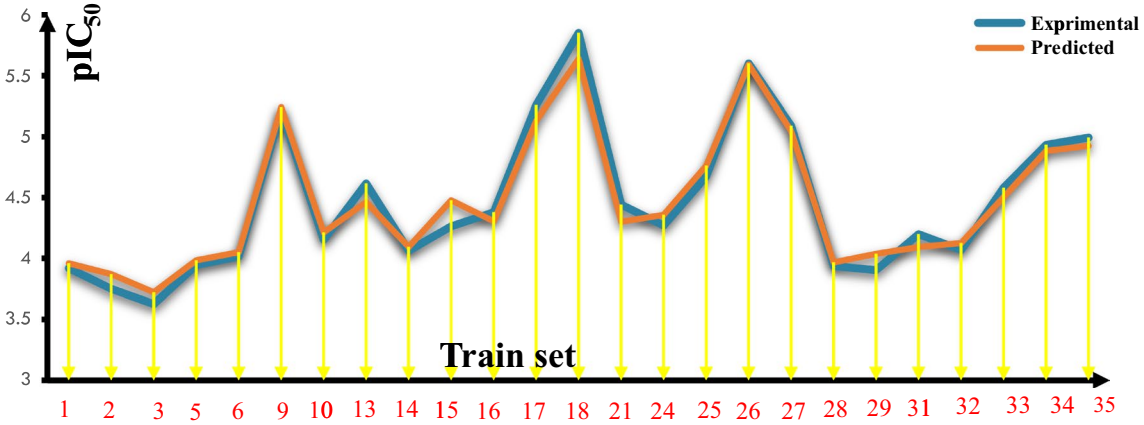


Fig. 3 Comparison of predicted and experimental values of train data set with their specific error prediction in the MLR model with 5 descriptors

crucial descriptor to build the model. So, a mono-descriptor model named simple model was made with descriptor DELS.

Evaluation of the models

GA-MLR model includes 5 final variables as the most influential descriptors. To assess the nonlinearity relation between the descriptors and pIC_{50} a reliable model was constructed based on 5 selected descriptors by use of the LS-SVM method. The results of this model were significant. So, it is a good solution to compare the predictive ability of the model through two different methods.

One of the evaluation methods is the comparison of statistical parameters related to QSAR models. In this case, parameters such as Q^2, R^2, RMSE, F and S were calculated for the MLR and LS-SVM models. These results for the GA-MLR model based on 5 descriptors and 10 test compounds in Table 4 represent a good prediction capacity. The model has a high multiple correlation coefficient (0.951) and a low prediction error. Figures 3 and 4 illustrate the calculated and experimental values of pIC_{50} for the train and test data set, respectively. The maximum prediction error was a 5.028% error which is acceptable.

The regression line indicates the comparison between predicted and experimental values in Fig. 5. Also, the residual graph of the MLR model with 5 final descriptors is shown in Fig. 6. As it is obvious, congestion of compounds either

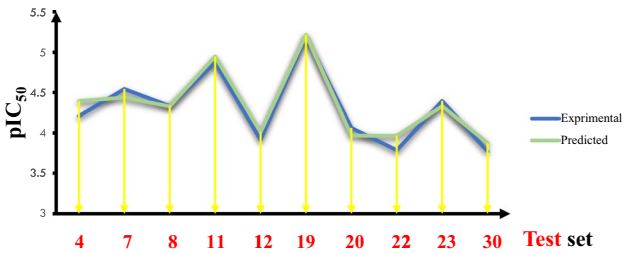


Fig. 4 Comparison of predicted and experimental values of test data set with their specific error prediction in the MLR model with 5 descriptors

for train or for test set shows they are well distributed, and none of them has unaccepted distance from the fitted lines.

According to Fig. 7, in the applicability domain analysis, one point (compound 12) with a residual more than 3 in William plot was predicted with slightly higher error. These errors may be due to an error in experimental data. The other points all stayed in the determined applicability domain by William plot.

On the other hand, the statistical results of the LS-SVM model with 5 descriptors in Table 4 describe that the model can predict appropriately and it is more useful than its MLR model. Like the MLR model, in the LS-SVM model based on 5 descriptors, the residual graph in Fig. 8 demonstrates the proper distribution of the data set.

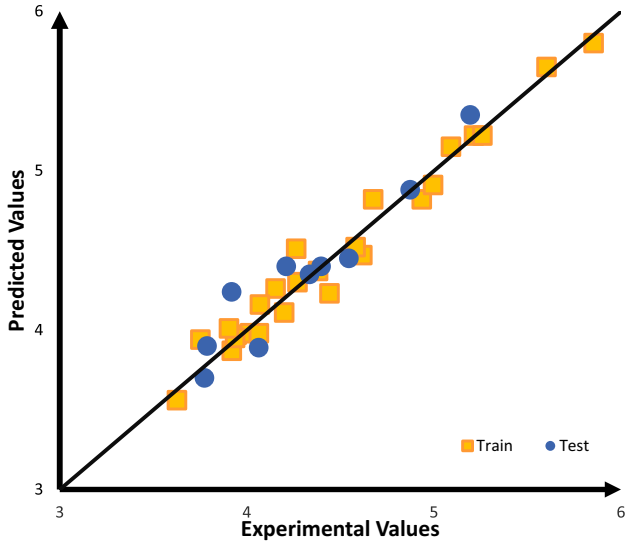


Fig. 5 The regression line of the MLR model with 5 descriptors

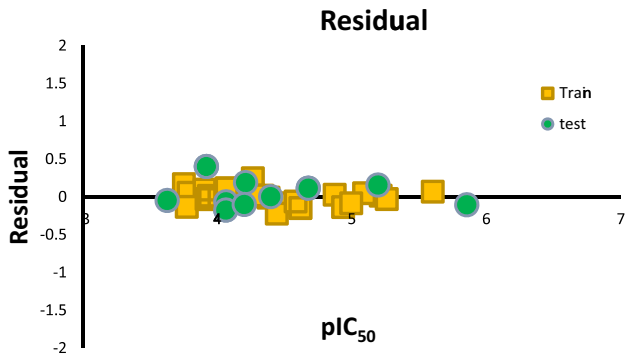


Fig. 6 The residual graph of the MLR model with 5 descriptors

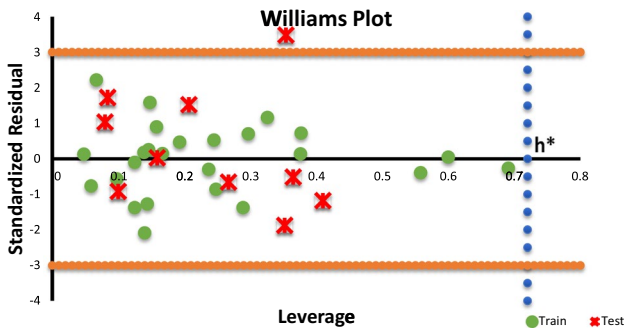


Fig. 7 The William plot of MLR model with 5 descriptors

As it is shown in Fig. 9, in the William plot of this model all compounds stand in the applicability domain. As a result, both linear and nonlinear models have an acceptable predictive capacity for inhibitory activity calculation. The

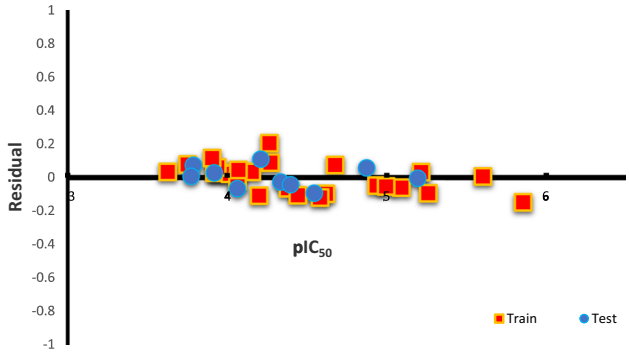


Fig. 8 The residual graph of the LS-SVM model with 5 descriptors

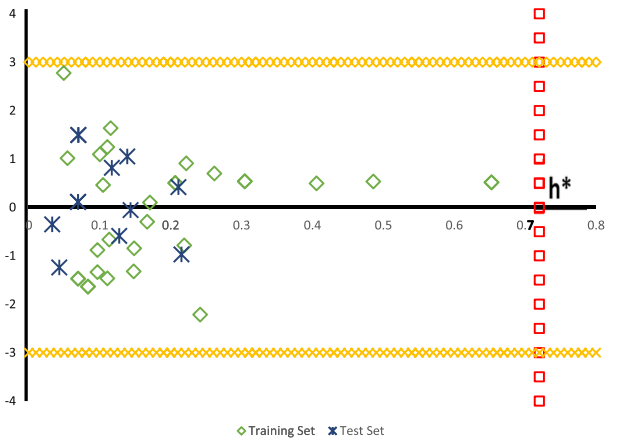


Fig. 9 The William plot of the LS-SVM model with 5 descriptors

predicted values of pIC_{50} using these models are displayed in Table 1.

To ensure the stability of these models, they were validated with different test groups and nearly all of them represented good results. Table 5 indicates the average statistical values of ten new LS-SVM models. These results prove that the fitness of the model is not dependent on the selected test set as by varying the test and train set it still can predict satisfactorily. Therefore the models present favorable statistical results to be trusted as reliable predictive models.

Descriptors analysis to explore a simple model

The effectiveness of each descriptor in the QSAR model is investigated with sensitivity analysis. In this method, a descriptor is eliminated and the difference between RMSE values in this state and the base case (with all descriptors) is observed. A greater difference means that the descriptor had a more profound role in the model [27].

Figure 10 describes the calculated sensitivity test values to find the most effective descriptors in the model.

Table 5 The average statistical results of ten LS-SVM models with various random test groups of compounds based on 5 descriptors in the model

Overall					Train set				Test set	
Q^2	R^2	F	RMSE	R^2	Q^2	F	RMSE	R^2	F	RMSE
0.96	0.97	235.24	0.09	0.96	0.90	95.69	0.10	0.94	10.89	0.13

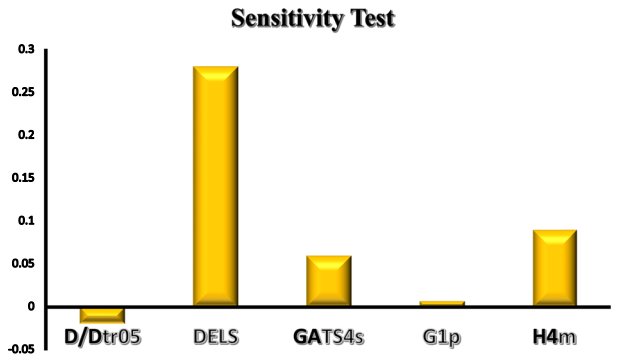


Fig. 10 Sensitivity test of model descriptors to find the most effective variable on the LS-SVM model based on 5 descriptors

Different descriptors as independent variables in the linear equation come from several categories in Dragon descriptors and thus convey different structural information about the compounds. DELS is a topological descriptor with a positive sign in the MLR equation that discloses basic information about the size of molecules, degree of branching, flexibility and the overall shape topological indices which are 2D descriptors based on graph theory concepts [19]. Another essential descriptor, GATS4s, is the Geary autocorrelation of lag 4 weighted by I-state, containing information about the distribution of inherent state along with the topological structure [33]. The higher value of this descriptor leads to higher pIC_{50}. H4m is H autocorrelation of lag 4/weighted by atomic masses which is a GETAWAY descriptor [34] whose lower values cause higher pIC_{50}. Descriptor G1p is the 1st component symmetry directional WHIM index/weighted by atomic polarizabilities [35]. It has a positive sign in the MLR equation, and thus, the pIC_{50} value increases at higher values of this descriptor (the higher value of pIC_{50} indicates a lower value of inhibitory activity). The last descriptor D/Dtr05 is a ring descriptor [36] with a negative sign in the linear equation and a negative effect on pIC_{50} based on the sensitivity test.

As expected (according to Table 3) DELS descriptor had the main role among all descriptors. Results show that H4m, GATS4s, D/Dtr05, and G1p stand in the next places, respectively.

Table 3 illustrates that the DELS descriptor has a high correlation (0.796) with pIC_{50} which is verified by sensitivity analysis too. Therefore, it seems necessary to make a comparative study between the base model and the model constructed with this descriptor. To evaluate the simple QSAR model constructed by the use of the LS-SVM method, the statistical results were derived, and as reported in Table 6, they present a satisfactory accuracy.

According to this table, $R^2 = 0.888$, $Q^2 = 0.872$, RMSE = 0.185 and $F = 221.459$, which means the simple model can be a favorable model to predict pIC_{50} values of the compounds with a high degree of reliability. Besides, the regression line in Fig. 11 and residual diagram in Fig. 12 show the acceptable dispersion of compounds by the simple model.

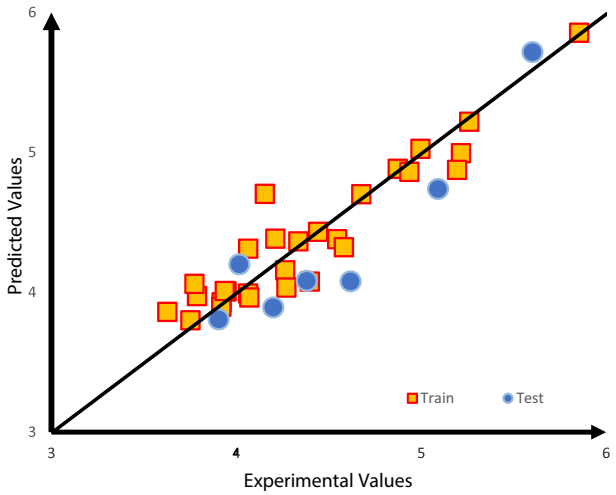


Fig. 11 The regression line of the simple model with one descriptor

Table 6 The statistical results of the simple model with DELS descriptor based on the LS-SVM method

	Overall					Train–test				
	R^2	Q^2	RMSE	F	S	R^2_{train}	R^2_{test}	Q^2_{train}	F_{train}	$RMSE_{train}$
LS-SVM	0.888	0.872	0.185	221.459	0.190	0.877	0.708	0.855	173.745	0.191

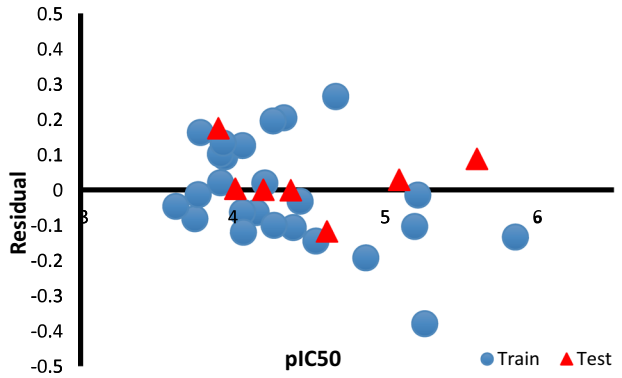


Fig. 12 The residual diagram of the simple model with DELS descriptor

Table 7 The average results of six LS-SVM models with different test groups for the simple model

Average overall results			Average train–test results	
Q^2	R^2	R^2_{train}	R^2_{test}	Q^2_{train}
0.85	0.83	0.78	0.75	0.74

Not only the selected model with a specific test group had good statistical results, but also other different test groups were studied and showed acceptable ability to predict the inhibitory activity. The final results of models based on the DELS descriptor with different test groups are summarized in Table 7.

As the final result, in this research, a simple model with only one descriptor (DELS) through the LS-SVM method was extracted to predict pIC_{50} values of α-glucosidase inhibitors with good statistical features. However, the other calculated models using GA-MLR and LS-SVM methods had better statistical results albeit with 5 variables and the nonlinear model had even better prediction capability. The best R^2 value in previous studies is 0.872, and most of these studies had used just a single linear or nonlinear method to build their QSAR models. Hence, it seems necessary to compare different linear and nonlinear models to find the best model for pIC_{50} prediction. Table 8 presents a summarized survey on various works in this scope, and it can be observed that this work has better results in comparison with recent studies in this field. Therefore, the presented models can be useful to predict the inhibitory activity of these special α-glucosidase inhibitors.

Homology modeling

The baker's yeast α-glucosidase was applied in the homology modeling approach. A suitable structural template was found for homology modeling in the Protein Data Bank (PDB) at the National Center for Biotechnology and Information (NCBI). The amino acid sequence of the α-glucosidase was inputted using BLAST and PSIBLAST algorithms and was retrieved with 72.51% identification to build the homology model that comprises 584 amino acid residues from the SWISS-PROT protein sequence data bank (http://www.expasy.org/sprot/; Accession No.). Figure 13 shows sequence alignment between yeast α-glucosidase and the template 3A47 taken from SWISS-MODEL site. The structure of the simulated protein was designed and is depicted in Fig. 14.

The Ramachandra server[a] was used to evaluate the accuracy of amino acid placements which was determined to be

Table 8 Comparison between recent QSAR and molecular docking studies on the α-glucosidase enzyme

Authors	Year	Molecular docking	Compound	Model type	Q^2	R^2	Des. no.	Data. no.	Data. no./des no.
Syahrul Imran [1]	2015	Yes	Flavone hydrazone	MLR	0.705	0.848	4	21	5.25
Leila Dinparast [4]	2015	Yes	Benzimidazole derivatives	MLR	0.69	0.600	2	14	7
Yan Liu [18]	2008	No	xanthone derivatives	MLR	0.839	0.872	3	33	11
Khairedine Kraim [19]	2009	No	xanthone and curcuminoid	MLR	0.815	0.857	5	44	8.8
Asadollahi-Baboli [7]	2018	Yes	tetracyclic oxindole derivatives	GA-PLS/SVM	0.871	0.837*	4	34	8.5
This work	2019	Yes	Arylated hydrazinyl thiazole based pyridine	MLR SVM	0.987	0.989	5	35	7

*R^2 prediction

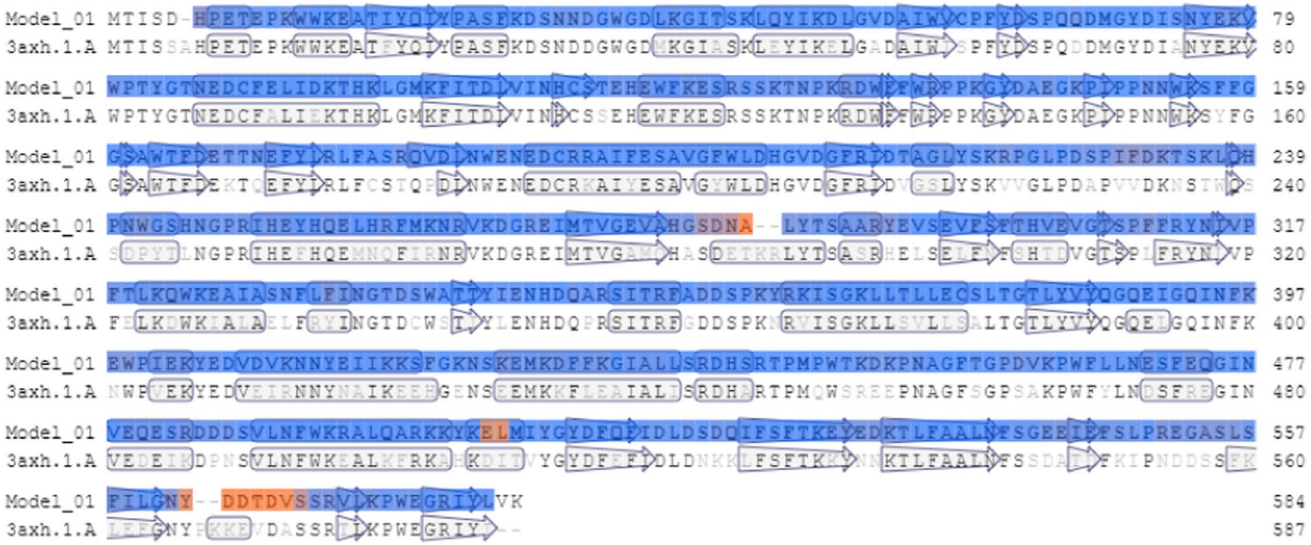


Fig. 13 Amino acid alignment in homology modeling of yeast α-glucosidase

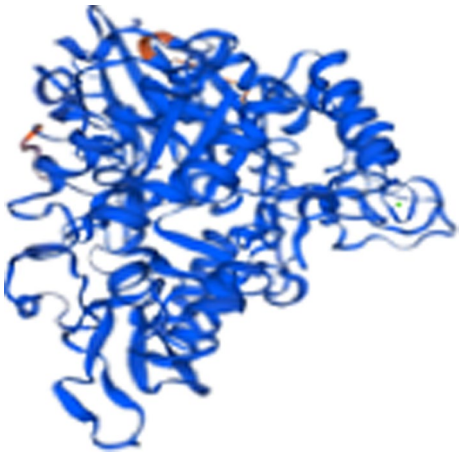


Fig. 14 Structure of the simulated protein with homology modeling method to use it in molecular docking study instead of real protein structure

equal to 97% according to Fig. 15. In other words 97% of the amino acids have been located in allowable zones which indicates the high quality of the forecasted structure.

Molecular docking

Molecular docking was exerted on the compounds to calculate useful descriptors and considering different orientations of ligands in the α-glucosidase active site. All docking features were obtained by the use of AutoDock tools and binana [37]. Different models were established by these descriptors, but none of them had good statistical results as good as Dragon descriptors to apply the significant effect in QSAR modeling.

The different binding mode of ligands with protein was considered. The impressive interaction of the inhibitors with the diversified residues in the active site of the enzyme was gained.

Finding a rational relation between these compounds and their structures to understand how some compounds had the most activity depends on their structural properties, and often it is hard work. In this study, three of the most active compounds are shown in Fig. 16: compound 9, 19 and 28. The common residuals in these compounds comprise from Phe (177, 311, 157) and Arg (312, 439) groups. They had an effective role to improve pIC_{50} values. Also, it demonstrated the hydrophobic interactions between the enzyme and ligands, different atoms in the structures and their positions, the residuals, hydrogen binding and the other connections. The best binding position of protein with ligands in the active site receptor is useful to design and produce some new drugs.

Different descriptors of the calculated QSAR model already described physical and topological properties, geometry, ring structures and atom binding position have a significant effect on the inhibitory activity. Also, information from molecular docking processes can be used to understand the structure of the compounds with more details which helps QSAR explain compounds structurally and find the best compounds to produce medicine. So, according to the QSAR and docking findings, it is necessary to notice how atoms are gathered to construct the complexes.

Analysis of designed compounds

New inhibitors have been designed based on arylated hydrazinyl thiazole-based pyridine scaffold by QSAR and

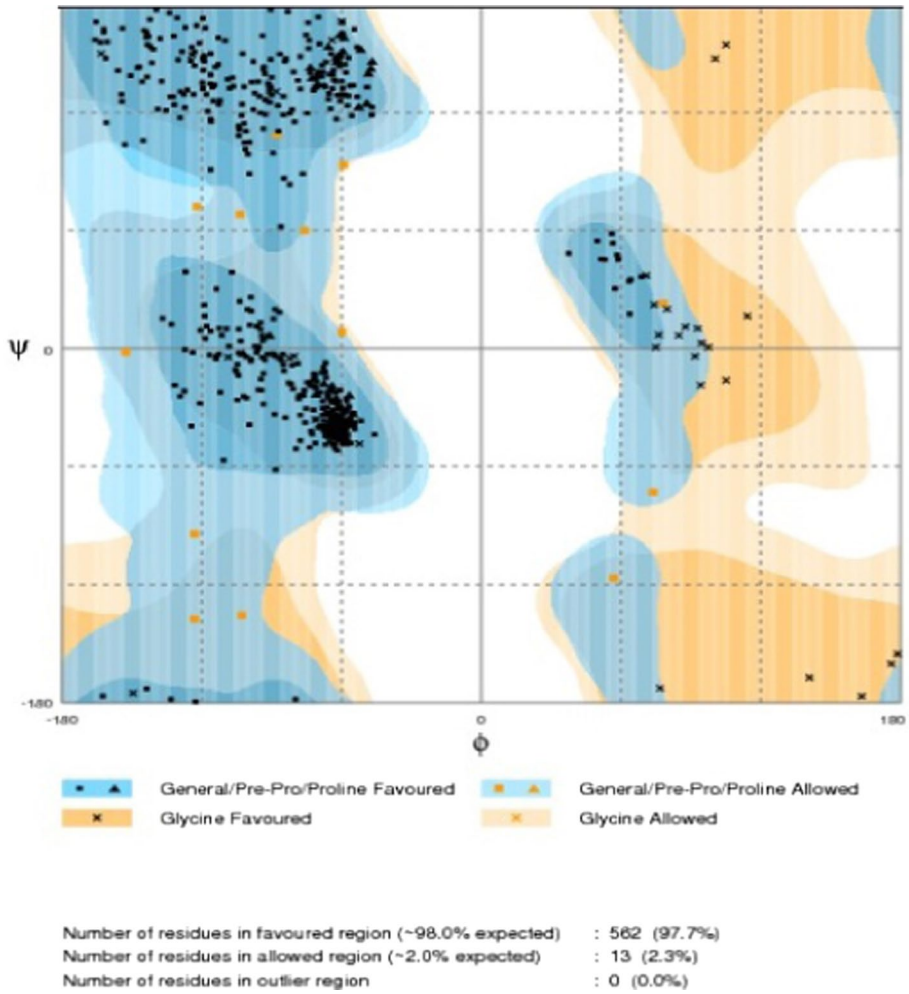


Fig. 15 The precision of amino acid replacement in the accurate zone via homology modeling

molecular docking approach. A study of the inhibitors reveals that halogen molecules (F and Cl atoms) and OH have a key role in increasing the inhibitory activity. The most active designed inhibitors are shown in Table S1 and Figure S1 (as supporting information in supplementary materials) with their structures and calculated pIC_{50} values using the presented MLR model based on 5 descriptors. All pIC_{50} values are better than the main inhibitors of the study. Of course, these values need to be verified experimentally after the synthesis of the designed compounds.

In the docking process, the correlation between free energy and pIC_{50} values was calculated for all designed compounds. Although the correlation has been improved in comparison with the main descriptors, it still does not have a significant value (−0.226). Also, the interaction of molecules by different amino acids was investigated. The common residuals that have been repeated almost in all inhibitors are the Phe (157, 177, 158 and 311). Two structures with high activity had a hydrophobic interaction of His 239, Arg 312 and Asp 349. 2D representation of the most active new structures A1, A2 and B3 is shown in Fig. 17.

Conclusion

In the present study, two different approaches, namely GA-MLR and LS-SVM methods, were applied to establish linear and nonlinear QSAR models to predict the biological activity of a set of arylated hydrazinyl thiazole-based pyridine derivatives. Among various descriptors calculated, the 5 most potent descriptors were selected via GA to build the final QSAR model. DELS descriptor among the selected descriptors had the highest correlation (0.796) with pIC_{50}. It was able to build a QSAR model solely with favorable

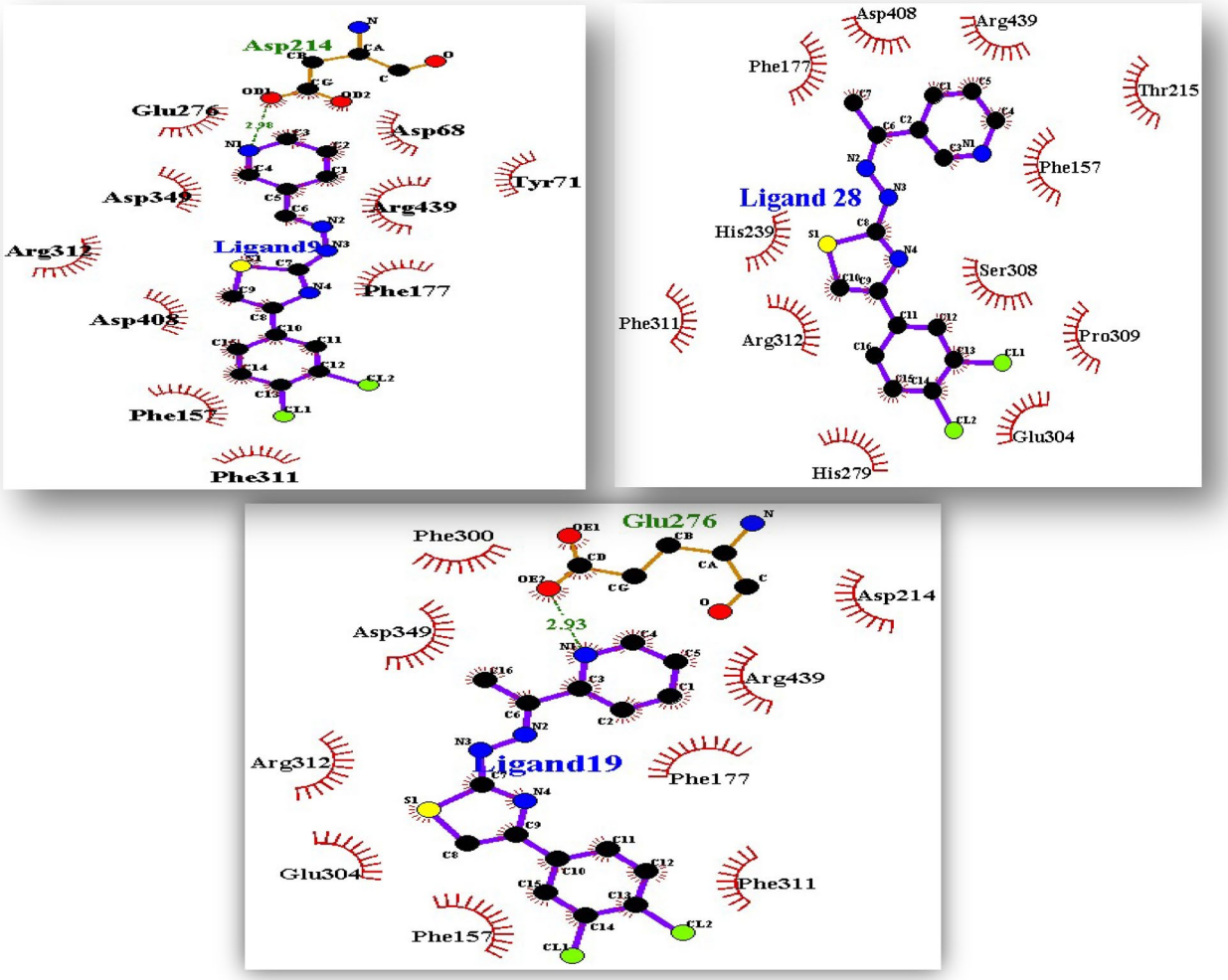


Fig. 16 2D representation of the most active docked structures in molecular docking study: compound 9, 19 and 28

prediction ability. In previous studies on α-glucosidase inhibition, the best-reported value for R^2 was about 0.872, while in the present study with a QSAR model with 5 final descriptors the value of R^2 is 0.989 in the nonlinear model and it is about 0.888 in the simple model (using descriptor DELS based on LS-SVM method). Thus the presented models even the simple model can forecast the inhibitory activity of the compounds with higher accuracy than the previous modeling studies. Also, branching information and the size of molecules that come from the DELS descriptor had been considered as the most effective subjects on inhibitory activities of the compounds. Three of the best predicted pIC_{50} values belong to compounds 9, 19 and 28 all have an aromatic ring connected to two branches of Cl atoms next to each other which reveals the fundamental role of halogen atoms in the inhibition of enzyme activities. Finally, the most active designed compounds (addressed as A1 in this study) had the best pIC_{50} value of 9.22 comparable to the basic data set.

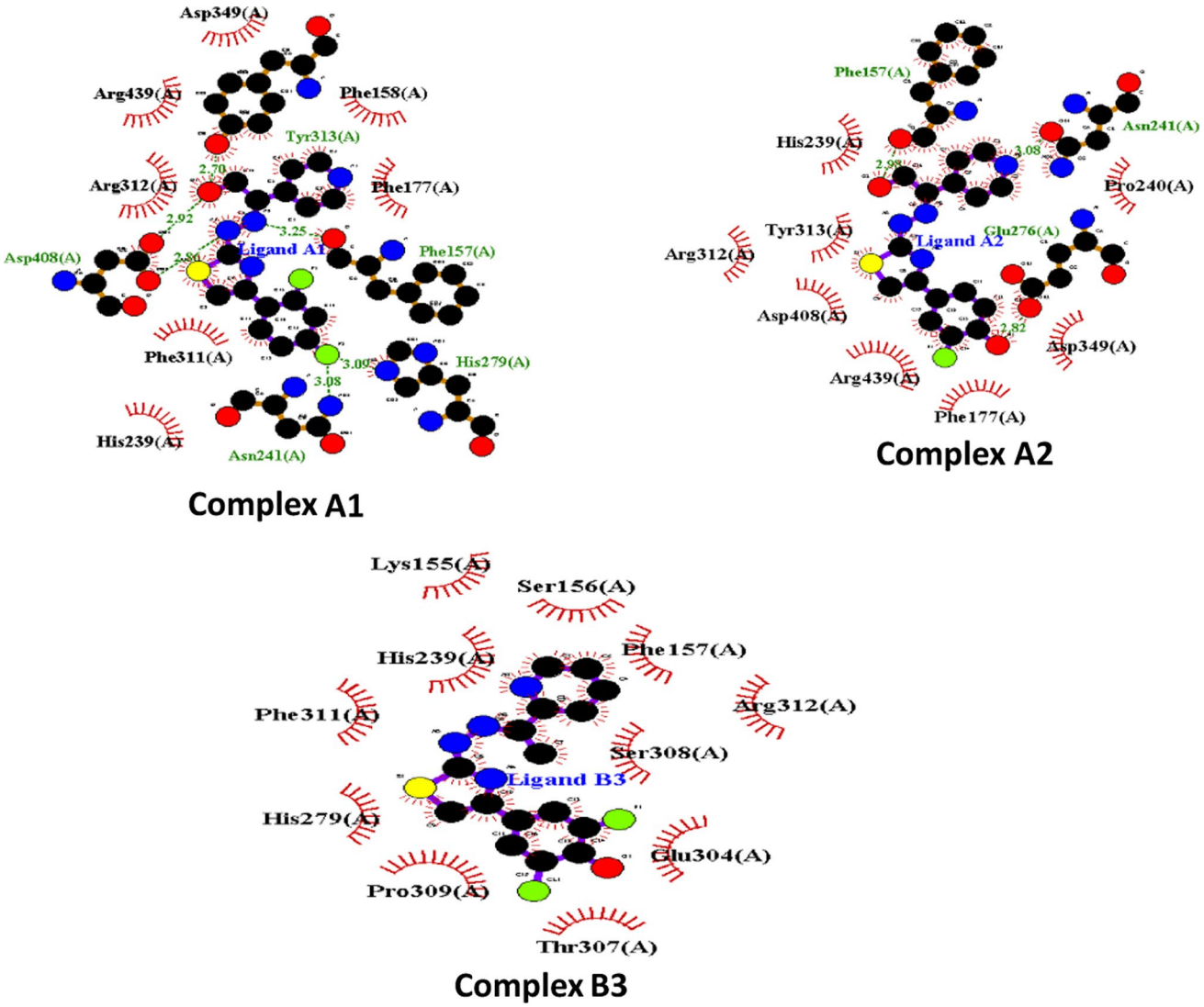


Fig. 17 The representation of the most active designed compounds

Supplementary Information The online version of this article (https://doi.org/10.1007/s11030-020-10164-5) contains supplementary material, which is available to authorized users.

Acknowledgments The authors would like to gratefully acknowledge the support from the Institute of Petroleum Engineering (IPE), University of Tehran.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Imran S, Taha M, Ismail NH, Kashif SM, Rahim F, Jamil W, Hariono M, Yusuf M, Wahab H (2015) Synthesis of novel flavone hydrazones: in-vitro evaluation of α-glucosidase inhibition, QSAR analysis and docking studies. Eur J Med Chem 105:156–170. https://doi.org/10.1016/j.ejmech.2015.10.017
2. Goldenberg RM (2011) Management of unmet needs in type 2 diabetes mellitus: the role of incretin agents. Can J Diabetes 35(5):518–527. https://doi.org/10.1016/S1499-2671(11)80008-0
3. Narender T, Madhur G, Jaiswal N, Agrawal M, Maurya CK, Rahuja N, Srivastava AK, Tamrakar AK (2013) Synthesis of novel triterpene and *N*-allylated/*N*-alkylated niacin hybrids as α-glucosidase inhibitors. Eur J Med Chem 63:162–169. https://doi.org/10.1016/j.ejmech.2013.01.053
4. Dinparast L, Valizadeh H, Bahadori MB, Soltani S, Asghari B, Rashidi MR (2016) Design, synthesis, α-glucosidase inhibitory activity, molecular docking and QSAR studies of benzimidazole

derivatives. J Mol Struct 1114:84–94. https://doi.org/10.1016/j.molstruc.2016.02.005
5. Park H, Hwang KY, Kim YH, Oh KH, Lee JY, Kim K (2008) Discovery and biological evaluation of novel α-glucosidase inhibitors with in vivo antidiabetic effect. Bioorg Med Chem Lett 18(13):3711–3715. https://doi.org/10.1016/j.bmcl.2008.05.056
6. Park H, Hwang KY, Oh KH, Kim YH, Lee JY, Kim K (2008) Discovery of novel α-glucosidase inhibitors based on the virtual screening with the homology-modeled protein structure. Bioorg Med Chem 16(1):284–292. https://doi.org/10.1016/j.bmc.2007.09.036
7. Asadollahi-Baboli M, Dehnavi S (2018) Docking and QSAR analysis of tetracyclic oxindole derivatives as α-glucosidase inhibitors. Comput Biol Chem 76:283–292. https://doi.org/10.1016/j.compbiolchem.2018.07.019
8. Scott LJ, Spencer CM (2000) Miglitol: a review of its therapeutic potential in type 2 diabetes. Drugs 59(3):521–549. https://doi.org/10.2165/00003495-200059030-00012
9. Wang SL (2018) New novel α-glucosidase inhibitors produced by microbial conversion. Process Biochem 65:228–232. https://doi.org/10.1016/j.procbio.2017.11.016
10. Channar PA, Saeed A, Larik FA, Rashid S, Iqbal Q, Rozi M, Younis S, Mahar J (2017) Design and synthesis of 2, 6-di (substituted phenyl) thiazolo [3, 2-b]-1, 2, 4-triazoles as α-glucosidase and α-amylase inhibitors, co-relative pharmacokinetics and 3D QSAR and risk analysis. Biomed Pharmacother 94:499–513. https://doi.org/10.1016/j.biopha.2017.07.139
11. Ali F, Khan KM, Salar U, Taha M, Ismail NH, Wadood A, Riaz M, Perveen S (2017) Hydrazinyl arylthiazole based pyridine scaffolds: synthesis, structural characterization, in vitro α-glucosidase inhibitory activity, and in silico studies. Eur J Med Chem 138:255–272. https://doi.org/10.1016/j.ejmech.2017.06.041
12. Ghaslani D, Gorji ZE, Gorji AE, Riahi S (2017) Descriptive and predictive models for Henry's law constant of CO2 in ionic liquids: a QSPR study. Chem Eng Res Des 120:15–25. https://doi.org/10.1016/j.cherd.2016.12.020
13. Hasanebrahimi G, Riahi S, Fini MF (2017) Exploring beneficial structural features of ionic surfactants for wettability alteration of carbonate rocks using QSPR modeling technique. J Mol Liq 240:196–208. https://doi.org/10.1016/j.molliq.2017.05.009
14. Mehraein I, Riahi S (2017) The QSPR models to predict the solubility of CO_2 in ionic liquids based on least-squares support vector machines and genetic algorithm-multi linear regression. J Mol Liq 225:521–530. https://doi.org/10.1016/j.molliq.2016.10.133
15. Abbasi-Radmoghaddam Z, Riahi S, Gharaghani S, Mohammadi-Khanaposhtanai M (2020) Design of potential anti-tumor PARP-1 inhibitors by QSAR and molecular modeling studies. Mol Divers. https://doi.org/10.1007/s11030-020-10063-9
16. Rezaei B, Riahi S (2016) Prediction of CO_2 loading of amines in carbon capture process using membrane contactors: a molecular modeling. J Nat Gas Sci Eng 33:388–396. https://doi.org/10.1016/j.jngse.2016.05.003
17. Liu Z, Liu Y, Zeng G, Shao B, Chen M, Li Z, Jiang Y, Liu Y, Zhang Y, Zhong H (2018) Application of molecular docking for the degradation of organic pollutants in the environmental remediation: a review. Chemosphere 203:139–150. https://doi.org/10.1016/j.chemosphere.2018.03.179
18. Liu Y, Ke Z, Cui J, Chen WH, Ma L, Wang B (2008) Synthesis, inhibitory activities, and QSAR study of xanthone derivatives as α-glucosidase inhibitors. Bioorg Med Chem 16(15):7185–7192. https://doi.org/10.1016/j.bmc.2008.06.043
19. Kraim K, Khatmi D, Saihi Y, Ferkous F, Brahimi M (2009) Quantitative structure activity relationship for the computational prediction of α-glucosidase inhibitory. Chemom Intell Lab Syst 97(2):118–126. https://doi.org/10.1016/j.chemolab.2009.03.006
20. Release H (2002) 7.5 for windows, molecular modeling system, Hypercube. Inc. http://www.hyper.com
21. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H (2009) Gaussian 09, Gaussian, Inc., Wallingford, CT, vol 32, pp 5648–5652
22. Todeschini R, Consonni V, Mauri A, Pavan M (2002) DRAGON software. Milano, Italy
23. Gagic Z, Nikolic K, Ivkovic B, Filipic S, Agbaba D (2016) QSAR studies and design of new analogs of vitamin E with enhanced antiproliferative activity on MCF-7 breast cancer cells. J Taiwan Inst Chem Eng 59:33–44. https://doi.org/10.1016/j.jtice.2015.07.019
24. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, London
25. Golbraikh A, Tropsha A (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers 5(4):231–243. https://doi.org/10.1023/A:1021372108686
26. Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. J Chem Inf Comput Sci 43(2):579–586. https://doi.org/10.1021/ci025626i
27. Gharaghani S, Khayamian T, Ebrahimi M (2013) Molecular dynamics simulation study and molecular docking descriptors in structure-based QSAR on acetylcholinesterase (AChE) inhibitors. SAR QSAR Environ Res 24(9):773–794. https://doi.org/10.1080/1062936X.2013.792877
28. Aouidate A, Ghaleb A, Ghamali M, Chtita S, Choukrad M, Sbai A, Bouachrine M, Lakhlifi T (2016) Combining DFT and QSAR studies for predicting psychotomimetic activity of substituted phenethylamines using statistical methods. J Taibah Univers Sci 10(6):787–796. https://doi.org/10.1016/j.jtusci.2016.07.001
29. Pitman MR, Menz RI (2006). Methods for protein homology modelling. In: Applied mycology and biotechnology, vol 6. Elsevier, pp 37–59
30. https://www.uniprot.org/uniprot/P53341.fasta. Accessed Sept 2018
31. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2):455–461. https://doi.org/10.1016/S1874-5334(06)80005-5
32. Laskowski RA, Swindells MB. (2011). LigPlot+: multiple ligand–protein interaction diagrams for drug discovery, pp 2778–2786. https://doi.org/10.1021/ci200227u
33. He J, Peng T, Yang X, Liu H (2018) Development of QSAR models for predicting the binding affinity of endocrine disrupting chemicals to eight fish estrogen receptor. Ecotoxicol Environ Saf 148:211–219. https://doi.org/10.1016/j.ecoenv.2017.10.023
34. Jouyban A, Shayanfar A, Ghafourian T, Acree WE Jr (2014) Solubility prediction of pharmaceuticals in dioxane + water mixtures at various temperatures: effects of different descriptors and feature selection methods. J Mol Liq 195:125–131. https://doi.org/10.1016/j.molliq.2014.02.012
35. Jukić M, Rastija V, Opačak-Bernardi T, Stolić I, Krstulović L, Bajić M, Glavaš-Obrovac L (2017) Antitumor activity of 3, 4-ethylenedioxythiophene derivatives and quantitative structure-activity relationship analysis. J Mol Struct 1133:66–73. https://doi.org/10.1016/j.molstruc.2016.11.074
36. Roy K, Das RN (2013) QSTR with extended topochemical atom (ETA) indices. 16. Development of predictive classification and regression models for toxicity of ionic liquids towards *Daphnia magna*. J Hazard Mater 254:166–178. https://doi.org/10.1016/j.jhazmat.2013.03.023
37. Durrant JD, McCammon JA (2011) BINANA: a novel algorithm for ligand-binding characterization. J Mol Gr Model 29(6):888–893. https://doi.org/10.1016/j.jmgm.2011.01.004