Biointerface Research in Applied Chemistry
Platinum Open Access Journal (ISSN: 2069-5837)

Article
Volume 12, Issue 3, 2022, 3090 - 3105
https://doi.org/10.33263/BRIAC123.30903105

Development of QSAR Models Using Singular Value Decomposition Method: A Case Study for Predicting Anti-HIV-1 and Anti-HCV Biological Activities

Ismail Hdoufane [1], Daoud Ounaissi [2,3], Azzouz Dermoune [3,*], Driss Cherqaoui [1]

[1] Department of Chemistry, Faculty of Sciences Semlalia, Cadi Ayyad University of Marrakech, BP 2390, Morocco; ismail.hdoufane@edu.uca.ma (I.H.); cherqaoui@uca.ma (D.C.);
[2] Clermont Auvergne University, INRAE, VetAgro Sup, UMR Herbivores, F-63122 Saint-Genès-Champanelle, France; daoud.ounaissi@gmail.com (D.O.) ;
[3] Laboratory of Probability and Statistics, UFR de Mathématiques, USTL, Bat. M2, 59655 Villeneuve d'Ascq Cédex, France; azzouz.dermoune@univ-lille1.fr (A.D.) ;
* Correspondence: ismail.hdoufane@edu.uca.ma (I.H.);

Scopus Author ID 57194186259
Received: 19.05.2021; Revised: 25.06.2021; Accepted: 1.07.2021; Published: 8.08.2021

Abstract: This study performed a detailed approach derived by coupling singular value decomposition (SVD) with multiple linear regression (MLR) methods on the performance and predictive capability of the quantitative structure-activity relationship (QSAR). The study was carried out on two different datasets of 128 HIV-1 attachment inhibitors and 115 HCV analogs. For both datasets, the structure of each compound was represented by suitable molecular descriptors. Then, the two datasets were divided into training and test sets employing the Kennard-Stone procedure (K-S). Both MLR and SVD-MLR models were developed to link the structure of the studied compounds to their reported biological activities. The selected models were subjected to the internal leave-one-out cross-validation method, and their predictive abilities were evaluated using the external test set. The developed SVD-MLR models were robust and reliable with an external determination coefficient (R^2_{test}) of 0.9755 and a mean-square error (MSE) of 0.0205, as well as an R^2_{test} of 0.9179 and MSE of 0.0298 for the HCV and the HIV set, respectively. In return, this model could be developed to predict the activities of a non-seen extra set of organic molecules for the purpose of either virtual screening or lead/hit optimization.

Keywords: HCV; HIV; Multiple Linear Regression (MLR); Singular Value Decomposition (SVD); QSAR.

1. Introduction

Quantitative structure-activity/property relationship (QSAR/QSPR) is considered a mathematical and statistical hypothesis that attempts to find models to accurately predict the biological activity of chemical compounds using their physicochemical properties [1–3]. The application of this approach on chemical datasets of molecules results in a mathematical model that indicates the relationship between the response endpoints (biological activity, property, toxicity, etc.) and the chemical structure information (described as molecular descriptors). QSAR/QSPR methodologies have been applied in various areas of chemistry for predicting the activities of chemical molecules, classifying active and non-active compounds, and serving as a powerful tool in dealing with issues related to the environmental and toxicological domains [4–9].

To begin with, as it is known, QSAR investigations cover three main sessions: data pre-processing, model construction, and model validation. Firstly, for the pre-processing session and after collecting the samples, the dataset must be split into a training set (used to construct the model) and a test set (used to assess the predictive ability) [10]. To better calibrate the constructed model, the training set must be constituted with representative samples and exhibit a well-balanced data distribution [11]. By the same token, the test set must represent the whole dataset [12].

Secondly, for the model construction session, several QSAR studies illustrate various linear and nonlinear techniques used for setting up regression and classification models, such as multiple linear regression (MLR), artificial neural network (ANN), support vector machine (SVM), partial least square (PLS) and a combination of particle swarm optimization with SVM namely (PSO-SVM) [13–20].

Finally, the model's validation, such as internal validation, external validation, and defining the applicability domain (AD), are the common techniques used to ensure the performance, fitting, and robustness of a QSAR model.

Herein, a method based on the combination of singular value decomposition (SVD) with the MLR methods was developed to set up robust and reliable models. The SVD topology has been established for complex square matrices by Autonne [21] and general rectangular matrices by Eckart and Young [22]. Its definition, a description of the use of this method in regression analysis, and its properties have been described in the literature [23,24].

In the current study, we shall see how the SVD topology combined with the MLR method can be used to improve the predictive ability and the performance of the constructed QSAR models reliably. Therefore, QSAR investigations were conducted on two different datasets of HIV-1 and HCV inhibitors. This work aims to assess and improve the predictive ability and robustness of the developed model based on the SVD topology. A comparison with the obtained results from the MLR model was performed as well.

2. Materials and Methods

2.1. Datasets preparation.

The current study includes two different datasets collected from the literature. A dataset of 128 HIV inhibitors [25–28] was evaluated in order to identify oral HIV-1 attachment inhibitors with the potential to improve the potency of BMS-488043 and/or BMS-378806 [29,30] and another dataset of 115 Hepatitis C virus inhibitors (HCV) [31–33] that was assessed for the ability to inhibit HCV RNA replication in the HCV replicon were used in this study. For both datasets, the biological activity is expressed by the concentration required to reduce 50% of the inhibitory activity (EC_{50} in µM). In order to use the activity data as the dependent variable, all the data were converted to their negative logarithm units $-\log_{10}(EC_{50})$ (i.e., pEC_{50}). The structures, the values of the biological activity, and the numbering of the compounds for HIV and HCV datasets are given in Tables S1 and S2, respectively.

To evolve and assess QSAR models, the Kennard and Stone (K-S) method was used to divide the data into training and test sets [34]. The training set comprising 70% of the compounds was used to set up the QSAR models, and the remaining 30% were used to evaluate the predictive ability of the resulting models.

2.2. Molecular descriptors generation.

For both datasets, all compounds were drawn and optimized using Gaussian 09 program [35]. The optimization was conducted in the ground state by the DFT method at the B3LYP level of theory with a 6-31G(d,p) basis set. Then, structural parameters (descriptors), which are mathematical values that describe the physical and chemical properties of a molecule, were computed using Dragon 7 program [36]. This program allows the calculation of 5225 data, including several descriptors such as constitutional indices, ring descriptors, topological indices, GETAWAY descriptors, etc.

As it is well-known, creating a predictive QSAR model is quite promising. Initially, we performed a pre-processing data method by removing (a) descriptors with a standard deviation of less than 0.0001, (b) descriptors with at least one missing value, (c) descriptors that are per-correlated with more than 0.95, and (d) descriptors with constant or almost constant values. Further, the stepwise multiple linear regression (stepwise-MLR) approach was applied to select the most important descriptors among the calculated ones. This technique has proved to be a suitable computational method in data analysis problems [37,38]. Finally, four selected descriptors in both HIV and HCV cases are illustrated in Tables 1 and 2.

Table 1. List of the selected molecular descriptors and their physical-chemical meaning for the HIV set.

Descriptor	Meaning
TDB04v	3D Topological distance-based descriptors - lag 4 weighted by the van der Waals volume
TDB10s	3D Topological distance-based descriptors – lag 10 weighted by the I-state
SpPosA_B(e)	Normalized spectral positive sum from Burden matrix weighted by Sanderson electronegativity
CATS2D_03_LL	CATS2D Lipophilic-Lipophilic at lag 3

Table 2. List of the selected molecular descriptors and their physical-chemical meaning for the HCV set.

Descriptor	Meaning
MATS5p	Moran autocorrelation of lag 5 weighted by polarizability
CATS2D_07_AL	CATS2D Acceptor-Lipophilic at lag 07
MATS1s	Moran autocorrelation of lag 1 weighted by I-state
F06[N-F]	Frequency of N - F at topological distance 6

For the HIV set, TDB04v and TDB10s are two parameters contemporarily based on the topological and geometric distances (also called 3D-TDB descriptors) [39]. TDB04v and TDB10s are related to the van der Waals volume and the intrinsic state, respectively [40]. SpPosA_B(e) is based on the topological shape and the electric state [41]. CATS2D_03_LL is based on the topological distance and provides further insight into the pattern of side-chain substituents in terms of lipophilic (L) character in a molecule. For the HCV set, MATS5p and MATS1s descriptors belong to the 2D-autocorrelation class. MATS5p is weighted by atomic polarizabilities, while MATS1s is calculated by applying the Moran coefficient to the molecular graph and weighted by intrinsic state. These two parameters are related to the dimension and shape of the molecules [42]. The third related descriptor is F06[N-F], which belongs to the substructure of atom pairs descriptors, consists of the Nitrogen and Fluorine atoms that are not directly connected and separated by the topological distance. The fourth parameter is CATS2D_07_AL, which is based on the topological distance and provides information between the hydrogen-bond acceptor (HBA) and lipophilic (L) pharmacophore points [40].

2.3. QSAR model development.

In this study, the developed MLR and SVD-MLR models were executed using R software version 3.5.1, which is a free software for statistical computing and graphics [43]. These models are based on MLR and SVD methods and implement rigorous internal and external validation based on different validation criteria. The main ideas about MLR and SVD developed approaches are given below:

Let us consider p descriptors $X1, \dots, Xp$ For each l^{-th} realization $x_{l1}, \dots, x_{lp}$ of the p descriptors, we calculate the activity y_l, with $l = 1, \dots, n$. We have the regression equation $\mathbf{y} = \mathbf{X}\beta + \epsilon$ - what estimate of β would best separate the systematic component $\mathbf{X}\beta$ from the random component ϵ. The problem is to find $\hat{\beta}$ such that $\mathbf{X}\hat{\beta}$ is close to y. The response predicted by the model is $\hat{y} = \mathbf{X}\hat{\beta}$ or **H**y, where H is an orthogonal projection matrix of y onto space, spanned by **X**. The difference between the actual response and the predicted response is denoted by $\hat{\epsilon}$ (the residuals). We build a family of predicting models using the learning data $(y_l; x_{l1}, \dots, x_{lp})$, with $l = 1, \dots, n$, then we select the best model using the cross-validation method.

2.3.1. MLR method: least-squares estimation.

We might define the best estimate of β as that which minimizes the sum of the squared errors, $\epsilon^{\mathrm{T}}\epsilon = \|\mathbf{y} - \mathbf{X}\beta\|^2$. That is to say that the least-squares estimate called β minimizes

$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta)$.

We find that $\hat{\beta}$ satisfies $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\beta} = \mathbf{X}^{\mathrm{T}}$y (the normal equations). Now provided $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible

$$\hat{\beta} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{X}\hat{\beta}\mathbf{y} = \mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is the orthogonal projection of y onto space spanned by **X**.

Finally, we can remember that the least-squares model is defined by

$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

2.3.2. SVD-MLR model.

Usually, if R^2 is small or equivalently if $\frac{\|\mathbf{H}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2}$ is small, then we reject the linear regression model. Here we show how to improve R^2 using singular value decomposition (SVD).

The column vectors $\boldsymbol{y}^{(n)} = (y_1, \dots, y_n)^{\mathrm{T}}$, $x_0^{(n)} = 1^n$, $x_1^{(n)} = (x_{11}, \dots, x_{n1})^{\mathrm{T}}$,…, $x_p^{(n)} = (x_{1p}, \dots, x_{np})^{\mathrm{T}}$ belong to the Euclidean space $\mathbb{R}^n$. The projector on the subspace spanned by the columns of the design matrix $\mathbf{X}^{(n)} = x_0^{(n)}, \dots, x_p^{(n)}$ has the matrix

$$\mathbf{H}^{(n)} = \mathbf{X}^{(n)}\left[\left(\mathbf{X}^{(n)}\right)^{\mathrm{T}}\mathbf{X}^{(n)}\right]^{-1}\left(\mathbf{X}^{(n)}\right)^{\mathrm{T}}.$$

SVD tells us that

$$\mathbf{H}^{(n)} = \mathbf{V}^{(n)}\boldsymbol{diag}(1, \dots, 1, 0, \dots, 0)(\mathbf{V}^{(n)})^{\mathrm{T}},$$

where the columns $v_1^{(n)}, \dots, v_n^{(n)}$ form an orthonormal basis of $\mathbb{R}^n$. It follows that

$$\boldsymbol{y}^{(n)} = \mathbf{H}^{(n)}\boldsymbol{y}^{(n)} + \sum_{k=p+1}^{n} (v_k^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\,v_k^{(n)}.$$

Observe that $\mathbf{H}^{(n)}\boldsymbol{y}^{(n)} = \sum_{k=0}^{p}(v_k^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\,v_k^{(n)}$ and

$$\left\|\boldsymbol{y}^{(n)}\right\|^2 = \left\|\mathbf{H}^{(n)}y^{(n)}\right\|^2 + \sum_{k=\mathrm{p}+1}^{n}\left|\mathrm{v}_k^{(n)\mathrm{T}}\boldsymbol{y}^{(n)}\right|^2$$

We sort in descending order the sequence $\left|(v_{p+1}^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right|^2, \dots, \left|\left(v_n^{(n)}\right)^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right|^2$:

$$\left|(v_{i_{p+1}}^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right|^2 \geq \cdots \geq \left|\left(v_{i_n}^{(n)}\right)^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right|^2,$$

with the integers $i_{p+1}, \dots, i_n \in \{p+1, \dots, n\}$.

By considering the new design matrix $\widetilde{\mathbf{X}}^n = \left[\mathbf{X}^n, v_{i_{p+1}}^{(n)}, \dots, v_{i_q}^{(n)}\right]$ *with* $q \in \{p+1, \dots, n\}$, and we regress $\boldsymbol{y}^{(n)}$ on the latter design matrix, we obtain

$$\hat{\mathbf{y}} = \widetilde{\mathbf{X}}^{(n)}\left[(\widetilde{\mathbf{X}}^{(n)})^{\mathrm{T}}\widetilde{\mathbf{X}}^{(n)}\right]^{-1}(\widetilde{\mathbf{X}}^{(n)})^{\mathrm{T}}\boldsymbol{y}^{(n)} = \mathbf{X}^{(n)}\left[(\mathbf{X}^{(n)})^{\mathrm{T}}\mathbf{X}^{(n)}\right]^{-1}(\mathbf{X}^{(n)})^{\mathrm{T}}\boldsymbol{y}^{(n)} + \sum_{k=p+1}^{n}\left[(v_{i_k}^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right]v_{i_k}^{(n)},$$

and then we improve the R^2. In order to predict the activity/property of a new molecule c, we proceed as follows. We form the design matrix $\widetilde{\mathbf{X}}^{(n+1)}$ such that $x_{ij}^{(n+1)} = x_{ij}$ with $i = 1, \dots, n$, $j = 0, \dots, p$, $x_{n+1j}^{(n+1)} = x_{n+1j}$ with $j = 0, \dots, p$, and $x_{n+1j}^{(n+1)} = v_{n+1i_j}^{(n+1)}$ with $j = p+1, \dots, q$. Here $\mathbf{V}^{(n+1)} = \left[v_1^{(n+1)}, \dots., v_{n+1}^{(n+1)}\right]$ is the orthogonal basis of $\mathbb{R}^{n+1}$ given be the SVD of $\mathbf{H}^{(n+1)}$, i.e.,

$$\mathbf{H}^{(n+1)} = \mathbf{V}^{(n+1)}\,\boldsymbol{diag}(1, \dots, 1, 0, \dots, 0)\left(\mathbf{V}^{(n+1)}\right)^{\mathrm{T}}.$$

Finally, we predict $\boldsymbol{y}_{(n+1)}$ by

$$\left(1, x_{n+11}, \dots, x_{n+1p}\right)v_{n+1i_{p+1}}^{(n+1)}, \dots, v_{n+1i_q}^{(n+1)}\left[(\widetilde{\mathbf{X}}^{(n)})^{\mathrm{T}}\widetilde{\mathbf{X}}^{(n)}\right]^{-1}(\widetilde{\mathbf{X}}^{(n)})^{\mathrm{T}}\boldsymbol{y}^{(n)}$$

$$= \left(1, x_{n+11}, \dots, x_{n+1p}\right)\left[(\mathbf{X}^{(n)})^{\mathrm{T}}\mathbf{X}^{(n)}\right]^{-1}(\mathbf{X}^{(n)})^{\mathrm{T}}y^{(n)} + \sum_{k=p+1}^{n}\left[(v_{i_k}^{(n)})^{\mathrm{T}}\,\boldsymbol{y}^{(n)}\right]v_{n+1i_k}^{(n+1)}.$$

2.4. QSAR model's validation.

Model validation is of pivotal importance in the development of the regression QSAR model. In most cases, two common validation methods are used for evaluating the performance of a generated QSAR model, including the leave-one-out cross-validation (*LOO-CV*) using the training set and the external validation using the test set. These methods play a critical role in the assessment of the stability and reliability of the constructed models. The established QSAR models are evaluated using the determination coefficient (R^2) and the Mean Square Error (*MSE*). These metrics are calculated as follows:

$$R^2 = 1 - \sum_{i=1}^{n}(Y_{i,exp} - Y_{i,pre})^2 \Big/ \sum_{i=1}^{n}(Y_{i,exp} - \bar{Y}_{exp})^2$$

$$MSE = \sum_{i=1}^{n}(Y_{i,exp} - Y_{i,pre})^2/n$$

where $Y_{i,exp}$ and $Y_{i,pre}$ are the experimental and the predicted values of pEC_{50}, respectively. $\bar{Y}_{exp}$ is the average of the experimental activities; n is the number of molecules in the dataset. $\bar{Y}_{exp}$ and $\bar{Y}_{pre}$ are the average of the experimental and predictive activities, respectively.

In the internal validation, *LOO-CV* is commonly applied to assess models as an internal validation. Cross-validated (q_{LOO}^2) and *MSE* are two explored parameters that verify the reliability of the model [44]. For instance, a higher value of q_{LOO}^2 and a lower value of *MSE*

imply that the internal predictive performance of models is effective. In the external validation, the predictive ability is verified by calculating the R^2_{test} determination coefficient using a test set and also the *MSE*. The q^2_{LOO} and the R^2_{test} are calculated in the same way as R^2.

Furthermore, several studies have been done to find the appropriate statistical parameters and criteria to evaluate the external predictive capabilities of a QSAR model. The commonly used internal and external validation criteria are those proposed by Golbraikh and Tropsha approaches [45,46]. Thus, the following metrics were adopted to validate all constructed models.

$q^2_{LOO} > 0.5$

$R^2_{test} > 0.6$

As an additional criterion for the external validation of QSAR models, Mean Average Error (*MAE*) was used along with a conceptually simpler statistical parameter that basically verifies the agreement of experimental and predicted data: the concordance correlation coefficient (*CCC*) of Lin et al. [47,48]. Moreover, Chirico and Gramatica had recommended the thresholds for the following parameters to be *MAE* < 0.6 and *CCC* > 0.85 [49,50].

$$MAE = \sum_{i=1}^{n_{ext}} \left|(Y_{i,exp} - Y_{i,pre})\right| / n_{Ext}$$

$$CCC = \frac{2 \sum_{i=1}^{n_{Ext}} (Y_{i,exp} - \bar{Y}_{exp})(Y_{i,pre} - \bar{Y}_{pre})}{\sum_{i=1}^{n_{Ext}} (Y_{i,exp} - \bar{Y}_{exp})^2 + \sum_{i=1}^{n_{ext}} (Y_{i,pre} - \bar{Y}_{pre})^2 + n_{Ext} (\bar{Y}_{exp} - \bar{Y}_{pre})^2}$$

where n_{Ext} is the number of molecules in the test set. Overall, the quality of the established models was checked using the statistical metrics described above, and a general schematic flowchart of this QSAR study is shown in Figure 1.

2.5. Applicability domain.

Applicability domain (AD), which represents the chemical space for reliable predictions, is defined as the hypothetical structural and functional domain within which the predictions of QSAR models are termed reliable. Each obtained model has its own space for the application. Only compounds falling within the model's space are expected to be accurately predicted. For this purpose, the approach based on leverage is quite recommended for defining the AD of a QSAR model [46]. The leverage value (h) of a query chemical used to evaluate the model's AD is proportional to its Mahalanobis distance measure from the centroid of the training set [51]. The leverages are calculated for a given dataset X by obtaining the leverage matrix (H) with the equation below:

$$h_i = x_i^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} x_i, \quad (i = 1, \ldots, n)$$

where x_i is the descriptor row-vector of the query compounds, and X is $(m \times p)$ matrix of the data set (m is the number of the training set samples and p is the number of descriptors). The superscript T refers to the transpose of the matrix X and the vector x_i. n is the number of the query compounds. The warning value defines the delimited scope of the model (h^*), generally, set as $3\,p/n$. Leverage greater than the warning value h^* means that the predicted response results from substantial extrapolation of the model and, therefore, may not be reliable.

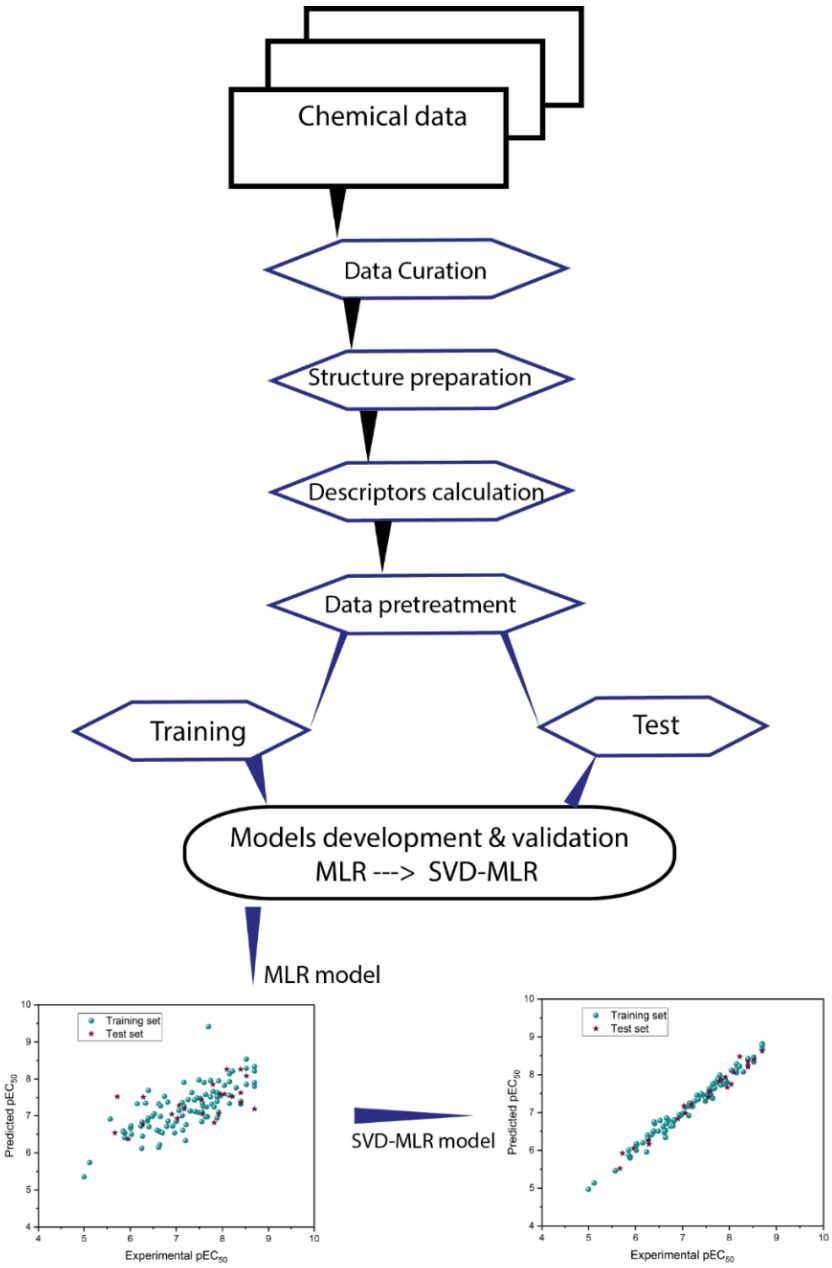

Figure 1. The general computational workflow of the present study.

3. Results and discussion

Several sophisticated nonlinear mathematical methods such as ANN and SVM have become quite efficient towards strengthening the basis of a given structure-activity relationship, which, in turn, evolves a more approximate quantitative rationale.

Apart from this, other hybrid methodologies providing good correlations and meaningful structure–activity-based regression models, including GA-MLR, GA-SVM, and PSO-SVM models, were performed to formulate an excellent predictive model [20,52–55]. In this regard, a method based on the combination of the SVD methodology and the classical MLR was developed to improve the predictive ability of a constructed MLR model.

This approach was subjected to the three principal sessions described previously in the introduction section. Furthermore, the QSAR investigation derived from this approach was evaluated by means of the standard criteria of the internal and external validations. The AD represented by the Williams plots was determined as well.

3.1. Computation (internal validation).

Validation of the QSAR models was required to test the prediction and the generalization of the methods. Any model needs to be validated before using it for predicting new trials. Tropsha et al. [46,56] have dealt with this problem by providing a set of procedures for developing and/or evaluating QSAR models. In order to be reliable and predictive, QSAR models should: (a) be statistically significant and robust, (b) be validated by making accurate predictions for external data sets, and (c) have their application scope defined.

3.1.1. MLR model.

The MLR models for the HCV and HIV datasets were established. The obtained linear equations are listed as follows:

pEC_{50} (HCV) = 8.78 + 10.84*(MATS1s) + 8.80*(MATS5p) + 0.25*(F06[N-F]) – 0.19*(CATS2D_07_AL) (Eq. 1)

pEC_{50} (HIV) = – 43.41 + 4.83*(TDB04v) – 0.07*(TDB10s) + 32.82*(SpPosA_B(e)) + 0.09*(CATS2D_03_LL) (Eq. 2)

For both datasets (HIV and HCV), the developed MLR models could not accurately predict the experimental activities. For instance, R^2, q^2_{LOO} and *MSE* statistical parameters for the training set were 0.5507, 0.5510, and 0.3249 for the HCV data, respectively. Similar to the HCV model, the corresponding statistical metrics R^2, q^2_{LOO} and *MSE* for the HIV set were 0.6132, 0.5741, and 0.5190, respectively. The obtained R^2 and q^2_{LOO} were slightly acceptable, while the *MSE* has a higher value.

The scatter plot of predicted versus experimental values for both sets is shown in Figures 2 and 3. These two figures show that the pEC_{50} values predicted by the MLR are far from the experimental ones. Additionally, the obtained statistics on these models show that the models are unacceptable for further validation steps. Nevertheless, we performed the external validation analysis to have a better comparison with the SVD-MLR models.

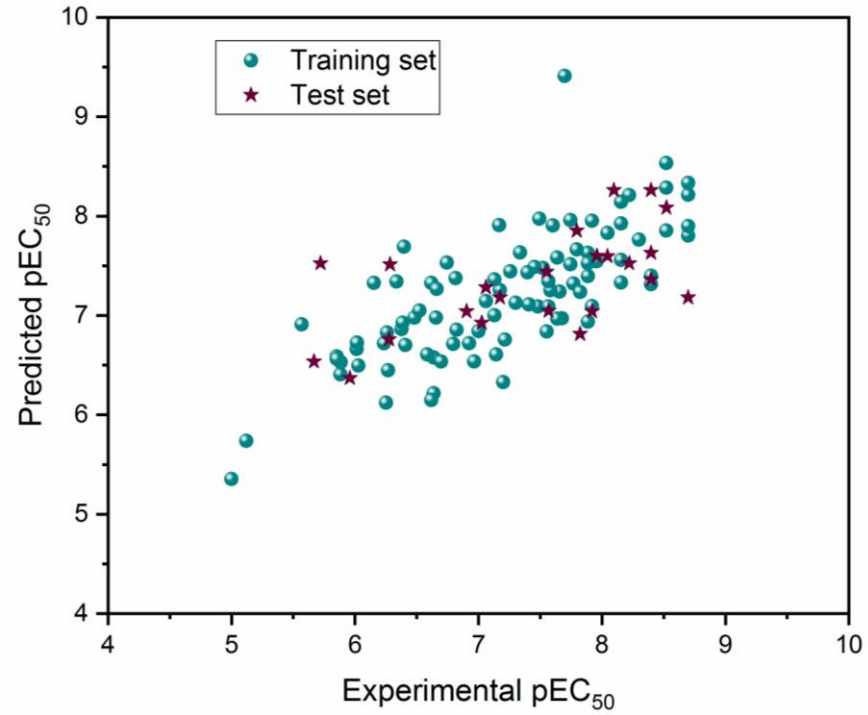

Figure 2. Predicted versus experimental pEC_{50} for the HCV dataset.

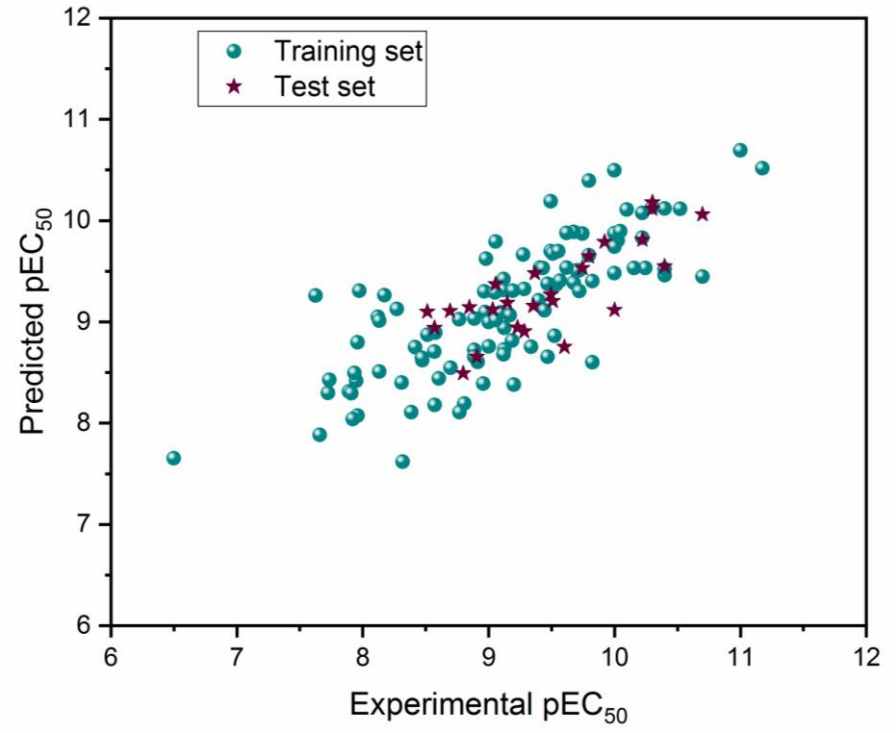

Figure 3. Predicted versus experimental pEC_{50} for the HIV dataset.

3.1.2. SVD-MLR model.

The SVD-MLR model consisting of the same descriptors was developed. The same criteria estimate the quality of the fitting. High R^2 coefficient and low *MSE* have been obtained by means of this method. For the HCV data, R^2, q^2_{LOO} and *MSE* statistical parameters for the training set were respectively 0.9802, 0.9855, and 0.0143. Similarly, the corresponding ones for the HIV data were 0.9810, 0.9728, and 0.0132.

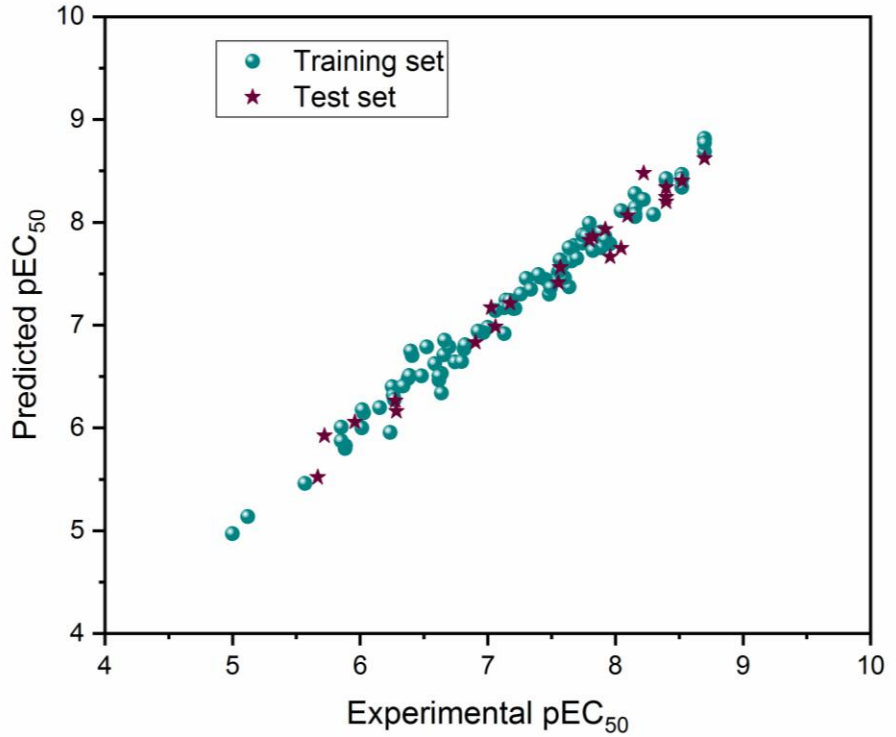

Figure 4. Predicted versus experimental pEC_{50} for the HCV dataset.

According to these results, it is clear that the results obtained from the SVD-MLR method are encouraging and better than those of the MLR method. Indeed, the SVD's determination coefficient is higher, and its standard deviation is lower than the MLR model's. Furthermore, the scatter plots of predicted versus experimental values for both sets are shown in Figures 4 and 5. These two figures show that the pEC50 values predicted by the SVD-MLR method are closer to the experimental ones. The residuals of the predicted pEC50 values plotted versus the predicted pEC50 are illustrated in Figures 6 and 7 for the HCV, and the HIV sets, respectively.

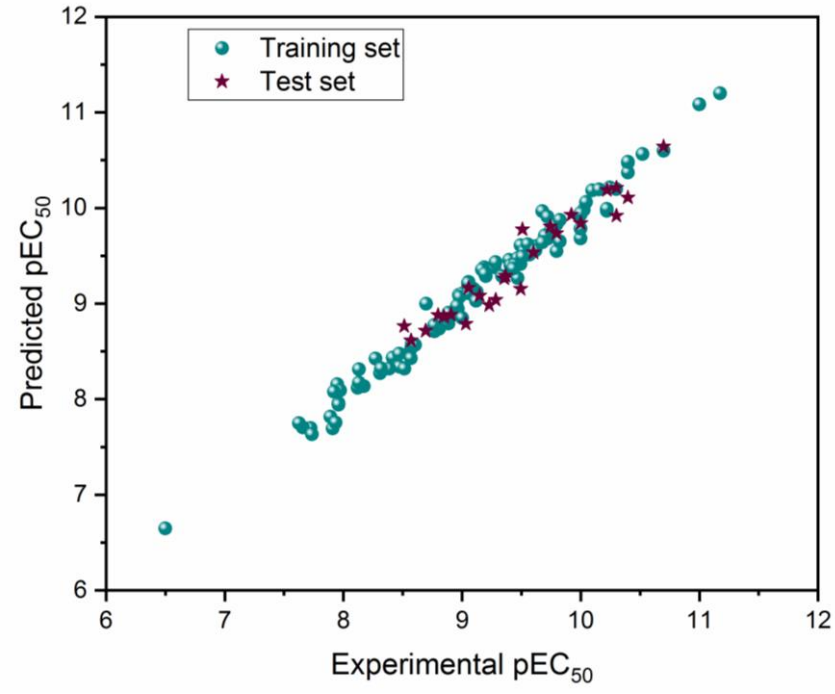

Figure 5. Predicted versus experimental pEC_{50} for the HIV dataset.

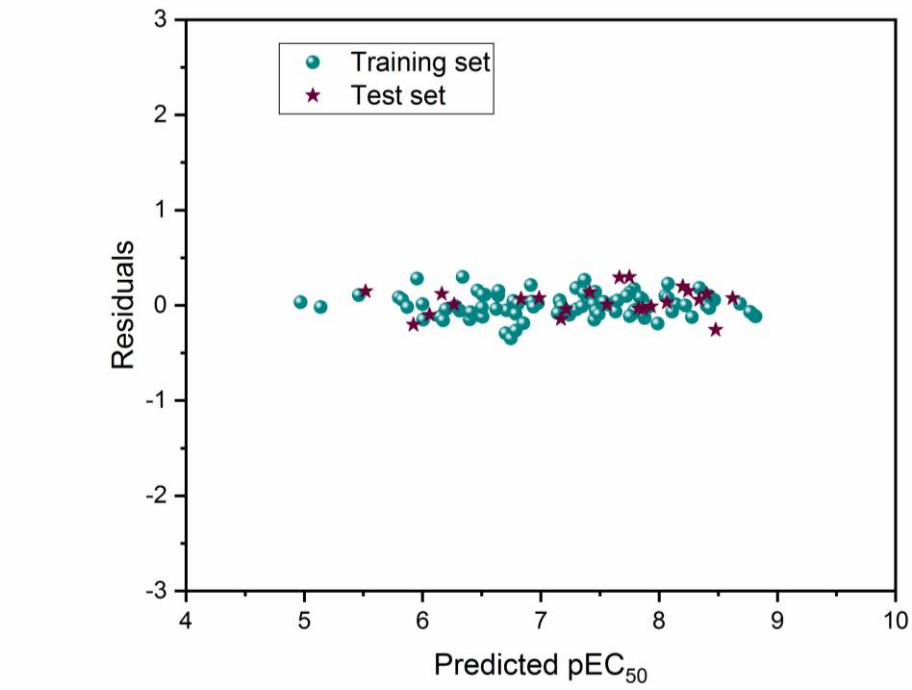

Figure 6. Distribution of the standard residuals by the SVD-MLR model for the HCV dataset.

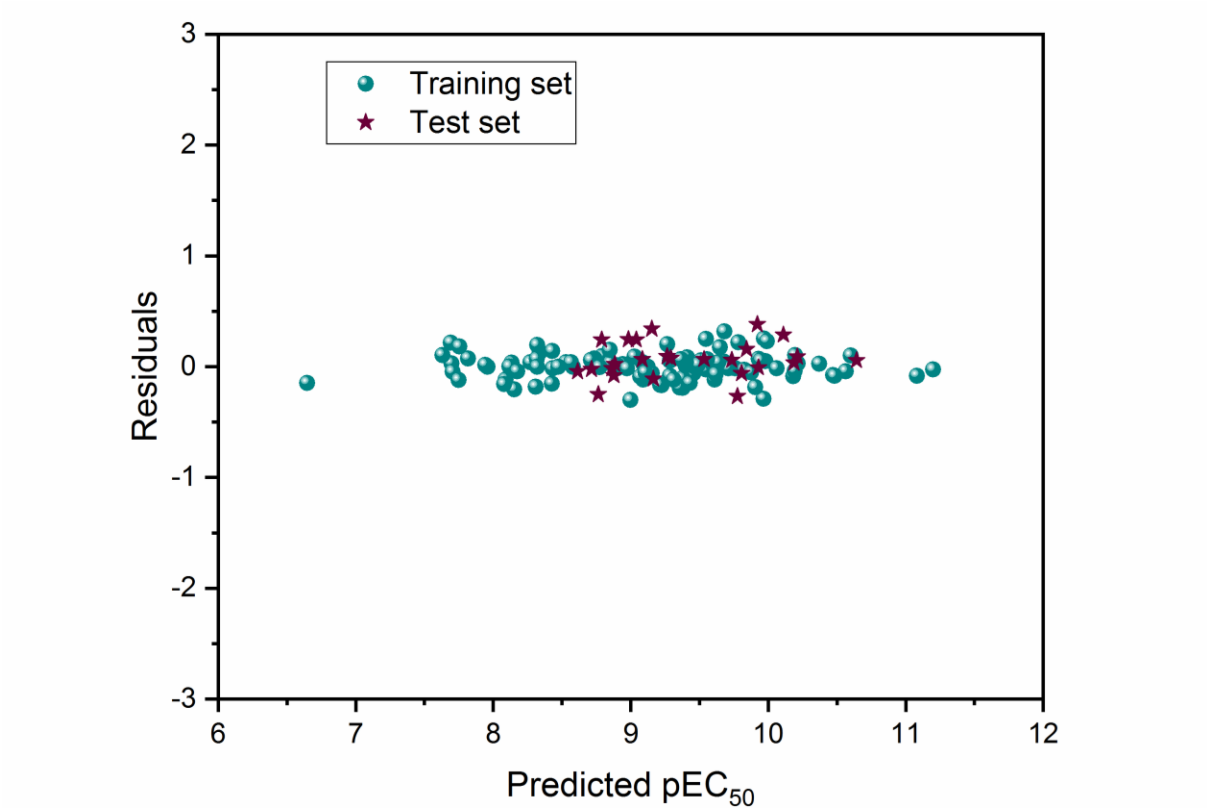

Figure 7. Distribution of the standard residuals by the SVD-MLR model for the HIV dataset.

3.2. Prediction (external validation).

Tables 3 and 4 show the associated statistical parameters for the HCV and the HIV datasets for the external prediction, respectively. It can be seen that the R^2_{test} of the SVD-MLR models is significantly higher than the one of the MLR models.

Table 3. Statistical parameters of different constructed QSAR models for the HCV dataset.

	Training set		*LOO-CV*		Test set	
Method	*R²*	*MSE*	q^2_{LOO}	*MSE*	R^2_{test}	*MSE*
MLR	0.5507	0.3249	0.5510	0.3247	0.3237	0.5699
SVD-MLR	0.9802	0.0143	0.9855	0.0104	0.9755	0.0205

Table 4. Statistical parameters of different constructed QSAR models for the HIV dataset.

	Training set		*LOO-CV*		Test set	
Method	*R²*	*MSE*	q^2_{LOO}	*MSE*	R^2_{test}	*MSE*
MLR	0.6132	0.5190	0.5741	0.5446	0.5975	0.4175
SVD-MLR	0.9810	0.0132	0.9728	0.0188	0.9176	0.0298

From the above results, it can be seen that the fitness and robustness of the hybrid SVD-MLR approach are both very good, and it was successfully externally validated as well. Additionally, from the scatter plots obtained by means of the SVD-MLR model of the experimental and predicted pEC50 activities (Figures 4 and 5), we can see that all the samples are distributed near the diagonal with a good fitting, which indicates that the predicted values are very close to the experimental ones. Furthermore, all the calculated internal validation metrics satisfy the conditions described previously.

On analysis, The MAE and CCC metrics, which are shown in Table 5, were determined to ensure the robustness of the established models.

Table 5. *MAE* and *CCC* statistical parameters for the external validation.

	HIV set		HCV set	
	MAE	CCC	MAE	CCC
MLR	0.3442	0.7003	0.5832	0.4660
SVD-MLR	0.1320	0.9598	0.1144	0.9874

It is interesting to note that the SVD-MLR approach provides better results comparing to the MLR method. All external validation metrics of the SVD-MLR models are sufficient to ensure its performance. These models were found to have smaller values of *MSE* for the internal and external verification in both cases. Thereby the constructed models by means of the hybrid SVD-MLR approach exhibits a higher predictive capability and robustness.

3.3. Applicability domain evaluation.

In this study, the method based on leverage values is employed to evaluate the practical applicability domain of the constructed models. The results, illustrated as Williams plot, are depicted in Figures 8 and 9. The cutoff value of ±3 standard deviation (s.d.) is fixed to determine the limits where the samples are considered outliers for both sets. The warning lines (i.e. leverage threshold (h*)) are, respectively, 0.145 and 0.163 for HIV and HCV sets. For the HIV data, except for compounds No. 30, 51, and 52, chemicals are within the AD. Even for these chemicals (30, 51, and 52), whose h values are beyond h*, the predicted pEC_{50} are close to their experimental values. Additionally, except compound No. 92, all compounds are within the standard residuals of ±3 (s.d). Except for compound No. 4, the remaining samples are within the application scope for the HCV data. Similar to the HIV model, except compounds 16 and 65, all entities are within the standard residuals of ±3 (s.d). Nonetheless, for these two compounds (16 and 65), which slightly exceed the cutoff value, their predicted values are close to the experimental ones. All established models showed a minimum of 90% coverage for the studied datasets. The coverage rate was 96.8% and 97.4% of the total chemical domain for HIV-1 and HCV datasets. Hence, the developed models for these sets are reasonable as well.

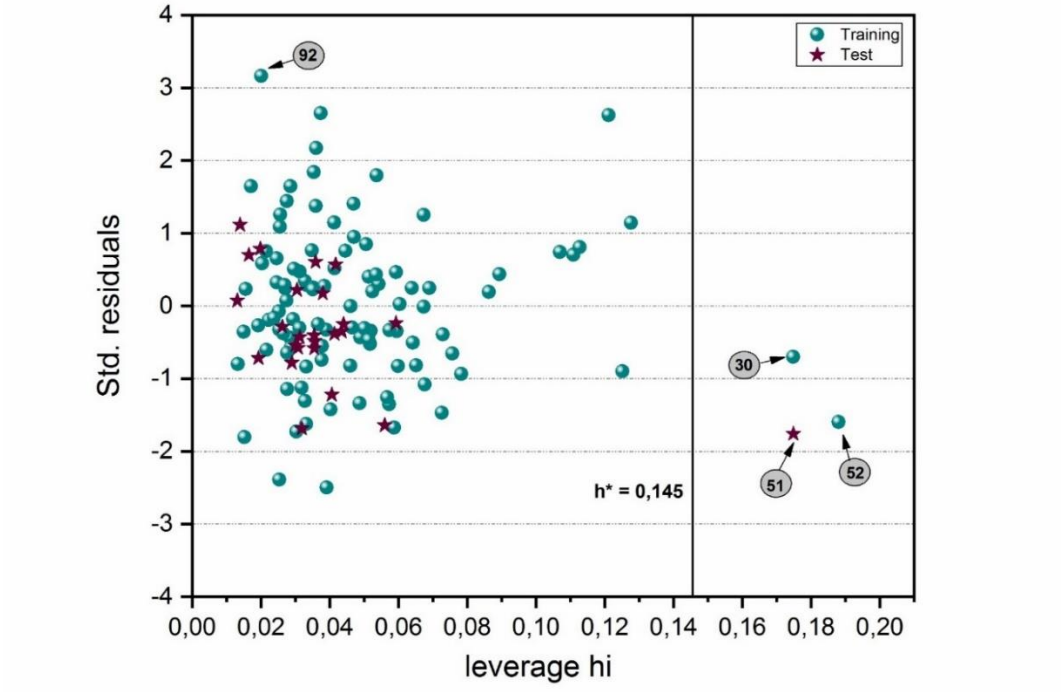

Figure 8. Williams plot of the developed model for HIV data.

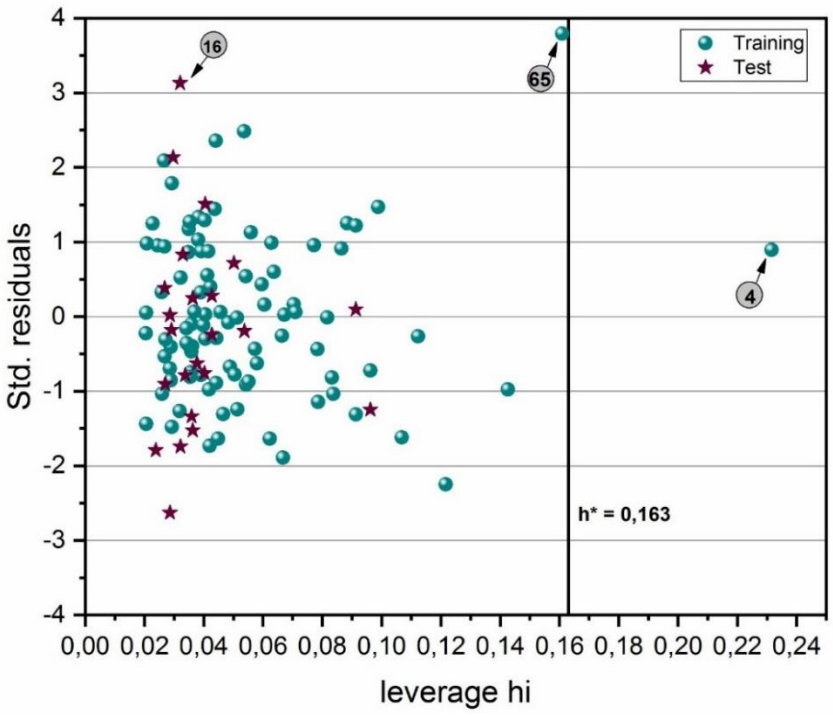

Figure 9. Williams plot of the developed models for HCV data.

4. Conclusions

In the present research, taking EC_{50} as the dependent variable and the four selected generated descriptors as the independent variable of 128 HIV-1 and 115 HCV derivatives, QSAR study of these series and their corresponding inhibitory activities were investigated using the MLR and the developed approach by combining the SVD and the MLR methods. It has been shown that the derived SVD-MLR models met the recommended internal and external validation metrics. These established models are robust, stable, and predictive with satisfactory performance comparing to the MLR models. Finally, this comparison between the established models shows that the SVD-MLR models give satisfactory results. Consequently, the SVD methodology could be integrated with the MLR method and could be used to develop QSAR/QSPR models for improving and predicting the biological activity, the property, and the toxicity of the chemical dataset of molecules.

Funding

This research received no external funding.

Acknowledgments

The authors declare no acknowledgments.

Conflicts of Interest

The authors report no conflict of interests.

References

1. Kubinyi, H. *QSAR : Hansch analysis and related approaches*; VCH, 1993.
2. Hoekman, D. Exploring QSAR Fundamentals and Applications in Chemistry and Biology, Volume 1.

Hydrophobic, Electronic and Steric Constants, Volume 2 J. Am. Chem. Soc. 1995, 117, 9782. **1996**, https://doi.org/10.1021/ja965433+.
3. Mombelli, E.; Pandard, P. Evaluation of the OECD QSAR toolbox automatic workflow for the prediction of the acute toxicity of organic chemicals to fathead minnow. *Regul. Toxicol. Pharmacol.* **2021**, *122*, 104893, https://doi.org/10.1016/j.yrtph.2021.104893.
4. Piir, G.; Sild, S.; Maran, U. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *Chemosphere* **2021**, *262*, 128313, https://doi.org/10.1016/j.chemosphere.2020.128313.
5. Miličević, A.; Šinko, G. Development of a simple QSAR model for reliable evaluation of acetylcholinesterase inhibitor potency. *Eur. J. Pharm. Sci.* **2021**, *160*, 105757, https://doi.org/10.1016/j.ejps.2021.105757.
6. Meftahi, N.; Walker, M.L.; Smith, B.J. Predicting aqueous solubility by QSPR modeling. *J. Mol. Graph. Model.* **2021**, *106*, 107901, https://doi.org/10.1016/j.jmgm.2021.107901.
7. Sun, Y.; Chen, M.; Zhao, Y.; Zhu, Z.; Xing, H.; Zhang, P.; Zhang, X.; Ding, Y. Machine learning assisted QSPR model for prediction of ionic liquid's refractive index and viscosity: The effect of representations of ionic liquid and ensemble model development. *J. Mol. Liq.* **2021**, *333*, 115970, https://doi.org/10.1016/j.molliq.2021.115970.
8. Toropov, A.A.; Raška, I.; Toropova, A.P.; Raškova, M.; Veselinović, A.M.; Veselinović, J.B. The study of the index of ideality of correlation as a new criterion of predictive potential of QSPR/QSAR-models. *Sci. Total Environ.* **2019**, *659*, 1387–1394, https://doi.org/10.1016/j.scitotenv.2018.12.439.
9. Algamal, Z.Y.; Qasim, M.K.; Lee, M.H.; Mohammad Ali, H.T. High-dimensional QSAR/QSPR classification modeling based on improving pigeon optimization algorithm. *Chemom. Intell. Lab. Syst.* **2020**, *206*, 104170, https://doi.org/10.1016/j.chemolab.2020.104170.
10. Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26*, 1–16, https://doi.org/10.3390/molecules26041111.
11. Andrada, M.F.; Vega-Hissi, E.G.; Estrada, M.R.; Garro Martinez, J.C. Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models. *SAR QSAR Environ. Res.* **2017**, *28*, 1011–1023, https://doi.org/10.1080/1062936X.2017.1397056.
12. Martin, T.M.; Harten, P.; Young, D.M.; Muratov, E.N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570–2578, https://doi.org/10.1021/ci300338w.
13. Mandal, A.S.; Roy, K. Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives. *Eur. J. Med. Chem.* **2009**, *44*, 1509–1524, https://doi.org/10.1016/j.ejmech.2008.07.020.
14. Hannongbua, S.; Pungpo, P.; Limtrakul, J.; Wolschann, P. Quantitative structure-activity relationships and comparative molecular field analysis of TIBO derivatised HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided. Mol. Des.* **1999**, *13*, 563–77, https://doi.org/10.1023/A:1008013917905.
15. Hdoufane, I.; Bjij, I.; Soliman, M.; Tadjer, A.; Villemin, D.; Bogdanov, J.; Cherqaoui, D. In Silico SAR Studies of HIV-1 Inhibitors. *Pharmaceuticals* **2018**, *11*, 69, https://doi.org/10.3390/ph11030069.
16. Chen, J.; Zhang, M.; Ma, Q.; Qin, D.; Zhang, L.; Lu, X. QSAR study of pyrazolo[1,5-a]pyrimidine derivative inhibitors of Chk1. *Chemom. Intell. Lab. Syst.* **2016**, *150*, 23–28, https://doi.org/10.1016/j.chemolab.2015.10.014.
17. Hdoufane, I.; Stoycheva, J.; Tadjer, A.; Villemin, D.; Najdoska-Bogdanov, M.; Bogdanov, J.; Cherqaoui, D. QSAR and molecular docking studies of indole-based analogs as HIV-1 attachment inhibitors. *J. Mol. Struct.* **2019**, *1193*, 429–443, https://doi.org/10.1016/j.molstruc.2019.05.056.
18. Mozafari, Z.; Arab Chamjangali, M.; Arashi, M. 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. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103998, https://doi.org/10.1016/j.chemolab.2020.103998.
19. Yu, X. Support vector machine-based model for toxicity of organic compounds against fish. *Regul. Toxicol. Pharmacol.* **2021**, *123*, 104942, https://doi.org/10.1016/j.yrtph.2021.104942.
20. Vahedi, N.; Mohammadhosseini, M.; Nekoei, M. QSAR Study of PARP Inhibitors by GA-MLR, GA-SVM and GA-ANN Approaches. *Curr. Anal. Chem.* **2020**, *16*, 1088–1105, https://doi.org/10.2174/1573411016999200518083359.
21. Autonne, L. Sur les groupes linéaires, réels et orthogonaux. *Bull. Soc. Math. Fr.* **1902**, *30*, 121–133.
22. C. Eckart; Young, G. A principal axis transformation for mathematics non-Hermitian matrices. **1939**, *45*, 118–121.
23. Klema, V.; Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Trans.*

Automat. Contr. **1980**, *25*, 164–176, https://doi.org/10.1109/TAC.1980.1102314.
24. Mandel, J. Use of the singular value decomposition in regression analysis. *Am. Stat.* **1982**, *36*, 15–24, https://doi.org/10.1080/00031305.1982.10482771.
25. Yeung, K.-S.; Qiu, Z.; Xue, Q.; Fang, H.; Yang, Z.; Zadjura, L.; D'Arienzo, C.J.; Eggers, B.J.; Riccardi, K.; Shi, P.-Y.; Gong, Y.-F.; Browning, M.R.; Gao, Q.; Hansel, S.; Kenneth, S.; Lin, P.-F.; Meanwell, N.A.; Kadow, J.F. Inhibitors of HIV-1 attachment. Part 7: Indole-7-carboxamides as potent and orally bioavailable antiviral agents. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 198–202, https://doi.org/10.1016/j.bmcl.2012.10.115.
26. Yeung, K.-S.; Qiu, Z.; Yin, Z.; Trehan, A.; Fang, H.; Pearce, B.; Yang, Z.; Zadjura, L.; D'Arienzo, C.J.; Riccardi, K.; Shi, P.-Y.; Spicer, T.P.; Gong, Y.-F.; Browning, M.R.; Hansel, S.; Santone, K.; Barker, J.; Coulter, T.; Lin, P.-F.; Meanwell, N.A.; Kadow, J.F. Inhibitors of HIV-1 attachment. Part 8: The effect of C7-heteroaryl substitution on the potency, and in vitro and in vivo profiles of indole-based inhibitors. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 203–208, https://doi.org/10.1016/j.bmcl.2012.10.117.
27. Wang, T.; Yang, Z.; Zhang, Z.; Gong, Y.-F.; Riccardi, K.A.; Lin, P.-F.; Parker, D.D.; Rahematpura, S.; Mathew, M.; Zheng, M.; Meanwell, N.A.; Kadow, J.F.; Bender, J.A. Inhibitors of HIV-1 attachment. Part 10. The discovery and structure–activity relationships of 4-azaindole cores. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 213–217, https://doi.org/10.1016/j.bmcl.2012.10.120.
28. Bender, J.A.; Yang, Z.; Eggers, B.; Gong, Y.F.; Lin, P.F.; Parker, D.D.; Rahematpura, S.; Zheng, M.; Meanwell, N.A.; Kadow, J.F. Inhibitors of HIV-1 attachment. Part 11: The discovery and structure-activity relationships associated with 4,6-diazaindole cores. *Bioorganic Med. Chem. Lett.* **2013**, *23*, 218–222, https://doi.org/10.1016/j.bmcl.2012.10.118.
29. Yang, Z.; Zadjura, L.M.; Marino, A.M.; D'Arienzo, C.J.; Malinowski, J.; Gesenberg, C.; Lin, P.; Colonno, R.J.; Wang, T.; Kadow, J.F.; Meanwell, N.A.; Hansel, S.B. Utilization of in vitro Caco-2 permeability and liver microsomal half-life screens in discovering BMS-488043, a novel HIV-1 attachment inhibitor with improved pharmacokinetic properties. *J. Pharm. Sci.* **2010**, *99*, 2135–2152, https://doi.org/10.1002/jps.21948.
30. Liu, T.; Huang, B.; Zhan, P.; De Clercq, E.; Liu, X. Discovery of small molecular inhibitors targeting HIV-1 gp120–CD4 interaction drived from BMS-378806. *Eur. J. Med. Chem.* **2014**, *86*, 481–490, https://doi.org/10.1016/j.ejmech.2014.09.012.
31. Zhang, X.; Zhang, N.; Chen, G.; Turpoff, A.; Ren, H.; Takasugi, J.; Morrill, C.; Zhu, J.; Li, C.; Lennox, W.; Paget, S.; Liu, Y.; Almstead, N.; Njoroge, F.G.; Gu, Z.; Komatsu, T.; Clausen, V.; Espiritu, C.; Graci, J.; Colacino, J.; Lahser, F.; Risher, N.; Weetall, M.; Nomeir, A.; Karp, G.M. Discovery of novel HCV inhibitors: Synthesis and biological activity of 6-(indol-2-yl)pyridine-3-sulfonamides targeting hepatitis C virus NS4B. *Bioorganic Med. Chem. Lett.* **2013**, *23*, 3947–3953, https://doi.org/10.1016/j.bmcl.2013.04.049.
32. Zhang, N.; Zhang, X.; Zhu, J.; Turpoff, A.; Chen, G.; Morrill, C.; Huang, S.; Lennox, W.; Kakarla, R.; Liu, R.; Li, C.; Ren, H.; Almstead, N.; Venkatraman, S.; Njoroge, F.G.; Gu, Z.; Clausen, V.; Graci, J.; Jung, S.P.; Zheng, Y.; Colacino, J.M.; Lahser, F.; Sheedy, J.; Mollin, A.; Weetall, M.; Nomeir, A.; Karp, G.M. Structure–Activity Relationship (SAR) Optimization of 6-(Indol-2-yl)pyridine-3-sulfonamides: Identification of Potent, Selective, and Orally Bioavailable Small Molecules Targeting Hepatitis C (HCV) NS4B. *J. Med. Chem.* **2014**, *57*, 2121–2135, https://doi.org/10.1021/jm401621g.
33. Chen, G.; Ren, H.; Zhang, N.; Lennox, W.; Turpoff, A.; Paget, S.; Li, C.; Almstead, N.; Njoroge, F.G.; Gu, Z.; Graci, J.; Jung, S.P.; Colacino, J.; Lahser, F.; Zhao, X.; Weetall, M.; Nomeir, A.; Karp, G.M. 6-(Azaindol-2-yl)pyridine-3-sulfonamides as potent and selective inhibitors targeting hepatitis C virus NS4B. *Bioorg Med Chem Lett* **2015**, *25*, 781–786, https://doi.org/10.1016/j.bmcl.2014.12.093.
34. Kennard, R.W.; Stone, L.A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148, https://doi.org/10.1080/00401706.1969.10490666.
35. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A. Gaussian 09, Wallingford CT, Available online: http://gaussian.com/.
36. Talete srl Dragon (Software for Molecular Descriptor Calculation, version 7.0) Available online: https://chm.kode-solutions.net/products_dragon.php.
37. Douali, L.; Villemin, D.; Zyad, A.; Cherqaoui, D. Artificial neural networks: Nonlinear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors. *Mol. Divers.* **2004**, *8*, 1–8, https://doi.org/10.1023/B:MODI.0000006753.11500.37.
38. Roy, K.; Pratim Roy, P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques.

Eur. J. Med. Chem. **2009**, *44*, 2913–2922, https://doi.org/10.1016/j.ejmech.2008.12.004.
39. Consonni, V.; Todeschini, R. Structure –Activity Relationships by Autocorrelation Descriptors and Genetic Algorithms. In *Chemoinformatics and Advanced Machine Learning Perspectives*; Lodhi, H., Yamanishi, Y., Eds.; IGI Global, 2011; 60–94.
40. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2009.
41. Consonni, V.; Todeschini, R. New Spectral Indices for Molecule Description. *Match* **2008**, *60*, 3–14.
42. Moran, P.A.P. Notes on Continuous Stochastic Phenomena. *Biometrika* **1950**, *37*, 17, https://doi.org/10.2307/2332142.
43. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria. Available online: http://www.r-project.org/.
44. Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504, https://doi.org/10.2174/138161207782794257.
45. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276, https://doi.org/10.1016/S1093-3263(01)00123-1.
46. Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77, https://doi.org/10.1002/qsar.200390007.
47. Lin, L.I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255, https://doi.org/10.2307/2532051.
48. Lin, L.I.-K. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics* **1992**, *48*, 599, https://doi.org/10.2307/2532314.
49. Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320–2335, https://doi.org/10.1021/ci200211n.
50. Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044–2058, https://doi.org/10.1021/ci300084j.
51. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791–4810, https://doi.org/10.3390/molecules17054791.
52. Chen, J.; Zhang, L.; Guo, H.; Wang, S.; Wang, L.; Ma, L.; Lu, X. Activity prediction of hepatitis C virus NS5B polymerase inhibitors of pyridazinone derivatives. *Chemom. Intell. Lab. Syst.* **2014**, *134*, 100–109, https://doi.org/10.1016/j.chemolab.2014.03.015.
53. Pourbasheer, E.; Aalizadeh, R.; Ganjali, M.R. QSAR study of CK2 inhibitors by GA-MLR and GA-SVM methods. *Arab. J. Chem.* **2015**, https://doi.org/10.1016/j.arabjc.2014.12.021.
54. Quang, N.M.; Mau, T.X.; Ai Nhung, N.T.; Minh An, T.N.; Van Tat, P. Novel QSPR modeling of stability constants of metal-thiosemicarbazone complexes by hybrid multivariate technique: GA-MLR, GA-SVR and GA-ANN. *J. Mol. Struct.* **2019**, *1195*, 95–109, https://doi.org/10.1016/J.MOLSTRUC.2019.05.050.
55. Van Tat, P.; Nhung, N.T.A. Insight prediction of receptor binding activity of a set of benzamide derivatives using hybrid QSAR models: GA-MLR and GA-SVR. *Vietnam J. Chem.* **2020**, *58*, 191–200, https://doi.org/10.1002/vjch.201900152.
56. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **2010**, *29*, 476–488, https://doi.org/10.1002/minf.201000061.