

Application of MIA-QSAR in Designing New Protein P38 MAP Kinase Compounds Using a Genetic Algorithm

Mitra Mirshafiei¹, Ali Niazi^{2*}, Atisa Yazdanipour²

¹ Department of Chemistry, Arak Branch, Islamic Azad University, P. O. B. 567-38135, 38 Arak, Iran

² Department of Chemistry, Central Tehran Branch, Islamic Azad University, P. O. B. 1955847881, 13 Tehran, Iran

* Corresponding author, e-mail: ali.niazi@iauctb.ac.ir

Received: 19 July 2022, Accepted: 31 August 2022, Published online: 11 January 2023

Abstract

Multivariate image analysis quantitative structure-activity relationship (MIA-QSAR) study aims to obtain information from a descriptor set, which are image pixels of two-dimensional molecule structures. In the QSAR study of protein P38 mitogen-activated protein (MAP) kinase compounds, the genetic algorithm application for pixel selection and image processing is investigated. There is a quantitative relationship between the structure and the pIC_{50} based on the information obtained. (The pIC_{50} is the negative logarithm of the half-maximal inhibitory concentration (IC_{50}), so $\text{pIC}_{50} = -\log \text{IC}_{50}$.) Protein P38 MAP kinase inhibitors are used in the treatment of malignant tumors. The development of a model to predict the pIC_{50} of these compounds was performed in this study. To accomplish this, the molecules were first plotted and fixed in the same coordinates in ChemSketch. Then, the images were processed in the MATLAB program. Partial least squares (PLS) model, orthogonal signal correction partial least squares (OSC-PLS) model, and genetic algorithm partial least squares (GA-PLS) model methods are used to generate quantitative models, and pIC_{50} prediction is performed. The GA-PLS model has the highest predictive power for a series of statistical parameters such as root mean square error of prediction (RMSEP) and relative standard errors of prediction (RSEP). Finally, the molecular junction (docking) was done for predicted molecules in quantitative structure activity relationship (QSAR) with an appropriate receptor and acceptable results were obtained. These results are good and proper for the prediction of compounds with better properties.

Keywords

multivariate image analysis (MIA), genetic algorithms, partial least squares

1 Introduction

Mitogen-activated protein kinases (MAPKs) are serine/threonine kinases that have been studied in many areas due to the high levels of conservation of various eukaryotic cells and have been shown to play a key role in signal transduction from cell to nucleus as well as cell survival and death [1, 2]. MAPKs are divided into three subgroups based on amino acids found in the activation loop between the serine and threonine subunits [3]. Kinases are involved in almost every aspect of physiology. P38 mitogen-activated protein kinase (MAPK) is one of the important enzymes that inhibit the immune system to treat autoimmune diseases [4–6].

The study of the relationship between the properties of molecules and their structure is one of the most important fields of application for chemometric methods. These studies, known as the "Quantitative structure-activity relationship (QSAR)", investigate the relationship between activity and the various properties of molecules with their

structural characteristics [7–10]. QSAR is widely used in drug design processes to improve the therapeutic indices of compound designation. QSAR models are mathematical equations written based on the chemical composition of compounds and their biological activity. The first component in the definition of a quantitative structure-activity relationship (QSAR) model is the calculation of structural descriptors based on the composition's molecular structure. In general, various descriptors are used in QSAR modeling [11]. These descriptors are divided into different groups, including structural, geometric, spatial, quantum, chemical. Descriptors containing information are useful in ensuring that the model's predictive power is acceptable at a satisfactory level. A QSAR study is a powerful tool for researching and analyzing the structure and activity of chemical compounds, and it is widely used in drug chemistry to investigate drug inhibition [12–14]. Linear

methods such as principal component regression and partial least squares are used in QSAR studies to model and predict the activity of drug compounds, as are nonlinear methods such as artificial neural networks.

Recently, computer-aided drug discovery has received attention and has been extensively employed in medical chemistry. Computational techniques are used in computer-aided drug design to find, create, and study medicines and other physiologically active compounds. The quantitative structure-activity relationship is among such strategies. QSAR is one of the most significant Chemometrics applications, providing essential knowledge for the development of novel compounds that operate on a particular target and have desirable features. The QSAR approach is now widely used in pharmaceuticals, drug design, toxicology, geology, as well as remote sensing [15]. The QSAR approach provides a mathematical relationship between a compound's chemical structure and its physical, chemical, or biological characteristics. Then, after analyzing the response between receptor and ligand, a novel compound is designed. Compounds with comparable physicochemical qualities have similar physiological activities. The QSAR technique for drug discovery establishes a relation between the structural features of possible drug candidates and their ability to block a given biological function. When compared to other QSAR approaches, multivariate image Computational modeling of analysis (multivariate image analysis quantitative structure-activity relationship (MIA-QSAR)) offered a fast analysis result as accurate as of the most advanced methods available today, while also being inexpensive and simple to manage and predict any modeled response for a congeneric series of chemical structures without the need for 3D alignment or conformational analysis. In this method, 2D images of pixels suggest topo-chemical characteristics of chemicals, and a model between such descriptors and a y-block comprised of independent variables is built. Multivariate image analysis QSAR is a non-invasive analysis that saves time and money while processing a large amount of data. The MIA-QSAR technique aims to correlate numerous columns of individual variables to a single column dependent variable, y . Various coordinates of pixels in the molecular drawing depict structural changes in the MIA-QSAR technique, and these changes were utilized to show variation in bioactivity for a congeneric group of drug-like compounds [16, 17]. In the modeling process, the substitution pattern, as well as the congeneric series of compounds, may be used to predict the

bioactivities of comparable compounds. The MIA-QSAR approach include the following steps:

- drawing molecule structures, creating images and aligning them;
- image denoising followed by image unfolding to a two-way array as well as descriptor generation;
- regression modeling and feature selection.

Modeling is one of the most critical procedures addressed. Multiple linear regression (MLR) [18], partial least squares [19], and artificial neural networks are some of the approaches that may be applied. MLR has been frequently used in QSAR studies, despite its relatively low accuracy. Furthermore, MLR works well when the number of rows is higher than the number of columns. In most circumstances, artificial neural networks (ANN) demonstrates enough accuracy; nevertheless, there is a risk of overfitting the training data and, as a result, not being able to extrapolate appropriate data information.

Genetic algorithm (GA) employs several fitness criteria and genetic functions [20]. It is shown that preprocessing prior to partial least squares (PLS) regression eliminates unnecessary information and gives adequate input for PLS, hence improving model quality. Prior studies [21, 22] have explored orthogonal signal correction (OSC).

Freitas et al. [23] proposed a simple and comprehensible approach established on 2D image analysis. A library of (S)-N-[(1-ethyl-2-pyrrolidiny)methyl]-6-methoxybenzamide with an affinity for the dopamine D2 receptor subtype was used to select 40 calibration compounds and 18 test compounds. The pixels of the 2D structures were used to build descriptors for each molecule. For regression, they used bi-linear PLS (conventional), and for leave-one-out cross-validation, they used the nonlinear iterative partial least square (NIPALS) approach. A Q^2 value of 0.58 for the test chemical series was predicted and exhibited a similar estimation ability for additional data sets [23].

Furthermore, Freitas [24] used a QSAR technique based on multivariate image analysis (MIA) descriptors to a series of anti-human immunodeficiency virus-1 (anti-HIV-1) active 2-amino-6-arylsulfonylbenzotrioles and thio and sulfinyl analogs. Two models, as well as a collection of molecules, were constructed utilizing a range of sketching tools to assess the technique's ability in modeling. Both models had sufficient predictive performance, with Q^2 values of 0.712 and 0.624 for cross-validation and Q^2 values of 0.823 and 0.747 for external validation. To anticipate absorption patterns for prospective new

therapies for the listed drugs, the topological polar surface area (TPSA) and variables originating from the rule of five were applied [24].

Goodarzi and Freitas [25] employed the MIA-QSAR coupling method principal component analysis adaptive-network-based fuzzy inference systems (PCA-ANFIS) to estimate the anti-HIV efficacy of transcriptase inhibitors in high-activity, well-absorbed chemicals in 2010. The MIA-QSAR/PCA-ANFIS model was compared to the N-PLS MIA-QSAR/PLS model using the PLS and N-PLS regression models. The multilinear PLS models are called N-PLS models in general. N-PLS is an algorithm of the PLS family adapted to multimodal data (tensor variables). The outcomes demonstrated that the aforementioned approach worked superior to the other two regression procedures [25].

Cormanich et al. [26] utilized the MIA-QSAR model to relate the 2,5-diaminobenzophenone's 2D chemical structure to its biological activity. They used 74 calibration series and 18 test series out of a total of 92 potential combinations. The calibration series was subjected to cross-validation to identify the optimal number of hidden variables for the partial least squares (PLS) regression calibration model. The image analysis technique model outperforms 3D QSAR methods with an R^2 score of 0.91 and a Q^2 score of 0.56 [26].

Nunes and Freitas (2013) [27, 28] used MIA-QSAR to mimic the sweetness of disaccharide molecules. They chose 40 images of disaccharides with similar molecular structures for the calibration series, 30 compounds for the test series, and 10 compounds for the test series. To forecast log Reed–Solomon ($\log(RS)$), they used a chemical structure model. The calibration set's R^2 was 0.97, the test set's R^2 was 0.94, and the test set's root mean square error of cross-validation (RMSECV) was 0.86 [27]. The MIA-QSAR approach [28], which uses pixels of two-dimensional chemical structures as descriptors, was used to simulate inhibitors of chemokine receptors.

In 2015, Duarte et al. [29] used a series of quinolone derivatives (an antimalarial drug) and two methods of normal image analysis and color image analysis to model the activity of the drug. They also performed a multivariate linear regression study between the two methods. The statistical results of the color image analysis method performed by PLS are R^2 equal to 0.807, R_{CV}^2 equal to 0.664 and R_{pred}^2 equal to 0.969, indicating that when using the color image analysis method compared to the conventional method used better prediction [29].

In 2017, Akrami and Niazi [30] linked the two-dimensional chemical structure to activity and used the MIA-QSAR model to predict the inhibitory activity of dihydropyridine (DHP) derivatives. Of the 35 compounds studied, 24 compounds were selected as training series and 11 compounds as test series. The value of Q^2 was obtained by the wavelet transform - genetic algorithm - partial least squares (WT-GA-PLS) method for the test series of 0.92, which indicates the suitability of the selected model for prediction [30].

Benzamide herbicides consists of a class of photosynthetic system II (PSII) inhibitors used to control weeds. Pereira et al. [31] used MIA-QSAR for modeling. These powerful and predictable models help to estimate pIC_{50} of new chemical options available in agriculture. The pIC_{50} is the negative logarithm of the half-maximal inhibitory concentration (IC_{50}), so $pIC_{50} = -\log IC_{50}$. Chemical properties affecting the activities of herbicides were analyzed using MIA contour maps [31].

This study aims to create an MIA-QSAR model for MAP KIMASE inhibitors, estimate their pIC_{50} using the GA-PLS model, and ultimately develop novel compounds based on this model. The pIC_{50} is one of the metrics used to describe a measure of a substance's potency in inhibiting a given biological or biochemical activity, demonstrating the lowest molar concentration required to inhibit 50% of the enzyme.

The QSAR model in this paper was built using 1409 descriptors. A model for defining kinase inhibitor activity was developed using a combination of GA, OSC, and PLS. The main objective of this study is to develop models for predicting the activity of pharmaceutical derivatives of P38 MAP kinase using the partial least squares (PLS) model, orthogonal signal correction partial least squares (OSC-PLS) model, and genetic algorithm partial least squares (GA-PLS) model methods. In addition, the predictive power of models has been investigated using the standard methods in studies using intersectional and external evaluation models.

2 Materials and methods

2.1 Instrumentation

ChemSketch [32] was used to draw the structure of the molecules in this work in 2022. Ratul Bhowmik et al. [33] studied the design of new florfenicol analogues for bacterial acetyltransferase which have better medicinal properties and fewer side effects. Using the base florfenicol framework, they applied primary and secondary modifications by ChemSketch software. All of these modified molecules were screened according to drug similarity rules.

Accordingly, absorption, distribution, metabolism, excretion and toxicity (ADMET) analysis was performed on the modified molecules. Based on findings of this study, the molecules are designed as main molecules in inhibition of bacterial chloramphenicol acetyltransferase enzymes, which are used to treat vibriosis [33].

Shivaleela et al. [34] studied the design of Thalidomide-based small molecule inhibitors for tumor necrosis factor alpha (TNF- α), which is mainly secreted by monocytes and macrophages, as an important therapeutic target for several diseases. They [34] also designed several thalidomide analogues by ChemSketch. This program is a comprehensive package for drawing, editing, and transforming two-dimensional structures into three-dimensional structures. The application of ChemSketch in the literature shows the good performance of the program. Paint was used to center the images and convert the structure of molecules to images. MATLAB was used to perform statistical calculations, create descriptors and models [35]. MATLAB is a type of computer program that performs mathematical computations. Arrays and matrices serve as the foundation of the data in this program. The same feature enables the user to solve numerical calculation and transform it into matrices and arrays.

2.2 Data set

The first step in the QSAR is to select the appropriate data set. A valid model can only be created with a proper data set. All compounds' experimental values of the studied quantity should be measured under the same conditions. The chemical compounds investigated in this work were obtained from the literature [36]. These data include 46 different protein P38 MAP kinase inhibitor combinations, with biological activity expressed as pIC_{50} . In this study, first, molecules are drawn in ChemSketch (Table 1). Then, the Paint program is used to select the pixel points common to all of the molecules and fix all of the molecules in points. The calibration set and the test set were used to divide the data from the Kennard-Stone algorithm [37], and 37 combinations were chosen as the calibration set and 9 combinations as the test set based on the algorithm. The selected calibration samples completely show the space of variables in the Kennard-Stone algorithm, and test samples are placed in space.

2.3 Multivariate image analysis

Analog image data must be converted to digital data to establish a relationship between biological activity and molecule images in QSAR modeling. In this study,

molecules are first drawn in ChemSketch, then saved as BMP files and opened in Paint. The images are adjusted so that the common point of all molecules is fixed at 80×80 coordinates. The two-dimensional images of the chemical structures of the 46 compounds were then placed in the unfolding phase (as seen in Fig. 1).

For each composition, 30400 image descriptors were calculated in MATLAB by converting pixels to binary numbers. Pixels were obtained for each of the 46 combinations investigated, resulting in a matrix with dimensions of 46×30400 . Preliminary processes must be performed to communicate quantitatively between the input data and the activity vector.

3 Results and discussion

3.1 Multivariate image analysis descriptors

The descriptors in multivariate image processing are pixels of the structural form of the molecules in question, with two or three dimensions. These pixels correspond to our dependent variables, which are used in the QSAR model's design. Statistical methods are used to select descriptors. The majority of variable selection methods are based on reducing the predicted error. As a result, it is critical to calculate the validity of a model derived from a data set that was ineffective in the model's construction and its predicted error. There are many methods for modeling, including multivariate calibration, multiple linear regression, principal component regression, partial least squares regression, and exploratory random search methods (genetic algorithm) [38].

3.2 Principal component analysis of the data set

The principal component analyses [39, 40] (PCA) is a strong method for extracting valuable data and lowering the quantity of data from PCA applications such as assessing the link between data, the potential of more remote data, the presence of categories between data, and scoring and downloading. PCA aims to reduce the number of variables by drawing a line along the axis with the largest data diversity. The first principal component (PC1) is the line that includes the greatest information. The first principal component determines the data's greatest variation. The second principal component (PC2), which is orthogonal to PC1, describes the largest amount of variation remaining in the data. As a result, the amount of variation of the data that is vertically aligned with all preceding PCs is defined by each new PC. The Score matrix on PC1 is produced by scoring the data points from the primary space X on PC1. These variations can be viewed on a second axis that is perpendicular to the first and on which the data is shown, termed Score on PC2.

Table 1 Structures and inhibitory activity of P38 MAP kinase derivatives (Structures were made by the authors with ChemSketch [32])

No.	Structure	pIC ₅₀	No.	Structure	pIC ₅₀
1		6.49	7		6.14
2		6.72	8		>5
3		7.03	9		5.89
4		6.06	10		6.80
5		6.05	11		>5
6		6.31	12		5.89
			13		6.14
			14		8.46

Table 1 Structures and inhibitory activity of P38 MAP kinase derivatives (Structures were made by the authors with ChemSketch [32]) (continued)

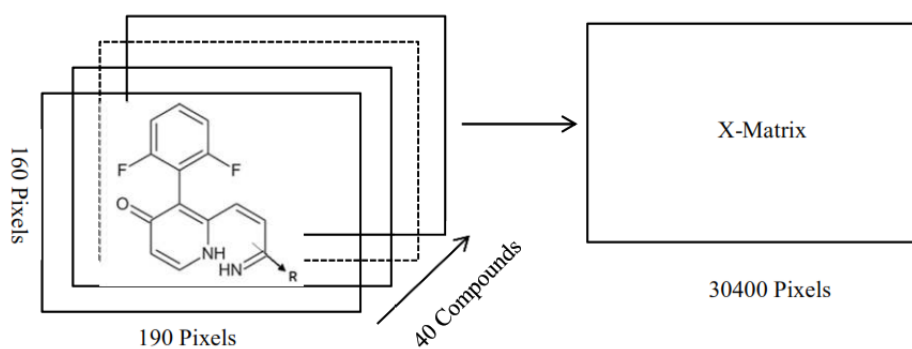
No.	Structure	pIC ₅₀	No.	Structure	pIC ₅₀
15		7.70	22		6.38
16		6.32	23		6.62
17		5.84	24		7.15
18		7.36	25		>5
19		7.04	26		>5
20		7.09	27		7.32
21		5.62	28		7.1

Table 1 Structures and inhibitory activity of P38 MAP kinase derivatives (Structures were made by the authors with ChemSketch [32]) (continued)

No.	Structure	pIC ₅₀	No.	Structure	pIC ₅₀
29		6.59	36		7.52
30		>5	37		6.89
31		>5	38		6.55
32		7.49	39		7.02
33		7.49	40		7.05
34		7.96	41		6.70
35		7.85	42		7.17

Table 1 Structures and inhibitory activity of P38 MAP kinase derivatives (Structures were made by the authors with ChemSketch [32]) (continued)

No.	Structure	pIC ₅₀	No.	Structure	pIC ₅₀
43		6.6	45		6.85
44		6.67	46		6.59

**Fig. 1** 2D-images and unfolding step of the 46 chemical structures to give the X-matrix

The verticality of variables indicates that the least amount of collinearity exists among them.

Collinearity classification was used to classify these PCs based on their strongest link to the activity of P38MAP KINASE derivatives. Then, starting with the PCs with the highest order, we insert them into the principle component regression (PCR) model one by one until the PC enters the model. Then, we put the PCs with the highest order into the PCR model and keep doing so until there is no change in the improvement of the statistical parameter under study when the PC is put into the model (R^2 was considered equal to 0.98). More information on variations in the activity of MAP KINASE P38 derivatives may be found in high collinearity PCs. The Kernard Stone algorithm was used to partition the data into calibration and test series, with 37 and 9 combinations were chosen as calibration and test series, respectively. Examining two different computational approaches (comparative molecular field analysis (CoMFA) and comparative molecular similarity indices

analysis (CoMSIA)) to identify the necessary structural conditions in a three-dimensional chemical space to regulate the inhibitory activity of dipeptidyl peptidase IV (DPP-IV) derivatives of trifluorophenyl. Sharma et al. [41] used data set models containing 87 compounds and experimental set consisting of 21 compounds of triforphenyl and obtained suitable results for the design of new compounds with DPP-IV inhibitory activity.

Bhattacharya et al. [42] used an integrated computational approach in relation to sodium/glucose cotransporter 2 (SGLT2) inhibitors to develop new adiabatic compounds. Using various drug design tools, they performed computational analysis to obtain the best possible molecules from a dataset of 90 C-aryl glucoside analogues. Atom-based analysis 3D-QSAR (CoMFA, CoMSIA) was performed using 63 molecules as a training set and another 27 molecules were used as experimental sets to determine the role of different fields and atoms in developing the model [40].

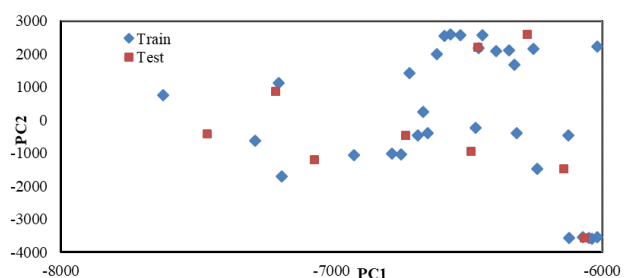
In development of 3D-QSAR models, El-Mernissi et al. [43] used 2-oxoquinoline Arylaminothiazole derivatives; to generate 3D QSAR model, they used a training set consisting of 20 compounds and an experimental set containing 5 compounds for validation. The developed QSAR models were effective in designing new compounds and predicting their pIC_{50} .

The new variables are defined solely in the space of the original variables when using the PCA technique to minimize the amount of the data. Data classification is one feature of score diagrams, which display the position of descriptors in the new coordinates.

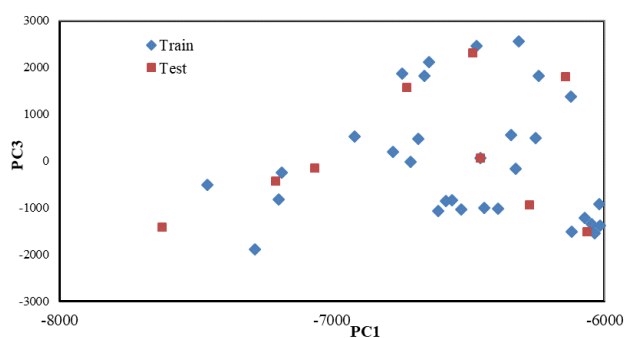
PCA was performed on the matrix of independent variables using the partial least squares regression approach, with the result having the best collinearity with the score of the dependent variable matrix. When the collinearity of a variable with the activity was less than 0.1, PCA was conducted on two-dimensional image descriptors in this study. The three primary components account for 90% of the change, according to PCA data (PC1 = 46.24%, PC2 = 28.5% and PC3 = 16.38%). As shown in Fig. 2, the compounds are not classified in any particular way.

3.3 PLS and OSC-PLS modeling

The partial least squares regression technique outperforms other multivariate calibration methods. This method has a wide range of applications in QSAR studies. The nonlinear



(a)



(b)

Fig. 2 Principal components analysis of the 2D image descriptors for the data set; (a) PC1 versus PC2; (b) PC1 versus PC3

iterative partial least square (NIPALS) algorithm is designed for the direct calculation of vectors and eigenvalues in this method. The main objective of linear regression is to determine the relationship between independent and dependent variables. PLS is used in MIA-QSAR to establish relationships between activity matrices as dependent variables and matrix pixels as independent variables. The leave-one-out method is used to validate the model, and the root mean square error is calculated (Tables 2 and 3).

Orthogonal signal correction (OSC) is a principal component analysis-based preprocessing method for removing information that does not depend on independent variables. Wold et al. [44] first proposed this method in 1998. To improve calibration efficiency, the OSC is an appropriate preprocessing for PLS calibration. As a result, different scientists offer various algorithms for reducing model complexity and eliminating orthogonal signals. Using OSC preprocessing eliminates the system's perpendicular noise.

3.4 GA-PLS modeling

The calibration model variables are selected with the least error in each generation and have the best characteristics in the genetic algorithm (GA) method [44]. These selective variables improve the model's ability to predict. One of the issues in this study is the selection of a set of pixel descriptors. The genetic algorithm was used as a variable selection technique to solve this problem. We choose the variables in the category to improve the efficiency of the genetic algorithm. The appropriate variables are chosen in each category, and then another genetic algorithm is run among the variables chosen to select the best variables. Although this method is a powerful technique for selecting variables [45–50], the main issue is the randomness of the variable selection process. As a result, if this algorithm is used only once to select a descriptor (pixel), there is a risk that the first-generation variables are not good and are trapped at the local minimum due to the randomness of the method. So, to solve the problem, this algorithm is run several times to eliminate the defect. As a result, GA was performed ten times, and the frequency of variable selection in the executions was calculated, and finally, 98 variables with higher frequencies advanced to the next stage. The genetic algorithm is optimized by changing and selecting the fitness function value, and the selected pixel descriptors are then used to run the partial least squares algorithm (PLS). As shown in Fig. 3, the number of latent variables in the GA-PLS model is reduced to three, and the number three is chosen as the optimal latent variable for the training set in the GA-PLS model. According to the

Table 2 Observation and calculation values of pIC₅₀ using PLS, OSC-PLS and GA-PLS models

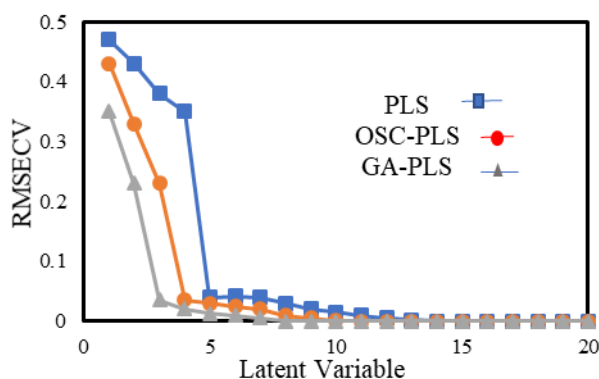
Number of compounds	Observation pIC ₅₀	PLS		OSC-PLS		GA-PLS	
		predicted	error	predicted	error	predicted	error
1	6.49	6.39	-0.09	6.58	0.09	6.73	0.24
5	6.05	5.93	-0.11	5.89	-0.15	5.81	-0.23
7	6.14	6.14	0	6.44	0.30	6.17	0.04
17	5.84	6.65	-0.38	7.42	0.03	7.10	0.06
21	5.62	7.24	0.62	6.90	0.28	5.41	-1.20
27	7.32	7.76	0.27	8.20	0.71	7.65	0.16
28	7.1	7.87	-0.08	8.40	0.44	7.86	-0.09
33	7.49	6.48	-0.53	7.19	0.17	7.08	0.06
36	7.52	6.24	-0.92	7.82	0.65	7.19	0.03
LVs*		5		4		3	

* Latent variables (LVs)

Table 3 Validation of models

Model	RMSEc*	RMSEp	R ²	Q ²	R ² _{pred}	r ² _m	cR ² _p
PLS	0.31	0.34	0.91	0.83	0.78	0.68	0.78
OSC-PLS	0.28	0.30	0.94	0.84	0.81	0.74	0.79
GA-PLS	0.19	0.20	0.98	0.96	0.90	0.79	0.82

* Root mean square error of calibration (RMSEc)

**Fig. 3** The RMSECV versus number of latent variables

findings, the genetic algorithm is an appropriate method for selecting variables in image analysis. The results of PLS and OSC-PLS and GA-PLS are shown in Fig. 3.

3.5 Model validation and prediction of pIC₅₀

Here, the predictive ability of the PLS, OSC-PLS, GA-PLS approaches was evaluated. Table 2 shows the structures of nine compounds whose inhibitory action has been predicted based on their predicted designs. The capacity to forecast must also be confirmed in QSAR investigations as an important step. The appropriateness of the models was assessed using various statistical metrics.

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{pred}} - y_{i,\text{obs}})^2}{n}} \quad (1)$$

$$\text{RSEP}(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^n (y_{i,\text{pred}} - y_{i,\text{obs}})^2}{\sum (y_{i,\text{obs}})^2}} \quad (2)$$

In Eqs. (1) and (2), $y_{i,\text{pred}}$ is the anticipated value of pIC₅₀ using various models, $y_{i,\text{obs}}$ is the actual value of pIC₅₀, and n is the number of evaluation or estimation sets. The root mean square error of prediction (RMSEP)/relative standard errors of prediction (RSEP) values for the prediction of pIC₅₀ for P38 MAP kinases are presented in. Table 3 presents additional statistical metrics (R^2 , Q^2) used to assess the models appropriateness to forecast the activity of the chemicals under study.

The R^2 parameter measures the models quality, whereas the Q^2 parameter assesses its external predictive potential. The linear regression of pIC₅₀ laboratory findings has a correlation coefficient based on the models anticipated education and screening set value. However, the minimum partial squares with error (-1.20, 0.24) were found in the genetic algorithm for the inhibitory effect of kinase derivatives. Other statistical variables, such as the cross-validation coefficient (Q^2 and R^2), were used to suppress the activity of kinase derivatives. The following is a list of the variables in Table 3.

Such variables have favorable statistical features. Fig. 4 shows the anticipated inhibitory activity for each model compared to the actual values.

Table 2 displays the anticipated pIC_{50} values and the relative error estimation using the PLS, OSC-PLS, and GA-PLS techniques. Fig. 4 indicate the model-predicted pIC_{50} values concerning the empirical data. The relationship between the actual result and the anticipated pIC_{50} by model GA-PLS is satisfactory, with R^2 equaling 0.97. The data in Table 3 indicates that the GA-PLS model provides a low relative error percentage and high statistical quality with appropriate statistical quality. In contrast, the other two models have more ideal hidden variables.

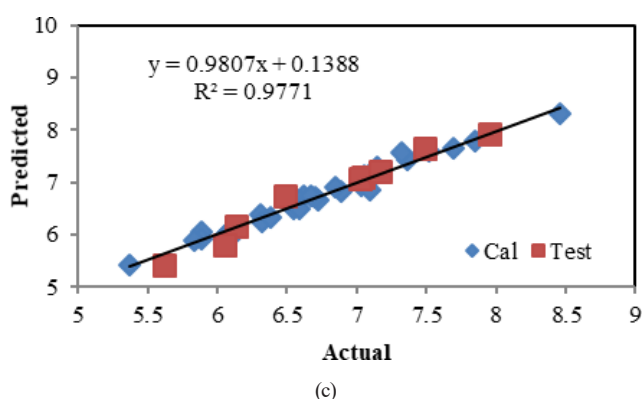
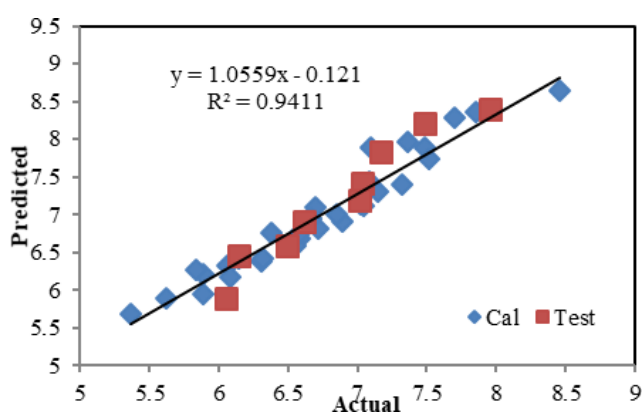
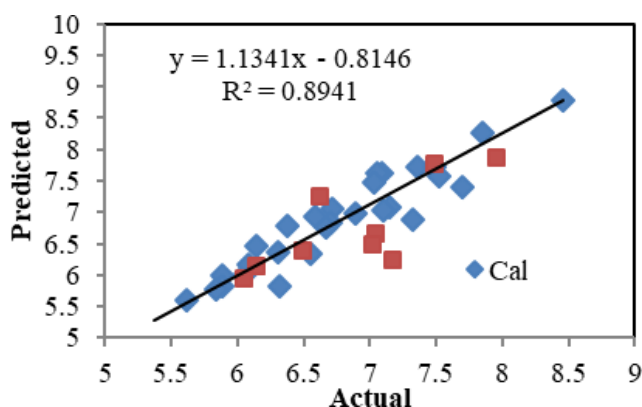


Fig. 4 Plots of predicted versus actual pIC_{50} with (a) PLS, (b) OSC-PLS and (c) GA-PLS

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,pred} - y_{i,obs})^2}{\sum_{i=1}^n (y_{i,obs} - \bar{y})^2} \quad (3)$$

$$Q^2 = 1 - \frac{\sum_{i=1}^n (y_{i,pred} - y_{loo})^2}{\sum_{i=1}^n (y_{i,obs} - \bar{y})^2} \quad (4)$$

3.5.1 Y-randomization test

This is a widely used technique to ensure the robustness of a QSAR model. In this test, the dependent-variable vector, Y-vector, is randomly shuffled and a new QSAR model is developed using the original independent-variable matrix. The process is repeated three times the average of the three measurements showed low R^2 values 0.214, 0.272 and 0.296 and Q^2 values 0.192, 0.202 and 0.261 for the PLS, OSC-PLS and GA-PLS, respectively. If all QSAR models obtained in the Y-randomization test have relatively high R^2 and Q^2 , it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modeling method [51, 52].

3.5.2 Validation by using r_m^2 criteria

We used Q^2 criterion to show the results of internal validity and predictable criterion of R^2 ($Q_{ext(F1)}^2$) to show the external validity. The calculated measure of R_{pred}^2 by using external validation, is used as a parameter for choosing of QSAR model with statistical sense. Significant part of the parameter is higher than the 0.5 threshold, however likely, it doesn't necessarily show the propinquity of predictable activity quantities with observed data. Probably this can be explained by the fact that the denominator term for calculating R_{pred}^2 equation is $Y_{test} - \bar{Y}_{training}$. This indicates as the difference between observed activity of test set combination and averaged quantity of the said set increases, R_{pred}^2 increases, too. If the difference is reasonably significant, regardless of predicted quantity of test set combination, the R_{pred}^2 amount increases. Therefore, it's likely that there is a significant difference between predicted quantity of activity and observed quantity of test set combinations; though it's possible that they are able to maintain their general correlation [53]. In order to devaluate the error and better demonstration of predicted activity of the observed test set, corrected amounts of r^2 (r_m^2) and the threshold of 0.5 have been calculated (as in Eq. (5)):

$$r_m^2 = r^2 \left(1 - \sqrt{(r^2 - r_0^2)} \right). \quad (5)$$

The evaluation of the QSAR model was also carried out by the mentioned statistical parameter and the results are shown in Table 3.

3.5.3 Randomization

To examine ability of advanced QSAR model, they have been validated by using randomization method. Y-randomization method is used by hashing the data in Y-column and descriptive matrix (X-Matrix) is remained intact. Each time, the models are developed with hashed data and correlation coefficient is calculated. If squared correlation coefficient of original QSAR model (R^2) is higher than averaged squared correlation coefficient of random models (R_r^2), then it is probable that advanced model is considered as a sufficient one. In this paper, a randomization model with 98% confidence is used. In randomization of model, Y-hashing examines specificity of advanced QSAR model by the descriptive element in the model. However, there is no strategy provided for showing the right difference between R^2 and R_r^2 to have a valid statistical model. So, to determine the difference between R^2 and R_r^2 , the criteria showing advanced QSAR model validity, we used another parameter, namely R_p^2 . This parameter eliminates R^2 because of the insignificant difference between R^2 and R_r^2 . The threshold for R_p^2 is 0.5 and, if QSAR model surpasses this threshold, it may result in considering the model as a sufficient one and this cannot happen by chance [54]. To show the R_p^2 quantity, so far, we have used Eq. (6):

$$R_p^2 = R^2 \sqrt{(R^2 - R_r^2)}. \quad (6)$$

Nonetheless, in ideal situation, average quantity of R^2 for random models must be zero (0); in fact, R_p^2 must be zero (0). As a result, in this position, R_p^2 must be equivalent to R^2 in advanced QSAR model. Therefore, the corrected equation of R_p^2 (${}^cR_p^2$) suggested by Todeschini is formulated (Eq. (7)):

$${}^cR_p^2 = R \sqrt{(R^2 - R_r^2)}. \quad (7)$$

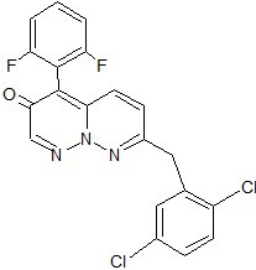
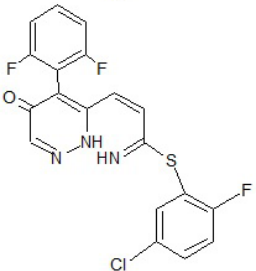
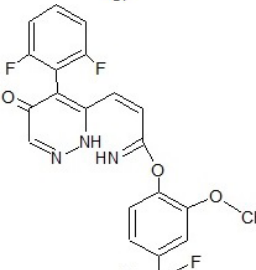
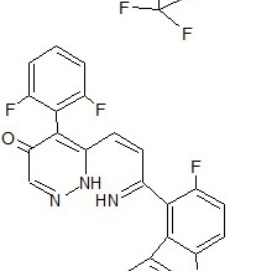
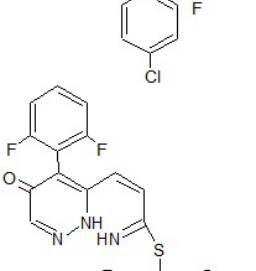
The QSAR model was evaluated by the statistical parameter, and the result are presented in Table 3.

3.6 Molecular design

The role of computation in molecular design has grown steadily since the late 1960s [55, 56]. In the early days emphasis was on statistical and computational approaches aimed at quantifying the relationship of chemical structure to biological properties. In addition, recent modeling by computational approaches has become a critical tool in the drug discovery process. As an application of proposed method,

we investigated GA-PLS model to predict the inhibitory activity of five new p38 MAP-KINASE compounds whose biological tests were not performed with them yet. Table 4 shows the chemical structure of five new compounds and their inhibitory activity calculated by this proposed method.

Table 4 Structural modification of P38 MAP kinase and predicted pIC_{50} by GA-PLS (Structures were made by the authors with ChemSketch [32])

Number of design	Chemical structure	Predicted pIC_{50} calculated by GA-PLS
1		7.16
2		7.23
3		7.89
4		6.91
5		5.88

3.7 Examination of compounds by Lipinski's rule

Lipinski and his co-workers in 2001, presented a guideline for the prediction of absorption of the orally active compounds through the definition of the rule of 5. The rule of 5, based on the calculation of the distribution of properties of several thousand drugs, predicts that low absorption or penetration happens in some cases when an intended molecule has the following properties [57, 58]. Extensive effort is progressing with the goal of discovery development and getting better of new drugs in the early stages of research and development processes for identification of drug-like properties in molecules. Although, there are different approaches for this problem, probably the most convenient and common approach is the very developed approach by Chris Lipinski and his co-workers in Pfizer Company that generally is known as Lipinski's rule or the rule of 5 (ROF). The rule of 5 is based on four features which include these items: molecular weight (MW), logarithm P ($\log P$), the number of hydrogen bond donors (HBD) which is equivalent to the number of OH and NH groups, and the number of hydrogen bond acceptors (HBA) which is equivalent to the number of oxygen and nitrogen atoms. This rule is true when the molecular weight of a molecule is greater than or equal to 500, its HDB number is greater than or equal to 5, its HBA number is greater than or equal to 10 and its $\log P$ is greater than or equal to 5 (the quantity $\log P$ is the logarithm of octanol/water partition coefficient, or the amount of water and oil solubility that is used to the prediction of the solubility rate) [59]. In this research, we investigated $M\log P$. Moriguchi's calculation always gives the right answer and this value is presented on the ADME Swiss server by Lipinski et al. [60]. ADME is an abbreviation in pharmacokinetics and pharmacology for "absorption, distribution, metabolism, and excretion". Sometimes, the potential or real toxicity of the compound is taken into account (ADME-Tox or ADMET). Since the values of the parameters for all these features are multiple of 5, the set of mentioned rules is known as the rule

of 5. Total values (ROF score) are variable between 0 to 4. The molecules with a ROF score larger than 1 are less considered for researches. As Lipinski and his colleagues have pointed out, these molecules do not have to be removed necessarily from investigations. Instead, they should have lower priority in the research and development processes. Finally, it should be noted that as you know, many drugs exceed the values of the rule of 5. But since the rule of 5 had been initially designed as a tool for the study of drug similarities, however, the application of this rule for this purpose has made it very practical and effective [61].

The polar surface area (PSA) of the molecule is another significant factor that has a direct role in the permeability of bioactive compounds. By definition, the polar surface area is the surface of the molecule that has oxygen, nitrogen, or hydrogen connected to these two. Based on studies on various structural banks, it has been specified that permeability of compounds increases with mass increase and also with reduced polar surface area. Based on the results of these studies, compounds with a polar surface area more than 140, do not show good permeability [62, 63]. Five suggested compounds in the present study follow Lipinski's rules. Therefore, they can be placed in the group of pharmaceutical compounds with proper absorption and penetration. Lipinski's parameters of these compounds are shown in Table 5. Lipinski's rules and physicochemical properties of compounds were predicted by the ADME Swiss server.

3.8 Molecular docking studies

In this research, in order to the investigation of anticancer properties of the suggested pharmaceutical compounds, the tendency of these compounds to the interaction with respective receptor was examined by molecular docking studies. According to Hadaji et al. [36], the studied compounds have an inhibitory function on P38 MAP kinase protein and by inhibition of this protein, they cause decreasing of kinase enzyme activity. Also, they are effective in the control of cancer progression and the growth of cancerous

Table 5 Lipinski's parameters of five suggested compounds

Compounds	Hydrogen bond donors (≤ 5)	Hydrogen bond acceptors (≤ 10)	Molecular mass < 500	$M\log P > 4.15$	High lipophilicity ($i \log P < 5$)	Lipinski
1	0	3	416.25	5.40	3.85	Yes; $M\log P > 4.15$
2	2	5	420.84	4.03	3.08	Yes; 0 violation
3	2	9	450.36	3.58	3.56	Yes; 0 violation
4	2	6	482.86	5.14	3.26	Yes; 1 violation: $M\log P > 4.15$
5	2	9	470.38	3.24	4.30	Yes; 1 violation: $M\log P > 4.15$

tumors. Expansion of artificial kinase inhibitors needs to use of the chemical syntheses for creating a small molecule that blocks kinase active site and stops its operation. Kinase inhibitors which are marked with tinib suffix in the bill of materials (BOM) can stop the kinase operation with several different mechanisms. The most prevalent mechanism is the blocking of an ATP connection point that prevents the binding of phosphate residues which is essential for phosphorylation. These kinase inhibitors utilize as anti-cancer agents for targeting tumor or vascular endothelial cells. This method is named Targeted Therapy. Because kinase inhibitors have a specific and well-known performance contrary to the conventional chemotherapy wherein, they make no difference relative to tumor tissues cells which are rapidly dividing. In view of the mentioned items above, this protein was considered as a receptor in our docking studies.

Diverse crystallographic structures of mitogen-activated protein kinase were investigated in the protein DataBank station and the desired structure was selected by pdb ID: 1OVE. Also, the structures of five proposed compounds were drawn by Marvin Beans software [64] and were considered as ligands. All molecular docking studies were performed by Schrodinger maestro software [65].

Investigation of molecular docking results exhibited that five proposed compounds have the acceptable ability in mitogen kinase enzyme inhibition and their inhibitory properties are different depending on the functional groups in the structure and the bonding energies are variable from -50.36 to -85.65 kcal mol⁻¹. Linked free energies determine the amount of tendency of the proposed compounds (ligands) to the interaction with the enzyme active site. It is clear that the more negative values are indicative of more tendencies and the formation of the more stable ligand-receptor complex. The obtained results from docking studies included docking scores and the binding free energies for different poses of each of the five compounds are displayed in Table 6. The best POSE with better and displayable interactions in 2D and 3D spaces have been shown in highlights of Table 6, and relevant 2D and 3D images of ligand-receptor complexes to them have been shown in Fig. 5.

3.8.1 Investigation of the results of molecular docking studies

Different obtained results of docking such as appropriate values of bonding energies and docking scores (Table 7), the suitable number of involved amino acids as well as the existence of various electrostatic interactions, hydrogen bonds, halogen bonds, and etc. between ligand and

Table 6 The obtained results from docking studies for different poses of five proposed compounds

Pose	Docking score (kcal mol ⁻¹)	MMGBSA* (kcal mol ⁻¹)
1	-10.898	-73.33
1	-10.095	-60.09
2	-13.495	-61.12
2	-12.815	-85.65
2	-10.627	-60.07
2	-10.146	-75.37
2	-10.040	-75.23
3	-10.093	-64.55
3	-10.058	-62.93
3	-10.032	-67.07
3	-9.875	-69.69
3	-9.730	-69.15
3	-9.289	-65.29
4	-9.471	-50.36
4	-9.113	-53.11
4	-8.891	-52.67
4	-8.600	-53.71
5	-12.676	-76.44
5	-10.588	-62.89
5	-10.187	-62.35
5	-9.650	-61.64
5	-9.645	-74.03

* Molecular mechanics generalized born surface area (MMGBSA)

receptor revealed that all suggested compounds make a good and convenient connection with mitogen kinase enzyme active site.

As shown in Table 7, the number of amino acids engaged in interaction for each of the proposed compounds has been very convenient and acceptable. Fig. 5 shows that all of these compounds are in the right position in the active site of the mitogen kinase enzyme. Based on the 2D figures of ligand-receptor complexes, amino acids GLY110, MET109, ASP168, LYS53 have the most role in the interaction between the proposed compound number 1 and kinase enzyme. Amino acids LYS53, MET109, GLY110, and TYR35 have a more important role in the case of the proposed compound number 2. Regarding compound number 3, more effective amino acids in interaction are LYS53 and TYR35. In compound number 4, amino acids LYS53, and ASP168 and in compound number 5, amino acids TYR35, LYS53, GLY110, and MET109 play an important role in the bond created between ligand and receptor.

Also the examination of ligand-receptor complexes of docking exposed that the oxygen of carbonyl group

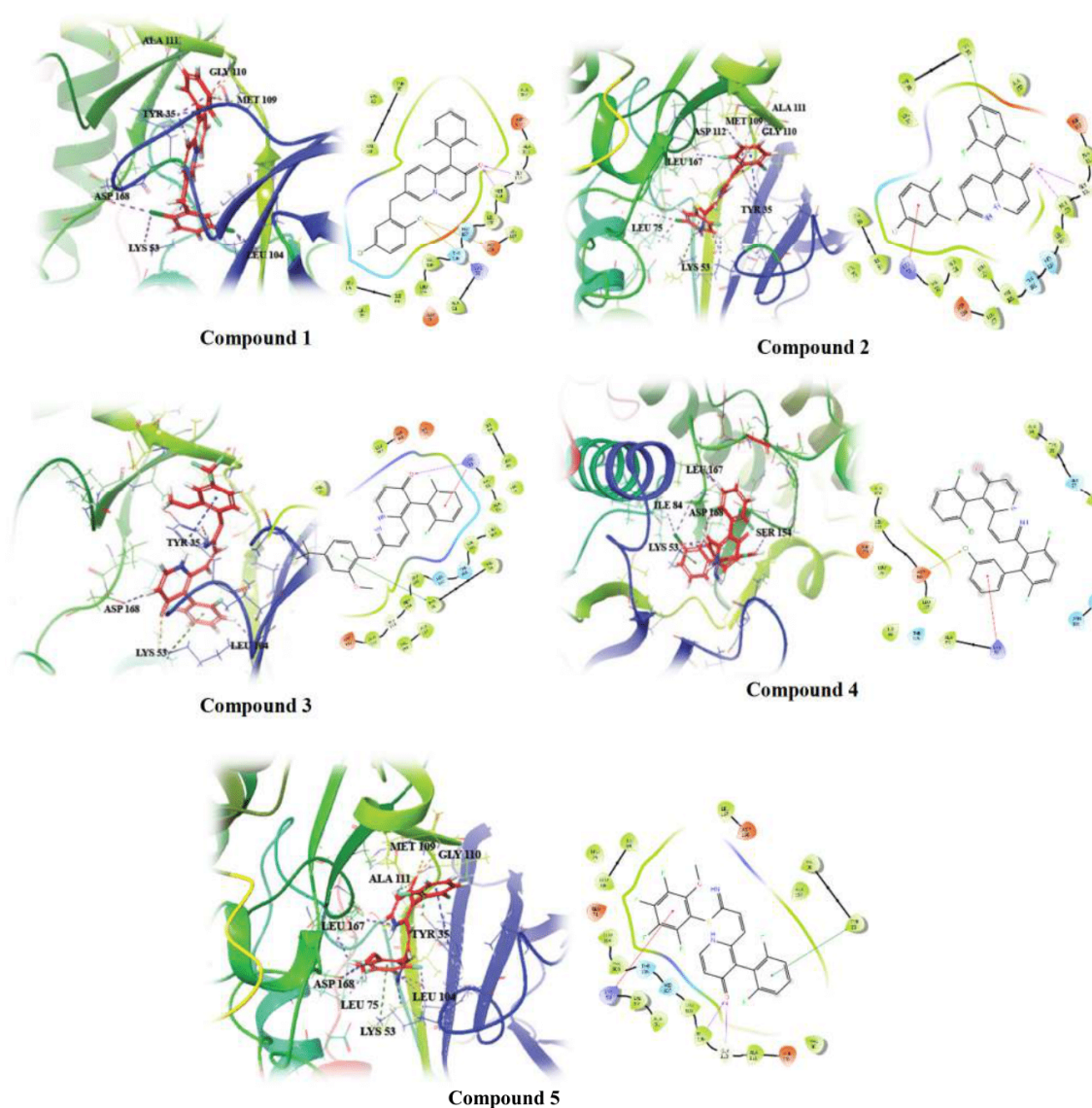


Fig. 5 2D and 3D images of ligand-receptor complexes related to highlighted poses in Table 6

Table 7 The interaction between amino acids and five proposed compounds

Compounds	Ligand-interacting amino acids	MMGBSA (kcal mol ⁻¹)
1	GLY110, MET109, ASP168, LYS53, TYR35, LEU104, ALA111	-73.33
2	LYS53, MET109, GLY110, LEU75, TYR35, ALA111, ASP112, LEU167	-85.65
3	LYS53, TYR35, ASP168, LEU104	-67.07
4	LYS53, ASP168, SER154, ILE84, LEU167	-52.67
5	TYR35, LYS53, GLY110, MET109, ALA111, LEU167, ASP168, LEU75, LEU104	-76.44

in the proposed compound number 1 establishes hydrogen bond (H-Bond) with amino acids MET109 and GLY110. Likewise, the chloride functional group in this compound participates in the halogen bonds formation with amino acids ASP168 and LYS53. The oxygen of carbonyl group in compound number 2, like compound number 1, gives hydrogen bond (H-Bond) with the amino acid MET109.

The ring with chlorine and fluorine functional groups gives Pi-Cation interaction with the amino acid LYS53. Also, the ring with two fluorine groups in this compound forms Pi-Pi stacking interaction with the amino acid TYR35. In the proposed compound number 3, amino acid LYS53 gives a hydrogen bond (H-Bond) with the oxygen of carbonyl group and participates in the Pi-Cation interaction

with a ring that has two fluorine groups. Also in this compound, the amino acid TYR35 forms Pi-Pi stacking interaction with a ring that has three fluorine groups. In compound number 4, amino acid ASP168 creates the halogen bond with the chloride functional group and amino acid LYS53 takes part in the Pi-Cation interaction with a ring that has a working group. In the proposed compound number 5, amino acids MET109 and GLY110 make hydrogen bonds with the oxygen of the carbonyl group. Also, amino acid LYS53 participates in the Pi-Cation interaction with a ring that has four fluorine groups. In this compound, the amino acid TYR35 institutes Pi-Pi stacking interaction with a ring that has two fluorine groups.

3.8.2 Investigation of physicochemical properties of the proposed compounds

Having the convenient physicochemical properties makes the compound more effective and picks it up as a suitable drug candidate. Then, in this research in addition to lipophilicity in Lipinski, other features such as water solubility, the amount of gastrointestinal absorption, and the polarity of the five proposed compounds were studied. The results of these investigations have been shown in Table 8 (the results were predicted by the ADME Swiss server).

In general, the results of this study indicate that the proposed compounds have appropriate physicochemical properties such as high gastrointestinal absorption, relatively good solubility, and a high degree of polarization. Also, the results of molecular docking studies for five proposed

compounds showed the good tendency of these compounds to interact with mitogen kinase protein. Therefore, more investigations regarding these compounds in the laboratory environment and the examination of the inhibitory effect of these compounds in the in vivo and in vitro conditions can confirm the high potential of these compounds in the control of cancer tumors progression and their treatment with more confidence.

3.8.3 Studying binding energy

The binding energy among combinations was examined, and the results for 8 structures (1–8) is showed in the table (Table 9). The pIC_{50} chart has been made drawing on binding energy (Fig. 6). The appropriate R^2 shows that these parameters have linear relation and our docking method is valid.

The free binding energy was calculated for 1 to 8 compounds, the values of which are shown in Table 10, and the free energy diagram was plotted in terms of pIC_{50} , as shown in the Fig. 7. Appropriate R^2 indicates that there is a linear relationship between the two parameters and the results show that these compounds have good performance and acceptable bond energy. Based on these diagrams and their line equations, pIC_{50} was calculated for the test series, the values of which are shown with the actual values of Observation in Table 11. The pIC_{50} was also calculated for the proposed compounds. Its values are shown in Table 12 according to the appropriate QSAR model (GA-PLS), docking score and free energy.

Table 8 Physicochemical properties of five proposed compounds

Compounds	Chemical formula	Polar surface area (PSA \leq 140 Å)	Solubility	Digestion
1	$C_{22}H_{13}C_{12}F_2NO$	22 Å ²	1.94e-04 mg ml ⁻¹	Low
2	$C_{20}H_{12}ClF_3N_2OS$	82.01 Å ²	1.02e-06 mg ml ⁻¹	Low
3	$C_{22}H_{15}F_5N_2O_3$	75.17 Å ²	2.07e-06 mg ml ⁻¹	Low
4	$C_{26}H_{15}ClF_4N_2O$	56.71 Å ²	2.78e-09 mg ml ⁻¹	Low
5	$C_{21}H_{12}F_6N_2O_2S$	91.24 Å ²	5.68e-07 mg ml ⁻¹	Low

Table 9 Docking score regarding pIC_{50} for the first 8 combinations

No.	pIC_{50} (nM)	Docking score (kcal mol ⁻¹)
1	6.49	-10.812
2	6.72	-10.702
3	7.03	-11.485
4	6.06	-10.142
5	6.05	-9.692
6	6.31	-10.461
7	6.14	-9.787
8	5.89	-9.602

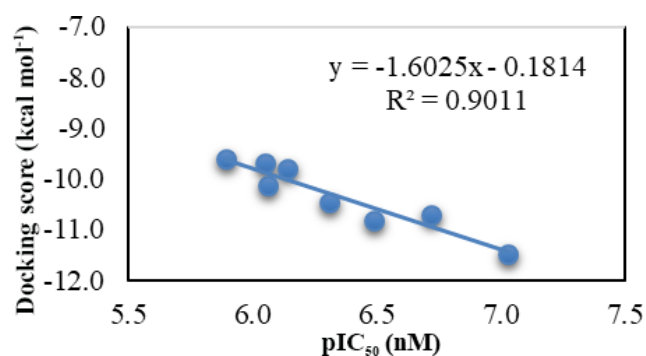
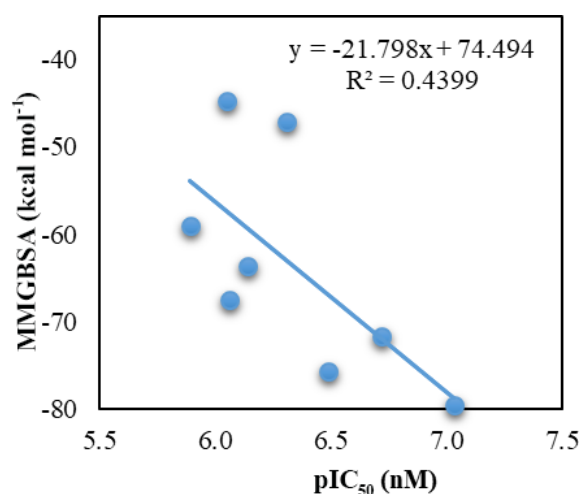


Fig. 6 Docking score regarding pIC_{50} for the first 8 combinations

Table 10 MMGBSA values and pIC₅₀ for the first 8 compounds

No.	pIC ₅₀ (nM)	MMGBSA (kcal mol ⁻¹)
1	6.49	-75.69
2	6.72	-71.69
3	7.03	-79.58
4	6.06	-67.55
5	6.05	-44.67
6	6.31	-47.16
7	6.14	-63.69
8	5.89	-58.96


Fig. 7 MMGBSA regarding pIC₅₀ plot for the first 8 compounds

4 Conclusion

Image processing has made remarkable progress in theoretical and practical aspects in recent decades, and its application can be seen in a wide range of sciences and industries. QSAR studies are an important application of

Table 11 Predicted pIC₅₀ values by docking scores for test set

No.	pIC ₅₀ (nM) observation	pIC ₅₀ (nM) predicted based on docking score
1	6.49	6.63
5	6.05	5.93
7	6.14	5.99
17	5.84	6.62
21	5.62	5.12
27	7.32	6.23
28	7.1	6.47
33	7.49	6.36
36	7.52	6.35

Table 12 Predicted pIC₅₀ table values for proposed compounds

No.	Docking score (kcal mol ⁻¹)	MMGBSA (kcal mol ⁻¹)	pIC ₅₀ (nM) predicted based on docking score	pIC ₅₀ (nM) predicted based on MMGBSA	pIC ₅₀ (nM) predicted by GA-PLS
1	-10.898	-73.33	6.68	6.78	7.16
2	-12.815	-85.65	7.88	7.35	7.23
3	-10.032	-67.07	6.15	6.49	7.89
4	-8.891	-52.67	5.44	5.83	6.91
5	-12.676	-76.44	7.8	6.92	5.88

image processing because they provide an acceptable prediction of compound activity, and the results can be chemically interpreted. The GA-PLS model was used in this study to investigate QSAR inhibitory activity. The model's RMSEC and RMSEP values are 0.19 and 0.20, respectively. As a result, incorporating the GA method into the PLS model improves the predictive validity of the QSAR model.

References

- [1] Cuenda, A. "Mitogen-Activated Protein Kinases (MAPK) in Cancer", In: Boffetta, P., Hainaut, P. (eds.) Encyclopedia of Cancer, 2019, pp. 472–480. ISBN 978-0-12-812485-7
<https://doi.org/10.1016/B978-0-12-801238-3.64980-2>
- [2] Milletti, F., Hermann, J. C. "Targeted Kinase Selectivity from Kinase Profiling Data", ACS Medicinal Chemistry Letters, 3(5), pp. 383–386, 2012.
<https://doi.org/10.1021/ml300012r>
- [3] Nakamura, S., Pourkheirandish, M., Morishige, H., Kubo, Y., Nakamura, M., Ichimura, K., ..., Komatsuda, T. "Mitogen-Activated Protein Kinase Kinase 3 Regulates Seed Dormancy in Barley", Current Biology, 26(6), pp. 775–781, 2016.
<https://doi.org/10.1016/j.cub.2016.01.024>
- [4] Mavropoulos, A., Orfanidou, T., Liaskos, C., Smyk, D. S., Spyrou, V., Sakkas, L. I., Rigopoulou, E. I., Bogdanos, D. P. "p38 MAPK Signaling in Pemphigus: Implications for Skin Autoimmunity", Autoimmune Diseases, 2013, 728529, 2013.
<https://doi.org/10.1155/2013/728529>
- [5] Dhabhar, F. S. "Psychological stress and immunoprotection versus immunopathology in the skin", Clinics in Dermatology, 31(1), pp. 18–30, 2013.
<https://doi.org/10.1016/j.clindermatol.2011.11.003>
- [6] Hosoyama, Y., Domae, E., Goda, S., Matsumoto, N. "Effects of gallotannin on osteoclastogenesis and the p38 MAP kinase pathway", Orthodontic Waves, 75(4), pp. 105–113, 2016.
<https://doi.org/10.1016/j.odw.2016.10.007>
- [7] De-Eknamkul, W., Umehara, K., Monthakantirat, O., Toth, R., Frecer, V., Knapic, L., Braiuca, P., Noguchi, H., Miertus, S. "QSAR study of natural estrogen-like isoflavonoids and diphenolics from Thai medicinal plants", Journal of Molecular Graphics and Modelling, 29(6), pp. 784–794, 2011.
<https://doi.org/10.1016/j.jmgm.2011.01.001>
- [8] Asirvatham, S., Dhokchawle, B. V., Tauro, S. J. "Quantitative structure activity relationships studies of non-steroidal anti-inflammatory drugs: A review", Arabian Journal of Chemistry, 12(8), pp. 3948–3962, 2019.
<https://doi.org/10.1016/j.arabj.2016.03.002>

- [9] Yamashita, M., Sawano, J., Umeda, R., Tatsumi, A., Kumeda, Y., Iida, A. "Structure–Activity Relationship Studies of Antimicrobial Naphthoquinones Derived from Constituents of *Tabebuia avellaneda*", Chemical and Pharmaceutical Bulletin, 69(7), pp. 661–673, 2021.
<https://doi.org/10.1248/cpb.c21-00208>
- [10] Jomehzadeh, N., Koolivand, Z., Dahdouh, E., Akbari, A., Zahedi, A., Chamkouri, N. "Investigating in-vitro antimicrobial activity, biosynthesis, and characterization of silver nanoparticles, zinc oxide nanoparticles, and silver-zinc oxide nanocomposites using *Pistacia Atlantica Resin*", Materials Today Communications, 27, 102457, 2021.
<https://doi.org/10.1016/j.mtcomm.2021.102457>
- [11] Mauri, A., Consonni, V., Todeschini, R. "Molecular Descriptors", In: Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., G. Papadopoulos, M., Reis, H., K. Shukla, M. (eds.) Handbook of Computational Chemistry, Springer, pp. 2065–2093, 2017. ISBN 978-3-319-27281-8
https://doi.org/10.1007/978-3-319-27282-5_51
- [12] Rahmouni, A., Touhami, M., Benaissa, T. "Fukui Indices as QSAR Model Descriptors: The Case of the Anti-HIV Activity of 1-2-[(Hydroxyethoxy) Methyl]-6-(Phenylthio) Thymine Derivatives", International Journal of Chemoinformatics and Chemical Engineering (IJCCE), 6(2), pp. 31–44, 2017.
<https://doi.org/10.4018/IJCCE.2017070103>
- [13] Mahama, O., Aboudramane, K., Soleymane, K., Sylvain, C., Sekou, D., Drissa, S. "Anticancer Activities and QSAR Study of Novel Agents with a Chemical Profile of Benzimidazolyl-Retrochalcone", Open Journal of Medicinal Chemistry, 10(3), pp. 113–127, 2020.
<https://doi.org/10.4236/ojmc.2020.103006>
- [14] Mostoufi, A., Chamkouri, N., Kordrostami, S., Alghasibahaahmadi, E., Mojaddami, A. "3-Hydroxypyrimidine-2, 4-dione Derivatives as HIV Reverse Transcriptase-Associated RNase H Inhibitors: QSAR Analysis and Molecular Docking Studies", Iranian Journal of Pharmaceutical Research, 19(1), pp. 84–97, 2020.
<https://doi.org/10.22037/2Fijpr.2020.1101004>
- [15] Wang, T., Tang, L., Luan, F., Cordeiro, M. N. D. S. "Prediction of the Toxicity of Binary Mixtures by QSAR Approach Using the Hypothetical Descriptors", International Journal of Molecular Sciences, 19(11), 3423, 2018.
<https://doi.org/10.3390/ijms19113423>
- [16] Gozalbes, R., Jacewicz, M., Annand, R., Tsaioun, K., Pineda-Lucena, A. "QSAR-based permeability model for drug-like compounds", Bioorganic & Medicinal Chemistry, 19(8), pp. 2615–2624, 2011.
<https://doi.org/10.1016/j.bmc.2011.03.011>
- [17] Daré, J. K., Freitas, M. P. "Different approaches to encode and model 3D information in a MIA-QSAR perspective", Chemometrics and Intelligent Laboratory Systems, 212, 104286, 2021.
<https://doi.org/10.1016/j.chemolab.2021.104286>
- [18] Pardoe, I. "Multiple Linear Regression", In: Applied Regression Modeling, John Wiley & Sons, Inc., 2020, pp. 95–158. ISBN 9781119615866
<https://doi.org/10.1002/9781119615941.ch3>
- [19] Roback, P., Legler, J. "Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R", Chapman and Hall/CRC, 2021. ISBN 9780429066665
<https://doi.org/10.1201/9780429066665>
- [20] Krishnan, A., Williams, L. J., McIntosh, A. R., Abdi, H. "Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review", NeuroImage, 56(2), pp. 455–475, 2011.
<https://doi.org/10.1016/j.neuroimage.2010.07.034>
- [21] Gallagher, N. B., Lawrence, L. "Introduction to Hyperspectral and Multivariate Image Analysis and Principal Components Analysis for Multivariate Images", 2020. [online] Available at: https://www.researchgate.net/publication/346731395_Introduction_to_Hyperspectral_and_Multivariate_Image_Analysis_and_Principal_Components_Analysis_for_Multivariate_Images [Accessed: 07 December 2020]
- [22] Kulkarni, S. A., Madhavan, T. "Hologram Quantitative Structure Activity Relationship Analysis of JNK Antagonists", Journal of the Chosun Natural Science, 8(2), pp. 81–88, 2015.
<https://doi.org/10.13160/ricns.2015.8.2.81>
- [23] Freitas, M. P., Brown, S. D., Martins, J. A. "MIA-QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis", Journal of Molecular Structure, 738(1–3), pp. 149–154, 2005.
<https://doi.org/10.1016/j.molstruc.2004.11.065>
- [24] Freitas, M. P. "MIA-QSAR modelling of anti-HIV-1 activities of some 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners", Organic & Biomolecular Chemistry, 4(6), pp. 1154–1159, 2006.
<https://doi.org/10.1039/b516396j>
- [25] Goodarzi, M., Freitas, M. P. "MIA-QSAR coupled to principal component analysis-adaptive neuro-fuzzy inference systems (PCA-ANFIS) for the modeling of the anti-HIV reverse transcriptase activities of TIBO derivatives", European Journal of Medicinal Chemistry, 45(4), pp. 1352–1358, 2010.
<https://doi.org/10.1016/j.ejmech.2009.12.028>
- [26] Cormanich, R. A., Freitas, M. P., Rittner, R. "2D chemical drawings correlate to bioactivities: MIA-QSAR modelling of antimalarial activities of 2,5-diaminobenzophenone derivatives", Journal of the Brazilian Chemical Society, 22(4), pp. 637–642, 2011.
<https://doi.org/10.1590/S0103-50532011000400004>
- [27] Nunes, C. A., Freitas, M. P. "aug-MIA-QSPR on the modeling of sweetness values of disaccharide derivatives", LWT - Food Science and Technology, 51(2), pp. 405–408, 2013.
<https://doi.org/10.1016/j.lwt.2012.11.019>
- [28] Nunes, C. A., Freitas, M. P. "Introducing new dimensions in MIA-QSAR: A case for chemokine receptor inhibitors", European Journal of Medicinal Chemistry, 62, pp. 297–300, 2013.
<https://doi.org/10.1016/j.ejmech.2013.01.005>
- [29] Duarte, M. H., Barigye, S. J., Freitas, M. P. "Exploring MIA-QSARs' for Antimalarial Quinolone-4(1H)-Imines", Combinatorial Chemistry & High Throughput Screening, 18(2), pp. 208–216, 2015.
<https://doi.org/10.2174/1386207318666141229123349>

- [30] Akrami, A., Niazi, A. "Application of MIA for a QSAR Study of Inhibitory Activity of DHP Derivatives and Design of New Compounds Using WT and GA for Pixel Processing", *Polycyclic Aromatic Compounds*, 37(5), pp. 442–455, 2017.
<https://doi.org/10.1080/10406638.2015.1129978>
- [31] Pereira, I. V., Daré, J. K., da Cunha, E. F. F., Freitas, M. P. "MIA-QSAR study of the structural merging of (thio)benzamide herbicides with photosynthetic system II inhibitory activities", *Journal of Biomolecular Structure and Dynamics*, 2022.
<https://doi.org/10.1080/07391102.2022.2055649>
- [32] Advanced Chemistry Development "ChemSketch (Version 5)", [computer program] Available at: www.acdlabs.com [Accessed: 07 December 2018]
- [33] Bhowmik, R., Roy, S., Sengupta, S., Ravi, L. "Computer aided drug design of florfenicol to target chloramphenicol acetyltransferase of vibriosis causing pathogens", *Journal of Applied Biology & Biotechnology*, 10(1), pp. 76–84, 2022.
<https://doi.org/10.7324/JABB.2021.100110>
- [34] Shivaleela, B., Srushti, S. C., Shreedevi, S. J., Babu, R. L. "Thalidomide-based inhibitor for TNF- α : designing and *Insilico* evaluation", *Future Journal of Pharmaceutical Sciences*, 8(5), pp. 1–10, 2022.
<https://doi.org/10.1186/s43094-021-00393-2>
- [35] MathWorks, Inc. "MATLAB (Version 7.13)", [computer program] R2011b.
- [36] Hadaji, E. G., Bourass, M., Ouammou, A., Bouachrine, M. "3D-QSAR models to predict anti-cancer activity on a series of protein P38 MAP kinase inhibitors", *Journal of Taibah University for Science*, 11(3), pp. 392–407, 2017.
<https://doi.org/10.1016/j.jtusci.2016.05.006>
- [37] Kennard, R. W., Stone, L. A. "Computer Aided Design of Experiments", *Technometrics*, 11(1), pp. 137–148, 1969.
<https://doi.org/10.1080/00401706.1969.10490666>
- [38] Andrade-Garda, J. M., Carlosena-Zubieta, A., Boqué-Martí, R., Ferré-Baldrich, J. "Partial Least-Squares Regression", In: Andrade-Garda, J. (ed.) *Basic Chemometric Techniques in Atomic Spectroscopy*, The Royal Society of Chemistry, 2013, pp. 280–347. ISBN 978-1-84973-796-8
<https://doi.org/10.1039/9781849739344-00280>
- [39] Diaz, V. F., De Ketelaere, B., Aernouts, B., Saeys, W. "Cost-efficient unsupervised sample selection for multivariate calibration", *Chemometrics and Intelligent Laboratory Systems*, 215, 104352, 2021.
<https://doi.org/10.1016/j.chemolab.2021.104352>
- [40] Ashoori, M., Nezhadali, M., Shiehmorteza, M. "The Relationship Between Visfatin Levels and Anthropometric Parameters, And Insulin Resistance In Women with Prediabetes and Type 2 Diabetes", *Yafteh*, 20(3), pp. 9–18, 2018.
- [41] Sharma, M. C., Jain, S., Sharma, R. "Trifluorophenyl-based inhibitors of dipeptidyl peptidase-IV as antidiabetic agents: 3D-QSAR COMFA, CoMSIA methodologies", *Network Modeling Analysis in Health Informatics and Bioinformatics*, 7(1), 1, 2018.
<https://doi.org/10.1007%2Fs13721-017-0163-8>
- [42] Bhattacharya, S., Asati, V., Mishra, M., Das, R., Kashaw, V., Kashaw, S. K. "Integrated computational approach on sodium-glucose co-transporter 2 (SGLT2) Inhibitors for the development of novel antidiabetic agents", *Journal of Molecular Structure*, 1227, 129511, 2021.
<https://doi.org/10.1016/j.molstruc.2020.129511>
- [43] El-Mernissi, R., El Khatibi, K., Khaldan, A., ElMchichi, L., Shahinozzaman, M., Ajana, M. A., Lakhli, T., Bouachrine, M. "2-Oxoquinoline Arylaminothiazole Derivatives in Identifying Novel Potential Anticancer Agents by Applying 3D-QSAR, Docking, and Molecular Dynamics Simulation Studies", *Journal of the Mexican Chemical Society*, 66(1), pp. 79–94, 2022.
<https://doi.org/10.29356/jmcs.v66i1.1578>
- [44] Wold, S., Antti, H., Lindgren, F., Öhman, J. "Orthogonal signal correction of near-infrared spectra", *Chemometrics and Intelligent Laboratory Systems*, 44(1–2), pp. 175–185, 1998.
[https://doi.org/10.1016/S0169-7439\(98\)00109-9](https://doi.org/10.1016/S0169-7439(98)00109-9)
- [45] Niazi, A., Leardi, R. "Genetic algorithms in chemometric", *Journal of Chemometrics*, 26(6), pp. 345–351, 2012.
<https://doi.org/10.1002/cem.2426>
- [46] Leardi, R. "Genetic algorithms in chemistry", *Journal of Chromatography A*, 1158(1–2), pp. 226–233, 2007.
<https://doi.org/10.1016/j.chroma.2007.04.025>
- [47] Guha, R., Ghosh, M., Kapri, S., Shaw, S., Mutsuddi, S., Bhateja, V., Sarkar, R. "Deluge based Genetic Algorithm for feature selection", *Evolutionary Intelligence*, 14(2), pp. 357–367, 2021.
<https://doi.org/10.1007/s12065-019-00218-5>
- [48] Too, J., Abdullah, A. R. "A new and fast rival genetic algorithm for feature selection", *The Journal of Supercomputing*, 77(3), pp. 2844–2874, 2021.
<https://doi.org/10.1007/s11227-020-03378-9>
- [49] Song, K., Li, L., Tedesco, L. P., Li, S., Clercin, N. A., Hall, B. E., Shi, K. "Hyperspectral determination of eutrophication for a water supply source via genetic algorithm–partial least squares (GA–PLS) modeling", *Science of The Total Environment*, 426, pp. 220–232, 2012.
<https://doi.org/10.1016/j.scitotenv.2012.03.058>
- [50] Nekoei, M. "Genetic Algorithm Based Wavelengths Selection Coupled with Partial Least Squares for Simultaneous Spectrophotometric Determination of Phosphate and Silicate in Detergent Products", *Current Analytical Chemistry*, 14(2), pp. 151–158, 2018.
<https://doi.org/10.2174/1573411013666170703162902>
- [51] Kawamura, K., Watanabe, N., Sakanoue, S., Lee, H. J., Lim, J., Yoshitoshi, R. "Genetic algorithm-based partial least squares regression for estimating legume content in a grass-legume mixture using field hyperspectral measurements", *Grassland Science*, 59(3), pp. 166–172, 2013.
<https://doi.org/10.1111/grs.12026>
- [52] Tropsha, A. "Best Practices for QSAR Model Development, Validation, and Exploitation", *Molecular Informatics: Models – Molecules – Systems*, 29(6–7), pp. 476–488, 2010.
<https://doi.org/10.1002/minf.201000061>

- [53] Mitra, I., Saha, A., Roy, K. "Quantitative Structure–Activity Relationship Modeling of Antioxidant Activities of Hydroxybenzalacetones Using Quantum Chemical, Physicochemical and Spatial Descriptors", *Chemical Biology & Drug Design*, 73(5), pp. 526–536, 2009.
<https://doi.org/10.1111/j.1747-0285.2009.00801.x>
- [54] Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., Das, R. N. "Some case studies on application of r_m^{2m} metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data", *Journal of Computational Chemistry*, 34(12), pp. 1071–1082, 2013.
<https://doi.org/10.1002/jcc.23231>
- [55] Richon, A. B. "An early history of the molecular modeling industry", *Drug Discovery Today*, 13(15–16), pp. 659–664, 2008.
<https://doi.org/10.1016/j.drudis.2008.03.012>
- [56] Singh, B., Mal, G., Gautam, S. K., Mukesh, M. "Computer-Aided Drug Discovery", In: *Advances in Animal Biotechnology*, Springer Cham, 2019, pp. 471–481. ISBN 978-3-030-21308-4
https://doi.org/10.1007/978-3-030-21309-1_44
- [57] Barret, R. "6 - Lipinski's Rule of Five", In: *Therapeutical Chemistry*, Elsevier Ltd., 2018, pp. 97–100. ISBN 978-1-78548-288-5
<https://doi.org/10.1016/b978-1-78548-288-5.50006-8>
- [58] Long, K., Kostman, S. J., Fernandez, C., Burnett, J. C., Hury, D. M. "Do Zebrafish Obey Lipinski Rules?", *ACS Medicinal Chemistry Letters*, 10(6), pp. 1002–1006, 2019.
<https://doi.org/10.1021/acsmchemlett.9b00063>
- [59] Sattarzadeh, M. B., Shojaii, A., Toosi, M. N., Abdollahi-Fard, M., Avanaki, F. A., Taher, M., Shieh-morteza, M., Hashem-Dabaghian, F. "Topical Mastic Oil for Treatment of Functional Dyspepsia: A Randomized, Triple-Blind Controlled Trial", *Galen Medical Journal*, 10, e1965, 2021.
<https://doi.org/10.31661/gmj.v10i0.1965>
- [60] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", *Advanced Drug Delivery Reviews*, 64(Supplement), pp. 4–17, 2012.
<https://doi.org/10.1016/j.addr.2012.09.019>
- [61] Petit, J., Meurice, N., Kaiser, C., Maggiora, G. "Softening the Rule of Five—where to draw the line?", *Bioorganic & Medicinal Chemistry*, 20(18), pp. 5343–5351, 2012.
<https://doi.org/10.1016/j.bmc.2011.11.064>
- [62] Chen, X., Li, H., Tian, L., Li, Q., Luo, J., Zhang, Y. "Analysis of the Physicochemical Properties of Acaricides Based on Lipinski's Rule of Five", *Journal of Computational Biology*, 27(9), pp. 1397–1406, 2020.
<https://doi.org/10.1089/cmb.2019.0323>
- [63] Ertl, P. "Polar Surface Area", In: Mannhold, R. (ed.) *Molecular Drug Properties: Measurement and Prediction*, Wiley-VCH Verlag GmbH & Co. KGaA, 2007, pp. 111–126. ISBN 9783527317554
<https://doi.org/10.1002/9783527621286.ch5>
- [64] Chemaxon Ltd. "Marvin Sketch (Version 15.6.29)", [computer program] Available at: <https://www.chemaxon.com> [Accessed: 29 June 2015]
- [65] Schrödinger LLC. "Schrödinger (Version 12.5)", [computer program] Available at: <https://www.schrodinger.com> [Accessed: 18 March 2021]