Int. J. Nonlinear Anal. Appl. In Press, 1–16
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2024.34613.5176

A QSAR study into pyrazolone inhibitors and design of new compounds using genetic algorithm

Mitra Mirshafiei[a], Ali Niazi[b,*], Atisa Yazdanipour[b]

[a]Department of Chemistry, Arak Branch, Islamic Azad University, Arak, Iran
[b]Department of Chemistry, Central Tehran Branch, Islamic Azad University, Tehran, Iran

(Communicated by Haydar Akca)

Abstract

This study employed multivariate image analysis to explore the quantitative relationship between molecular structure and the inhibitory activity of pyrazolone. MIA-QSAR, a data mining technique based on two-dimensional images (descriptors), was utilized to establish a quantitative relationship between structure and pIC50. Descriptors are pixel images of two-dimensional molecular structures. The study aimed to establish the relationship between IC50 activity-dependent variables and independent variables (pixels or hidden variables). Principal component analysis (PCA) was conducted on the resulting descriptors, and principal components (PCs) were extracted. Using a genetic algorithm (GA) partial least squares (PLS) with vertical preprocessing, the descriptors obtained were screened and the inhibitory activities of these compounds were modelled as a function of molecular structures by different chemometric methods in pyrazolone derivatives, namely PLS, orthogonal signal correction (OSC), GA, PLS, and GA-PLS. The statistical parameters of the most suitable model for pyrazolone derivatives, RMSEP (0.19) and RSEP (1.21), demonstrated high predictive power for the OSC-GA-PLS model. The proposed QSAR models were utilized to predict the inhibitory activity of new compounds.

Keywords: genetic algorithms, multivariate image processing, partial least squares, principal component analysis, pyrazolone, quantitative structure-activity relationship
2020 MSC: 74E40

1 Introduction

Pyrazoles, which are heterocycles with five members and contain two adjacent nitrogen atoms, were discovered and classified by Ludwig Knorr and Eduard Buchner. These bioactive nitrogen compounds exhibit several important pharmacological properties such as antibacterial activity, appetite suppression, anti-leukemia, anti-diabetic, antihypertensive, anti-viral, platelet inhibition, and antitumor effects [31, 42]. Additionally, pyrazoles have been recognized for their photophysical properties and applications, such as UV light stabilization for polystyrene and highly selective fluorescence sensors [46].

Pyrazole was synthesized from ethylene and diazomethane. Pyrazoles are a significant category of heterocyclic compounds that have garnered substantial attention due to their broad applications in fields such as pharmacology,

*Corresponding author
Email addresses: m.mirshafiee96@gmail.com (Mitra Mirshafiei), ali.niazi@iauctb.ac.ir (Ali Niazi), ateesa.yazdanipour@gmail.com (Atisa Yazdanipour)

Received: June 2024 *Accepted:* September 2024

soil chemistry, and photography. In the solid state and concentrated solutions, pyrazoles form dimers through intramolecular hydrogen bonds. Consequently, the melting and boiling points of this compound are higher than those of pyrrole and pyridine [14].

Pyrazolones are an especially important subgroup of pyrazoles, known for their extensive biological and medicinal properties. Specifically, pyrazolone, with its five-membered lactam ring, is considered an active core in pharmacology and the dye industry [6]. Pyrazolone compounds exhibit a wide range of biological activities. Additionally, these compounds are utilized in the production of agricultural pesticides and in the dye industry [7, 33, 49]. Therefore, the development of new synthetic methods for these compounds is of paramount importance.

Pyrazolone is derived from the hydrazone segment, which has biological properties and exhibits activities such as potential enzyme inhibition. Historically, heterocycles have garnered the attention of scientists. Among various heterocyclic compounds, pyrazolones and their derivatives have been studied more extensively, as they act as both products and intermediates in fields such as chemicals, analytical chemistry, agriculture, biology, and pharmacy.

Prabhakaran and Velmathi [39] investigated the synthesis of diphenyl pyrazolone piperidinium salt derivatives from the reaction between aromatic aldehyde, 3-methyl phenyl pyrazolone, and piperidine in ethanol. This synthesis method demonstrated advantages such as shorter reaction times, excellent yields, cost-effectiveness, user-friendly procedures, structural diversity, and purification exclusively through filtration. Dhawan [11] explored pyrazolone derivatives. Recognizing their importance, he developed synthetic strategies for both existing and new pyrazolone derivatives and investigated their biochemical applications. His research aimed to compare the chemical synthetic designs with the biological advancements of pyrazolone derivatives. Gouault et al. [23] conducted a study on pyrazolone derivatives and discovered that they could be synthesized from perfluorinated allylic alcohols using a direct two-step process. This process involved a Heck coupling/isomerization reaction, followed by condensation with hydrazine. By combining a cyclopropanation approach with F-SPE purification, this method facilitated the acquisition of final products with high purity. Due to the significant importance of pyrazolone derivatives, our current research delves into this compound and aims to predict new structures that have not yet undergone biological testing.

Chemometrics, which is an important tool in analytical chemistry, encompasses several developed techniques that can be used in various areas of chemistry [37]. Therefore, chemometrics is considered a powerful tool, especially in complex chemical devices. The branches of chemometrics include methods such as pattern recognition, classification, signal processing, multivariate calibration, and experimental design. In simpler systems, algorithms like Multiple Linear Regression (MLR) [40], Principal Component Analysis (PCA) [15], and Partial Least Squares (PLS) [38] are commonly employed [27, 43].

Quantitative structure-activity relationship (QSAR) techniques encompass a wide range of approaches, ranging from chemical measurements and biological experiments to statistical techniques and the interpretation of their findings. The QSAR method can be extended and applied in various molecular design applications, including environmental science, predicting different types of biological activity by linking similar series of compounds, optimizing leader compounds, classifying and diagnosing drugs, explaining the mechanism of drug reactions, and predicting new leader structures for medicinal design [8, 10]. In summary, QSAR models are capable of predicting the biological activities of untested or inaccessible compounds. Moreover, they assist in identifying the chemical properties that influence the biological activities of compounds [5].

The field of computational molecular design was introduced in the late 1960s. Computer-Aided Drug Designing (CADD) has gained popularity as a beneficial tool for the logical development of drugs. It minimizes the time required for drug analysis, characterization, and structure optimization. Over the course of 20 years, this method has been developed and evolved, finding widespread applications in the field of chemistry. It is utilized for advanced drug design, as well as for designing vaccines, antibodies, proteins or peptides, and other macro/micro molecules and substances.

A mathematical association between the chemical structure of compounds and their chemical, biological, or physical features is provided by the QSAR technique. Following the analysis of the response between the ligand and receptor, a new compound is then designed. The compounds featuring comparable physicochemical qualities have the same physiological functions. An association is established between the capability of potential drug candidates to block a certain biological function and their structural characteristics via the QSAR approach. The comparison of the multivariate image computational modelling of analysis (MIA-QSAR) with the alternative QSAR techniques indicated that MIA-QSAR presented fast results of analysis, which are as precise as those of the most advanced techniques at hand. Furthermore, it is simple and inexpensive when used to predict and manage all modelled responses for a congeneric collection of chemical structures without requiring conformational analysis or three-dimensional alignment. In the present technique, a two-dimensional representation of pixel indicates topo-chemical features of chemicals, and a model is built between these descriptors and a y-block that includes the independent variables. As a non-invasive

analysis, MIA-QSAR leads to saving money and time while it is capable of processing a large sum of data. The objective of such a technique is correlating a great number of columns of separate variables to y as a single-column dependent variable. Different coordinates of pixels observed in the molecular drawing reflect the structural changes that occurred in the MIA-QSAR approaches. Such changes were employed to depict bioactivity variations for a congeneric group categorized under the title of drug-like compounds [9, 24]. During the modelling procedure, one may employ the congeneric series of compounds and the substitution pattern in order to anticipate the bioactivities related to comparable compounds. Below are the steps of the MIA-QSAR technique:

- Drying the structures of the molecule, creation of images and their alignment
- Image denoising and subsequently, image unfolding to a two-way array and also descriptor generation
- Regression modeling, as well as selecting the features

One of the most crucial processes addressed is modeling. A number of the techniques that can be adopted include partial least squares [35], artificial neural networks, and multiple linear regression [30]. Even though the accuracy of MLR is relatively low, it has been frequently employed in QSAR investigations. In addition, in cases where the number of rows exceeds that of the columns, MLR shows a good function. Although artificial neural networks present enough accuracy in most conditions, they present a risk of overfitting the training data. Thus, they are not capable of extrapolating proper data information.

In genetic algorithms, a number of genetic functions and fitness criteria are utilized [19]. It is demonstrated that by eliminating the redundant data before the PLS regression, preprocessing presents enough inputs for the purpose of PLS. As a result, it improves the model quality. Earlier investigations [32] have scrutinized OSC.

In a review paper, Geladi [16] thoroughly described various strategies for image regression and investigated the application of PCA, PCR, and PLS in 2-D image analysis. Antonelli et al. [3] utilized multivariate image analysis and a wavelet-based classification algorithm to evaluate food color. Noordam and Geladi [34] proposed a novel approach to modeling and representing classes for image analysis techniques. Freitas et al. [18] examined a series of derivatives of (1-ethyl-2-pyrrolidinyl)methyl]-6- methoxybenzamides (Dopamine D_2 receptors subtype) and applied 2D image analysis to QSAR assays. Antunes et al. [4] employed the MIA-QSAR technique to model biological activities related to certain phosphodiesterase inhibitors. Tessier et al. [44] also utilized multivariate image analysis to evaluate the alumina content of anode cover material. Freitas et al. [21] employed the MIA-QSAR approach to predict the electrophoretic enantiomer modeling of specific amino-aromatic acids and esters. Goodarzi and Freitas [22] modeled antifungal activities of a series of azole derivatives using multivariate image analysis descriptors. Additionally, Garkani-Nejad et al. [20] employed the MIA-QSAR strategy to predict 13 C chemical shifts of naphthalene derivatives. Fritz [17] estimated the potential of anxiolytic agents using MIA-QSAR. In 2009, Goodarzi and Freitas [22] modeled the antifungal activity of several azole derivatives using multivariate image analysis descriptors. In 2011, Saghaei et al. [41] used PLS and RBF neural network models in MIA-QSAR for kinase inhibitors. Dorate et al. [13] utilized a series of quinolone derivatives (antimalarial drugs) and developed both normal image and color image analysis models to predict the activity of anti-malaria drugs. Akrami and Niazi [1] established a relationship between the two-dimensional chemical structure and activity, indicating that the two-dimensional chemical structure is related to the activity. They predicted the inhibitory activity of DHP derivatives using the MIA-QSAR model. Amri et al. [2] used the OSC-GA-PLS model in MIA-SAR for 5α-reductase inhibitors. This study aimed to establish a relationship between the pixels present in the studied compounds as independent variables and the inhibitory activity of the compounds as dependent variables.

In this study, image analysis is conducted in two phases. The first phase involves extracting data from images, while the second phase focuses on performing regression analysis on the extracted data. The pixels within the images serve as dependent variables that are used in designing the QSAR model. The objective of this study was twofold. Firstly, it aimed to develop a model using PLS, GA, and OSC-GA-PLS to predict the activity of drug derivatives of pyrazolone. The developed model was utilized to predict the inhibitory activities of new pyrazolones. Secondly, it sought to establish a relationship between the pixels present in the studied compounds, considered as independent variables, and the inhibitory activity of the compounds, which were considered to be dependent variables.

2 Materials and methods

2.1 Instrumentation

The ChemOffice package (2013 version) was used to perform calculations and draw molecular structures. Paint software was utilized to convert the molecular structures into images and centralize them. Statistical calculations for

creating descriptors and modeling were conducted using MATLAB software (version 7.13, Mathworks Inc). MATLAB is a computer software designed for mathematical calculations, utilizing arrays and matrices for data representation. This feature enables users to numerically solve various types of calculations and convert them into matrices and arrays. The selection of variables and PLS modeling was carried out using the PLS program from PLSToolboxVersion4.

2.2 Data set

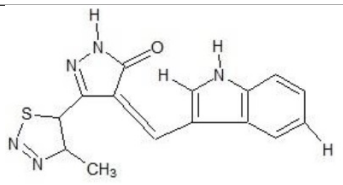
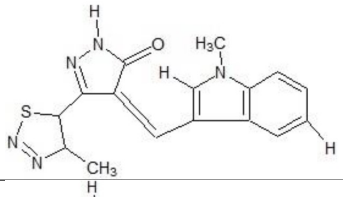
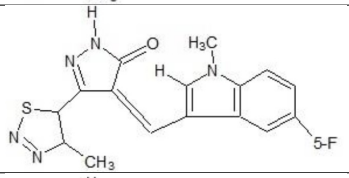
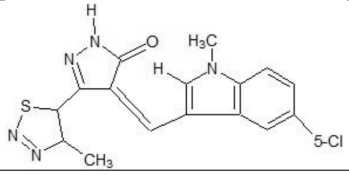
The first step in QSAR is the selection of an appropriate dataset to create a valid model. Valid results and models can only be obtained from suitable data. Therefore, it is important to obtain data from reputable research groups or journals. The experimental values for the quantity under study must be measured for all compounds under identical conditions. These compounds should have the same structure, with only a variable part of the molecule that affects their activity or inhibition. In this study, all chemicals used, including 33 inhibitor pyrazolone compounds with biological activity as pIC50, were sourced from the literature [36]. The structures of the pyrazolone compounds and their inhibitor activity values are presented in Table 1. Using these pyrazolone derivatives and the most suitable computational model, we predicted compounds with similar structures and inhibition ranges.

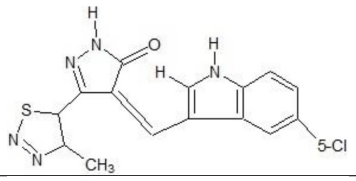
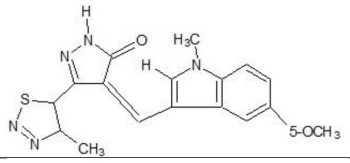
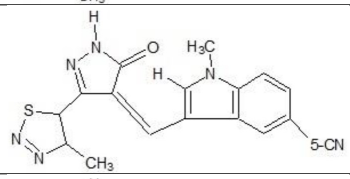
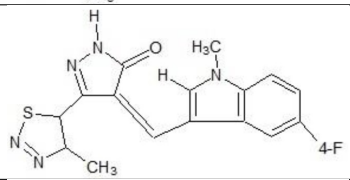
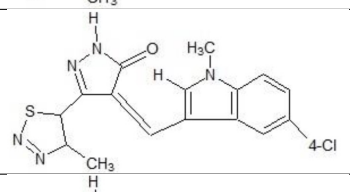
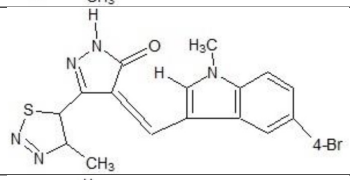
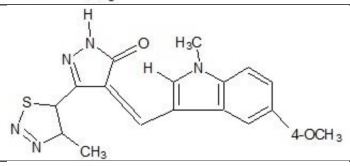
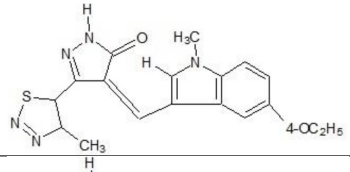
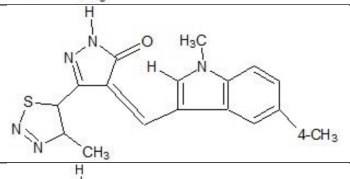
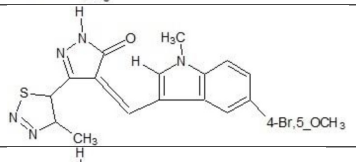
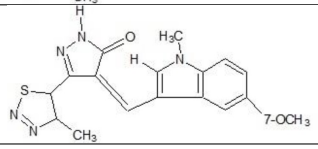
For this work, the molecules were initially drawn using ISIS Draw software. Subsequently, all molecules were converted into images using paint software and fixed at a specific coordinate point (the same point for all images). The molecule frames were integrated in such a way that, when opened in Microsoft Office Picture Manager, the molecules overlapped and even single-pixel movements were prevented. This was the only change made to the different parts of the molecules. The Kennard-Stones algorithm [29] was implemented in MATLAB to ensure that the training and prediction sets encompassed all the main data. Therefore, 27 compounds were selected as the calibration set, while 6 compounds were chosen as the test set. The Kennard-Stones algorithm is widely used as the best method for splitting data into training and calibration sets in many QSAR/QSPR studies. It selects a set with a homogeneous distribution in the definition space.

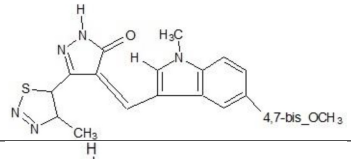
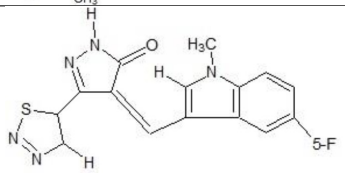
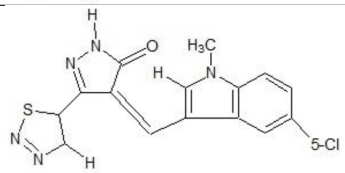
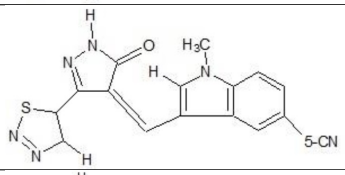
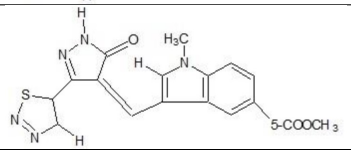
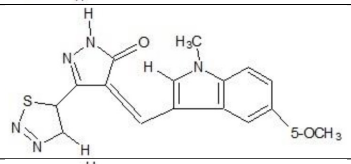
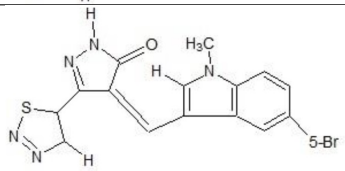
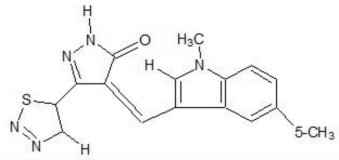
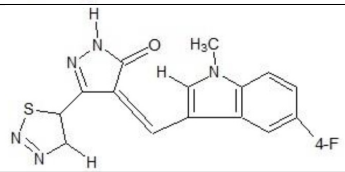
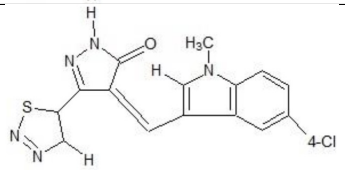
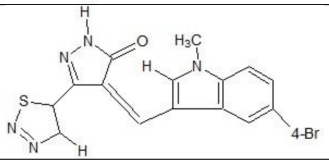
2.3 Analysis of multivariable image

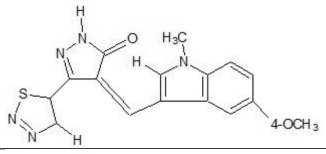
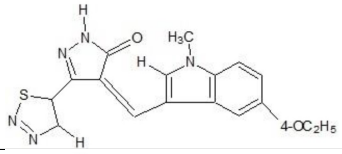
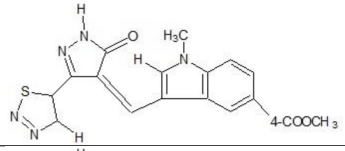
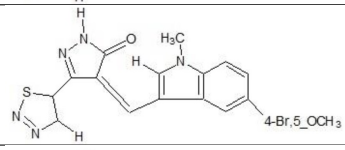
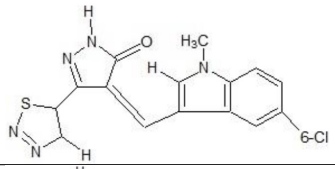
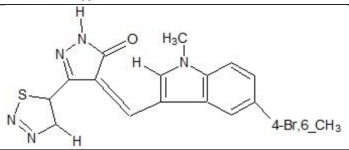
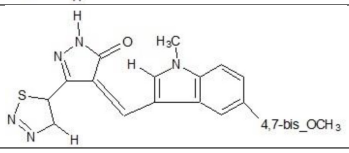
In order to establish a relationship between biological activity and molecular images, it is necessary to convert analog data into digital data as a prerequisite to QSAR modeling. Therefore, the molecules presented in Table 1 were systematically drawn with the same font and size as in ISISDRAW software. The images were saved in BMP format and subsequently opened in paint software.

Tabela 1: Chemical structures of pyrazolone derivatives and their inhibiting activity

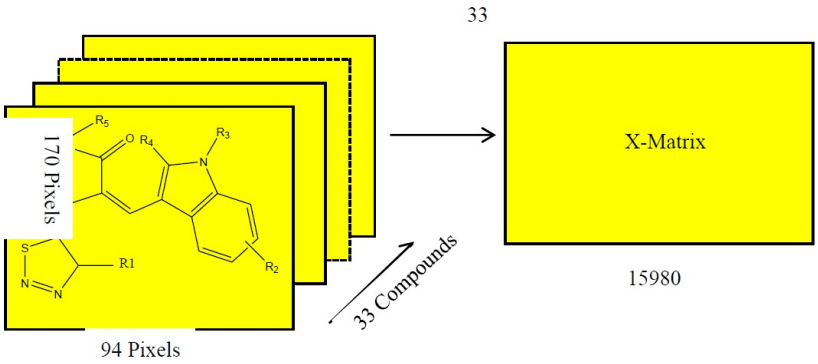
No.	Structure	IC_{50}	PIC_{50}
1		1000	-3
2		95	-1.98
3		99	-2
4		152	-2.18

5		300	-2.48
6		90	-1.95
7		300	-2.48
8		137	-2.14
9		45	-1.65
10		34	-1.53
11		19	-1.28
12		38	-1.58
13		58	-1.76
14		23	-1.36
15		109	-2.04

16		20	-1.30
17		48	-1.68
18		34	-1.53
19		174	-2.24
20		73	-1.86
21		28	-1.45
22		110	-2.04
23		35	-1.54
24		16	-1.20
25		16	-1.20
26		8	-0.90

27		13	-1.11
28		25	-1.40
29		22	-1.34
30		6	-0.78
31		73	-1.86
32		12	-1.08
33		10	-1.00

To modify the molecular images, common points for all molecules were set at 170×94 coordinates. Subsequently, the 2D images of chemical structures were defined for 33 compounds by converting pixels into 15984 binary numbers as image descriptors. In MIA-QSAR studies, image descriptors are usually the pixels of 2D or 3D images. The obtained pixels for these 33 compounds formed a 33×15980 matrix. (as seen in Fig. 1). Additionally, primary processes need to be performed on the data to develop a quantitative relationship between input data and activity vector. In this study, the PLS method was used for modeling. As the dataset was large, GA was employed for pixel selection.



Rysunek 1: Two-dimensional images and unfolding stage for 33 chemical structures so as to present the X-matrix

2.4 Molecular docking study

In the research continuation, to study the effectiveness of the four proposed drug compounds, the compatibility of these compounds with the corresponding receptor was investigated through molecular docking studies. According to the article by Pasha et al., the studied compounds demonstrated good performance in inhibiting the target protein and consequently controlling the progression of cancer [36].

Since the target receptor for our structurally derived compounds is 9C2 450P Cytochrome Human, various crystallographic structures from the target receptor were examined in the Protein Data Bank. Finally, a suitable structure with a resolution of 2.0 angstroms, along with the cofactor Flurbiprofen and PDB ID O9R1, was selected as the target in the molecular docking study [47].

Subsequently, the structures of the four proposed compounds were drawn using Marvin Beans software (version 15.6.29) and considered as ligands. All molecular docking studies were conducted using Schrodinger's Maestro software (version 12.5) and the Dock Ligand mechanism.

3 Results and discussion

3.1 Descriptors of multivariable image analysis

Descriptors in multivariable image processing refer to the pixels of 2D or 3D structural images of desired molecules. These pixels serve as dependent variables in the design of QSAR models. The selection of descriptors is typically based on statistical methods that aim to minimize prediction errors. It is therefore crucial to validate the model derived from the dataset and calculate its prediction error. Various modeling methods can be employed, including multivariable calibration, multi-linear regression, principal component regression, partial least squares regression, and random exploratory search approaches [26].

3.2 Principal component analysis

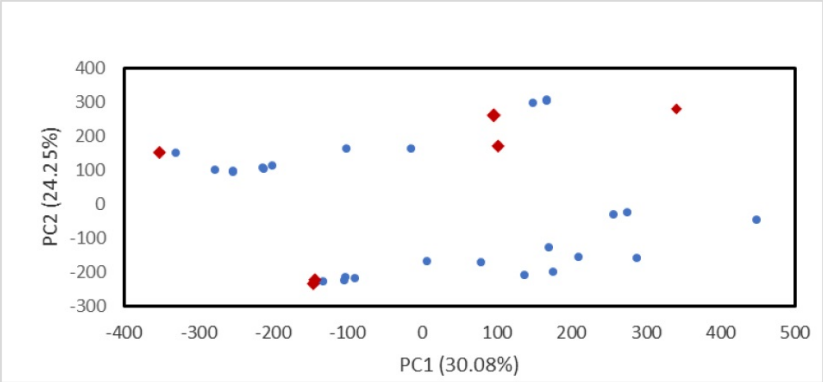
PCA [12] is a powerful tool used to extract valuable data and reduce the amount of information. It aims to decrease the number of variables while capturing the highest variance in the data. The first Principal Component (PC1) represents a line along the axis that contains the most information and accounts for the greatest data variance. PC2, which is perpendicular to PC1, explains the maximum residual variance in the data. Each Principal Component (PC) is orthogonal to the others, and together they define the variances of the data. Transforming the data points from the original space X onto PC1 results in a score matrix on PC1. Similarly, applying these operations on an axis perpendicular to the first yields the score matrix on PC2.

To produce a model with high prediction power, it is important to orthogonalize the variables, which requires minimizing collinearity between them. The Principal Components (PCs) were classified based on their maximum dependence on the activity of pyrazolone derivatives, a step referred to as collinearity classification. The PCs with the highest order were then imported into the PCR model until the desired statistical parameter (in this study, $R^2 = 0.98$) was no longer improving. PCs with higher collinearity contained more information about changes in the activity of pyrazolone derivatives.

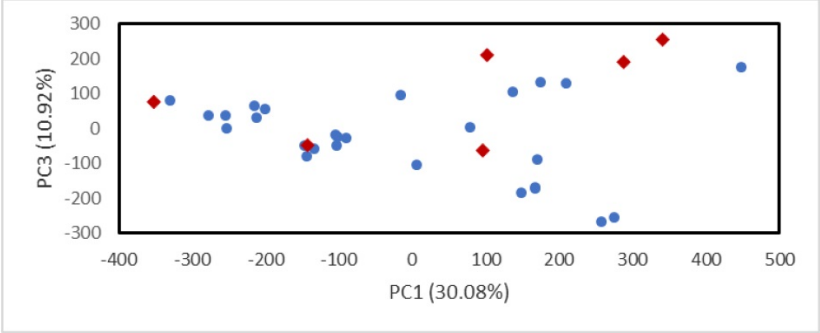
In the partial least squares method, PCA is performed on the independent variables matrix in such a way that the obtained score exhibits the highest collinearity with the score of the dependent variables matrix. In this study, PCA was performed on 2D image descriptors, and variables associated with an activity value less than 0.1 were considered. The PCA results showed that three principal components accounted for 65% of the changes (PC1 = 30.08%, PC2 = 24.25%, and PC3 = 10.92%). As depicted in Figs. 2—4, there is no distinct classification between compounds.

3.3 PLS

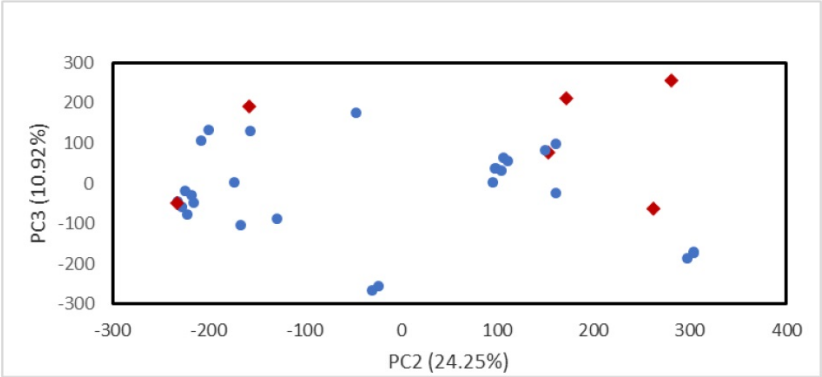
PLS regression is a widely used and highly capable multivariable calibration method that has gained significant interest in recent years. Compared to other methods, it is more efficient and effective. The main objective of linear regression is to establish a quantitative relationship between several independent variables and a dependent variable. In this study, the independent variables are the pixels of molecule image descriptors, their Principal Components (PCs) or Latent Variables (LVs), while the dependent variable is IC50 activity. PLS linear regression is a powerful method specifically designed to solve high-dimensional problems. Regression is one of the most common data analysis problems in science and technology. For example, in chemistry, Y and X could represent the properties and composition



Rysunek 2: An analysis of PC1 versus PC2 for 2D image descriptors.



Rysunek 3: An analysis of PC1 versus PC3 for 2D image descriptors.



Rysunek 4: An analysis of PC2 versus PC3 for 2D image descriptors.

of a chemical, respectively, or the quality and quantity of the products and the chemical structure of the molecules, respectively.

Multiple linear regression is typically used to model Y in terms of X when the variables are relatively few and independent. However, in multiple linear regression, only a small number of variables are selected at each step, resulting in a simpler model that ignores the error in the matrix Y. Models constructed with more variables have several advantages, including higher sensitivity in detecting outlier data and a reduction in standard deviation by a factor of $\sqrt{n}$ when an average of n measurements is used instead of a single measurement.

In general, Y is modeled in terms of X using multiple linear regression when there are relatively few and independent variables. However, multiple linear regression ignores the errors in the Y matrix. Although a small number of variables is selected at each step in multiple linear regression, leading to a simpler model, it fails to account for the error in the Y matrix. Models constructed with more variables have certain benefits. Firstly, they have higher sensitivity in detecting

outlier data. Secondly, the standard deviation decreases by a factor of $\sqrt{n}$ when an average of n measurements is used instead of a single measurement.

The PLS method is more appropriate for regression modeling in larger datasets. It relies on latent principal component-like variables. Additionally, PLS converts the dependent variables into orthogonal latent variables (LVs), allowing it to model multiple dependent variables and resolve collinearity problems among the independent variables. However, PLS is sensitive to the number of variables not included in the prediction model. This sensitivity worsens when a larger number of variables is provided to the model, as it diminishes the predictive power below that of other regression models. The PLS method selects quantities that not only describe changes in the independent variable X but also express the variations in the dependent variable Y.

LV is similar to principal components (PCs). Principal component analysis (PCA) is utilized to reduce the number of initial variables to fewer reduced variables. Specifically, it maintains the most significant PCs while eliminating the rest. This results in reduced matrix dimensions and noise elimination; however, this noise may contain information from the Y matrix. In other words, PCA is performed on both the X and Y matrices. Therefore, LV is used instead of PC, and the PC with the highest covariance with Y is selected from the X matrix.

The least squares method is a multivariable approach that reduces data dimensions and creates a latent variable to establish a relationship between the dependent variables (dependent on IC50 activity) and the independent variable (pixels or latent variables).

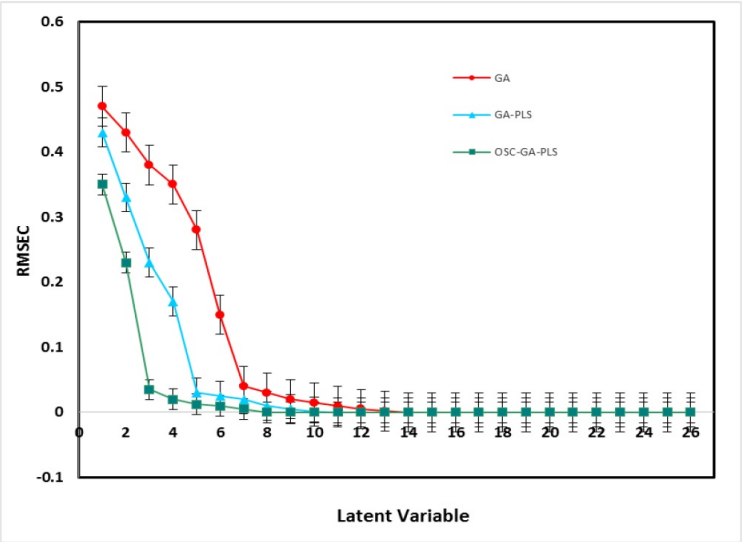
3.4 GA-PLS modeling

A genetic algorithm mimics the principles of genetic science and applies them to generate offspring with improved characteristics, closer to the desired problem target [45]. The competition between genes and the dominance of the selected dominant gene (chosen by the algorithm to generate the next generation) while eliminating less successful genes (those with responses far from the problem target) is an effective method for solving complex and challenging problems [25, 50]. QSAR analyses involve the use of a wide variety of variables. The success of QSAR/QSPR models heavily relies on the selection of variables (molecule descriptors) in a way that they contain the most informative content without overlapping. The genetic algorithm method can be employed as an interesting and flexible approach to achieve this goal. Consequently, this study utilized the genetic algorithm to select descriptors (pixels), which were then used for Partial Least Squares (PLS) calibration. PLS is commonly used in quantitative analysis and MIA-QSAR studies due to its ability to orthogonalize variables and exhibit high correlation between pixel descriptors. In PLS calibration models, the optimal number of latent variables or principal components (PCs) is determined using the Cross-validation algorithm (Leave-one-out) to identify the best model, and the Root Mean Square Error of Prediction (RMSEP) value is calculated. The optimal number of latent variables is identified by examining the RMSEP chart, which indicates the minimum corresponding value. As depicted in Figure 5, the minimum RMSEP value corresponds to the number of latent variables (LVs) being 7. Therefore, the optimal value for PLS is 7, and increasing the number of LVs beyond that does not result in significant changes in the RMSEP values.

Based on the results presented in Figure 5, GA-PLS successfully reduced the number of latent variables to 5. Consequently, 5 was chosen as the optimal number of latent variables in the GA-PLS model for training, as it demonstrated the highest stability. This study showcased the effectiveness of this method for selecting variables (pixels) in image processing. Raw data often contains substantial noise, outliers, and significant variations in dynamic and interval sampling. Utilizing raw data without any form of preprocessing can weaken subsequent analyses and modeling. Hence, preprocessing involves modifying the raw data to simplify and enhance its effectiveness for further processing, such as modeling.

The Orthogonal Signal Correction (OSC) method, initially proposed by [28, 48], is a preprocessing approach for multivariable and 3D calibration methods. This technique removes the information pertaining to independent variables and is also employed in image processing to preprocess descriptors before data analysis. Following data preprocessing, the genetic algorithm determines the optimal function, and ultimately, the selected pixel descriptors are utilized in the PLS method.

Finally, the combination of OSC as a preprocessing method and the genetic algorithm was utilized to determine the optimal number of latent variables in the training set by identifying the number associated with the minimum RMSEC value. As shown in Figure 5, the RMSE was minimized when the number of LVs was equal to 3. Hence, the number 3 was selected as the optimal LV for the training set using the OSC-GA-PLS approach. The findings demonstrate that this method is a robust technique for variable selection in image processing. These modifications are illustrated with respect to each other in Figure 5.



Rysunek 5: A comparison of RMSECV and latent variables.

3.5 Model evaluation methods

Various statistical measures, such as correlation coefficient (R^2), are employed to assess the adequacy of the model. R^2 represents the ratio of the regression sum of squares to the total sum of squares. It reflects the consistency between experimental and calculated data and serves as an indicator of the explanatory power of the regression model. Additionally, R^2 evaluates the model's quality, while prediction capability (Q^2) measures its external prediction accuracy. Higher R^2 and Q^2 values indicate a more optimal and reliable model. This study aimed to explore the predictive capabilities of three models: PLS, GA-PLS, and OSC-GA-PLS. The statistical parameters RMSEP and RSEP were used to evaluate the model's performance in predicting the pIC50 of pyrazolone compounds. Table 2 presents additional statistical parameters, namely R^2 and Q^2, for model evaluation. The results revealed that the OSC-GA-PLS model demonstrated a more favorable correlation between actual and predicted pIC50 values, with an R^2 value of 0.91. Furthermore, considering the data presented in Table II, the OSC-GA-PLS model exhibited superior statistical quality, with a lower percentage of relative error and a smaller number of latent variables compared to the other two models. Determining the appropriate number of latent variables is crucial in data analysis since subsequent principal components or latent variables may contain noise. The R^2 of the model typically increases with an increase in the number of latent variables. However, the Q^2 initially improves, reaches a maximum value, and then plateaus. This maximum value of Q^2 corresponds to the most suitable number of latent variables. The value of R^2 is calculated using Equation (3.1), whereas Q^2 is determined based on Equation (3.2).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,pred} - y_{i,obs})^2}{\sum(y_{i,obs} - \overline{y})^2} \tag{3.1}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,pred} - y_{loo})^2}{\sum(y_{i,obs} - \overline{y})^2} \tag{3.2}$$

in the above equations, $y_{i,pred}$ is the value of pIC50 predicted using different models, $y_{i,obs}$ is the actual value of pIC50, and n is the number of combinations of the test or prediction set.

3.6 Molecular design

In this study, the OSC-GA-PLS model was employed as a CADD method to predict the pIC50 values of four newly synthesized pyrazolone compounds that had not been subjected to prior biological experiments. Table 3 displays the molecular structures of these four compounds along with their corresponding pIC50 values calculated using the OSC-GA-PLS method.

3.7 Molecular docking studies

The results of molecular docking studies indicated that the four proposed compounds have adequate capabilities in forming suitable and stable interactions with the receptor. Therefore, they can serve as suitable inhibitors for the target

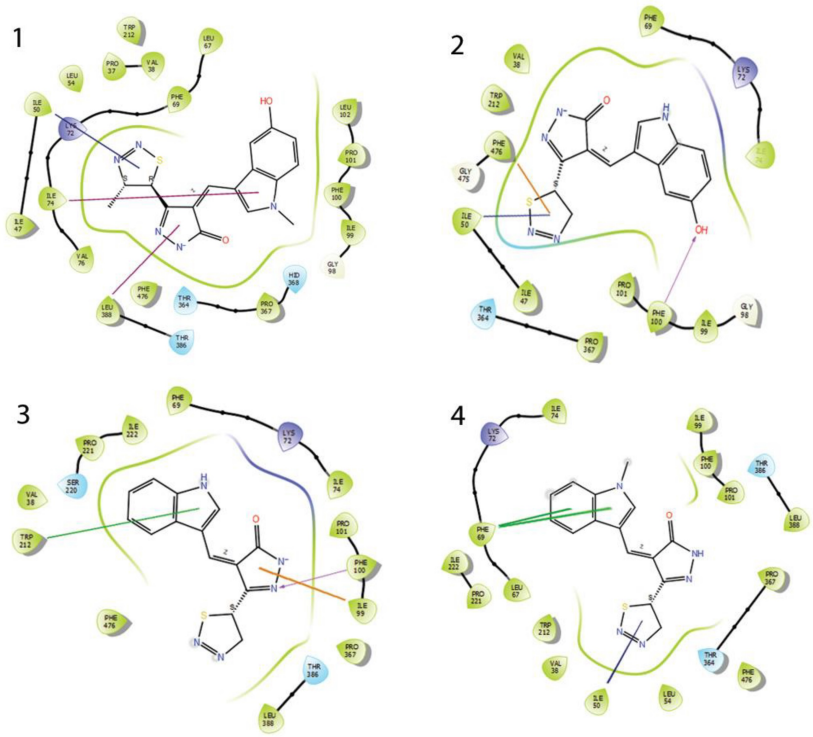
Tabela 2: Observed and calculated Values of pIC50 and prediction error percentage using PLS, GA-PLS and OSC-GA-PLS methods.

Number of compounds	PIC_{50}	**Predicted** PIC_{50}		
		PLS	**GA-PLS**	**OPC-GA-PLS**
1	-3	-2.3476	-2.514	-2.8643
12	-1.58	-1.877	-1.56	-1.4467
14	-1.36	-1.2714	-1.2032	-1.0602
16	-1.3	-1.3157	-1.4285	-1.388
18	-1.53	-1.7065	-1.6756	-1.6958
32	-1.08	-1.051	-1.0894	-0.82713
LVs		7	5	3
Q^2		0.70	0.82	0.89
R^2		0.77	0.87	0.91
RMSEP		0.30	0.22	0.19
RSEP (%)		2.98	1.61	1.21

Tabela 3: The molecular structure of new pyrazolone compounds and the pIC50 value by OSC-GA-PLS method.

No	R_1	R_2	R_3	R_4	R_5	**Observed**
1	CH_3	OC_3H_7	CH_3	H	H	-2.19
2	H	$4-Br, 5-OCH_3$	H	H	H	-2.29
3	H	H	H	H	H	-1.90
4	H	H	CH_3	H	H	-2.33

under investigation. The docking scores in this study ranged from -7.112 to -8.350 kilojoules per mole, demonstrating favorable negative values, which indicate a strong tendency for ligands to interact and form stable complexes with the receptor. The docking scores and the amino acids involved in interactions for each of the four compounds are presented in Table 4.



Rysunek 6: Interactions of each of the four ligand-receptor complexes under study

As shown in Table 4, compound 3 exhibits a more favorable docking score and a higher number of involved amino acids, which suggest that it may have better interaction with the receptor compared to the other three structures. The interaction images of the ligand-receptor complexes for all four compounds are depicted in Figure 6.

The analysis of the docking results revealed interactions such as hydrogen bonds, hydrophobic interactions, electrostatic interactions, π- π stacking, and more between the ligands and the receptor. All proposed compounds exhibited

Tabela 4: Docking scores and amino acids involved in interactions

Structure	**Receptor**	**Docking Score** ($kcal\ mol^{-1}$)	**Amino acids involved in the interaction**
1	1R9O	-7.296	ILE74, LEU388, ILE50, LEU67, PRO101, PRO367, PHE69
2	1R9O	-7.137	PHE100, TRP212, PHE476, ILE50
3	1R9O	-8.350	PHE100, LYS72, PHE476, TRP212, ILE74, PRO101, PRO367, ILE99, PRO101
4	1R9O	-7.112	ILE50, PHE69, TRP212

appropriate binding affinity with the active site of the receptor.

Regarding the first structure based on ligand-receptor complexes, amino acids LEU388 and 450PHE played significant roles. Leucine LEU388 formed a hydrophobic ($\sigma-$ π) interaction with the imidazole ring, while isoleucine ILE50 and ILE74 created hydrophobic ($\sigma-$ π) and π-alkyl bonds, respectively. Proline PRO101 and PRO367 formed π-alkyl bonds. Additionally, phenylalanine PHE69 formed a π-alkyl bond with the ligand forming an energy freeing binding with a -7.296 kcal/mol.

In compound number 2, phenylalanine PHE100, isoleucine ILE50, and phenylalanine PHE476 were the key amino acids. Phenylalanine PHE100 established a hydrogen bond, isoleucine ILE50 created a Hydrophobic Alkyl bond, and tryptophan 212TRP formed electrostatic (π– π), (π-onion), hydrophobic (π-Alkyl), and π- π stacking interactions. The binding energy for this structure was -7.296 kcal/mol.

In compound number 3, the more effective amino acids in interaction are PHE100, ILE99, and TRP212. Amino acid ILE99, isoleucine, forms an Alkyl- π hydrophobic bond with the imidazole ring, leading to a suitable interaction that enhances the stability of the ligand-receptor complex. Amino acid TRP212, tryptophan, engages in hydrophobic interaction in the form of stacking π- π with the pyrrole. Phenylalanine PHE100 forms a hydrogen bond with the ligand, contributing to proper stability of the ligand-receptor bond.

Amino acid LYS72, lysine, establishes an electrostatic bond with the ligand, and amino acid PHE476, phenylalanine, forms a Sulfur- π bond. Proline PRO101 and PRO367 contribute to Alkyl Hydrophobic bonding along with Isoleucine ILE99, forming the best structure with a binding free energy of -8.350 kcal/mol. The presence of a strong hydrogen bond and two hydrophobic bonds and two π-alkyl bonds contributes to the enhanced stability of the ligand-receptor bond, forming the optimal structure with a binding free energy of -8.350 kcal/mol.

In compound number 4, isoleucine ILE50 and phenylalanine 69PHE had the most significant impact on ligand-receptor binding. Isoleucine ILE50 formed an alkyl-hydrophobic bond, while phenylalanine PHE69 created a stacked π- π bond. Tryptophan TRP212 formed a sulfur- π bond. The binding energy for this structure was -7.112 kcal/mol.

In conclusion, the results obtained in this study confirm the high potential of the four proposed structures in controlling the progression of cancer tumors and treating them.

4 Conclusions

In the present era, there is significant attention focused on the development of algorithms, modeling codes, and the enhancement of computational power. This is primarily because experimental studies are both costly and time-consuming. Consequently, QSAR techniques have gained prominence as they enable the progression of numerous models with high reliability and predictive power. The main advantages of these models are their ease of use, simplicity, high accuracy and reliability, and strong predictive power. Notably, image processing stands out as one of the most important applications of the QSAR technique. The results obtained from predicting compound activity can be chemically interpreted. In this study, the OSC-GA-PLS model is employed to investigate the inhibitory activity of pyrazolone derivatives. The model's RMSEP and RMSEC values were found to be 0.19 and 1.21, respectively. Consequently, applying the preprocessing method and incorporating the PLS model into the GA model significantly improves the validity of QSAR predictions. As evident from the aforementioned results, the OSC-GA-PLS method proves to be a highly powerful and useful tool for predicting the inhibitory activity of pyrazolone derivatives by utilizing pixels as a prediction tool.

Acknowledgments

The authors appreciate the Islamic Azad University, Arak, Iran for their support.

Author Contributions

Conceived and extracted data: MM, AN; Analyzed the data: MM, AN, AY, Wrote the paper MM, AN, AY.

Conflict of interest

The authors declare that they have no conflict of interest.

Author Contributions

Upon request, the data and material will be made available.

Literatura

[1] A. Akrami and A. Niazi, *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*, Polycycl. Aromat. Compd. **37** (2017), no. 5, 442–455.

[2] F. Amri, A. Niazi, A. Yazdanipour, and T. Momeni-Isfahani, *Computational modelling of the* 5α*-reductase inhibitors based on the MIA-SAR approach and design of new compounds*, J. Sci. Arts. **19** (2019), no. 4, 1011–1020.

[3] A. Antonelli, M. Cocchi, P. Fava, G. Foca, G.C. Franchini, D. Manzini, and A. Ulrici, *Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm*, Anal. Chim. Acta. **515** (2004), no. 1, 3–13.

[4] J.E. Antunes, M.P. Freitas, and R. Rittner, *Bioactivities of a series of phosphodiesterase type 5 (PDE-5) inhibitors as modelled by MIA-QSAR*, Eur. J. Med. Chem. **43** (2008), no. 8, 1632–1638.

[5] S. Asirvatham, B.V. Dhokchawle, and S.J. Tauro, *Quantitative structure activity relationships studies of non-steroidal anti-inflammatory drugs: A review*, Arab. J. Chem. **12** (2019), no. 8, 3948–3962.

[6] M. Ayaz, M. Ayaz, F. Ali, A. Saeed, A. Khurshid, G. Shabir, T. Ahmad, S.A. Raza Kazmi, and H.A. Khan, *Synthesis of symmetric bridged bis-pyrazolone based metal complex acid dyes and their applications on leather*, Journal of Fluorescence, **28** (2018), 1181–1193.

[7] B. Borah, K.D. Dwivedi, and L.R. Chowhan, *Applications of pyrazolone in multicomponent reactions for the synthesis of dihydropyrano [2, 3-c] pyrazoles and spiro-pyrano [2, 3-c] pyrazoles in aqueous medium*, ARKIVOC: Online J. Organ. Chem. **2021** (2021).

[8] A.P. da Silva, R.M. de Angelo, H. de Paula, K.M. Honório, and A.B. da Silva, *Drug design of new 5-HT 6 antagonists: a QSAR study of arylsulfonamide derivatives*, Struct. Chem. **31** (2020), 1585–1597.

[9] J.K. Dare and M.P. Freitas, *Different approaches to encode and model 3D information in a MIA-QSAR perspective*, Chemomet. Intell. Lab. Syst. **212** (2021), p. 104286.

[10] P.I.C. de Oliveira, Pa.H. de Santana Miranda, E.M.G. Lourenço, P.S. de Santana Nogueira Silverio, and E.G. Barbosa, *Planning new Trypanosoma cruzi CYP51 inhibitors using QSAR studies*, Molecular Diver. **25** (2021), 2219–2235.

[11] S. Dhawan, R. Narang, G.L. Khatik, H.K. Chopra, and S.K. Nayak, *Strategies for chemical synthesis of pyrazolone derivatives and their bio-significance*, J. Chem. Pharm. Res. **8** (2016), 969–981.

[12] V.F. Diaz, B. De Ketelaere, B. Aernouts, and W. Saeys, *Cost-efficient unsupervised sample selection for multivariate calibration*, Chemometr. Intell. Lab. Syst. **215** (2021), p. 104352.

[13] H.M. Duarte, J.S. Barigye, and P.M. Freitas, *Exploring MIA-QSARs' for antimalarial quinolon-4 (1H)-imines*, Comb. Chem. High Throughput Screen. **18** (2015), no. 2, 208–216.

[14] T. Eicher, S. Hauptmann, and A. Speicher, *The Chemistry of Heterocycles: Structures, Reactions, Synthesis, and Applications*, John Wiley & Sons, 2013.

[15] S. Elbahi, Y. Koubi, M. Boutalaka, H. Hajji, M. Chokrad, M.A. Ajana, and T. Lakhlifi, *The c-Src kinase inhibitors: 2D-QSAR study by multiple linear regression method*, RHAZES: Green Appl. Chem. **13** (2021), 29–42.

[16] K.H. Esbensen, P.L. Geladi, and H.F. Grahn, *Strategies for multivariate image regression*, Chemometr. Intell. Lab. Syst. **14** (1992), no. 1–3, 357–374.

[17] M.P. Freitas, *Multivariate image analysis applied to QSAR: Evaluation to a series of potential anxiolytic agents*, Chemometr. Intell. Lab. Syst. **91** (2008), no. 2, 173–176.

[18] M.P. Freitas, S.D. Brown, and J.A. Martins, *MIA-QSAR: A simple 2D image-based approach for quantitative structure–activity relationship analysis*, J. Mol. Struct. **738** (2005), no. 1–3, 149–154.

[19] N.B. Gallagher and L. Lawrence, *Introduction to hyperspectral and multivariate image analysis and principal components analysis for multivariate images*, https://www.researchgate.net/profile/Neal-Gallagher-2/publication/346731395_Introduction_to_Hyperspectral_and_Multivariate_Image_Analysis_and_Principal_Components_Analysis_for_Multivariate_Images/links/5fcfd0b245851568d14d60ee/Introduction-to-Hyperspectral-and-Multivariate-Image-Analysis-and-Principal-Components-Analysis-for-Multivariate-Images.pdf, (2020).

[20] Z. Garkani-Nejad and M. Poshteh-Shirani, *Application of multivariate image analysis in QSPR study of 13C chemical shifts of naphthalene derivatives: A comparative study*, Talanta **83** (2010), no. 1, 225–232.

[21] M. Goodarzi and M.P. Freitas, *Prediction of electrophoretic enantioseparation of aromatic amino acids/esters through MIA-QSPR*, Sep. Purific. Technol. **68** (2009), no. 3, 363–366.

[22] M. Goodarzi and M.P. Freitas, *On the use of PLS and N-PLS in MIA-QSAR: azole antifungals*, Chemometr. Intell. Lab. Syst. **96** (2009), no. 1, 59–62.

[23] N. Gouault, A. Cariou, J.-F. Cupif, P. de Aguiar Amaral, and M. David, *Fluorous supported synthesis of pyrazolone derivatives from allylic alcohols using a palladium-catalyzed strategy*, Química Nova **39** (2016), 825–831.

[24] R. Gozalbes, M. Jacewicz, R. Annand, K. Tsaioun, and A. Pineda-Lucena, *QSAR-based permeability model for drug-like compounds*, Bioorg. Med. Chem. **19** (2011), no. 8, 2615–2624.

[25] R. Guha, M. Ghosh, S. Kapri, S. Shaw, S. Mutsuddi, V. Bhateja, and R. Sarkar, *Deluge based genetic algorithm for feature selection*, Evol. Intell. **14** (2021), no. 2, 357–367.

[26] G. Huang, X. Chen, L. Li, X. Chen, L. Yuan, and W. Shi, *Domain adaptive partial least squares regression*, Chemometr. Intell. Lab. Syst. **201** (2020), p. 103986.

[27] J. Jaumot and C. Bedia, *Mass spectrometry imaging: chemometric data analysis*, S. Brown, R. Tauler and B. Walczak, (eds.) Comprehensive chemometrics, 2nd eds., Elsevier, Oxford, 2020.

[28] C. Ju, T. Cai, W. Li, G. Sun, C. Lei, X. Di, and X. Man, *Improving forest burn severity estimations with partial least squares regression and orthogonal signal correction methods in Daxing'an Mountains*, China J. For. Res. **32** (2021), no. 3, 1157–1165.

[29] R.W. Kennard and L.A. Stone, *Computer aided design of experiments*, Technometrics **11** (1969), no. 1, 137–148.

[30] A. Krishnan, L.J. Williams, A.R. McIntosh, and H. Abdi, *Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review*, Neuroimage **56** (2011), no. 2, 455–475.

[31] R. Kshatriya, P. Shelke, S. Mali, G. Yashwantrao, A. Pratap, and S. Saha, *Synthesis and evaluation of anticancer activity of pyrazolone appended triarylmethanes (TRAMs)*, ChemistrySelect. **6** (2021), no. 24, 6230–6239.

[32] S.A. Kulkarni and T. Madhavan, *Hologram quantitative structure activity relationship analysis of JNK antagonists*, Journal of the Chosun Natural Science, **8** (2015), no. 2, 81–88.

[33] G. Neelima, T. Asha, S. Rukmini, G.K. Brundha, and S. Rajasekhar, *Recent trend on known pyrazolone endowed derivatives: An overview*, Int. J. Current Pharmac. Res. **14** (2022), no. 4, 28–35.

[34] J.C. Noordam, W.H. van den Broek, and L.M. Buydens, *Detection and classification of latent defects and diseases on raw French fries with multispectral imaging*, J. Sci. Food Agric. **85** (2005), no. 13, 2249–2259.

[35] I. Pardoe, *Applied Regression Modeling*, John Wiley & Sons, 2020.

[36] F.A. Pasha, M. Muddassar, M.M. Neaz, and S.J. Cho, *Pharmacophore and docking-based combined in-silico study of KDR inhibitors*, J. Mol. Graph Model. **28** (2009), no. 1, 54–61.

[37] M.D. Peris-Díaz and A. Krężel, *A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry*, TrAC Trends Anal. Chem. **135** (2021), p. 116157.

[38] N. Polson, V. Sokolov, and J. Xu, *Deep learning partial least squares*, arXiv preprint arXiv:2106.14085, (2021).

[39] S. Prabhakaran and S. Velmathi, *Domino synthesis of bis phenyl pyrazolone piperidinium salt derivatives*, Chem. Select **6** (2021), no. 33, 8696–8699.

[40] R. Rivera, *Principles of Managerial Statistics and Data Science*, John Wiley & Sons, 2020.

[41] L. Saghaie, M. Shahlaei, A. Madadkar-Sobhani, and A. Fassihi, *Application of partial least squares and radial basis function neural networks in multivariate imaging analysis-quantitative structure activity relationship: Study of cyclin dependent kinase 4 inhibitors*, J. Mol. Graph Model. **29** (2010), no. 4, 518–528.

[42] M.A. Shabaan, A.M. Kamal, S.I. Faggal, A.E. Elsahar, and K.O. Mohamed, *Synthesis and biological evaluation of pyrazolone analogues as potential anti-inflammatory agents targeting cyclooxygenases and 5-lipoxygenase*, Arch Pharm. **353** (2020), no. 4, 1900308.

[43] N. Shakour, F. Hadizadeh, P. Kesharwani, and A. Sahebkar, *3D-QSAR studies of 1, 2, 4-Oxadiazole derivatives as sortase a inhibitors*, BioMed. Res. Int. **2021** (2021), no. 1, p. 6380336.

[44] J. Tessier, C. Duchesne, C. Gauthier, and G. Dufour, *Estimation of alumina content of anode cover materials using multivariate image analysis techniques*, Chem. Eng. Sci. **63** (2008), no. 5, 1370–1380.

[45] J. Too and A.R. Abdullah, *A new and fast rival genetic algorithm for feature selection*, J. Supercomput. **77** (2021), no. 3, 2844–2874.

[46] R.J. Weekes and C.S. Hawes, *Synthesis, coordination chemistry and photophysical properties of naphtho-fused pyrazole ligands*, Cryst. Eng. Comm. **21** (2019), no. 34, 5152–5163.

[47] M.R. Wester, J.K. Yano, A. Schoch, C. Yang, K.J. Griffin, C.D. Stout, and E.F. Johnson, *The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-Å resolution*, J. Biol. Chem. **279** (2004), no. 34, 35630–35637.

[48] S. Wold, H. Antti, F. Lindgren, and J. Öhman, *Orthogonal signal correction of near-infrared spectra*, Chemometr. Intell. Lab. Syst. **44** (1998), no. 1–2, 175–185.

[49] Z. Zhao, X. Dai, C. Li, X. Wang, J. Tian, Y. Feng, J. Xie, C. Ma, Z. Nie, P. Fan, and M. Qian, *Pyrazolone structural motif in medicinal chemistry: Retrospect and prospect*, Eu. J. Medic. Chem. **186** (2020), 111893.

[50] X. Zou, Z. Hao, R. Yi, L. Guo, M. Shen, X. Li, Z. Wang, X. Zeng, and Y. Lu, *Quantitative analysis of soil by laser-induced breakdown spectroscopy using genetic algorithm-partial least squares*, Chin. J. Anal. Chem. **2015** (2015), 181–186.