



High predictive QSAR models for predicting the SARS coronavirus main protease inhibition activity of ketone-based covalent inhibitors

Bakhtyar Sepehri¹ · Mohammad Kohnehpoushi¹ · Raouf Ghavami¹

Received: 2 July 2021 / Accepted: 7 October 2021 / Published online: 26 October 2021
© Iranian Chemical Society 2021

Abstract

In this research, a dataset including 29 ketone-based covalent inhibitors with SARS-CoV-1 3CL^{pro} inhibition activity was used to develop high predictive QSAR models. Twenty-two molecules were put in train set and seven molecules in test set. By using stepwise MLR method for molecules in train set, four molecular descriptors including Mor26p, Hy, GATS7p and Mor04v were selected to build QSAR models. MLR and ANN methods were used to create QSAR models for predicting the activity of molecules in both train and test sets. Both QSAR models were validated by calculating several statistical parameters. R^2 values for the test set of MLR and ANN models were 0.93 and 0.95, respectively, and RMSE values for their test sets were 0.24 and 0.17, respectively. Other calculated statistical parameters (especially Q_{F3}^2 parameter) show that created ANN model has more predictive power with respect to developed MLR model (with four descriptor). Calculated leverages for all molecules show that predicted pIC₅₀ (by both QSAR models) for all molecules is acceptable, and drawn residuals plots show that there is no systematic error in building both QSAR models. Also, based on developed MLR model, used molecular descriptors were interpreted.

Keywords QSAR · SARS-CoV-1 · SARS-CoV-2 · 3CL^{pro} inhibition activity · COVID-19

Introduction

Coronavirus disease 19 (COVID-19) is a pandemic disease that has affected the health of peoples in the whole world. Until May 6, 2021, the World Health Organization (WHO) had reported 155,506,494 infected cases to COVID-19 (including 3,247,228 deaths) [1, 2]. The disease has spread from Wuhan in China (in late 2019) by a virus that has called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Since some coronaviruses had been transmitted from animals to humans, probably, the similar event has happened for SARS-CoV-2 [3–6]. Before COVID-19 pandemic, two coronaviruses including severe acute respiratory syndrome coronavirus (SARS-CoV-1) and Middle East respiratory syndrome coronavirus (MERS-CoV) had been transmitted to human from animals [7, 8]. Although SARS-CoV-2 has lower mortality rate (2.3%) with respect to SARS-CoV-1

(mortality rate 10%) and MERS-CoV (mortality rate 35%), it has higher reproductive number (2.0–2.5) with respect to SARS-CoV-1 (1.7–1.9) and MERS-CoV (< 1) [9–11]. Despite the lower mortality rate of SARS-CoV-2, it has killed more people with respect to SARS-CoV-1 and MERS-CoV because of its global pandemic outbreak. SARS-CoV-2 virus is present in body fluids such as cerebrospinal fluid and blood and usually is transmitted through respiratory droplets [12, 13]. So, from the beginning of COVID-19 outbreak, social distancing and closing mask have been suggested to reduce the number of infected cases [14]. Infected people show a variety of symptoms such as fever, difficulty breathing, taste or smell loss, headache, muscle ache, sore throat, runny nose and nausea [15]. Most of the patients show mild symptoms (~80%), and just the smaller proportion of them (~5%) have severe disease [16]. There are four subfamilies of coronaviruses including α -coronaviruses, β -coronaviruses, δ -coronaviruses and γ -coronaviruses, in which α - and β -coronaviruses infect mammals. SARS-CoV-1, MERS-CoV and SARS-CoV-2 are belonging to β -coronaviruses subfamily [17–19]. SARS-CoV-2 is a positive-sense, single-stranded RNA virus (+ ssRNA) that has been packed in an envelope. Spike membrane glycoproteins in the surface

✉ Bakhtyar Sepehri
sepehribakhtyar@yahoo.com

¹ Chemometrics Laboratory, Department of Chemistry, Faculty of Science, University of Kurdistan, P. O. Box 416, Sanandaj, Iran

of virus bind to angiotensin-converting enzyme 2 (ACE2) receptor in the membrane of human cells and enters virus to our cells [20–23]. Generally, designed drugs for COVID-19 treatment can be classified into four groups including drugs that prevent the replication and synthesis of RNA by targeting critical enzymes for the replication of the virus, drugs that block the binding of spike protein to ACE2 receptor on human cells, drugs that inhibit coronavirus virulence factors and drugs that inhibit a receptor or enzymes in human cells [24]. 3C-like cysteine protease (3CL^{pro}) is the main protease of SARS-CoV-2 that catalyzes the cleavage of polypeptides to their effector forms and has essential enzymatic role for virus life cycle [25, 26]. So it can be considered as a target for design drugs in COVID-19 treatment [27–29]. Quantitative structure–activity relationship (QSAR) is a computer-assisted drug design method that relates the structural features of molecules to their activities. QSAR models are useful in drug design process because they predict the activity of molecules quantitatively and determine structural features that increase the activity of molecules [30]. In this research, we have used a series of new synthesized compounds including 29 ketone-based molecules as covalent inhibitors of SARS-CoV-1 3CL^{pro} (that had been synthesized by Hoffman et al.) [31] to develop QSAR models with high predictive power for predicting their 3CL^{pro} inhibition activities. Hoffman et al. had shown that the greatest active compound in their research (compound 4 in their published paper and compound m15 in this research) is the covalent inhibitor of 3CL^{pro} SARS-CoV-1 (IC₅₀: 0.004 μM) and 3CL^{pro} SARS-CoV-2 (IC₅₀: 0.00027 μM) enzymes. The crystallographic structure of the complex of this compound with 3CL^{pro} SARS-CoV-2 is available in protein data bank (PDB ID: 6XHM). Also, performed researches by other groups show that the derivatives of available molecules in this dataset are covalent inhibitors for the 3CL^{pro} enzymes of MERS-CoV and SARS-CoV-2 [32–37]. Since SARS-CoV-1 and SARS-CoV-2 have high similarity in their genome [38] and the derivatives of molecules in this dataset are active against the 3CL^{pro} enzymes of SARS-CoV-1 and SARS-CoV-2, designed and optimized inhibitors by using developed QSAR models in this research help to design new drugs for treating COVID-19.

Materials and methods

Materials

A series of molecules including 29 ketone-based covalent inhibitors of 3CL^{pro} SARS-CoV-1 were selected from published paper by Hoffman et al. [31]. The chemical structure and activity of molecules are listed in Table 1. The activity of molecules was IC₅₀ in nano-molar unit. In the first step,

IC₅₀ values in nano-molar unit were converted to IC₅₀ values at molar unit and then they are converted to pIC₅₀ by using the following equation:

$$pIC_{50} = -\log(IC_{50}) \quad (1)$$

pIC₅₀ values had a wide range from 5.97 to 8.40. This dataset has suitable features that make it unique for developing QSAR models including the following:

- Dataset has the wide range of activities (more than 2 log unit);
- 3CL^{pro} SARS-CoV-1 inhibition activity in nano-molar level;
- Molecule m15 in the dataset shows potent inhibition activity against 3CL^{pro} SARS-CoV-1 (IC₅₀: 0.004 μM) and 3CL^{pro} SARS-CoV-2 (IC₅₀: 0.00027 μM);
- Molecule m15 is a covalent inhibitor of 3CL^{pro} SARS-CoV-2 (PDB ID: 6XHM);
- Molecule m15 in the dataset shows good selectivity against other proteases [31];
- Several researches have indicated that the derivatives of molecules in this dataset are covalent inhibitors of 3CL^{pro} enzymes in SARS-CoV-1, SARS-CoV-2 and MERS-CoV [32–37], so the developed model can help to design new drugs for treating COVID-19.

To develop QSAR models, the dataset was divided into a train set containing 22 molecules for developing QSAR models and a test set including 7 molecules (molecules m3, m8, m13, m14, m17, m21 and m23) for validating them. Molecules with low, moderate and high activities were put in both train and test sets manually, and molecules with the lowest and greatest activities were put into the train set.

Programs

The three-dimensional chemical structure of all molecules was built in HyperChem (version 7.1) software and optimized by using AMBER force field (the root-mean-square gradient was set to 0.0001 kcal mol⁻¹ Å⁻¹) [39]. Dragon software (version 5.5) was used to calculate molecular descriptors for the optimized structures of molecules [40]. SPSS software (version 16) was used to select informative descriptors by using stepwise multiple linear regression (stepwise MLR) [41]. All other chemometrics methods for building and validating models were performed in R software (version 3.6.3) [42]. RStudio software (Version 1.1.463) was used as integrated development environment (IDE) for R programming language [43]. MLRQSAR package (version 0.1.0) was used to develop multiple linear regression (MLR) model and validate it by performing leave-one-out cross-validation and Y-randomization test on MLR model. Also, it was used to compute descriptor

Table 1 (continued)

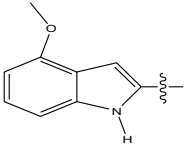

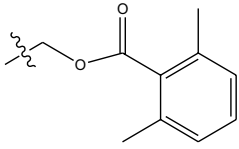
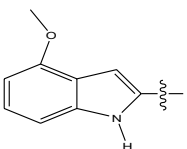

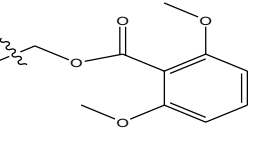
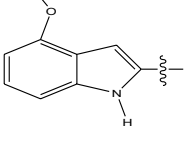

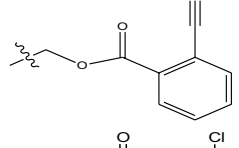

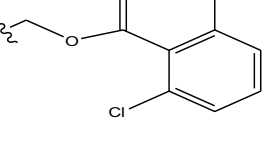
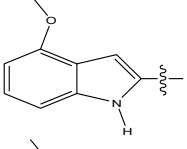

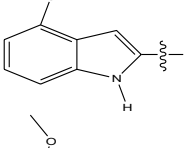

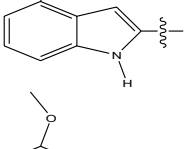

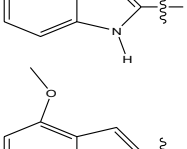

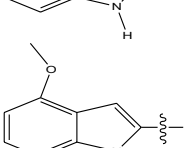

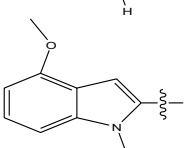



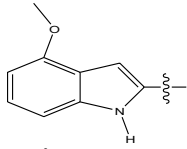
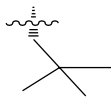
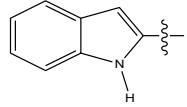
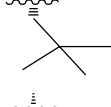
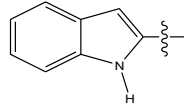
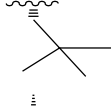
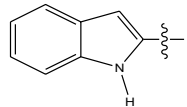
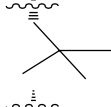
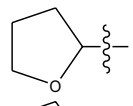
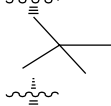
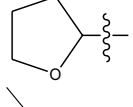
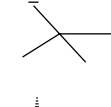
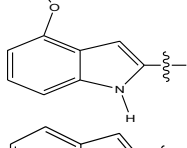
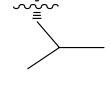
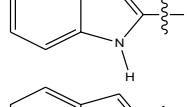
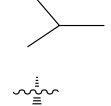
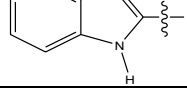
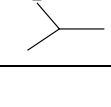
m10		H			74	7.13
m11		H			205	6.69
m12		H			17	7.77
m13	CH ₃	H			1028	5.99
m14		H		-CH ₂ OH	7	8.15
m15		H		-CH ₂ OH	4	8.40
m16		CH ₃		-CH ₂ OH	83	7.08
m17		H		-CH ₂ OH	20	7.70
m18		H		-CH ₂ OH	34	7.47
m19		H		-CH ₂ OH	44	7.36
m20		H		-CH ₂ OH	103	6.99

Table 1 (continued)

m21		H		- CH ₂ OMe	35	7.46
m22		H		- CH ₂ OH	20	7.70
m23		H		- CH ₂ OMe	105	6.98
m24		H		- CH ₂ OEt	112	6.95
m25		H		- CH ₂ OH	91	7.04
m26		H		- CH ₂ OMe	1076	5.97
m27		H		- CH ₂ OMe	53	7.28
m28		H		- CH ₂ OH	38	7.42
m29		H		- CH ₂ OMe	131	6.88

contribution for MLR model, calculate variance inflation factor (VIF) for descriptors, calculate several statistical parameters for validating both train and test sets of developed QSAR models and compute the applicability domain of created QSAR models based on the calculation of the leverage matrix [44, 45]. For building artificial neural network (ANN) model, h2o package (version 3.32.1.2) was used [46]. Also, ggplot2 package (version 3.3.3) was used to draw plots [47].

Methods

MLR modelling and validation

A MLR model has the following form:

$$pIC_{50} = \beta_0 + \beta_1 MD_1 + \beta_2 MD_2 + \dots + \beta_n MD_n \quad (2)$$

where β_0 is constant coefficient and β_1 to β_n are corresponding coefficients to the molecular descriptors MD_1 to MD_n .

Coefficients are obtained so that the sum of squared residuals (between predicted pIC_{50} and experimental pIC_{50}) is minimum. Also, leave-one-out cross validation (LOOCV) and Y-randomization tests were performed on this model to indicate that the created model is robust and has not been obtained by chance [48, 49].

ANN modelling

To create an ANN model in h2o package, h2o.deeplearning option was used. Although this package is able to build both shallow feedforward ANN model (ANN model with one hidden layer) and deep feedforward ANN model (ANN model with more than one hidden layer), we built a shallow feedforward ANN model due to the small size of dataset. In deep ANN model, the number of trainable parameters increases and the small size of dataset leads to overfitting. To solve overfitting in created model, dropout technique

was applied to network during its training and regularization terms were used in its cost function. Dropout removes some neurons from input and hidden layers during the training process, randomly. L1 (lasso) regularization, L2 (ridge) regularization and max_w2 (an upper limit for the (squared) sum of the incoming weights to a neuron) were added to loss function as regularization terms. The loss function in h2o.deeplearning has the following form that it is minimized for each training example j :

$$\text{Lossfunction} = L(W.B|j) + \lambda_1 R_1(W.B|j) + \lambda_2 R_2(W.B|j) \quad (3)$$

In Eq. 3, W is the collection $\{W_i\}_{1:N-1}$, where W_i denotes the weight matrix connecting layers i and $i + 1$ for a network of N layers and B is the collection $\{b_i\}_{1:N-1}$, where b_i denotes the column vector of biases for layer $i + 1$. In loss function, $L(W.B|j)$ was set to absolute that is the sum of residuals. $R_1(W.B|j)$ is the sum of all L1 norms for the weights and biases in the network, and L2 regularization is presented via $R_2(W.B|j)$ that is the sum of squares of all the weights and biases in the network. λ_1 and λ_2 are constant variables that generally they are set to a very small value (for example 10^{-5}). Also, maxout activation function was used for neurons in the hidden layer [50–53].

Applicability domain

The applicability domain of built QSAR models was investigated by calculating leverage matrix (H):

$$H = X(X^T X)^{-1} X^T \quad (4)$$

where X is descriptors matrix and the diagonal elements of H matrix are the leverages for objects (molecules). Critical leverage value was considered $3p/n$, where p is the number of descriptors in model plus one and n is the number of molecules in the train set. If calculated leverage (h) for a molecule is larger than critical leverage value, its predicted activity (by created model) is not acceptable [54, 55].

Statistical parameters for validating QSAR models

For validating created QSAR models, several statistical parameters have been calculated for both train and test sets including:

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \times \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \quad (5)$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^n (y_i - k \times \hat{y}_i)^2}{\sum_{i=1}^n \sum (y_i - \bar{y})^2} \quad (6)$$

$$r_0'^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - k' \times y_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \quad (7)$$

$$k = \frac{\sum_{i=1}^n (y_i \times \hat{y}_i)}{\sum_{i=1}^n (\hat{y}_i)^2} \quad (8)$$

$$k' = \frac{\sum_{i=1}^n (y_i \times \hat{y}_i)}{\sum_{i=1}^n (y_i)^2} \quad (9)$$

$$\bar{r}_m^2 = (r_m^2 + r_m'^2)/2 \quad (10)$$

$$\Delta r_m^2 = |r_m^2 - r_m'^2| \quad (11)$$

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (12)$$

$$r_m'^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0'^2}\right) \quad (13)$$

$$CCC^2 = \frac{2\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)}{\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2} \quad (14)$$

$$MAE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n} \quad (15)$$

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{Test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{Test}} (y_i - \bar{y}_{TR})^2} \quad (16)$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{Test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{Test}} (y_i - \bar{y}_{Test})^2} \quad (17)$$

$$Q_{F3}^2 = 1 - \frac{\frac{\sum_{i=1}^{n_{Test}} (y_i - \hat{y}_i)^2}{n_{Test}}}{\frac{\sum_{i=1}^{n_{Test}} (y_i - \bar{y}_{TR})^2}{n_{TR}}} \quad (18)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (19)$$

where y_i and \hat{y}_i are, respectively, the experimental and the predicted activity of molecule and \bar{y} and $\bar{\hat{y}}$ are the mean of the experimental and the predicted activities, respectively. \bar{y}_{TR} and \bar{y}_{Test} are the mean of the activity for train and test sets, respectively. Also, n , n_{TR} and n_{Test} are the number of compounds, the number of compounds in train set and the number of compounds in test set, respectively. CCC^2 is the squared concordance correlation coefficient, RMSE is the root-mean-squared error, and MAE is the mean absolute error [44, 56, 57].

Results and discussion

Model building and validation

Molecular descriptors that belong to all 22 descriptors blocks in Dragon software were calculated for all molecules. In the first step, molecular descriptors with few repeated values (fewer than 5) across samples and many zero values (with more than 10 zero values) across samples were removed. After this preprocessing step, 1203 molecular descriptor were remained. Stepwise MLR in SPSS software was used to select informative variables based on molecules in the train set. Four molecular descriptors were selected to develop QSAR models whose name and definition are listed in Table 2, and their values for all molecules are listed in Table 3. VIF values for Mor26p, Hy, GATS7p and Mor04v molecular descriptors were 1.06, 1.28, 1.12 and 1.21 which indicate that these descriptors have no collinearity and multi-collinearity problems and are suitable for creating QSAR models. Mor26p has the largest correlation with the activities of molecules ($R^2=0.59$), but a predictive model cannot be created just by using this descriptor, so Hy descriptor was added by stepwise MLR and the following model was created:

$$pIC_{50} = 6.769(\pm 0.253) - 3.236(\pm 0.588) \text{ Mor26p} + 0.563(\pm 0.119) \text{ Hy} \quad (20)$$

R^2 and RMSE values for the train set of this model were 0.77 and 0.22, respectively, and those for the test set were 0.79 and 0.32, respectively. R^2 value for LOO-CV on the train set was 0.72 which indicates that the created model is robust, and the maximum value of R^2 for ten runs of Y-randomization test was 0.17 which shows that the created model has not been obtained by chance. By adding another descriptor (GATS7p), a model with three descriptors was built:

$$pIC_{50} = 3.837(\pm 0.633) - 3.542(\pm 0.372) \text{ Mor26p} + 0.751(\pm 0.082) \text{ Hy} + 2.936(\pm 0.602) \text{ GATS7p} \quad (21)$$

R^2 and RMSE values for the train set of this model were 0.85 and 0.17, respectively, and those for the test set were 0.82 and 0.28, respectively. R^2 value for LOO-CV on the train set was 0.84 which indicates that the created model is robust, and the maximum value of R^2 for ten runs of Y-randomization test was 0.29 which shows that the created model has not been obtained by chance. As seen, adding GATS7p has increased the predictive power of QSAR model. For increasing the predictive power of model, another descriptor (Mor04v) was added to the model, and according to Topliss and Costello rule (the ratio of molecules in train set to used descriptors for building model should be at least 5 to 1) [58], this is the last descriptor that we can use for developing QSAR models. By using all four descriptors, the following equation was obtained in R software:

$$pIC_{50} = 3.837(\pm 0.633) - 3.542(\pm 0.372) \text{ Mor26p} + 0.751(\pm 0.082) \text{ Hy} + 2.936(\pm 0.602) \text{ GATS7p} + 0.245(\pm 0.065) \text{ Mor04v} \quad (22)$$

R^2 values for the train and test sets of this model were 0.92 and 0.93, respectively, and RMSE values for the train and test sets were 0.13 and 0.24, respectively. R^2 value for LOO-CV was 0.90 which shows that the created model is robust, and the maximum R^2 value for ten runs of Y-randomization test was 0.37 which indicates that the created model has not been obtained by chance. R^2 and RMSE values for the test set of created MLR models show that the created MLR model with all four descriptors has the highest predictive power. For further validation of the MLR model (MLR model with four descriptors), several statistical parameters

Table 2 The definition of selected descriptors by stepwise MLR

Descriptor	Type	Descriptor block	Definition
Mor26p	3D	3D-MoRSE descriptors	3D-MoRSE—signal 26/weighted by atomic polarizability
Hy	Others	Molecular properties	Hydrophilic factor
GATS7p	2D	2D autocorrelations	Geary autocorrelation-lag 7/weighted by atomic polarizability
Mor04v	3D	3D-MoRSE descriptors	3D-MoRSE—signal 04/weighted by atomic van der Waals volume

Table 3 Experimental and predicted pIC_{50} , descriptors values and leverage values for molecules (critical leverage value is 0.68)

Train set								
Molecule	Experimental pIC_{50}	Predicted pIC_{50} by		Descriptor values				Leverage
		MLR model	ANN model	Mor26p	Hy	GATS7p	Mor04v	
m1	6.66	6.80	6.66	0.231	1.525	0.96	- 0.759	0.0419
m2	6.74	6.94	6.76	0.198	1.478	0.973	- 0.671	0.0460
m4	7.07	7.13	7.09	0.155	1.415	1.038	- 1.095	0.0770
m5	7.10	7.14	7.11	0.199	1.399	1.078	- 0.833	0.0766
m6	7.06	6.94	6.97	0.224	1.395	1.04	- 0.841	0.0592
m7	7.28	7.36	7.27	0.061	1.399	1.012	- 1.166	0.1362
m9	7.01	7.22	7.05	0.11	1.418	0.992	- 0.826	0.0807
m10	7.13	6.92	7.11	0.25	1.377	1.102	- 1.235	0.1022
m11	6.69	6.52	6.63	0.229	1.385	0.957	- 1.457	0.1103
m12	7.77	7.76	7.65	0.002	1.399	1.069	- 1.049	0.2476
m15	8.40	8.12	8.22	0.022	2.336	0.968	- 0.962	0.1625
m16	7.08	7.11	6.94	0.193	1.548	0.987	- 0.431	0.0783
m18	7.47	7.55	7.37	0.224	2.304	1.036	- 1.108	0.1097
m19	7.36	7.42	7.32	0.281	2.243	1.049	- 0.754	0.1550
m20	6.99	7.03	6.90	0.246	2.243	1.049	- 2.854	0.5575
m22	7.70	7.63	7.59	0.129	2.341	0.9	- 0.606	0.2084
m24	6.95	6.90	6.87	0.131	1.52	0.896	- 1.006	0.0330
m25	7.04	6.88	6.91	0.298	1.697	1.001	- 0.493	0.1029
m26	5.97	6.00	5.97	0.379	0.933	0.995	- 0.469	0.1676
m27	7.28	7.36	7.16	0.18	2.336	0.968	- 1.777	0.1678
m28	7.42	7.55	7.37	0.118	2.376	0.926	- 1.544	0.1414
m29	6.88	6.74	6.80	0.185	1.573	0.933	- 1.464	0.0791
Test set								
Molecule	Experimental pIC_{50}	Predicted pIC_{50} by		Descriptor values				Leverage
		MLR model	ANN model	Mor26p	Hy	GATS7p	Mor04v	
m3	6.64	6.95	6.79	0.193	1.456	0.981	- 0.745	0.0444
m8	7.09	7.06	6.98	0.123	1.418	0.972	- 1.054	0.0644
m13	5.99	5.54	5.85	0.601	0.899	1.06	0.184	0.5221
m14	8.15	8.19	8.40	- 0.012	2.304	0.939	- 0.741	0.2222
m17	7.70	7.77	7.50	0.146	2.336	1.02	- 1.244	0.0943
m21	7.46	7.25	7.29	0.066	1.522	0.943	- 1.074	0.0762
m23	6.98	7.00	6.97	0.123	1.546	0.91	- 0.943	0.0357

were calculated for the train and test sets that are listed in Tables 4 and 5. Calculated values for these statistical parameters show that the created model is acceptable and has high predictive power. Predicted pIC_{50} for all molecules (in both train and test sets) by this model (MLR model with four descriptors) is listed in Table 3. Calculated leverages for all molecules (that are listed in Table 3) are smaller than critical leverages which show that the predicted pIC_{50} for all molecules (by MLR model with four descriptors) is acceptable. The plot of predicted pIC_{50} versus experimental pIC_{50} ,

William plot and residuals plot for the MLR model (MLR model with four descriptors) are shown in Fig. 1. The William plot in Fig. 1 shows that the created model has no outlier and the predicted pIC_{50} for all molecules (in both train and test sets) is acceptable, and the residual plot shows that there is no systematic error in creating MLR model with four descriptors. To develop more predictive power QSAR model, these four descriptors were used as input variables for training an ANN model. In the first step, a network with one hidden layer and 10 neurons was created. For optimizing the

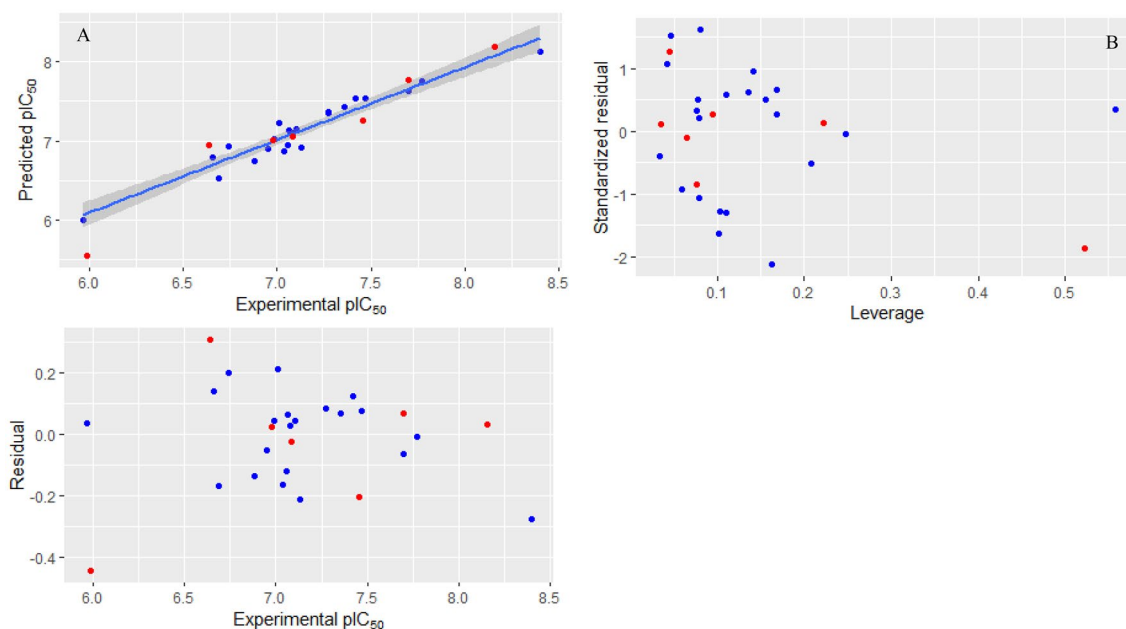
Table 4 Calculated statistical parameters for validating created QSAR models

Statistical parameters	Threshold values	MLR		ANN	
		Train set	Test set	Train set	Test set
CCC^2	> 0.6	0.92	0.91	0.96	0.95
R^2	> 0.6	0.92	0.93	0.99	0.95
$RMSE$	–	0.13	0.24	0.06	0.17
k	≤ 1.15 and ≥ 0.85	1.00	1.00	1.01	1.00
k'	≤ 1.15 and ≥ 0.85	1.00	1.00	0.99	1.00
r_0^2	> 0.6	0.92	0.89	0.98	0.94
$r_0'^2$	> 0.6	0.91	0.92	0.98	0.95
r_m^2	> 0.5	0.92	0.74	0.93	0.85
$r_m'^2$	> 0.5	0.85	0.83	0.92	0.90
$\overline{r_m^2}$	> 0.5	0.88	0.79	0.92	0.87
Δr_m^2	< 0.2	0.08	0.09	0.01	0.05
$(r^2 - r_0^2)/r^2$	< 0.1	0.00	0.04	0.00	0.01
$(r^2 - r_0'^2)/r^2$	< 0.1	0.01	0.01	0.00	0.00
$ r^2 - r_0'^2 $	< 0.3	0.01	0.03	0.00	0.01
MAE	–	0.00	0.03	0.06	0.03

Table 5 Calculated Q^2 -based statistical parameters for validating created QSAR models

Parameter	MLR	ANN
Q_{F1}^2	0.89	0.94
Q_{F2}^2	0.89	0.94
Q_{F3}^2	0.77	0.88

trainable parameters of ANN model, k-fold cross-validation test was used. In this method, molecules in train set were divided into three sets, and each time, both of them were used for training ANN model and other for its validation and this process was repeated for each fold. The R^2 value for each

**Fig. 1** Plots for created MLR model (train set with blue color and test set with red color): (A) the plot of predicted pIC_{50} versus experimental pIC_{50} ; (B) William plot (critical leverage is 0.68); (C) residuals plot

fold and their mean were calculated. The activation function for neuron in the hidden layer was set to maxout activation function. By increasing the number of neurons in the hidden layer to 100 (each time, 10 neurons were added to the hidden layer of previous network architecture), the average of R^2 values for all three folds was increased. Increasing the number of neurons in the hidden layer to more than 100 neurons did not increase the average of R^2 values for k-fold cross-validation test, significantly, so an ANN architecture with one hundred neurons in its hidden layer was selected as the best architecture. Also, L1 and L2 regularization terms were set to 0.00001 and max_w2 was set to its default value. Dropout ratio from 0 to 0.5 was examined for both input and hidden layers, and the best results were obtained when dropout ratio for the input layer and hidden layer was set to 0.1 and 0.3, respectively. Other parameters were set to their default. So created ANN model had four neurons in its input layer and one hundred neurons in its hidden layer (with max-out activation function) and one neuron in its output layer (with linear activation function). The predicted pIC_{50} for all molecules (in both train and test sets) is listed in Table 3, and the calculated statistical parameters for the train and test sets are listed in Tables 4 and 5. R^2 and RMSE values for the train set of ANN model were 0.99 and 0.06, respectively, and R^2 and RMSE values for the test set were 0.95 and 0.17, respectively. R^2 values for folds 1, 2 and 3 were 0.89, 0.69 and 0.68, respectively, and their mean was 0.75 which indicates that the created ANN model is robust. The plot of predicted pIC_{50} versus experimental pIC_{50} , William

plot and residuals plot for ANN model are shown in Fig. 2. Drawn residuals plot shows that there is no bias (systematic error) in creating this ANN model. William plot shows that molecule m15 is outlier, and based on this plot, predicted pIC_{50} by the ANN model for all molecules (in both train and test sets) is acceptable.

Descriptors interpretation

The contribution of Mor26p, Hy, GATS7p and Mor04v molecular descriptors in the building of MLR model with four descriptors was 11.70%, 23.10%, 52.60% and 3.72%, respectively, and this MLR model (with four descriptors) was used for descriptors interpretation. Negative coefficient sign for Mor26p shows that smaller values (negative values) for this descriptor are favorable for increasing the activities of molecules. For example, molecules m14 and m15 which have smaller values for this descriptor have the most potent activities among others. Among all molecules, the value of this descriptor is negative only for molecule m14. Mor26p is a descriptor that belongs to 3D molecular representations of structure based on electron diffraction (3D-MorSE) descriptors family that has been weighted by atomic polarizability. A study by Devinyak et al. [59] shows that the weighting of these descriptors by atomic polarizability decreases the effect of hydrogen significantly and diminishes the roles of nitrogen, oxygen and fluorine atoms. Also, they found that although these descriptors have information about the whole molecule, their final values are derived mostly from

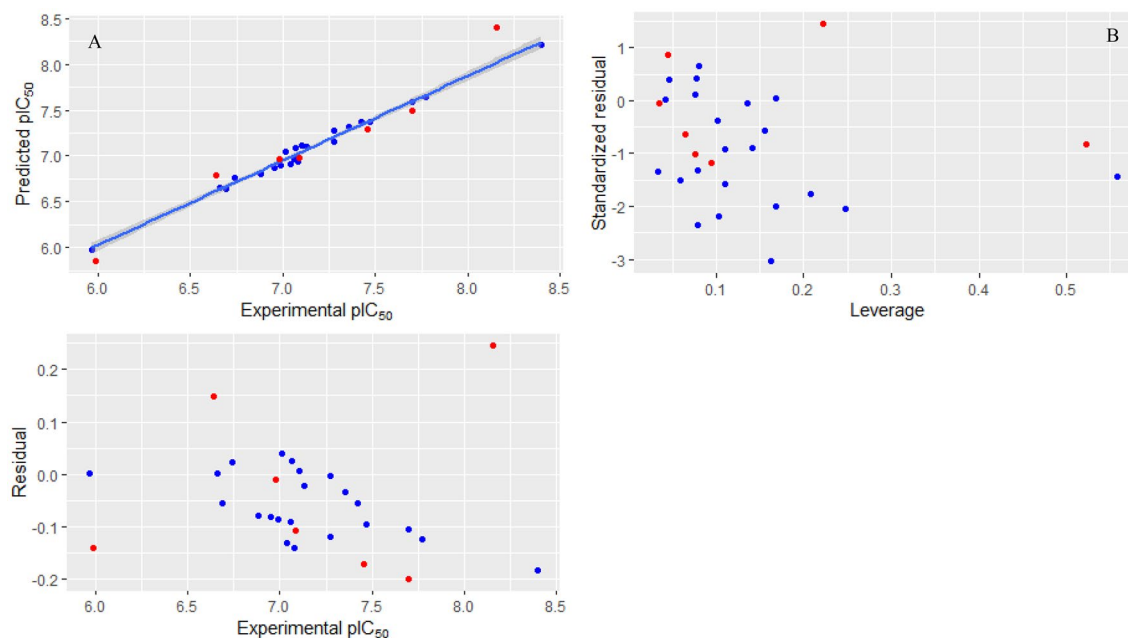


Fig. 2 Plots for created ANN model (train set with blue color and test set with red color): (A) the plot of predicted pIC_{50} versus experimental pIC_{50} ; (B) William plot (critical leverage is 0.68); (C) residuals plot

short-distance atomic pairs [59]. The presence of methoxy group on phenyl ring in the R_1 substituent of molecule m21 has decreased its Mor26p value and increased its activity with respect to molecule m23. The comparison of molecules m21, m23 and m26 shows that R_1 substituent with two fused rings is favorable for increasing the activity of molecule with respect to R_1 substituent with a ring because two fused rings decrease the Mor26p descriptor value for molecule. Replacing hydrogen atom in R_2 substituent with methyl group increases the Mor26p descriptor value and decreases the activity of molecule, so bulky groups in R_2 substituent are not favorable. Comparing molecules m15 and m17 with m20 shows that longer and bulky groups for R_3 substituent increase the value of Mor26p descriptor, so cyclic and long-chain groups for R_3 substituent are not favorable for increasing the activity of molecules. The presence of nitrile group on phenyl ring in R_4 substituent in molecules m7 and m12 has decreased the value of Mor26p descriptor for these two molecules, but comparing all molecules does not reveal a specific relationship between the size of R_4 substituent and Mor26p descriptor values for molecules. Hy is the hydrophilic factor for molecule, and MLR model shows that larger Hy descriptor values improve the activity of molecules. The R^2 value between Hy descriptor values and the activities of molecules is 0.52. Available data in Table 2 show that molecules with greater activities such as molecules m14, m15, m17 and m22 have larger Hy value. Hydrophilic groups such as hydroxyl group are favorable for increasing Hy descriptor value. Also, the presence of atoms with negative partial charge in R_1 substituent and less bulky groups in R_3 substituent increases the value of Hy descriptor. The developed MLR model shows that the larger value of GATS7p descriptor is favorable for increasing the activity of molecule. Mor04v descriptor belonging to 3D-MoRSE descriptors has been weighted by atomic van der Waals volume. Weighting descriptor by atomic van der Waals volume has similar effect with the weighting of 3D-MoRSE descriptor by atomic polarizability that decreases the effect of hydrogen significantly and diminishes the roles of nitrogen, oxygen and fluorine atoms [59]. In MLR model, Mor04v descriptor has a coefficient with positive sign, so larger values of this descriptor are favorable for increasing the activity of molecules. Except for molecule m13, the values of this descriptor are negative for other molecules (Table 3). Comparing molecules m1 to m14 shows that the larger value of Mor04v descriptor for molecule m13 is related to less bulky group for R_1 substituent in molecule m13. This situation is seen for molecules m25 and m26. Less bulky groups for R_1 substituent increase the value of Mor26p descriptor and decrease the activity of molecules. Since Mor04v descriptor has less contribution in creating model with respect to Mor26p descriptor, less bulky groups for R_1 substituent are not favorable for increasing the activity of molecules. The

contribution of Mor26p, Hy, GATS7p and Mor04v molecular descriptors in the building of ANN model was 26.27%, 26.09%, 25.62% and 21.99%, respectively, that show different values in comparison with the MLR model. Although GATS7p shows the largest contribution in the building of MLR model, in ANN model all four descriptors show comparable contribution in the building of model. Also, it should be considered that Mor26p has the largest correlation with the activities of molecules ($R^2 = 0.59$).

Comparing QSAR models

Calculated statistical parameters for the train and test sets of both models in Tables 4 and 5 show both QSAR models are acceptable and have high predictive power. Calculated CCC^2 , R^2 , $RMSE$, r_0^2 , $r_0'^2$, r_m^2 , $r_m'^2$, r_m^2 , $r_m'^2$ and Q^2 -based parameters (especially Q_{F3}^2 parameter) show that ANN model has more predictive power with respect to MLR model. William plot in Fig. 2 shows that molecule m15 is outlier in ANN model, but as seen from Table 2, ANN model has better prediction for its activity, and probably, it has happened because of the small standard deviation value of residuals for molecules in the train set of ANN model ($SD = 0.06$) with respect to MLR model ($SD = 0.13$).

Conclusions

The results of this research show the building of MLR and ANN models based on using Mor26p, Hy, GATS7p and Mor04v molecular descriptors which are suitable for predicting the SARS-CoV-1 3CL^{pro} inhibition activity of these ketone-based molecules. Although both created models are acceptable and show high predictive power, calculated R^2 - and Q^2 -based parameters and RMSE for both train and test sets of MLR model with four descriptors and ANN model show that the ANN model has more predictive power. The interpretation of descriptors (based on the developed MLR model with four descriptors) shows that groups with two fused rings in R_1 substituent are favorable for increasing the activity of molecule, bulky groups for R_2 substituent are not favorable for improving the activity of molecules, and the presence of cyclic groups and long-chain groups for R_3 substituent decreases the activity of molecules.

References

1. <https://covid19.who.int/>
2. T. Tuncer, F. Ozyurt, S. Dogan, A. Subasi, Chemometr. Intell. Lab. Syst. **210**, 104256 (2021)
3. G. Parsafar, V. Reddy, J. Iran. Chem. Soc. (2021). <https://doi.org/10.1007/s13738-021-02299-5>
4. S. Serte, H. Demirel, Comput. Biol. Med **132**, 104306 (2021)

5. A.T. Ton, F. Gentile, M. Hsing, F. Ban, A. Cherkasov, *Mol. Inf.* **39**, 2000028 (2020)
6. Y. Zhang, R.A. Greer, Y. Song, H. Praveen, Y. Song, *Eur. J. Pharm. Sci.* **160**, 105771 (2021)
7. V.M. Alves, T. Bobrowski, C.C. Melo-Filho, D. Korn, S. Auerbach, C. Schmitt, E.N. Muratov, A. Tropsha, *Mol. Inf.* **40**, 2000113 (2021)
8. M. Ciotti, M. Ciccozzi, A. Terrinoni, W.C. Jiang, C.B. Wang, S. Bernardini, *Crit. Rev. Clin. Lab. Sci.* **57**, 365–388 (2020)
9. E. Duverger, G. Herlem, F. Picaud, *J. Mol. Graph. Model.* **104**, 107834 (2021)
10. C.N. Cavasotto, J.I. Di Filippo, *Mol. Inf.* **40**, 2000115 (2021)
11. N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, E. Petersen, *Clin. Microbiol. Infect.* **26**, 729–734 (2020)
12. S. Mills, *Judic. Rev.* **25**, 71–79 (2020)
13. M.A. Kabir, R. Ahmed, R. Chowdhury, S.M. Asher Iqbal, R. Paulmurugan, U. Demirci, W. Asghar, *Microbes Infect* (2021). <https://doi.org/10.1016/j.micinf.2021.104832>
14. M. Hartt, *Cities and Health* (2020). <https://doi.org/10.1080/23748834.2020.1788770>
15. <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>
16. C.N. Cavasotto, M.S. Lamas, J. Maggini, *Eur. J. Pharmacol.* **890**, 173705 (2021)
17. F. Li, *Annu. Rev. Virol.* **3**, 237–261 (2016)
18. K. Kucukoglu, N. Faydal, D. Bul, *Med. Chem. Res.* **29**, 1935–1955 (2020)
19. A. Sattari, A. Ramazani, H. Aghahosseini, *J. Iran. Chem. Soc.* (2021). <https://doi.org/10.1007/s13738-021-02235-7>
20. A.G. Wrobel, D.J. Benton, S. Hussain, R. Harvey, S.R. Martin, C. Roustan, P.B. Rosenthal, J.J. Skehel, S.J. Gamblin, *Nat. Commun.* **11**, 5337 (2020)
21. R. Yousefi, A. Moosavi-Movahedi, *J. Iran. Chem. Soc.* **17**, 1257–1258 (2020)
22. S. Barge, D. Jade, G. Gosavi, N.C. Talukdar, J. Borah, *Eur. J. Pharm. Sci.* **162**, 105820 (2021)
23. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, *Nature* **581**, 215–220 (2020)
24. Y. Muhammed, *Biosaf. Health* **2**, 210–216 (2020)
25. K. Ghosh, S. Abdul Amin, S. Gayen, T. Jha, *J. Mol. Struct.* **1237**, 130366 (2021)
26. S. Zhang, M. Krumberger, M.A. Morris, C. Marie, T. Parrocha, A.G. Kreutzer, J.S. Nowick, *Eur. J. Med. Chem.* **218**, 113390 (2021)
27. P. Chellapandi, S. Saranya, *Med. Chem. Res.* **29**, 1777–1791 (2020)
28. C.K. Chang, S.M. Lin, R. Satange, S.C. Lin, S.C. Sun, H.Y. Wu, K. Kehn-Hall, M.H. Hou, *Comput. Struct. Biotechnol. J.* **19**, 2246–2255 (2021)
29. M.S. Mirtaleb, A.H. Mirtaleb, H. Nosrati, J. Heshmatnia, R. Falak, R. Zolfaghari Emameh, *Biomed. Pharmacother.* **138**, 111518 (2021)
30. R. Ahmadi, B. Sepehri, R. Ghavami, *J. Recept. Signal Transduct.* **39**, 264–275 (2019)
31. R.L. Hoffman, R.S. Kania, M.A. Brothers, J.F. Davies, R.A. Ferre, K.S. Gajiwala, M. He, R.J. Hogan, K. Kozminski, L.Y. Li, J.W. Lockner, J. Lou, M.T. Marra, L.J. Mitchell Jr., B.W. Murray, J.A. Nieman, S. Noell, S.P. Planken, T. Rowe, K. Ryan, G.J. Smith III., J.E. Solowiej, C.M. Steppan, B. Taggart, *J. Med. Chem.* **63**, 12725–12747 (2020)
32. L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, *Science* **368**, 409–412 (2020)
33. W. Dai, B. Zhang, X.M. Jiang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, J. Peng, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.K. Zhang, Y. Xu, H. Yang, H. Liu, *Science* **368**, 1331–1335 (2020)
34. S. Tomar, M.L. Johnston, S.E.S. John, H.L. Osswald, P.R. Nyalapatla, L.N. Paul, A.K. Ghosh, M.R. Denison, A.D. Mesecar, *J. Biol. Chem.* **290**, 19403–19422 (2015)
35. W. Dai, D. Jochmans, H. Xie, H. Yang, J. Li, H. Su, D. Chang, J. Wang, J. Peng, L. Zhu, Y. Nian, R. Hilgenfeld, H. Jiang, K. Chen, L. Zhang, Y. Xu, J. Neyts, H. Liu, *J. Med. Chem.* (2021). <https://doi.org/10.1021/acs.jmedchem.0c02258>
36. B. Bai, A. Belovodskiy, M. Hena, A.S. Kandadai, M.A. Joyce, H.A. Saffran, J.A. Shields, M.B. Khan, E. Arutyunova, J. Lu, S.K. Bajwa, D. Hockman, C. Fischer, T. Lamer, W. Vuong, M.J. van Belkum, Z. Gu, F. Lin, Y. Du, J. Xu, M. Rahim, H.S. Young, J.C. Vederas, D.L. Tyrrell, M.J. Lemieux, J.A. Nieman, *J. Med. Chem.* (2021). <https://doi.org/10.1021/acs.jmedchem.1c00616>
37. W. Vuong, M.B. Khan, C. Fischer, E. Arutyunova, T. Lamer, J. Shields, H.A. Saffran, R.T. McKay, M.J. van Belkum, M.A. Joyce, H.S. Young, D.L. Tyrrell, J.C. Vederas, M.J. Lemieux, *Nat. Commun.* **11**, 4282 (2020)
38. Z. Chen, S.S. Boon, M.H. Wang, R.W.Y. Chan, P.K.S. Chan, *J. Virol. Methods* **289**, 114032 (2021)
39. HyperChem 7.1. Gainesville, USA: Hypercube, Inc. Available from: <http://www.hyper.com>
40. Milano chemometrics and QSAR research group, 2007. Available from <http://www.taletе.mi.it/dragon.htm>
41. <http://www.spss.com>
42. <https://www.r-project.org/>
43. <https://rstudio.com/>
44. B. Sepehri, R. Ghavami, S. Farahbakhsh, R. Ahmadi, *Int. J. Environ. Sci. Technol.* (2021). <https://doi.org/10.1007/s13762-021-03271-9>
45. https://www.researchgate.net/publication/350459619_MLRQS_AR_package_version_010_for_R_programming_language
46. <https://cloud.r-project.org/web/packages/h2o/index.html>
47. <https://cran.r-project.org/web/packages/ggplot2/index.html>
48. R. Ghavami, B. Sepehri, *J. Iran. Chem. Soc.* **13**, 519–529 (2016)
49. R. Ghavami, B. Sepehri, *J. Chromatogr. A* **1233**, 116–125 (2012)
50. K. Phil, *Matlab Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence* (Apress, New York, 2017)
51. D. Cook, *Practical Machine Learning with H2O* (O'Reilly Media Inc, Massachusetts, 2017)
52. J. Moolayil, *Learn Keras for deep neural networks*, (Jojo Moolayil, 2019)
53. A. Candel, E. LeDell, *Deep learning with H₂O*, (H₂O.ai, Inc, 2020)
54. B. Sepehri, R. Ghavami, *Med. Chem.* **14**, 439–450 (2018)
55. B. Sepehri, R. Ghavami, *J. Mol. Struct.* **1130**, 922–928 (2017)
56. B. Sepehri, Z. Rasouli, Z. Hassanzadeh, R. Ghavami, *Med. Chem. Res.* **25**, 2895–2905 (2016)
57. B. Sepehri, R. Ghavami, *SAR QSAR Environ. Res.* **30**, 21–38 (2019)
58. B. Sepehri, *J. Mol. Liq.* **297**, 112013 (2020)
59. O. Devinyak, D. Havrylyuk, R. Lesyk, *J. Mol. Graph. Model.* **54**, 194–203 (2014)