Chemometrics and Intelligent Laboratory Systems 222 (2022) 104510

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Application of the LAD-LASSO as a dimensional reduction technique in the ANN-based QSAR study: Discovery of potent inhibitors using molecular docking simulation

Zeinab Mozafari [a,*], Mansour Arab Chamjangali [a], Mohammad Arashi [b], Nasser Goudarzi [a]

[a] Department of Chemistry, Shahrood University of Technology, Shahrood, Semnan, Iran
[b] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Khorasan Razavi, Iran

ARTICLE INFO

Keywords:
LAD-LASSO
Artificial neural network
Molecular docking
Cancer
HIV

ABSTRACT

In this study, the combination of the least absolute deviation-least absolute shrinkage and selection operator (LAD-LASSO) was introduced as a new variable selection method for the artificial neural network (ANN)-based quantitative structure-activity relationship (QSAR) studies. The biological activity of various chemical compounds was predicted using an ANN-based QSAR model combined with the efficient LAD-LASSO variable selection method. In this study, 3224 computed DRAGON descriptors were reduced to a smaller number using pre-processing methods. The descriptors with the most significant relevance to biological activities were chosen using the LAD-LASSO variable selection method. The selected descriptors were defined as ANN inputs and optimized the designed models. The biological activity of the test set compounds was predicted using the optimum ANN models. The coefficients of determination (R^2) for the test data in the different datasets were equal to 0.87, 0.84, and 0.87. Also, the MSE value of the test set is equal to 0.13, 0.07, and 0.11, respectively. The high R^2 and low MSE values demonstrate the good prediction ability of the constructed QSAR models. The applicability domain (AD) and Y-randomization test also proved the efficiency of the developed models. Finally, The performance of the QSAR model was evaluated by the identification of novel compounds with high potency. As a result, the weak structure of the dataset was identified and modified using the effect of selected descriptors on the biological activity, resulting in the establishment of new compounds with significant potency. The response value of the new suggested compounds was predicted using the optimum ANN models. Receptor-ligand interactions were extracted for all proposed compounds. The presence of different hydrophilic and hydrophobic interactions in the active site of the respective receptor indicates the high potential of suggested chemical compounds.

1. Introduction

Quantifying the physicochemical properties of synthetic chemical drugs and finding the relationships governing the measured quantities is always a unique and evolving issue in computer-aided drug design (CADD) [1]. Pharmaceutical chemists have been considering the use of computational chemistry and molecular modeling to build effective chemical compounds using computers in the last years [2,3]. The advantages of computational methods are reducing the synthesized chemical compounds in finding the potent compounds, acceleration of calculations and experiments with reliable prediction through the pharmacological properties of the molecular structure, and reducing the use of animal experiments and clinical trials [4]. Therefore, presenting a logical, quantitative structure-activity relationship (QSAR) model with high prediction ability and interpretability has always been considered. A high-performance QSAR model should demonstrate strong prediction ability for novel chemical compounds with the least amount of physicochemical parameters or descriptors. As a result, developing new and efficient variable selection methods to reduce data dimension in this study area is essential. Variable selection procedures improve the interpretability of the produced model by eliminating unnecessary and duplicate descriptors. In the last years, different variable selection methods such as classical, penalized, and several other new approaches such as genetic algorithms, particle swarm optimization, and ant colony optimization have been used to select the most relevant descriptors in QSAR studies [1,2,5–16]. Classical variable selection methods such as

* Corresponding author. ; .
E-mail addresses: Zeinab.Mozafari@shahroodut.ac.ir, chemist.mozafari1991@gmail.com (Z. Mozafari).

https://doi.org/10.1016/j.chemolab.2022.104510
Received 11 October 2021; Received in revised form 24 January 2022; Accepted 27 January 2022
Available online 1 February 2022

Table 1
LAD-LASSO selected descriptors and their meaning and the computed standardized beta values.

	No	Name	Description	Sub- Category	β_{Std}
Dataset A	1	Ms	Electro-topological State	Constitutional indices	0.65
	2	Mor28p	signal 28/weighted by polarizability	3D-MoRSE descriptors	−0.51
	3	MATS2m	Moran autocorrelation of lag 2 weighted by mass	2D autocorrelations	0.38
	4	B09NO	Presence/absence of N − O at topological distance 9	2D binary fingerprints	−0.37
	5	nSO2	number of sulfites (thio-/dithio-)	Functional group counts	0.27
	6	F03OO	Frequency of C − O at topological distance 3	2D frequency fingerprints	−0.24
	7	N069	Ar-NH2/X–NH2	Atom-centered fragments	0.22
	8	Mor13u	signal 13/unweighted	3D-MoRSE descriptors	0.19
	9	Mor32 m	signal 32/weighted by mass	3D-MoRSE descriptors	0.17
	10	Mor04 m	signal 04/weighted by mass	3D-MoRSE descriptors	−0.14
Dataset B	1	nArCO	number of ketones (aromatic)	Functional group counts	−0.65
	2	AMW	average molecular weight	Constitutional indices	0.52
	3	Mor21 m	signal 21/weighted by mass	3D-MoRSE descriptors	−0.42
	4	GGI9	topological charge index of order 9	2D autocorrelations	0.39
	5	nArX	number of X on aromatic ring	Functional group counts	0.35
	6	R4m	R autocorrelation of lag 4/weighted by mass	GETAWAY descriptors	−0.27
	7	MATS2m	Moran autocorrelation of lag 2 weighted by mass	2D autocorrelations	0.17
	8	RDF135 m	Radial Distribution Function - 135/weighted by mass	RDF descriptors	0.16
	9	CIC2	Complementary Information Content index (neighborhood symmetry of 2-order)	Information indices	0.14
	10	C001	CH3R/CH4	Atom-centered fragments	0.14
	11	Mor29 m	signal 29/weighted by mass	3D-MoRSE descriptors	0.10
	12	RDF020 m	Radial Distribution Function - 020/weighted by mass	RDF descriptors	−0.09
	13	RDF105 m	Radial Distribution Function - 105/weighted by mass	RDF descriptors	−0.08
	14	RDF130 m	Radial Distribution Function - 130/weighted by mass	RDF descriptors	0.06
Dataset C	1	AMW	average molecular weight	Constitutional indices	0.61
	2	RDF110 m	Radial Distribution Function - 110/weighted by mass	RDF descriptors	0.56
	3	Mor12e	signal 12/weighted by Sanderson electronegativity	3D-MoRSE descriptors	−0.54
	4	nCp	number of terminal primary C(Sp^3)	Functional group counts	0.44
	5	Mor29 m	signal 29/weighted by mass	3D-MoRSE descriptors	0.33
	6	F084	F attached to C1(Sp^2)	Atom-centered fragments	0.31
	7	RDF130 m	Radial Distribution Function - 130/weighted by mass	RDF descriptors	0.15
	8	nArOR	number of ethers (aromatic)	Functional group counts	−0.13
	9	RDF070 m	Radial Distribution Function - 070/weighted by mass	RDF descriptors	−0.08

forward selection, backward elimination, and stepwise regression methods have been considered in QSAR studies [17–20]. Classical approaches are based on the least square (LS) method with well-known limitations. The LS method fails to give significant results dealing with high-dimensional data (a small number of samples and a large number of variables). This deficiency is due to the high correlation between the variables, multicollinearity, and rank deficiency of the design in the linear predictor term.

"Several penalized regression methods have been developed, and have a unique framework and benefits, such as the ridge regression (Frank and Friedman 1993) [21], the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) [22], the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) [23], or the adaptive LASSO (Zou 2006) [24]. It is worth noting that these penalized regression methods strongly connect to the LS approach. Many researchers have recently worked on LS regression in conjunction with the LASSO. In both "small p, large n" and "large p, large n," LS-based LASSO produces a variety of intriguing results in terms of variable selection, estimation, and prediction properties in QSAR studies [25–29]. Although LS-based LASSO has good efficiency as either variable selection or modeling techniques in the QSAR studies."

"In the context of variable selection, LS-based LASSO has drawbacks such as low stability and sparsity and high bias in estimating large coefficients. Also, it is well known that outliers can cause severe issues for LS-based methods, such as LASSO, because of their sensitivity to outliers in finite samples. As a result, in the presence of outliers, a robust criterion should be used instead of the LS one. Accordingly, Wang and colleagues introduced a LAD criterion coupled with the L1 norm penalty function, which may choose relevant variables while compensating for vertical outlier observations [30]. The LAD approach is robust against heavy-tailed errors and severe outliers. Adding the L1 norm penalty function to the LAD estimator can obtain a spars model (LAD-LASSO) with a low bias for robust analysis, select the variables, and estimate the parameters simultaneously [15]. So applying such a robust method causes better prediction against the LS-based methods [30,31]. Due to the drawbacks of LASSO as an LS-based penalized method, LAD-LASSO benefits from both inherent advantages of the LAD and LASSO at the same time. So, following in the footsteps of our earlier research [32,33], the robust LAD-LASSO was used as the new variable selection method in the QSAR studies." So, the LAD-LASSO variable selection method was combined with the artificial neural network (ANN) modeling method to predict the biological activity of numerous datasets containing a variety of chemical structures. The ANN, as a powerful modeling method, can predict the biological activity of similar external compounds by establishing a relationship between the selected independent variables of the LAD-LASSO method and the dependent variable. The used training algorithm of the ANN in this study is a Levenberg-Marquardt (LM) training function. The LM algorithm is a hybrid of gradient descent and Gauss-Newton that has been utilized in QSAR studies to generate nonlinear ANN-based models. LM is the standard method that solves nonlinear least-squares and seeks to find the least multivariate nonlinear function. Therefore, the constructed ANN models of this study were trained using the LM algorithm due to several intrinsic advantages.

This study tries to present QSAR models with acceptable prediction ability and appropriate interpretability using an ANN model coupled with the LAD-LASSO variable selection method. "To the best of our knowledge, there is recently only one report on the combination of LAD and penalized method in QSAR studies, where AlDabbagh and et al. used a combination of LAD and bridge methods to predict the biological activity of 108 influenza virus neuraminidase. Results of this study revealed that the proposed LAD-bridge model exhibits superior predictability and robustness compared to alternative penalized modeling approaches [34]." So far, no research has been published on combining the LAD-LASSO variable selection method with the ANN-based QSAR modeling method. As a result, three datasets containing human immunodeficiency virus (HIV) and cancer inhibitors were subjected to the

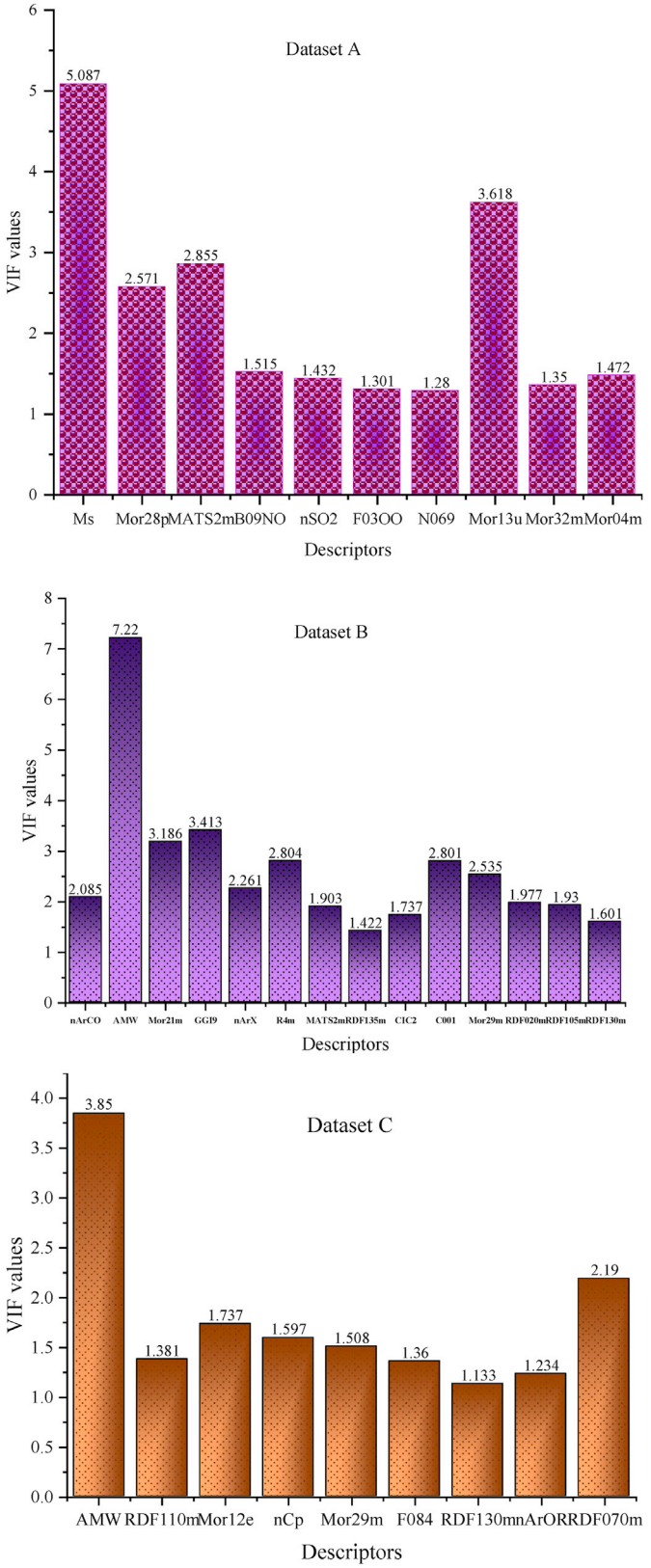


Fig. 1. The VIF values of the LAD-LASSO selected descriptors in different datasets A, B, and C.

LAD-LASSO approach, which is highly capable of variable selection. As inputs to the ANN modeling method, selected variables at the specified tuning parameter of λ were used. The ANN was trained using the LM training function, and the biological responses of the external chemical compounds were predicted after optimizing the ANN parameters. The results indicate that the LAD-LASSO-ANN models can produce reliable predictions across all studied datasets. In order to conduct a more thorough investigation, numerous statistical procedures were used, including applicability domain, Y-randomization, and the computation of different statistical parameters. All of the evaluation techniques demonstrate that the models effectively predict the biological activities of chemical compounds. Finally, several new potent chemical compounds with appropriate predicted biological activity were suggested using the model descriptors. Furthermore, the new compound-receptor interaction was also investigated using a molecular docking study. Additionally, the compounds are accepted in terms of drug-likeness rules [35–37], bioavailability score, and also, due to the ease of synthetic accessibility degree, the synthesis of compounds is easily possible.

2. | Theory

Consider the following multiple linear regression model:

$$y_i = x_i^{'}\beta + \in_i \quad \mathrm{i} = 1, \ \&., \mathrm{n} \tag{1}$$

where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})^{'}$ is the ith p-dimensional vector of molecular descriptors, and $\beta = (\beta_1, \ldots, \beta_p)^{'}$ is the vector of regression coefficients, and $\in_i$ is the ith random error component with the median equal to 0. As per the OLS technique for estimating the regression coefficients, one minimized $\sum_{i=1}^{n}(y_i - x_i^{'}\boldsymbol{\beta})^2$ with respect to $\boldsymbol{\beta}$. However, in the presence of outliers, the OLS estimates are sensitive, and instead, we may use a robust method such as the least absolute deviations (LAD) obtains as

$$\widehat{\boldsymbol{\beta}}^{LAD} = \underset{\boldsymbol{\beta}\in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - x_i^{'}\boldsymbol{\beta}| \tag{2}$$

when there are insignificant variables in the data, we can select important variables using automatic selectors such as LS-based LASSO, assuming sparsity. The LASSO has the following form:

$$LASSO = \sum_{i=1}^{n}\left(y_i - x_i^{'}\beta\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{3}$$

where $\lambda > 0$ is the tuning parameter. The larger the value of λ, the greater shrinking and lower the number of regression coefficients; otherwise, the smaller the value of λ, the greater the non-zero coefficients. Since LASSO uses a unique λ value for shrinking all coefficients, the same adjustment parameter is applied to all parameters, and the model suffers from bias. Adding the L1 norm penalty function to the LAD estimator can obtain a spars model with a low bias for robust analysis. In comparison to LAD, LAD-LASSO can select the variables and estimate the parameters simultaneously [15]. It is given by

$$LAD - LASSO = \sum_{i=1}^{n}|y_i - x_i^{'}\beta| + \lambda \sum_{j=1}^{p}|\beta_j| \tag{4}$$

The LAD-LASSO estimator can be considered as a Bayesian estimator so that each regression coefficient βj has a normal prior distribution with scale parameter $n\lambda_j$ and so $\widehat{\lambda_j} = \frac{1}{n|\beta_j|}$, which is estimated using the ordinary LAD estimation technique. The LAD-LASSO can be obtained using the "quantreg" package in R-program without complex computing programming.

Table 2
The optimized parameters for the best constructed LAD-LASSO-ANN models for various datasets.

Dataset	ANN Topology	Transfer Function	Training algorithm	Epoch	$MSE_{validation}$	$MAE_{validation}$	$R^2_{validation}$
Dataset A	5-6-1	logsig	LM	5	0.12	0.33	0.91
Dataset B	14-2-1	logsig	LM	5	0.09	0.24	0.90
Dataset C	7-4-1	tansig	LM	40	0.08	0.21	0.86

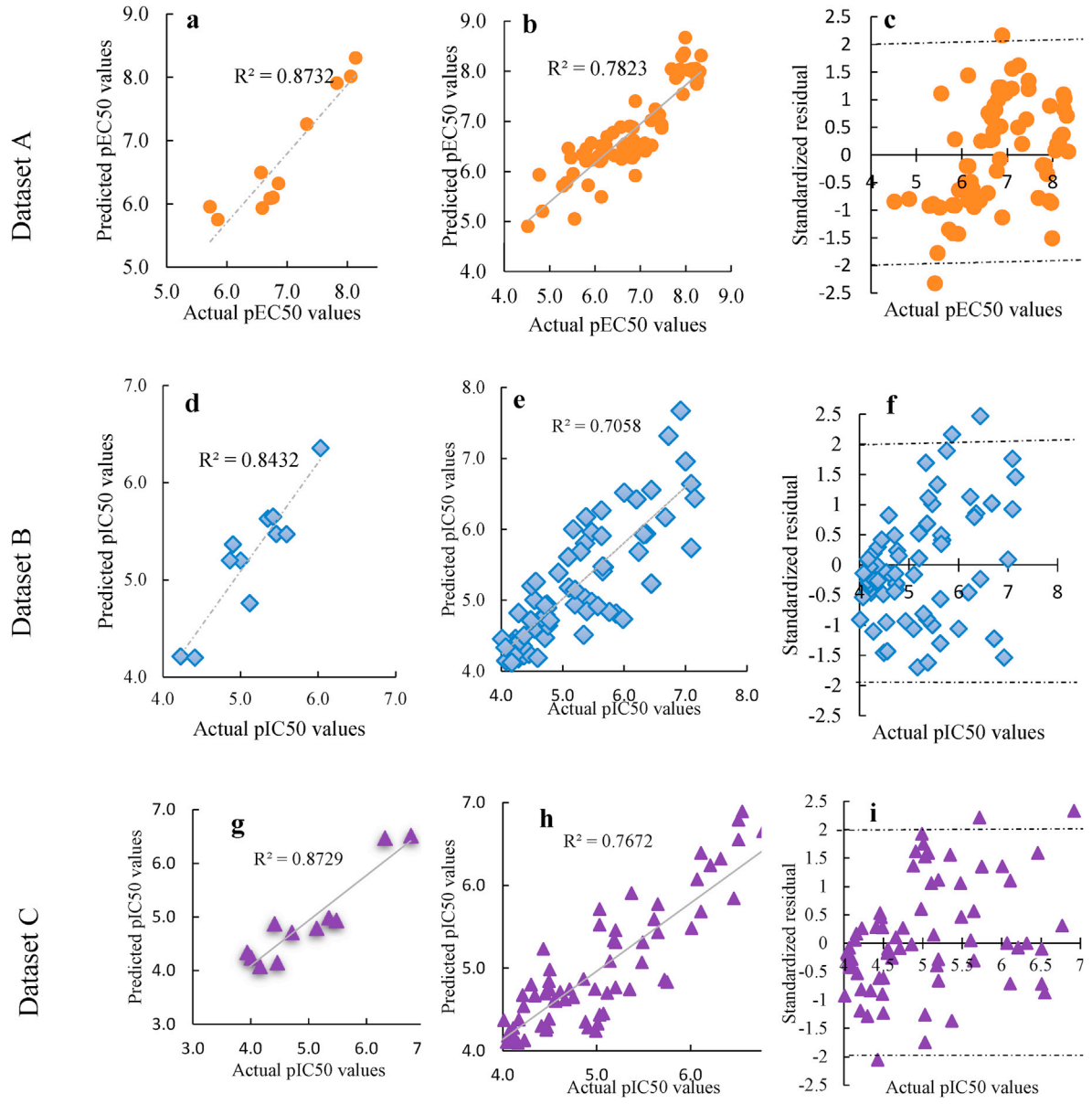


Fig. 2. The graph of predicted against to the actual response for all datasets (a) external set (b) LOO technique (c) Standardized residual graph for LOO.

3. | Computational study

3.1. | Datasets

To evaluate the efficacy of the LAD-LASSO approach, three datasets were examined, including 73 HIV inhibitors (dataset A) [38–42], 72 human colorectal carcinoma inhibitors (dataset B), and 70 human lung cancer inhibitors (dataset C) [43–45]. The structural details and the biological activities are given in Tables S1, S2, and S3 in the supplementary material. EC_{50} is half the maximum effective concentration of the studied compound and refers to the compound concentration to achieve a 50% effect. IC_{50} is also a half-maximum inhibitory concentration indicating the compound required to inhibit the biological process. EC_{50} and IC_{50} values were converted to the p-function (pEC_{50} and pIC_{50}) and used as informative dependent variables.

3.2. | Draw and optimization

It is necessary to optimize the structure of the investigated compounds in order to calculate accurate molecular descriptors with precise values for the compounds under investigation. As a result, the two-dimensional structures of all examined compounds were created in the

Table 3
Computed statistical factors for the best LAD-LASSO-LM-ANN models for the prediction of response values.

NO.	Formula	Dataset A		Dataset B		Dataset C		Acceptable range
		Test set	LOO	Test set	LOO	Test set	LOO	
		5-6-1 LAD-LASSO-LM-ANN		14-2-1 LAD-LASSO-LM-ANN		7-4-1 LAD-LASSO-LM-ANN		
1	$MAE = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$	0.29	0.36	0.23	0.30	0.28	0.29	$<0.1 \times Range_{Train}$
2	$REP(\%) = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	5.62	6.61	5.14	9.37	6.54	7.76	
3	$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$	0.13	0.20	0.07	0.24	0.11	0.15	–
4	$MRE = \frac{\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert}{n} \times 100$	5.24	5.92	4.42	6.84	5.89	5.93	–
5	$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	0.87	0.78	0.84	0.71	0.87	0.77	>0.6
6	R_0^2	0.86	0.72	0.77	0.66	0.85	0.73	Close to R^2
7	$Relative R_0^2 = \frac{(R^2 - R_0^2)}{R^2}$	0.01	0.08	0.08	0.07	0.02	0.05	<0.3
8	$R_m^2 = R^2 \times [1 - (R_0^2 - R^2)]$	0.78	0.59	0.62	0.55	0.75	0.62	Close to R^2
9	$R_0'^2$	0.80	0.78	0.84	0.70	0.87	0.76	>0.5
10	$Relative R_0'^2 = \frac{(R^2 - R_0'^2)}{R^2}$	0.08	0.00	0.00	0.01	0.00	0.01	<0.3
11	$R_m'^2 = R^2 \times [1 - (R_0'^2 - R^2))^{1/2}]$	0.65	0.72	0.57	0.59	0.73	0.66	Close to R^2
12	$R\text{-}R = \lvert R_0^2 - R_0'^2\rvert$	0.06	0.06	0.07	0.01	0.02	0.03	<0.1
13	k	0.97	1.00	0.98	0.99	0.98	1.00	$0.85 \le k \le 1.15$
14	k'	1.03	1.00	1.02	1.00	1.01	1.00	$0.85 \le k \le 1.15$

Table 4
Calculated model fit estimates, Q_{F1}^2, Q_{F2}^2 and Q_{F3}^2, for the datasets A, B, and C.

Dataset	Q_{F1}^2	Q_{F2}^2	Q_{F3}^2
A	0.77	0.81	0.87
B	0.75	0.80	0.86
C	0.99	0.99	0.99

Hyperchem software for all three datasets. The optimization process was carried out using the Polak–Ribière (conjugate gradient) algorithm, with the RMS gradient of 0.001 as the termination criterion for each molecule. The optimized structures were stored with the *.hin extension and used as input to DRAGON computing software [46].

3.3. | *Descriptor generation and screening*

Molecular descriptors were calculated in the DRAGON program utilizing optimal structures. By utilizing the caret package in R, descriptors with constant values and descriptors with variances close to zero were eliminated from datasets. These descriptors provided no meaningful information to the model and were therefore discarded [47]. In addition, among the two high correlated descriptors (above 0.9), the descriptor with the high relevance to the corresponded response remained, and the other was deleted. After preprocessing, the total descriptors were reduced from 3224 descriptors to 335 (Dataset A), 438 (Dataset B), and 429 (Dataset C) descriptors. These descriptors were separately used as independent variables in the LAD-LASSO method.

3.4. | *Selection of the effective descriptors*

Following the preprocessing step, the LAD-LASSO variable selection approach was applied to use the computed molecular descriptors as independent variables and the corresponding biological responses as the dependent variable. The datasets were divided into three parts using the Kennard-stone algorithm and the Euclidean distance technique to implement the variable selection method: training set data (about 70% of the total data), validation set data (about 15% of the total data), and test set data (about 15% of the total data). The LAD-LASSO variable selection method was used to choose the most relevant descriptors associated with the biological responses using the whole training and validation sets. Furthermore, the test data sets were omitted from the beginning in order to choose effective descriptors without the effect of the external groups of chemical compounds. Finally, in evaluating the ANN modeling, the absence of test sets data at any stage from variable selection to modeling

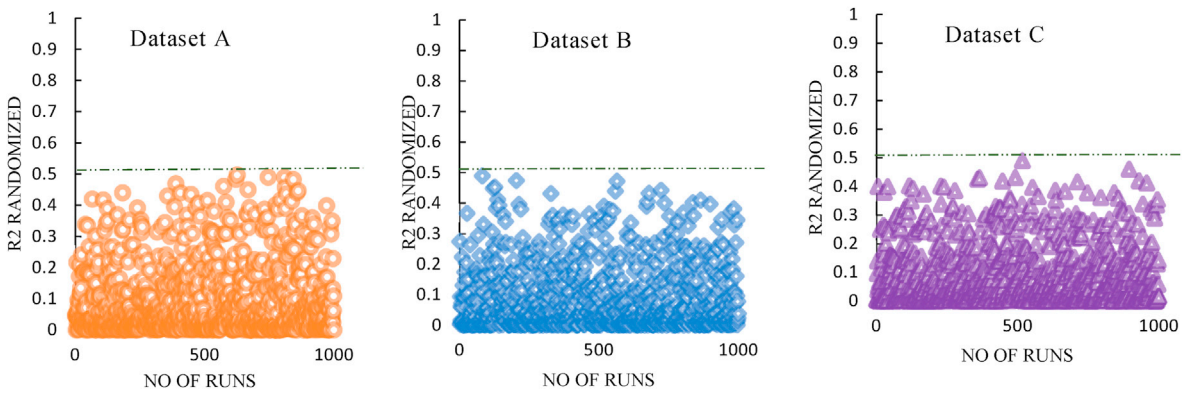


Fig. 3. The plot of Y-randomization for all studied datasets A, B, and C.

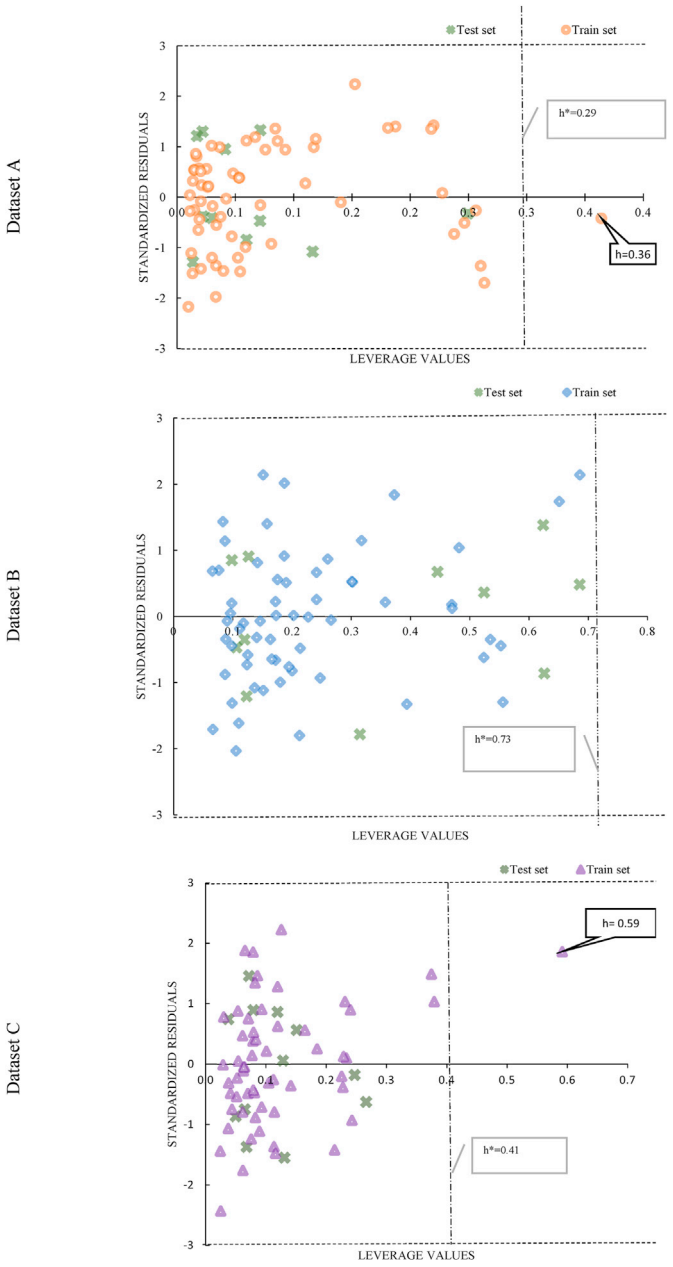


Fig. 4. Plot illustrating the AD for all datasets for the suggested ANN-based models.

is a suitable proof of model superiority. Therefore, descriptors with a non-zero coefficient (the numbers of 10, 9, and 14 descriptors for datasets A, B, and C, respectively) in the corresponding λ were extracted and used as input to the nonlinear ANN modeling. The name of descriptors, the category of each selected descriptor, and the standardized coefficient values were summarized in Table 1.

3.5. | *ANN modeling*

The relationship between the LAD-LASSO selected descriptors and the appropriate biological response was established using an artificial neural network (ANN) model. As a result, a three-layer ANN model with a backpropagation error technique was employed. The Levenberg–Marquardt (LM) training procedure was used to create ANN models for all three datasets (trainlm in Matlab Toolbox). Several ANN parameters were optimized simultaneously, and the optimal architectures of the ANN models are 5-6-1, 14-2-1, and 7-4-1 for datasets A, B, and C, respectively. The entire modeling process was carried out using the MATLAB program on a Windows operating system with a personal computer with processor properties of Intel Core i7-4790 K 4.0 GHz, RAM 8 GB.

3.6. | *Docking study*

The interaction between the active site amino acids of corresponding receptors and the suggested chemicals was determined using molecular docking. The molecular docking procedure was carried out with the assistance of the Autodock4.2 program [48]. The desired receptors for each dataset were selected according to the recommendations of the previous studies [38–45]. So the 3MEC (resolution equal to 2.30 Å) and the 3HHM receptors (resolution equal to 2.80 Å) were extracted from the protein data bank site for HIV and cancer inhibitors, respectively [49]. The downloaded receptor file was called in viewer lite software, and further preparation such as removing water molecules and cofactors and sub-chain without crystallographic ligand (cognate ligand) was done. Finally, the remaining structure was considered as studied receptors structures and stored with *.pdb extension. The prepared file of the receptor was called in the Autodock4.2 software. The required hydrogen was added, the non-polar hydrogens were merged, and the Kollman charge was added to balance the system charge, and finally, the receptor was saved as the pdbqt format file. The docking process was done in two steps. First, the evaluation process (dock-redock) was implemented, and the cognate ligand was docked into the active site to extract the optimum numbers of genetic algorithm runs (GA runs). The ligand was saved as the pdbqt. A grid box with 60 * 60 * 60 A^{o3} dimensions and with the coordinates of the center of gravity of the cognate ligands of each receptor (X = 50.431, Y = 63.204, and Z 13.716, = for pdb code of 3MEC and X = 60.092, Y = 62.473, and Z = 112.509 for pdb code of 3MHH) was generated. Then, the suggested compounds were separately docked into the active site of corresponding receptors using 150 runs of the Lamarckian genetic algorithm (LGA) [50]. Finally, the structure of each compound with the best conformation was extracted from the molecular docking, and the ligand-receptor interactions were further investigated using discovery studio software [51].

4. | Results and discussions

4.1. | *Selection of the significant descriptors*

The LAD-LASSO variable selection strategy was used to choose the optimal subset of all variables in this investigation. The DRAGON 5.5 software was used to calculate the molecular descriptors initially. Descriptors with the constant and near-constant values (descriptors with a near-zero variance) were deleted. Then, to reduce the effect of correlated descriptors, the correlations between the descriptors were calculated using the corecoeff command in MATLAB software [52]. Then, among the two descriptors with a correlation above 0.9, one descriptor with the highest correlation with the biological response remained, and the other was deleted. Next, the preprocessing procedure was performed for all three datasets, and the total 3224 molecular descriptors were reduced to 335, 438, and 429 for datasets A, B, and C. Then, the decreased descriptors were defined as the input of the LAD-LASSO variable selection method. Finally, the most relevant descriptors of different datasets were obtained after applying the LAD-LASSO variable selection method to the training and validation datasets. The most significant descriptors selected using the LAD-LASSO variable selection method were equal to 10, 9, and 14 molecular descriptors for datasets A, B, and C. Table 1 summarizes the names and categories of these descriptors. The most effective descriptors were used as ANN modeling inputs. The MTE package was used to implement the LAD-LASSO code in R [53].

The existence of collinearity was also examined among selected LAD-

Table 5
Physiochemical properties of the new designed compounds and studied compound.

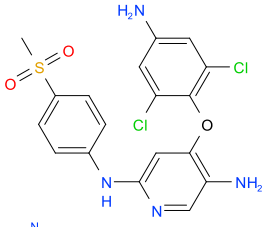
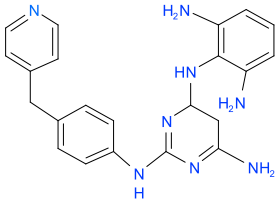
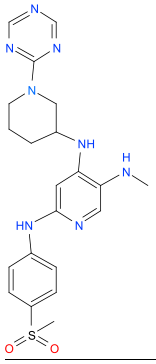
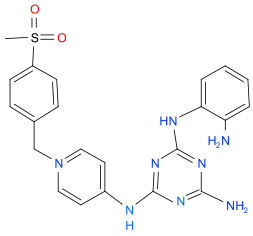
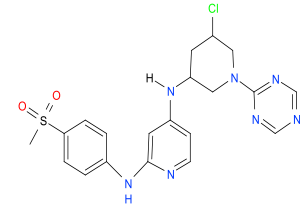
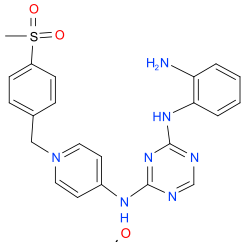
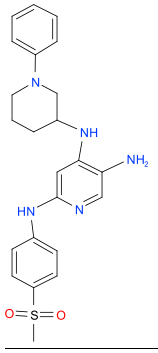
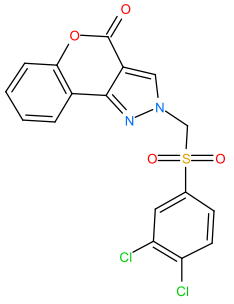
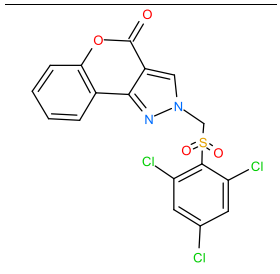
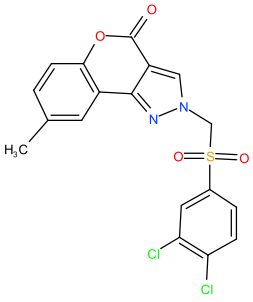
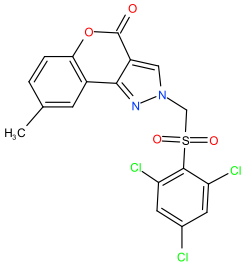
No.	Chemical Structures	Biological activity	Leverage	No.	Chemical Structures	Biological activity	Leverage
NC1 (A)		8.45	0.12	NC2 (A)		7.63	0.06
NC3 (A)		7.50	0.04	NC4 (A)		7.23	0.03
NC5 (A)		7.11	0.04	NC6 (A)		6.94	0.02
NC7 (A)		6.63	0.02	NC8 (B)		8.92	0.66
NC9 (B)		8.78	0.60	NC10 (B)		7.52	0.28
NC11 (C)		6.43	0.25				

Table 6
Drug-likeness and bioavailability parameters of the new designed compounds and studied compound.

NO	MW	#H-bond acceptors	#H-bond donors	MLOGP	Lipinski #violations	MR	WLOGP	nAT	Ghose #violations	TPSA	#Rot. bonds	Veber #violations	SA	BAS
NC1	404.9	4	3	2.05	0	107.2	4.94	44	0	128.71	5	0	3.1	0.55
NC2	406.5	5	5	-0.45	0	122.9	1.13	56	0	156.92	6	1	3.47	0.55
NC3	454.6	6	3	0.85	0	127.8	2.85	56	0	133.41	7	0	3.98	0.55
NC4*	464.5	5	4	0.45	0	134.3*	2.95	57	1	169.65*	7	1	4.34	0.55
NC5	460.0	6	2	1.35	0	123.3	3.22	53	0	121.38	6	0	4.17	0.55
NC6	453.6	6	3	1.16	0	129.4	3.04	59	0	134.5	7	1	3.54	0.56
NC7	437.6	3	3	2.4	0	129.6	4.41	58	0	108.73	6	0	3.89	0.55
NC8	409.2	5	0	3.35	0	99.67	4.81	36	0	90.55	3	0	3.51	0.55
NC9	443.7	5	0	3.58	0	104.68	5.46	36	0	90.55	3	0	3.58	0.55
NC10*	457.7	5	0	4.07	0	109.65	5.77	39	1	90.55	3	0	0.55	3.58
NC11	423.2	5	0	3.58	0	104.64	5.12	39	0	90.55	3	0	0.55	3.51

* The starred highlighted items are related to the compound that has been rejected by at least one of the Ghose or Veber rules

LASSO descriptors for all three datasets. The following equation was used to calculate the value of the variance inflation factor (VIF parameter):

$$VIF = \frac{1}{1 - R_i^2} \qquad i = 1, 2, 3, \&, p \qquad 5$$

R_i^2 is the square of multiple correlation coefficient generated from the regression of variable i on other variables, where p is the number of LAD-LASSO selected descriptors. VIF values less than ten indicate no collinearity between LAD-LASSO selected descriptors. Therefore, the results summarized in Fig. 1 show that the VIF values of all descriptors are less than 10 [54], and there is no collinearity between the LAD-LASSO selected descriptors of all three datasets.

4.2. | *Optimization of the ANN parameters*

The association between the LAD-LASSO chosen descriptors and the corresponding biological response was established using a nonlinear feed-forward ANN model with an error backpropagation technique. A multi-layer perceptron ANN model with one input layer, hidden layer, and one output layer was created for this purpose. According to the literature, using one hidden layer is sufficient for most prediction problems [55,56]. The number of neurons in the input layer, the number of the hidden layer inputs, the number of epochs, and the types of activation functions have all been optimized to find the best ANN architecture. The linear purlin function was used as the output layer, and the hyperbolic tangent sigmoid (tansig) and logarithmic sigmoid (logsig) were used as transfer functions. "It is clear that having a small and optimal number of ANN inputs for optimization is preferable. Obviously, 10, 14, and 9 selected descriptors using the LAD-LASSO method are not necessarily optimal subsets. Therefore, the best subset of descriptors should be selected among all possible subsets with 2 to entire selected descriptors, which are greater than 10^{10} subsets (according to the multiplication of descriptors modes by the number of nodes and the number of training epochs). The examination of the performances of such a lot number of subsets as the ANN inputs is more time-consuming. To consider the importance of each descriptor in the nonlinear ANN model, the selected descriptors were arranged according to the magnitude of the standardized coefficients (Table 1). Among all created subsets, a subset consisting of 5, 14, and 7 descriptors with the highest importance was selected as the optimal subset during the ANN optimization." So, different ANN structures (900, 1260, and 810 ANN structures for diverse datasets A, B, and C) were designed with various inputs (LAD-LASSO selected descriptors arranged based on the magnitude of the standardized coefficient). The training of the different architectures of ANN models was performed using training set data with the LM training function. The optimal structure of the ANN model was selected according to the minimum value of the $MSE_{validation}$. The optimal conditions of the best ANN structure for all three datasets are summarized in Table 2. In order to predict the test set data, the optimal ANN structure of each relevant dataset was used.

"Therefore, to compare the proposed method, the penalized LS-based LASSO variable selection method was used to evaluate the efficiency of LAD-LASSO. After applying the LS-based LASSO variable selection method to all three datasets (A, B, and C), 22, 55, and 27 molecular descriptors were obtained. Since the use of a large number of descriptors in ANN modeling causes overfitting [57], it is not suggested to utilize this number of descriptors to optimize the ANN approach. Therefore, the selected LS-based LASSO descriptors for all three datasets were arranged individually based on the magnitude of the standardized coefficients. An equal number of descriptors of the optimum LAD-LASSO-ANN model were selected (5, 14, and 7 for A, B, and C datasets, respectively) from LS-based LASSO selected descriptors based on the order of magnitude of the standardized coefficients and defined as ANN inputs. After optimizing the ANN parameters, the superior ANN model was obtained based on the minimum MSE value of the validation set for each dataset. The LASSO-ANN with an optimal structure of 5-2-1 and 20 training epochs, LM as a training function, and a logarithmic sigmoid as the transfer function has a minimum $MSE_{validation}$ of 0.17 for dataset A. The optimum LASSO-ANN for the dataset B is 14-2-1 with training epochs equal to 5, LM as a training function, and a tangent hyperbolic sigmoid as the transfer function with the $MSE_{validation}$ equal to 0.11. LASSO-ANN with

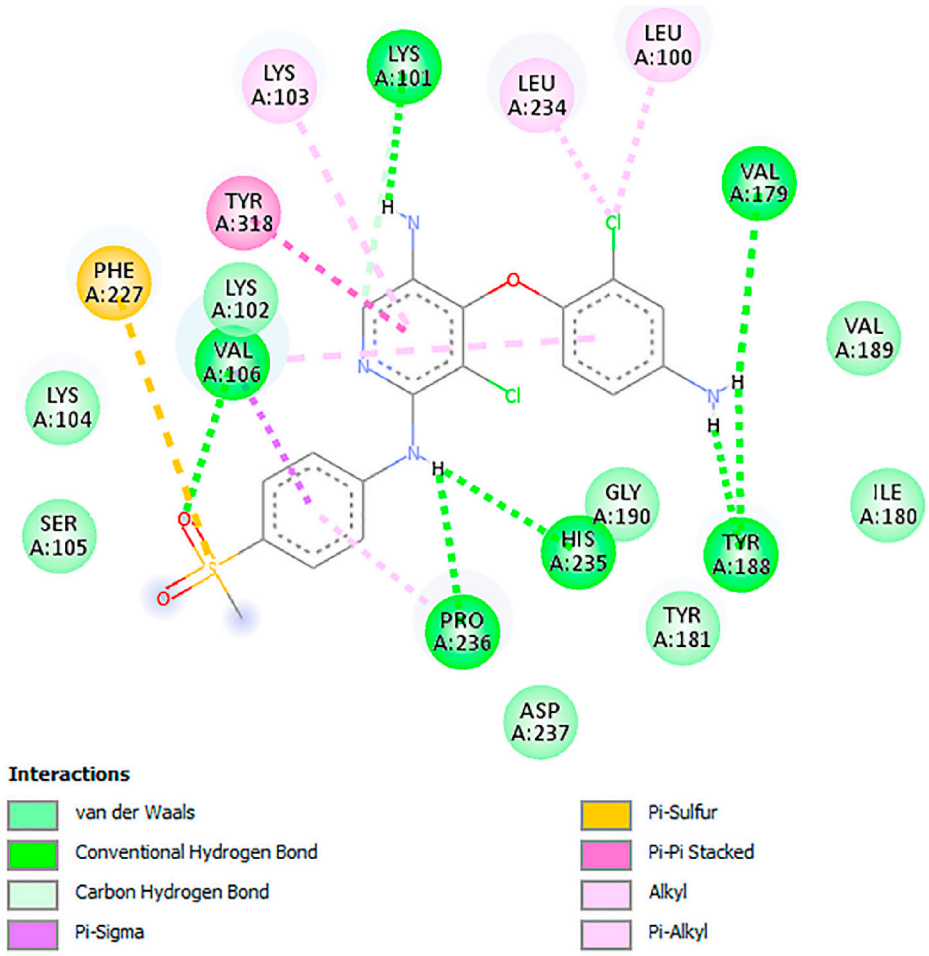


Fig. 5. The interaction of NC1 with 3M8Q.

7-3-1 architecture, 5 training epochs, LM as a training function, and a tangent hyperbolic sigmoid as the transfer function has a minimum $MSE_{validation}$ of 0.15 for dataset C."

4.3. | *Validation of the model*

The main goal of any QSAR modeling is to create a robust model with a high capacity to predict the biological activity of new compounds (test set data as the external set) accurately and reliably. Therefore, different criteria can be used to evaluate the developed ANN models, such as using validation (internal test) set, test (external test) set, and leave one out (LOO) technique for the prediction of the whole dataset. In addition, for the evaluation of produced QSAR models, the Y-randomization test, the applicability domain (AD), and the calculation of statistical parameters are commonly utilized.

In order to evaluate the prediction potency of the optimal ANN models for each data set, the biological activities of both validation and test sets data were predicted under the optimal conditions. The predicted values are given in Tables S1–S3. In addition, the predicted values were plotted in terms of actual values (Fig. 2a,d, and 2g). The determination coefficient (R^2) value above 0.6 indicates the acceptable prediction ability of the LAD-LASSO-ANN model in all three studied datasets.

The LOO technique was also used to evaluate the goodness of the optimum ANN models. To achieve this, the optimal ANN model was used, with the difference being that in this technique, one data point was extracted once as a test data point. After that, the model was trained with only the remaining data, the extracted data point was predicted. This procedure was repeated for the total number of data to predict the biological response of whole data once as a test set. Then the graph of the predicted values was plotted in terms of the actual values. The value of Q^2_{LOO} for all three studied data is greater than the acceptable value of 0.5, which indicates the high prediction ability and generalizability of the optimal ANN models with the selected descriptors of the LAD-LASSO method.

The prediction results of the model were also evaluated using standardized residual plots. So, the standardized residuals (r_i) were computed for the predicted values of the LOO technique according to the following equation:

$$r_i = \frac{e_i}{s_{e_i}} = \frac{(y_i - \hat{y}_i)}{s_{e_i}} \quad \text{Eq 6}$$

in which e_i is the difference between the actual and expected responses for each observation $i = 1, \ldots, n$, and s_{ei} is the standard deviation of residual values. The obtained standard residuals are plotted in terms of actual response values in Fig. 2b, e, and 2h. According to the obtained standardized residual graph, less than approximately 5% (at 95% confidence level) of the whole dataset (equal to four chemical compounds for all datasets) are out of the significant range of ± 2. Thus, the obtained results (Fig. 2c, f, and 2i) show a reasonably random pattern and prove that the developed QSAR models provide good fitting.

Along with the methodologies discussed previously, numerous commonly used statistical parameters were employed to evaluate the best LAD-LASSO-ANN models. The statistical calculations show that all statistical parameters have an acceptable value. The parameter name, calculation formula, calculated values for different datasets, and tolerable range are summarized in Table 3.

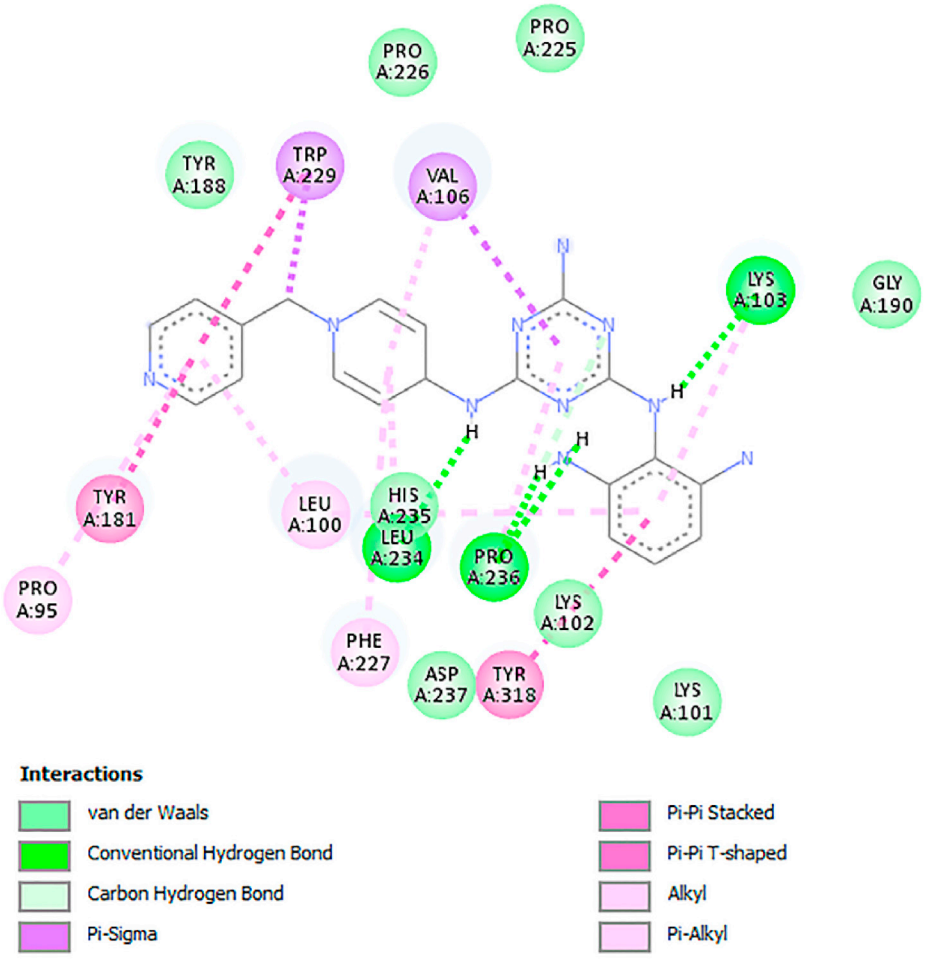


Fig. 6. The interaction of NC2 with 3M8Q.

"The results obtained confirm that the LAD-LASSO is an efficient variable selection method. However, a comparison between the efficiencies of LS-based LASSO and LAD-LASSO was conducted using LS-based LASSO selected descriptors as inputs of the optimized ANN model (LASSO-ANN denoted in section 4.2). Then, the biological activities of test compounds of all data sets (A, B, and C) were predicted using the optimum LASSO-ANN models. The MSE values of 0.19, 0.11, and 0.27 were obtained for test compounds of data sets A, B, and C, respectively. Comparing the obtained results with those of LAD-LASSO-ANN (Table 3), it can be seen that the selected LAD-LASSO descriptors have provided ANN models with good prediction ability and accuracy rather than LS-based LASSO as a penalized variable selection method."

"In addition to the calculated statistical parameters in Table 3, several other Q^2 dependent parameters such as Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} were calculated using the following equations:

$$Q^2_{F1} = 1 - \frac{\mathring{a}_{i=1}^{n_{Ext}} (\hat{y}_i - y_i)^2}{\mathring{a}_{i=1}^{n_{Ext}} (y_i - \bar{y}_{Tr})^2} = 1 - \frac{PRESS}{SS_{Ext}(\bar{y}_{Tr})} \quad (7)$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{Ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{Ext}} (y_i - \bar{y}_{Ext})^2} = 1 - \frac{PRESS}{SS_{Ext}(\bar{y}_{Ext})} \quad (8)$$

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{Ext}} (\hat{y}_i - y_i)^2\right]/n_{Ext}}{\left[\sum_{i=1}^{n_{Tr}} (y_i - \bar{y}_{Tr})^2\right]/n_{Tr}} = 1 - \frac{PRESS/n_{Ext}}{TSS/n_{Tr}} \quad (9)$$

where, $\bar{y}_{Tr}$ and $\bar{y}_{Ext}$ indicate the response means of the training set and the external test set, respectively. PRESS is the predicted residual error sum of squares, TSS is the total sum of squares, the sum of squared deviations from the data set. PRESS is the sum of squares of the $SS_{Ext}(\bar{y}_{Tr})$ and $SS_{Ext}(\bar{y}_{Ext})$ are the total sum of squares of the external set calculated using the mean of the training set and external set responses, respectively. Also, the n_{Tr} and n_{Ext} are the number of training and external sets objects [58, 59]. Thus, the Q^2 value should always be accompanied by descriptive statistics of the test set used to compute it. The calculated Q^2 parameters for all three data sets are given in Table 4. The values of Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} parameters are greater than 0.50 and close to 1.0, which indicate that the external test data are uniformly distributed over the range of the training set and appropriate some basic mathematical properties of the model, such as ergodic and associative properties."

1-Mean Absolute Error, 2- Relative Error of Prediction, 3- Mean Square Error, 4- Mean Relative Error. y_i is observed (experimental) value, $\hat{y}_i$ is predicted value and $\bar{y}$ is the average value of observed values, p is descriptors numbers, and n is compounds numbers. R^2 Squared correlation coefficient between the observed and predicted value of compounds with intercept R_0^2 the squared correlation coefficient between the observed and predicted value of compounds without intercept. $R_0'^2$ Bears the same meaning as R_0^2, but uses the reversed axis. k is the slopes of predicted vs. actual and $\mathbf{k}'$ are is vise versa. The conditions of each optimum ANN model was written above each model for example 5-6-1 is the optimum conditions of LAD-LASSO-LM-ANN model for dataset A.

The Y-randomization test was used to determine the model robustness and whether or not there was a probability that the association between the independent and dependent variables was a coincidence [8]. Hence, the biological response values were randomized 1000 times in the

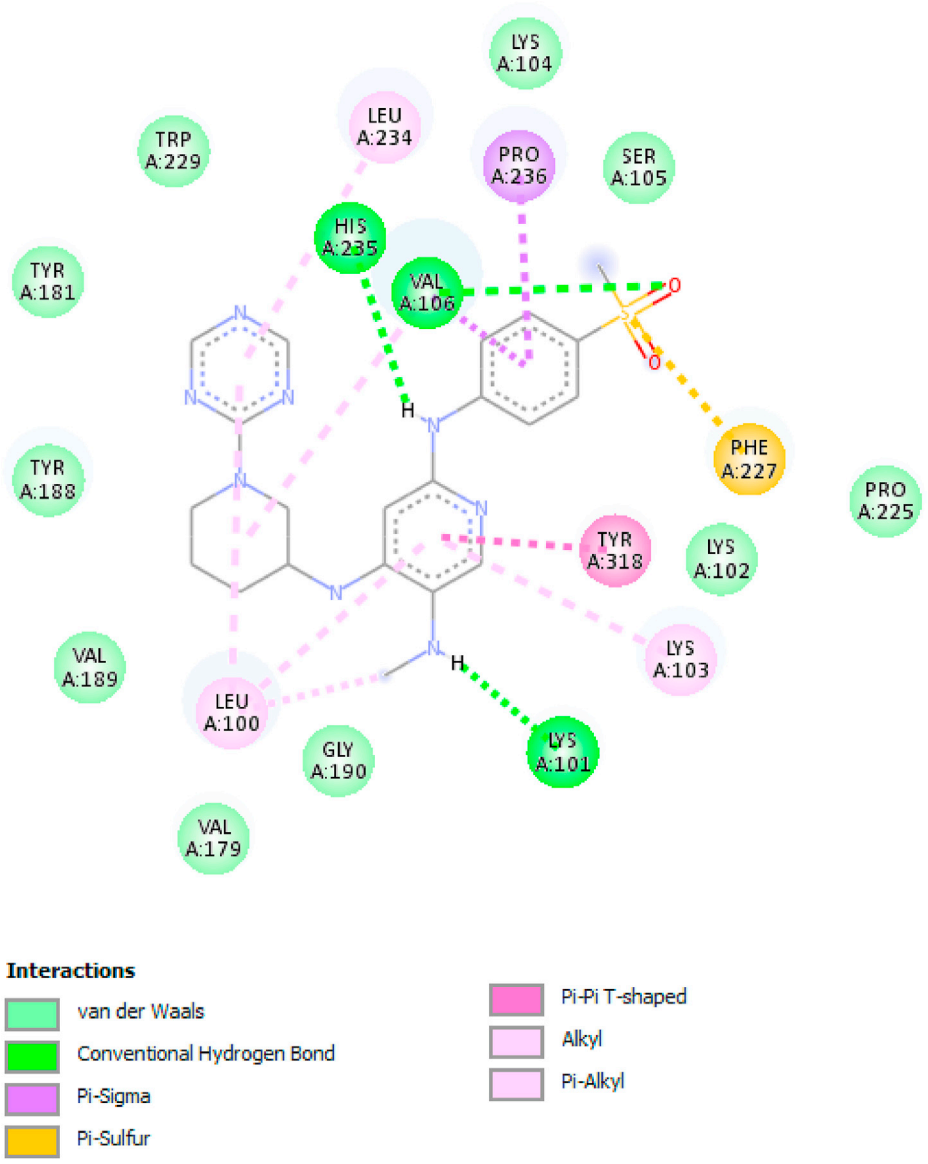


Fig. 7. The interaction of NC3 with 3M8Q.

range of its changes. The optimum LAD-LASSO-ANN models were developed with random biological responses. The biological activity values of the test set were predicted under the optimal conditions with a random response. This process was repeated 1000 times for each dataset in its own optimal LAD-LASSO-ANN model. The reliable QSAR model must have R^2 values much smaller than the acceptable value of 0.6 in random models. The predicted values were plotted in terms of actual values for every 1000 models, and the corresponding determination coefficient was obtained. The R^2 values for 1000 implementations in datasets A, B, and C are given in Fig. 3. As can be seen, R^2 values are smaller than 0.5, and the results show the robustness and goodness of the LAD-LASSO-ANN models.

Another QSAR model evaluation technique is the applicability domain (AD) analysis [60]. AD is a theoretical chemical space that is created using molecular descriptors of the training set and its relevant biological responses. Therefore, if the new chemical data is placed in this theoretical chemical space of AD, it indicates that the model has been able to predict it well in the face of the data that it has not seen (external test set). AD was obtained by calculating the leverage of the training dataset. "The leverage (H) of a query chemical is proportional to its Mahalanobis distance measure from the centroid of the training set. The leverages values are calculated for a given dataset using the following formula:"

$$h = x_i (X^T X)^{-1} x_i^T \quad (10)$$

where X is the matrix of molecular descriptors associated with the training set data and x_i is the vector of descriptors associated with each chemical data row. T also denotes the transposition of the matrix. The chemical space of AD was represented using the Williams diagram (Fig. 4). Hence the standardized residual values versus h data were plotted. For AD analysis, chemical data must be within the two confidence limits of the Williams plot. First, the chemical data should not exceed the standard deviation of 3 times the standardized residual. In addition, the h values of the chemical data should be less than the warning value of h*, which can be calculated using 3p/n. In this equation, p equals the number of model descriptors plus one, and n is the number of training set data [60]. As Fig. 4 shows, the h values of dataset B are in both acceptable ranges. In the Williams diagrams of datasets, A and C, only one data is more than h *, and the rest computed h values are satisfactory. Therefore, the robustness and reliability of the LAD-LASSO-ANN models were proven for all studied datasets. "QSAR

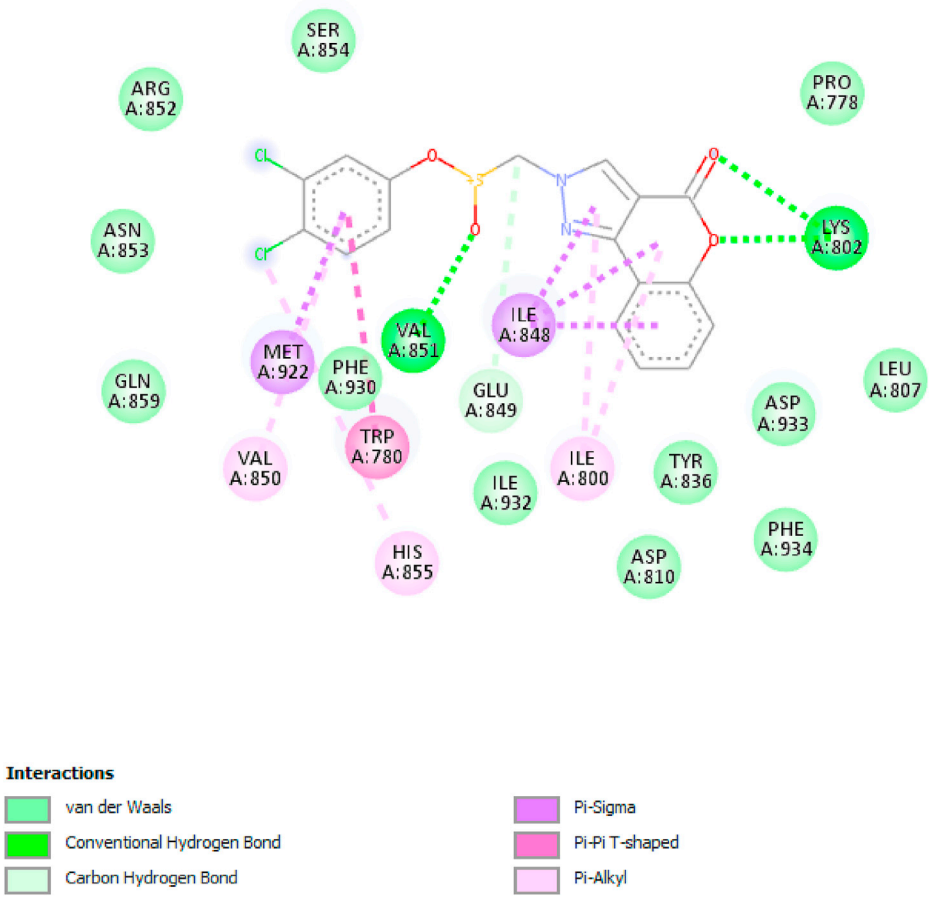


Fig. 8. The interaction of NC8 with 3HHM.

models that have been thoroughly established should be applicable in discovery of novel compounds located within the chemical space of the studied compounds [61]. It is well known that compounds located far away from the centroid value have higher leverage (h) values, indicating a general difference between the studied compounds [61]. Therefore, the similarities of the proposed compounds to the studied compounds were evaluated by calculating the h values of new compounds and examining their position within a defined domain. The calculated h values of new proposed compounds of each data set are listed in Table 5. According to the data obtained and the applicability domain of the model for all data sets (William's plots in Fig. 4), it is clear that the H values of the proposed compounds fall within the chemical range of studied data sets. Therefore, all proposed compounds have lower h values than h*, which means that the suggested compounds are accurate and reliable [61–63]."

4.3.1. The suggestion of new compounds

In this study, an attempt was made to propose new compounds with appropriate biological activity according to the selected LAD-LASSO descriptors in the model. Therefore, due to the effect of interpretable descriptors, the structures of weak compounds were modified, and several new compounds (NC) were suggested for each dataset. Then the biological activity of the proposed compounds was predicted using the superior model of each dataset. All proposed compounds were also tested for pharmacokinetic features and the accuracy of drug-likeness rules. For further exploration, compounds with appropriate biological activity were docked into the active site of the respective receptor. Finally, the docked complex (new compound- receptor) was analyzed using BIOVIA Discovery Studio Visualizer [64] to show the type of interactions (hydrophobic and hydrophilic) between the ligands and corresponding receptors. For all datasets A, B, and C, a discussion of how to perform structural modifications, how to implement molecular docking simulations, and the types of interactions is provided.

So, according to the descriptor coefficients of dataset A in Table 2, structural modifications such as adding NH_2 and SO_2 substitutions to the weak compounds (compound 40 with $pIC_{50} = 4.52$ and compound 36 with $pIC_{50} = 4.77$) cause increasing the biological activity. In addition, the negative coefficient of descriptor *F03[O–O]* indicates that compounds without this group have better biological activity. Therefore, there is no such group in the proposed structures. Due to the effects of explained descriptor, some modified structures (NC1 to NC7) have been suggested. Then, the LAD-LASSO descriptors were computed using optimized structures. Finally, the biological activities of the proposed structures were predicted using the optimum ANN modeling. As shown in Table 5, the biological activity of weak compounds had a significant improvement.

Due to the impact of the selected LAD-LASSO descriptors (Table 2), new compounds were also proposed using modifications of the weak compounds of dataset B compounds (compounds 26 and 27 with biological activity equal to 4.17 and 4.2, respectively). Obviously, the presence of halogenated groups on the aromatic ring and the absence of ketone in the proposed structures cause increasing biological activity. So, using HyperChem software, the proposed structures (NC8 to NC10) were designed and optimized, and then the selected LAD-LASSO descriptors were derived using DRAGON software. Finally, the best ANN model was used to determine the biological activity of the suggested chemicals. Table 5 summarizes the findings, which reveal that the weak structural modifications were successful, and the biological activity of the new compounds was greatly improved. The new structures of dataset C were also suggested using the influence of selected LAD-LASSO selected descriptors. As shown in Table 1, the absence of F attached to Carbon Sp^2

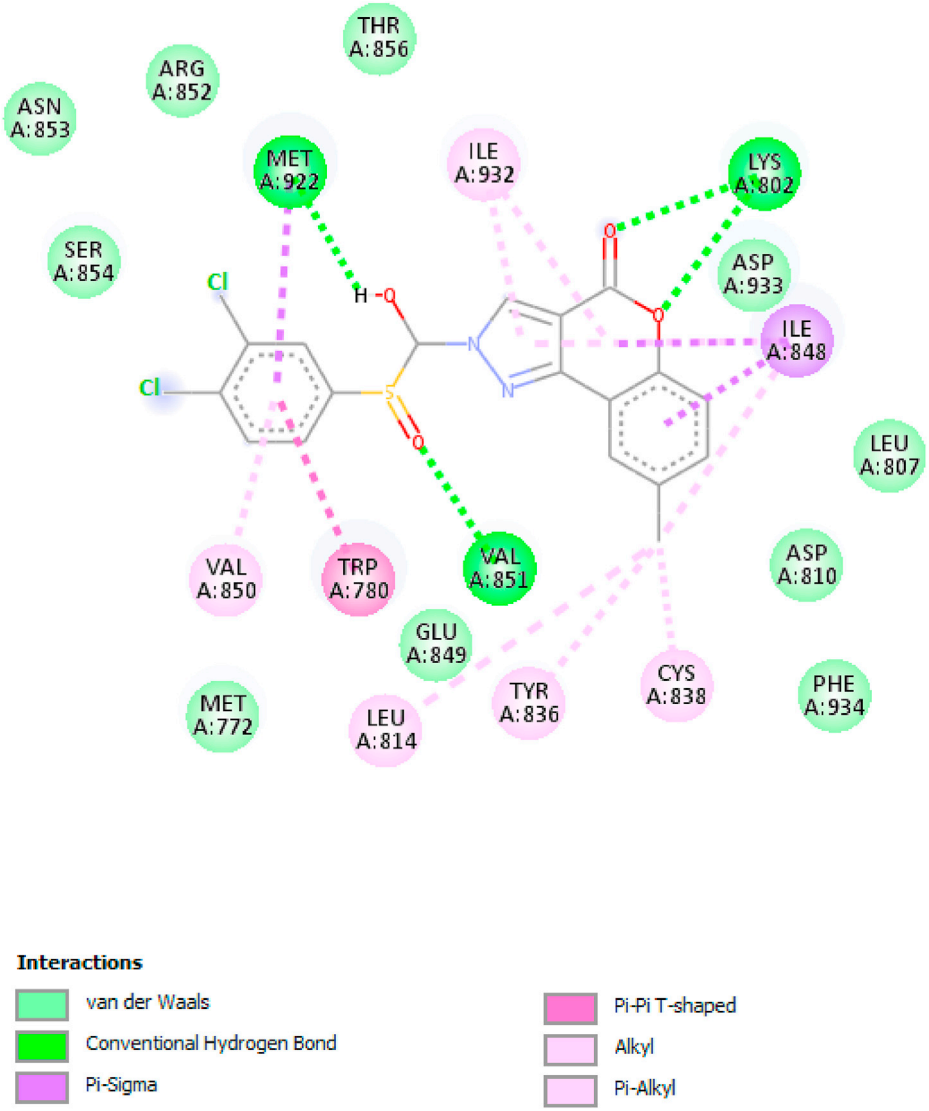


Fig. 9. The interaction of NC10 with 3HHM.

and the presence of terminal Carbon Sp^3 in the chemical structure increases the biological activity. Therefore, the proposed compounds should not have such a group. In addition, the proposed structures without alkoxy groups attached to the aromatic ring will have more biological activity. Due to the effect of the descriptions mentioned above, the weak compounds in dataset C were identified (compounds with biological activities equal to 4.01 and 4.1). Sp^3 carbon groups were added to the weak compounds, and structures without F attached to Sp^2 carbon and alkoxy attached to the ring were proposed. The LAD-LASSO model descriptors were derived using DRAGON software, and a modified structure (NC11) was designed and optimized. Once the optimal ANN was implemented, the biological activities of the proposed compounds could be predicted. Table 5 shows that the pharmacological activity of the recommended chemical compounds has improved. "An ADMET (absorption, distribution, metabolism, excretion, and toxicity) analysis was performed on the newly developed compounds to establish their drug-likeness and pharmacokinetics, respectively. The ADMET predictions can be obtained through the use of a web-based tool that is easily accessible such as SwissADME [65]. The physical and chemical qualities of a drug-like compound are critical in the progression of that molecule from an investigational new drug into a successful drug candidate. The physicochemical and pharmacokinetic properties of the drug candidate can be calculated using various rules of drug-likeness such as Lipinski (Rule of 5) [35], Ghose [36], and Veber [37]. A good drug candidate should be in excellent agreement with the most important drug-likeness rules. Rule of 5 states that the proposed drug candidate must have a number H-bond donor (expressed as the sum of OHs and NHs) lower than 5, the molecular weight (MW) lower than 500, the MLogP lower than 4.15, and H-bond acceptors (expressed as the sum of Ns and Os) lower than 10. As shown in Table 6, Lipinski's parameters were calculated for the suggested new compounds (NC1 to NC8 for dataset A, NC9-NC10, and NC11-NC12 for Datasets B and C, respectively). All the mentioned compounds are in agreement with the rule of 5 terms. As a result, to overcome the limitations of the rule of 5, some extensions have been introduced through the Ghose filter and Veber's modification to improve and access the qualitative possibility of the molecule becoming a more efficient oral drug [66]. According to the Ghose rule, the partition coefficient (WLogP) of an orally active drug should be in ranges between −0.4 and 5.6, the molar refractivity (MR) should range from 40 to 130, molecular weight (MW) should be range from 180 to 480, and the number of atoms (nAT) present in an orally active drug range between 20 and 70 [36]. Veber's modification includes the number of rotatable bonds (#Rot. bonds) lower than 10, and also, the total polar surface area (TPSA) of the drug should not be greater than 140 Area square [37]. As it can be seen in Table 6, most of the suggested compounds are in agreement with the drug-likeness rules. The results (Table 6) show that most proposed

compounds passed the Ghose and Veber filters, and two of them have a Ghose violation (NC4 and NC10). A bioavailability score (BAS) is formulated as the probability of a compound having bioavailability higher than 10% in the rat [67]. All the compounds have a bioavailability score of 0.55 or 0.56, which means excellent pharmacokinetic properties [67]. Additionally, the selected compounds were evaluated for synthetic accessibility (SA), using a scale ranging from 1 (extremely simple to synthesize) to 10 (extremely difficult and complex to synthesize)) [68]. All suggested compounds have a SA close to 3 (Table 6), indicating that they are relatively simple to synthesize." The proposed compounds approved by drug-likeness rules were used for docking simulation. After generating the proposed active compounds with acceptable pharmacokinetic properties, the interactions of the suggested chemical compounds with the active site of the receptor were also investigated. As a result, the structures of the proposed compounds were optimized and then saved as a pdb file and then used as input to the Autodock4.2 software. Under ideal conditions, all recommended compounds were docked to the active site (LGA runs equal to 150). The docking software was used to extract the optimal conformation of the proposed compounds with the lowest binding energy. The discovery studio software was used to acquire the ligand-receptor interactions. The 2D interaction figures are given for the most active proposed compounds in Figs. 5–9. The results indicated that the proposed compounds have a suitable hydrophilic, hydrophobic, and van der Waals relationship with the receptor, indicating that they will be stable in their receptor-binding sites. The interaction graphs of the suggested potential compounds can be seen in the supplementary file (Figs. S1–S6). The chemical structures of new suggested compounds and their predicted biological activity were summarized in Table 5.

The starred highlighted items are related to the compound that has been rejected by at least one of the Ghose or Veber rules.

5. Conclusion

In three datasets, the created LAD-LASSO-ANN model was utilized to predict the biological activity of chemical substances. The first step was to use the DRAGON software to calculate molecular descriptors. After preprocessing the data, the most successful LAD-LASSO descriptors were defined as ANN input. The ANN parameters were optimized, and the best ANN models were chosen based on the minimum value of the MSE of the validation sets. The biological activities of the test set compounds were predicted based on the optimum ANN model. Statistical parameters such as R^2 and MSE values were computed for the test sets. The results were equal to 0.87 and 0.13 for dataset A, 0.84 and 0.07 for dataset B, and 0.87 and 0.11 for dataset C. The AD test was also performed for all three data sets, and the results of the Williams diagram confirm the presence of the majority of data within an acceptable range. This study examined the significance of numerous descriptors and proposed several new highly active chemicals. The proposed compound structures were drawn and optimized. Afterward, the descriptors of the LAD-LASSO model were computed. Predicting the biological activity of the proposed drugs was accomplished using the optimum ANN. Table 5 demonstrates the relevant consequences for the estimated biological response of the suggested potent chemical compounds. The ligand-receptor interaction was extracted via molecular docking research. The different hydrophilic and hydrophobic interactions of the proposed compounds with the active site amino acids indicate the stability of the proposed compounds in the respective proteins. It should be noted that all the proposed compounds are acceptable in terms of the drug-likeness rules (Table 6). In addition, the ease of synthesis indicates that the synthesis of compounds is possible on a laboratory scale.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. Authorship contribution statement

Zeinab Mozafari: Software, Methodology, Writing - original draft, Investigation, Editing. **Mansour Arab Chamjangali**: Supervision, Review, and editing. **Mohammad Arashi**: Data curation, Methodology, Validation, Editing. **Nasser Goudarzi**: Review and editing.

Declaration of Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to express their gratitude to the Shahrood University of Technology Research Council for supporting this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104510.

References

[1] S. Zhong, J. Hu, X. Yu, H. Zhang, Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: transfer learning, data augmentation and model interpretation, Chem. Eng. J. 408 (2021), 127998.
[2] P.G. Achary, Applications of quantitative structure-Activity relationships (QSAR) based virtual screening in drug design: a review, Mini Rev. Med. Chem. 20 (2020) 1375–1388.
[3] J. Tabeshpour, A. Sahebkar, M.R. Zirak, M. Zeinali, M. Hashemzaei, S. Rakhshani, S. Rakhshani, Computer-aided drug design and drug pharmacokinetic prediction: a mini-review, Curr. Pharmaceut. Des. 24 (2018) 3014–3019.
[4] B.J. Neves, R.C. Braga, C.C. Melo-Filho, J.T. Moreira-Filho, E.N. Muratov, C.H. Andrade, QSAR-based virtual screening: advances and applications in drug discovery, Front. Pharmacol. 9 (2018) 1275.
[5] A.M. Alharthi, M.H. Lee, Z.Y. Algamal, A.M. Al-Fakih, Quantitative structure-activity relationship model for classifying the diverse series of antifungal agents using ratio weighted penalized logistic regression, SAR QSAR Environ. Res. 31 (2020) 571–583.
[6] S. D'Souza, K. Prema, S. Balaji, Feature Selection and Modeling Using Statistical and Machine Learning Methods, 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), IEEE, 2020, pp. 18–22.
[7] J.-N. Gong, L. Zhao, G. Chen, X. Chen, Z.-D. Chen, C.Y.-C. Chen, A novel artificial intelligence protocol to investigate potential leads for diabetes mellitus, Mol. Divers. (2021) 1–19.
[8] H. Hadni, M. Elhallaoui, 2D and 3D-QSAR, molecular docking and ADMET properties in silico studies of azaaurones as antimalarial agents, New J. Chem. 44 (2020) 6553–6565.
[9] V. Kumar, P. De, P.K. Ojha, A. Saha, K. Roy, A multi-layered variable selection strategy for QSAR modeling of butyrylcholinesterase inhibitors, Curr. Top. Med. Chem. 20 (2020) 1601–1627.
[10] D.M. Rajathei, S. Parthasarathy, S. Selvaraj, Combined QSAR model and chemical similarity search for novel HMG-CoA reductase inhibitors for coronary heart disease, Curr. Comput. Aided Drug Des. 16 (2020) 473–485.
[11] A. Gandhi, V. Masand, M. Zaki, S. Al-Hussain, A.B. Ghorbal, A. Chapolikar, QSAR analysis of sodium glucose co–transporter 2 (SGLT2) inhibitors for anti-hyperglycaemic lead development, SAR QSAR Environ. Res. 32 (2021) 731–744.
[12] H. Labjar, M. Al-Sarem, M. Kissi, Feature selection using a genetic algorithms and fuzzy logic in anti-human immunodeficiency virus prediction for drug discovery, J. Inf. Technol. Manag. 14 (2022) 23–36.
[13] E. Shamsi, A. Rahati, E. Dehghanian, A modified binary particle swarm optimization with a machine learning algorithm and molecular docking for QSAR modelling of cholinesterase inhibitors, SAR QSAR Environ. Res. 32 (2021) 745–767.
[14] M. Mahmoodi-Reihani, F. Abbasitabar, V. Zare-Shahabadi, In silico rational design and virtual screening of bioactive peptides based on QSAR modeling, ACS Omega 5 (2020) 5951–5958.
[15] S. Yousefinejad, B. Hemmateenejad, Chemometrics tools in QSAR/QSPR studies: a historical perspective, Chemometr. Intell. Lab. Syst. 149 (2015) 177–204.
[16] N.A. Al-Thanoon, O.S. Qasim, Z.Y. Algamal, A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics, Chemometr. Intell. Lab. Syst. 184 (2019) 142–152.

[17] B.A. Baviskar, S.L. Deore, A.I. Jadhav, 2D and 3D QSAR studies of saponin analogues as antifungal agents against Candida albicans, J. Young Pharm. 12 (2020) 48.
[18] S. Chtita, A. Belhassan, M. Bakhouch, A.I. Taourati, A. Aouidate, S. Belaidi, M. Moutaabbid, S. Belaaouad, M. Bouachrine, T. Lakhlifi, QSAR study of unsymmetrical aromatic disulfides as potent avian SARS-CoV main protease inhibitors using quantum chemical descriptors and statistical methods, Chemometr. Intell. Lab. Syst. 210 (2021), 104266.
[19] L. Elmchichi, A. Belhassan, A. Aouidate, A. Ghaleb, T. Lakhlifi, M. Bouachrine, QSAR study of new compounds based on 1, 2, 4-triazole as potential anticancer agents, Phys. Chem. Res 8 (2020) 125–137.
[20] Y. Huang, T. Li, S. Zheng, L. Fan, L. Su, Y. Zhao, H.-B. Xie, C. Li, QSAR modeling for the ozonation of diverse organic compounds in water, Sci. Total Environ. 715 (2020), 136816.
[21] L.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–135.
[22] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series. B. Stat. Methodol. 58 (1996) 267–288.
[23] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statistical Assoc. 96 (2001) 1348–1360.
[24] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2005) 301–320.
[25] M. Eklund, U. Norinder, S. Boyer, L. Carlsson, Benchmarking variable selection in QSAR, Mol. Inform. 31 (2012) 173–179.
[26] G. Ghasemi, S. Arshadi, A.N. Rashtehroodi, M. Nirouei, S. Shariati, Z. Rastgoo, QSAR investigation on quinolizidinyl derivatives in alzheimer's disease, J. Comput. Med. (2013) 2013.
[27] Z. Li, M.J. Sillanpää, Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection, Theor. Appl. Genet. 125 (2012) 419–435.
[28] Z. Mozafari, M.A. Chamjangali, M. Arashi, Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors, Chemometr. Intell. Lab. Syst. 200 (2020) 103998–104011.
[29] S. Wacker, S.Y. Noskov, Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel, Comput. Toxicol 6 (2018) 55–63.
[30] H. Wang, G. Li, G. Jiang, Robust regression shrinkage and consistent variable selection through the LAD-Lasso, J. Bus. Econ. Stat. 25 (2007) 347–355.
[31] A.M.E. Saleh, M. Arashi, B.G. Kibria, Theory of Ridge Regression Estimation with Applications, John Wiley & Sons, 2019.
[32] Z. Mozafari, M. Arab Chamjangali, M. Arashi, N. Goudarzi, Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network-based QSAR studies, J. Chemom. 35 (2021) e3338.
[33] Z. Mozafari, M.A. Chamjangali, M. Arashi, Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors, Chemometr. Intell. Lab. Syst. 200 (2020), 103998.
[34] Z.T. Al-Dabbagh, Z.Y. Algamal, Least absolute deviation estimator-bridge variable selection and estimation for quantitative structure–activity relationship model, J. Chemom. 33 (2019) e3139.
[35] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.
[36] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases, J. Comb. Chem. 1 (1999) 55–68.
[37] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, J. Med. Chem. 45 (2002) 2615–2623.
[38] W. Chen, P. Zhan, D. Rai, E. De Clercq, C. Pannecouque, J. Balzarini, Z. Zhou, H. Liu, X. Liu, Discovery of 2-pyridone derivatives as potent HIV-1 NNRTIs using molecular hybridization based on crystallographic overlays, Biorg. Med. Chem. 22 (2014) 1863–1872.
[39] X. Chen, X. Liu, Q. Meng, D. Wang, H. Liu, E. De Clercq, C. Pannecouque, J. Balzarini, X. Liu, Novel piperidinylamino-diarylpyrimidine derivatives with dual structural conformations as potent HIV-1 non-nucleoside reverse transcriptase inhibitors, Bioorg. Med. Chem. Lett 23 (2013) 6593–6597.
[40] X. Chen, P. Zhan, X. Liu, Z. Cheng, C. Meng, S. Shao, C. Pannecouque, E. De Clercq, X. Liu, Design, synthesis, anti-HIV evaluation and molecular modeling of piperidine-linked amino-triazine derivatives as potent non-nucleoside reverse transcriptase inhibitors, Biorg. Med. Chem. 20 (2012) 3856–3864.
[41] D. Li, P. Zhan, H. Liu, C. Pannecouque, J. Balzarini, E. De Clercq, X. Liu, Synthesis and biological evaluation of pyridazine derivatives as novel HIV-1 NNRTIs, Biorg. Med. Chem. 21 (2013) 2128–2134.
[42] J. Wang, P. Zhan, Z. Li, H. Liu, E. De Clercq, C. Pannecouque, X. Liu, Discovery of nitropyridine derivatives as potent HIV-1 non-nucleoside reverse transcriptase inhibitors via a structure-based core refining approach, Eur. J. Med. Chem. 76 (2014) 531–538.
[43] L. Lu, S. Sha, K. Wang, Y.-H. Zhang, Y.-D. Liu, G.-D. Ju, B. Wang, H.-L. Zhu, Discovery of chromeno [4, 3-c] pyrazol-4 (2H)-one containing carbonyl or oxime derivatives as potential, selective inhibitors PI3Kα, Chem. Pharm. Bull. 64 (2016) 1576–1581, c16-00388.
[44] Y. Yin, J.-Q. Hu, X. Wu, S. Sha, S.-F. Wang, F. Qiao, Z.-C. Song, H.-L. Zhu, Design, synthesis and biological evaluation of novel chromeno [4, 3-c] pyrazol-4 (2H)-one derivates containing sulfonamido as potential PI3Kα inhibitors, Biorg. Med. Chem. 27 (2019) 2261–2267.
[45] Y. Yin, X. Wu, H.-W. Han, S. Sha, S.-F. Wang, F. Qiao, A.-M. Lu, P.-C. Lv, H.-L. Zhu, Discovery and synthesis of a novel series of potent, selective inhibitors of the PI3Kα: 2-alkyl-chromeno [4, 3-c] pyrazol-4 (2 H)-one derivatives, Org. Biomol. Chem. 12 (2014) 9157–9165.
[46] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon Software: an Easy Approach to Molecular Descriptor Calculations, Match, 2006, pp. 237–248.
[47] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Software 28 (2008) 1–26.
[48] G.M. Morris, M. Lim-Wilby, Molecular Docking, Molecular Modeling of Proteins, Springer, 2008, pp. 365–382.
[49] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[50] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J. Comput. Chem. 19 (1998) 1639–1662.
[51] BIOVIA Dassault Systèmes, Discovery Studio Visualizer Software, AccelrysSan Diego, 2021.
[52] Matlab, Matlab the Mathworks, Inc., Natick, 2017.
[53] Y. Qin, S. Li, Y. Li, Y. Yu, Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection, arXiv preprint arXiv:1708.05439, 2017.
[54] L. Douali, D. Villemin, D. Cherqaoui, Neural networks: accurate nonlinear QSAR model for HEPT derivatives, J. Chem. Inf. Comput. Sci. 43 (2003) 1200–1207.
[55] P. Coulibaly, F. Anctil, B. Bobée, Prévision hydrologique par réseaux de neurones artificiels: état de l'art, Can. J. Civ 26 (1999) 293–304.
[56] F. Othman, M. Naseri, Reservoir inflow forecasting using artificial neural network, Int. J. Phys. Sci. 6 (2011) 434–440.
[57] F. Burden, D. Winkler, Bayesian Regularization of Neural Networks, Artificial neural networks, 2008, pp. 23–42.
[58] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q 2 parameter for QSAR validation, J. Chem. Inf. Model. 49 (2009) 1669–1678.
[59] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemom. 24 (2010) 194–201.
[60] P. Gramatica, Principles of QSAR modeling: comments and suggestions from personal experience, IJQSPR 5 (2020) 1–37.
[61] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (2012) 4791–4810.
[62] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[63] R. Darnag, B. Minaoui, M. Fakir, QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression, Arab. J. Chem. 10 (2017) S600–S608.
[64] D.S. Biovia, Discovery Studio Visualizer, 2017, p. 936. San Diego, CA, USA.
[65] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (2017) 1–13.
[66] O. Steve, N. Swainston, J. Handl, D.B. Kell, A 'rule of 0.5'for the metabolite-likeness of approved pharmaceutical drugs, Metabolomics 11 (2015) 323–339.
[67] Y.C. Martin, A bioavailability score, J. Med. Chem. 48 (2005) 3164–3170.
[68] P. Ertl, T. Schuhmann, A systematic cheminformatics analysis of functional groups occurring in natural products, J. Nat. Prod. 82 (2019) 1258–1263.