Egypt. J. Chem. **Vol. 65,** No. 9 pp. 763 - 775 (2022)

Egyptian Journal of Chemistry

http://ejchem.journals.ekb.eg/

Quantitative Structure Retention/Activity Relationship Study of some Dialkoxybenzamide Phosphodiesterase-4B Inhibitors

Nouran K. Al-Khalil Bek [1], Asmaa A. El-Zaher [1], Marwa A. Fouad [1,2 *], Mohammad Sammany[3], and Mohamed K. El-Ashrey [1]

[1] *Pharmaceutical Chemistry Department, Faculty of Pharmacy, Cairo University, Kasr El-Eini Street, P.O. Box 11562, Cairo, Egypt*

[2] *Pharmaceutical Chemistry Department, School of Pharmacy, Newgiza University, Newgiza, km 22 Cairo–Alexandria Desert Road, Cairo, Egypt*

[3] *Pharmacy Practice Department, Faculty of Pharmacy, Heliopolis University for Sustainable Development, 3 Cairo Belbes Desert Road, P.O. Box 2834, Cairo, Egypt*

Abstract

Recently synthesized dialkoxybenzamide derivatives, structurally analogous to Roflumilast with more selectivity against phosphodiesterase 4B, were highlighted in the present work. To find optimum chromatographic conditions for the elution of these compounds, a central composite experimental design was carried out by varying the stationary phase type as a categorical factor and mobile phase composition including the percentage of acetonitrile and the pH of the buffered water as continuous factors; the obtained retention times were utilized in the quantitative structure retention relationship (QSRR) studies. Furthermore, quantitative structure activity relationship (QSAR) studies for these promising compounds were performed. QSAR and QSRR models were built by different techniques namely multiple linear regression (MLR), principal component regression (PCR) for linear modelling and Principal Component-Artificial Neural Networks (PC-ANN) for nonlinear modelling. Internal validation (leave many out method), and external validation were used to evaluate the performance of the generated models. Depending on the calculated statistical parameters, PC-ANN QSAR model showed the best predictive power for the biological activities of the test set compounds. Whereas MLR technique was more suitable to build QSRR model, this model can help in understanding how the chemical structure and the lipophilicity of the compounds can affect the retention time and chromatographic behavior.

Keywords: Dialkoxybenzamide; PDE-4B inhibitors; Quantitative Structure-Activity Relationship; Quantitative Structure--Retention Relationship; Principal Component Analysis

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is defined as a long-term lung disease marked by progressive airflow limitations [1] due to abnormal chronic inflammation in the airways and lung tissues. It was classified as the third leading cause of death worldwide in 2016 and the fourth leading cause of death in the United States [2]. The available medications for the management of COPD include β2 agonists, anticholinergic, methylxanthines and inhaled corticosteroids, but all these medications can only reduce and alleviate the symptoms without disease-modifying properties [3].Therefore, there is a persistent need to develop new therapeutics for this respiratory disease.

Roflumilast (ROF), Figure 1, the PDE-4 inhibitor was launched in 2012 in the USA for COPD treatment, but this drug has many side effects such as headache, weight loss and gastrointestinal disturbance [4].

PDE-4 inhibitors prevent the hydrolysis of cAMP (cyclic adenosine monophosphate) to AMP (adenosine monophosphate), this inhibition will lead to protein kinase A activation, reduction of inflammatory mediators release, reduction of cytokine release, reduction of fibrotic lung remodeling and oxidative stress [5]. PDE-4 family encompasses four subtypes PDE-4A, PDE-4B, PDE-4C, and PDE-4D. Researchers believed that the PDE-4B subtype can open new doors in COPD treatment, due to its role in the inflammatory process and the selective inhibition

*Corresponding author e-mail: marwa.fouad@pharma.cu.edu.eg.; (Marwa A. Fouad).

Receive Date: 15 January 2022, Revise Date: 04 February 2022, Accept Date: 07 February 2022

DOI: 10.21608/EJCHEM.2022.115690.5263

of this subtype does not cause emesis which is highly related to PDE-4D inhibition [6,7]. This study will focus on newly synthesized dialkoxybenzamide derivatives Figure 1, that are structurally analogous to the known phosphodiesterase (PDE) inhibitor Roflumilast (ROF) with more selective properties against PDE-4B enzyme, these compounds could be promising and good candidates for the treatment of chronic obstructive pulmonary disease (COPD) [8,9].

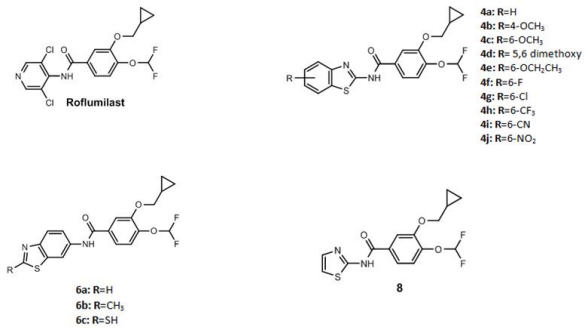

Figure 1. Structures of Roflumilast and the reported dialkoxybenzamide PDE4B inhibitors

QSAR and QSRR approaches play essential role in drug discovery field by reducing time and cost to get lead drugs. QSAR depends on a mathematical relationship between the chemical structure of the compound represented by descriptors (independent factors) and the activity (dependent factors). QSAR approach helps in identifying hits from a large library of compounds then these hit molecules can be further studied and activity experiments can be started [10]. Hence QSAR studies can avoid synthesis and testing large number of nonrelevant compounds. QSRR represents a specific type of quantitative structure-property relationship QSPR. QSRR model can mathematically relate the chemical structure of the compound represented by descriptors (independent factors) to the chromatographic retention (dependent factors) [10,11]. QSRR models can be utilized in: understanding the chromatographic behavior of molecules and separation mechanism in a specific chromatographic system, prediction the retention of new analytes, determination the most molecular descriptors that impact in the retention prediction and recently QSRR could predict analytes lipophilicity being highly related to the retention in reversed phase chromatographic system [11,12] .

Generating robust QSAR/QSRR models depends mainly on the selection of the most relevant molecular descriptors; noisy and irrelevant descriptors should be removed, this step can be performed by using variable selection techniques or feature extraction methods [13].

Feature extraction appears to be a promising technique specially in cases with a huge number of descriptors and the presence of multicollinearity problem, in this case, variable selection methods may drop in determination of the optimal variables [14]. Feature extraction depends on reducing dimensionality from high dimensional data to a lower dimensional feature space. One of the most common feature extraction approaches is Principal Component Analysis (PCA). PCA can produce a new matrix with new orthogonal variables (PCs) from the original one by dimension reduction, this reduction can simplify the analysis process [15].

After the selection step is performed, meaningful variables will be correlated to the target property by linear methods such as Multi Linear Regression (MLR), Principal Component Regression (PCR) or Partial Least Square (PLS) or by nonlinear methods such as artificial neural networks (ANN), Support Vector Machine (SVM) [16]. ANN modelling is one of the most attractive technique due to its simplicity and it is not so complicated comparatively to other machine learning techniques [17].

The aim of this research is to continue and complete a comprehensive study of a promising series of fourteen dialkoxybenzamide PDE4B inhibitors analogous to ROF that were previously synthesized [8] and to build robust QSAR and QSRR models by investigating three of the most widely applied and powerful techniques MLR, PCR and Principal component-Artificial Neural Networks PC-ANN. The study was concerned with PCA application as a data mining approach to overcome the redundant and multicollinear descriptors problem.

It is also worth to mention that central composite experimental design (CCD) was used to optimize the RP-HPLC elution of the compounds and achieve the desired chromatographic separation with sharp and symmetric peaks. In the contrast with traditional development of HPLC methods which is based on trial and error, this approach can avoid the large consumption of chemicals, time and efforts by modelling the data and using limited number of experiments. The retention times obtained from the optimized separation method will be used in the building of QSRR model.

2. Experimental

2.1 Chemicals, solvents, instruments and chromatographic analysis:

Fourteen compounds were in house synthesized according to the reported method, Figure 1 [8], roflumilast was purchased from Shanxi Jinjin chemicals co. ltd, Shanxi, China. HPLC-grade acetonitrile and methanol were purchased from sigma Aldrich, Germany. HPLC bi-distilled water was produced in house. Potassium dihydrogen phosphate from Sigma-Aldrich, Germany. Formic acid was

supplied from El Nasr Pharmaceutical Chemicals Co., Cairo, Egypt. Phosphoric acid and sodium hydroxide were used for adjusting pH of the mobile phases. pH meter Jenway3510, Essex-UK, England. Sonicator (power sonic).

Stock solutions were prepared by dissolving 5 mg of each compound in a mixture of (50:50) acetonitrile: methanol and diluted to 10 mL in the same mixture. The stock solutions were stored in 4 °C. One mL of each stock solution was diluted to 10 mL with the mobile phase to attain sample concentrations 50 µg mL^{-1}. Retention times of the compounds were recorded using Agilent 1100 series isocratic system, the used columns in this study were ZORBAX C8 (4.6 × 150 mm, 5µm) and C18 (4.6 × 250, 5µm). Injection volume was 10 µl. The mobile phase was pumped at a flow rate was 1 mL min^{-1} using UV detection at 205 nm. The UV detection of the compounds was at 205 nm. All the injections were carried out at room temperature.

2.2 Design of experiment DOE

Design-Expert7.0.0 (Stat-Ease, Inc., Minneapolis, USA) software, response surface methodology, central composite design (CCD) tool, face centered type was adopted for the determination of the optimum RP-HPLC analytical conditions, the three used factors were: pH, organic modifier ratio and columns. The pH of the mobile phase was varied to 3 different levels: 3, 5 and 7. Three different ratios of buffered water: acetonitrile (organic modifier): 70:30, 60:40 and 50:50 were used. Two types of columns were used either C8 or C18 column. Eighteen designed experiments were performed in random order and the retention time of each compound was recorded and used as a response, then the variance of the retention times at each run was calculated (Table 1) and utilized to compare between the runs.

2.3 QSAR and QSRR modelling

2.3.1 Calculation of molecular descriptors

SMILES (Simplified Molecular Input Line Entry System) of the studied structures were imported to Molecular Operating Environment (MOE, 10.2008) software [18] then inverted to 3D structures by MOE builder tool. The molecules with minimized energy geometry were optimized by using Hamiltonian MMFF94x forcefield until RMS gradient reached 0.05 $kcal.mol^{-1}Å^{-1}$ and partial charges were calculated. Finally, the descriptors for each compound were calculated. The number of these calculated molecular descriptors was 334.

2.3.2 Variable selection

Most influential variables were obtained by Stepwise Selection as a traditional variable selection method or by principal component analysis (PCA) as a feature reduction method. PCA was carried out by SPSS 25, dimension reduction toolbox. Before the application of PCA analysis, preprocessing data step was performed by excluding descriptors with constant values and descriptors with standard deviation equals 0, after that, 282 descriptors were obtained. These descriptors were scaled by standardization procedure and z-score was calculated ($z-score=\frac{xi-\bar{x}i}{\sigma i}$ where xi: nonstandardized value of the descriptor, $\bar{x}i$: mean value and σi: standard deviation) for each descriptor to ensure that descriptors with larger range will not take over smaller ones. The extraction of PCs based on eigen values greater than 1.

Table 1: Experimental runs designed by CCD

Run	Point space type	Factor1 A: pH	Factor 2 B: Organic modifier%	Factor 3 C: Column type	Response variance
1	Center	5	40	C18	2.569
2	Factorial	3	50	C8	3.106
3	Axial	5	30	C18	0.803
4	Axial	3	40	C8	1.411
5	Axial	7	40	C8	0.799
6	Factorial	7	30	C8	1.788
7	Factorial	3	30	C18	0.762
8	Axial	7	40	C18	0.640
9	Factorial	7	50	C18	0.124
10	Axial	5	50	C8	0.906
11	Factorial	7	50	C8	0.107
12	Factorial	3	30	C8	0.369
13	Axial	5	30	C8	0.951
14	Factorial	7	30	C18	0.200
15	Center	5	40	C18	2.569
16	Center	5	40	C8	0.727
17	Axial	5	50	C18	2.857
18	Center	5	40	C8	0.727

2.3.3 QSAR and QSRR models development

QSAR and QSRR models were built by SS-MLR, PCR as linear methods and PC-ANN as a nonlinear method. Multilayer perceptron neural network (executed on NeuroSolutions version 5) trained with back-propagation algorithm, the activation function was $\tanh^{-1}$, the training data set was shuffled, weight decay parameter was optimized and used to avoid overfitting and improve generalization. CVE (Cross Validation Error) was monitored during training process to reach the minimum value and the training of the NN stopped directly when this value begins to increase.

The architecture of NN consists of three layers: input layer which receive the vital independent variables and send them to the hidden layer, the number of neurons in this layer should be optimized, then these neurons will apply transformations to the input data and send them to the output layer to handle and calculate the response.

Model validation, the critical step which can assess the robustness and efficacy of the generated QSAR/QSRR models, was evaluated by Leave Many Out Cross Validation (LMO-CV) method. In LMO-CV method, each time 20% of the compounds were omitted randomly from the original training set, then the built model was used to predict the activity/retention time for the rest removed compounds. LMO-CV is more robust and reliable method to evaluate models with small data set [19] as in our case, we have only 15 compounds.

Defining Applicability Domain (AD) of QSAR/QSRR models is very important step, there is no model predictable and suitable for entire chemicals, the predictability of a model is only for new chemicals that fall in its domain[20]. AD can be investigated by utilizing one of the following methods: probability density distribution method, range-based method and distance-based method, this method commonly uses leverage approach, which has wide application in QSAR/QSRR studies. Warning leverage or threshold value h* can be calculated by following equation:

$$h^* = \frac{3(k+1)}{n} \qquad equation\ 1$$

k: the number of descriptors in the model
n: the number of training set compounds
to calculate leverage value of each compound:

$$h_i = x_i^T (X^T X)^{-1}\ x_i \qquad equation\ 2$$

x_i: the descriptor row vector calculated for the ith compound
X: the matrix of descriptors calculated for the training set

The leverage values for the studied compounds were calculated by Enalos node in KNIME [21] analytics platform (version 4.2.3).Williams plot, which display standardized residuals versus leverage values h, was applied. The squared area established within ± 3 standard deviations and a threshold leverage value (h*=1). Prediction considered unreliable for compounds with a high leverage value (hi>h*) [22].

3. Results and discussion

3.1 Optimization of chromatographical conditions by central composite design

Response surface methodology (RSM) is a widely used and powerful technique of design of experiments (DOE). It depends on mathematical and statistical relationships between factors (inputs) and target response. Central composite design (CCD) and Box-Behnken are the major tools of RSM.

CCD is an effective design can fit second-order polynomial models and requires only a minimum number of experiments for modeling [23, 24]. CCD building depends on three types of points [25, 26]: **factorial points** represent +1,−1 levels of factor. **Star (axial) points** this points location depends on Alpha value. Alpha value can also determine the type of the design, rotatable when (Alpha=1.41), orthogonal (Alpha=1.26) or face centered (Alpha=1) this design is preferred because it ensures that the axial point location in the factorial space [26]. **Center points** represent replicate terms and can assess the experimental error.

Central composite design (CCD) was our suitable choice as we have two numeric factors and one categorical, while in Box-Behnken method at least three factors should be numeric (continuous).

The three independent factors are: **Factor A**: pH of the aqueous part of the mobile phase, the studied levels were at 3, 5 and 7. Factor A is considered to be a numeric factor. **Factor B:** the organic modifier (acetonitrile) ratio also categorized as a numeric factor, was evaluated at three levels which were 30, 40 and 50%. **Factor C:** the column type used was either C8 or C18 as a categorical factor. Alpha value was selected equal to 1 and face centered design was constructed; the response was the natural logarithm (ln) of variance of retention times.

The number of designed experiments was 18, calculated by the following equation:

$$N = 2^k + 2k + c \qquad equation\ 3$$

N: the number of experiments
k: the number of factors
c: the number of replicates

Analysis of variation (ANOVA) was applied to check the significantly of the model and the included terms, Table 2. The obtained model was quadratic. The p-value of the obtained model was 0.0001, which means that the model is appropriate and significant. Model reduction was done to get rid of the nonsignificant terms; this step can be performed by different techniques such as backward, stepwise and forward selection. Backward was the preferred and the chosen method being more robust than the others. Herein, the nonsignificant term A^2 was removed. Finally, the obtained model after reduction was as follows:

$$\ln(variance) = 0.44 - 1.08A + 0.38B - 0.36C - 1.24AB + 0.69AC - 0.5BC - 0.49B^2 \quad equation\ 4$$

A: pH
B: organic modifier
C: column type

The model F-value 15.57 implies the model is significant. Values of Prob>F less than 0.05 indicates model terms are significant, in this case A, B, C, AB, AC, BC, B^2 are the significant model terms. The model correlation coefficient R^2 , Adj. R^2 and Pred. R^2 values were: 0.915, 0.8571 and 0.6776 respectively. Pred. R^2 value is in reasonable agreement with Adj. R^2 and accepted because the difference between them not greater than 0.2 or 20% [27]. Adaquate precision (Signal to Noise ratio) which refine the error between the predicted values and the average prediction was 13.207 this indicates adequate signal of the model. As R^2 and Adj. R^2 values were more than 0.7 and S/N more than 4, so this model can be used to navigate the design space and will be able to find the optimum chromatographic conditions for the studied compounds.

It can be concluded from Figure 2, the experimental and predicted values appear to be in a good agreement, Figure 2A. From the normal probability plot of residuals, it can be clearly noticed that the residuals are placed on a straight line and a normal distribution appears, Figure 2B. The plot of residuals vs the predicted response, Figure 2C, can prove the suitability and the fitness of the model for the used data. 3-D surface plots showed the impact of the factors on the response. The variance increased whenever pH values reduced and organic modifier increased, Figure 3. Perturbation plot, Figure 4, was employed to show the effect of the factors on the response. The curvature shape of factor B (organic modifier ratio) indicate that ln(variance) of the retention times could be more sensitive to this factor more than factor A (pH).

Finally, numerical optimization was performed under the following constraints: pH was in range, organic modifier was in range, the selected column was C8 as it showed more symmetric and sharper peaks than C18, the response ln(variance) was maximized. Different solutions were gotten after numerical optimization and maximization of the response and it was found that Run 2 (pH=3, organic modifier ratio = 50% using C8 column) is the optimum run with 0.988 desirability, Figure 5 and then used in QSRR studies.

Table 2: ANOVA for response surface reduced quadratic model

Source	Sum of squares	Df	Mean square	F-value	P-value		
Model	15.85	7	2.26	15.57	0.0001		
A-pH	9.02	1	9.02	62.01	<0.0001		
B-Organic modifier	1.37	1	1.37	9.44	0.0118		
C-Column	1.97	1	1.97	13.56	0.0042		
AB	8.93	1	8.93	61.39	<0.0001		
AC	3.66	1	3.66	25.17	0.0005		
BC	2.41	1	2.41	16.57	0.0022		
B^2	1.00	1	1.00	6.67	0.0256		
Residual	1.45	10	0.15				
Pure error	0.000	2.00	0.000				
Cor Total	17.30	17					
Std. Dev.	0.38						
Mean	-0.22						
PRESS	5.58						
R^2	0.9159	**Adj. R^2**	0.8571	**Pred.R^2**	0.6776	**Adeq. Precision**	13.207

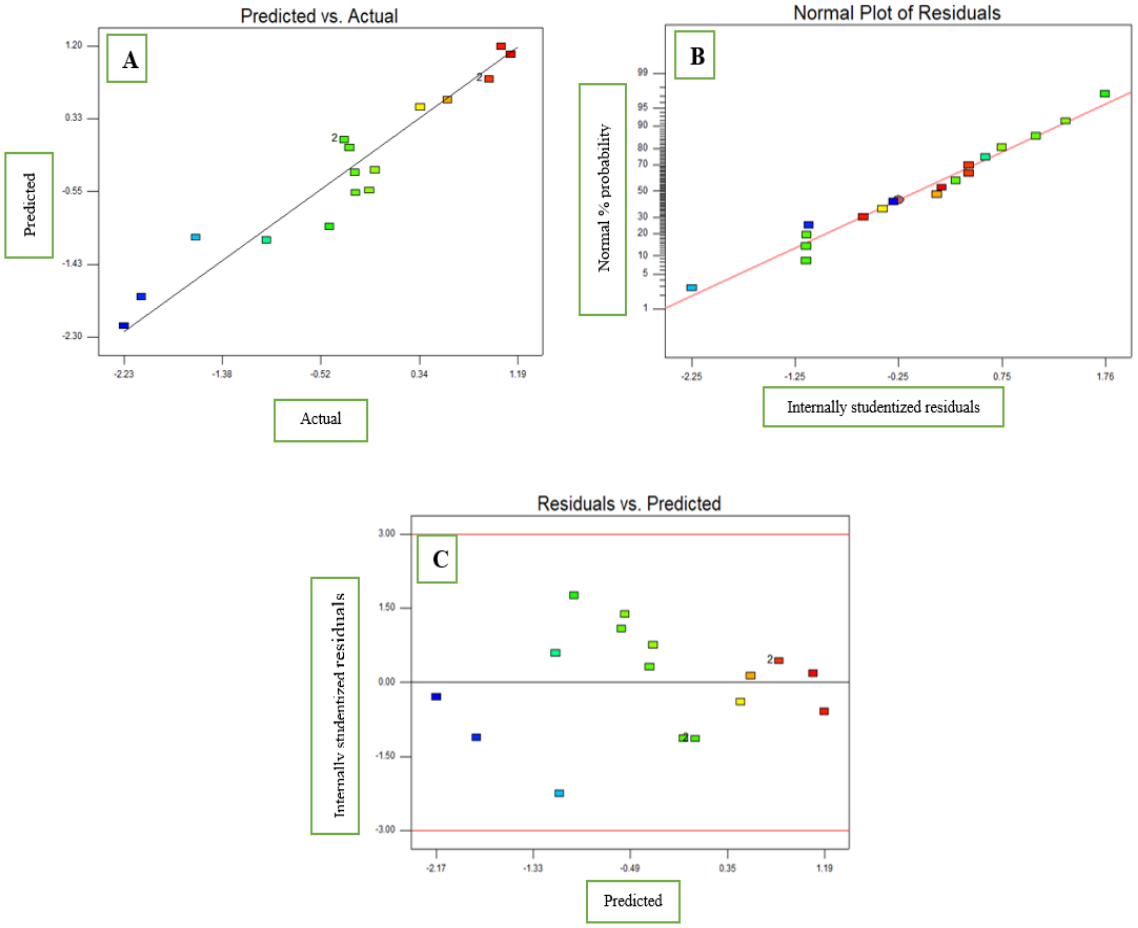

Figure 2. A: plot of actual vs predicted response **B:** normal probability plot **C**: plot of residuals vs predicted response

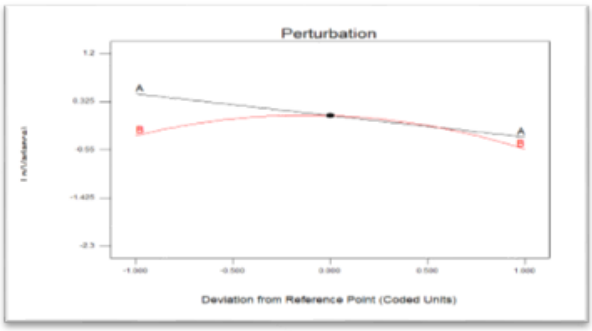

Figure 3. Perturbation plot of ln(variance)

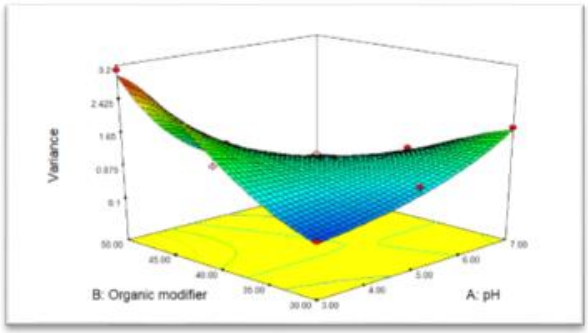

Figure 4. 3D surface plot of variance as function of pH and organic modifier ratio

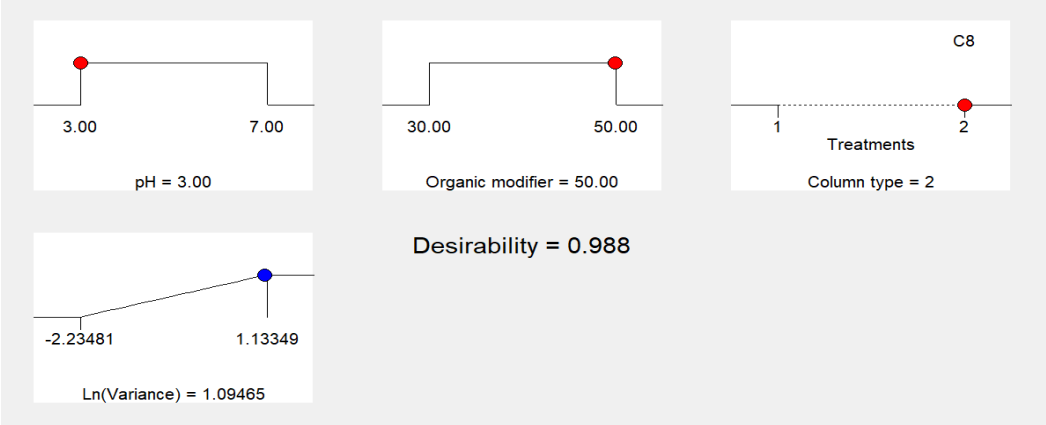

Figure 5. predicted solution for maximum ln (variance) of retention times by CCD numerical optimization

3.2 QSAR/QSRR models:

3.2.1 Linear models:

3.2.1.1 Stepwise Selection-MLR

Stepwise selection (which combines forward selection and backward elimination) was run on the set of calculated descriptors after excluding descriptors with constant values, descriptors with standard deviation equals 0 and descriptors with high correlation > 0.95 one of them was deleted.

QSRR study: logarithmic values of retention times (log t_R) were utilized as dependent variables. on the other hand, the selected descriptors, which represent the independent variables in the model, were: Q-VSA-FPNEG (the fractional negative polar van der waals surface area) with positive correlation coefficient, SlogP-VSA6 (log of the octanol/water partition coefficient calculated from the structure including implicit hydrogens). VSURF-HL2 (represent molecular hydrophilic-lipophilic balance) this descriptor inversely related to the retention. The standardized regression coefficients which reflect the importance and the significance of each descriptor in the model [12] were 0.708, 0.442 and 0.417 for Q-VSA-FPNEG, SlogP-VSA6 and VSURF- HL2 respectively, Figure 6A.

$$log\ t_R = 0.205 + 0.026(\mathrm{SlogP} - \mathrm{VSA6}) + 3.363(QVSA - FPNEG) - 5.826\ (VSURF - HL2) \quad equation\ 5$$

(R^2=0.82, R^2_{adj}=0.767, Q^2_{LMO}=0.783, R^2_{pred}= 0.83)

From the previous QSRR model, we have a sight on the factors that can affect chromatographic behavior. It is clearly apparent that the QSRR model was affected by the lipophilicity and the charges of the eluted compounds. The influence of functional groups such as presence of halogen could increase the lipophilicity and thus can cause increasing of the retention of the compound in reversed phase chromatography as seen in compound **4h** which has the highest retention time. The methoxy group could increase the polarity of the compound so reduction in the retention is noticed (Table 4).

QSAR study: pIC_{50} (-log scale of $1/IC_{50}$ in molar units) values were used in the model development (Table 5). The most influential descriptors were PM3-HOMO (a quantum-chemical descriptor represent the energy of the highest occupied molecular orbital and calculated by using PM3 Hamiltonian) with a positive coefficient, when this descriptor value increase this means the ability of the compound to donate electrons will increase and the nucleophilic reactions could occur easily [28,29]. From the obtained model it can be revealed that molecules with lower electron accepting properties could have higher activity. BCUT-SMR2 (BCUT descriptors use the atomic contribution to the molar refractivity calculated according to Wildman and Crippen SMR method) and FASA (water accessible surface area of all atoms with positive/negative partial charge). The importance of each descriptor in the model were PM3_HOMO > FASA > BCUT-SMR2, Figure 6B.

$$pIC50 = 8.086 + 4.33(\mathrm{BCUT} - \mathrm{SMR2}) + 0.596(PM3 - HOMO) + 5.9\ (FASA)\ equation\ 6$$

(R^2=0.83, R^2_{adj}=0.765, Q^2_{LMO}=0.745, R^2_{pred}= 0.7)

The robustness of the generated QSAR/QSRR models by SS-MLR was estimated by internal validation, leave-many out (L20%O) cross-validation method. The cross-validation correlation coefficient (Q^2 or R^2cv) values for each model were more than 0.5 [30]. Root mean-square error $RMSE_{CV}$ and $RMSE_{pred}$ values were low and did not exceed 0.5 (Table 6&7). External validation was carried out by prediction of the test set (non-seen before compounds), R^2_{pred} (it also can be termed $Q^2_{(F1)}$) value was more than 0.6, this indicates the predictability of the model, also additional statistical parameters for

external predictability were evaluated, such as the value of $\frac{R^2-R_0^2}{R^2}$ where R^2 and R_0^2 obtained from the plot of predicted values vs actual without and with zero intercept respectively, $\frac{R^2-\acute{R}_0^2}{R^2}$ where $\acute{R}_0^2$ obtained from the plot of actual values vs predicted with zero intercept, k the slope of predicted values vs actual and $\acute{k}$ the slope of actual values vs predicted (Table 6&7).

In Figure 7A, the plot of predicted log t_R by the generated QSRR model versus experimental values were in concurrence and showed significant correlation. In Figure 7B, it can be noticed that the plot of calculated residuals versus experimental values of log t_R was distributed around zero, this also can be an indicator of the predictive power of the generated model.

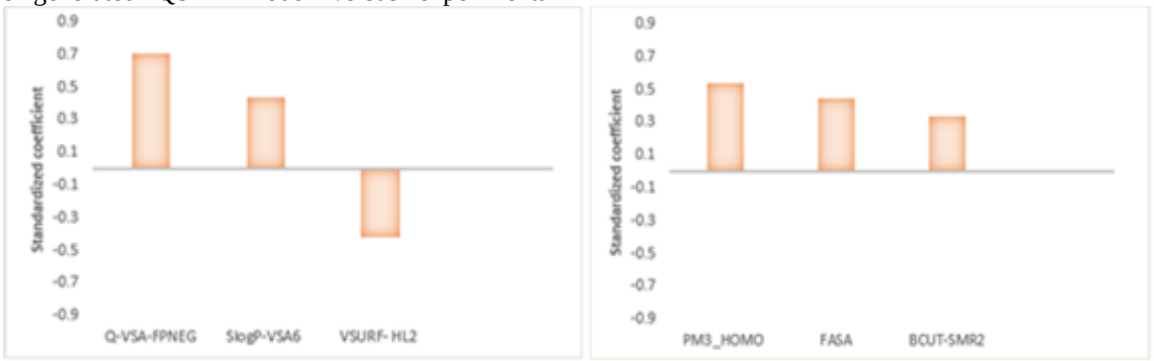

Figure 6. A: Importance of the contributed descriptors in MLR QSRR model, **B:** Importance of the contributed descriptors in MLR QSAR model

Table 4: experimental and predicted values of retention times in chromatographic conditions C8, pH=3 and 50:50 % (buffered water: ACN) and residuals

Compound	Experimental log t_R	Predicted log t_R					
		MLR	Res.	**PCR**	Res.	**PC-ANN**	Res.
4a	0.52	0.31	0.21	0.54	-0.02	0.41	0.11
6a	0.51	0.51	0	0.6	-0.09	0.43	0.08
Rof	0.63	0.71	-0.08	0.61	0.02	0.61	0.02
4b	0.52	0.45	0.07	0.53	-0.01	0.58	-0.06
4d	0.30	0.35	-0.05	0.3	0	0.38	-0.08
6b	0.49	0.6	-0.11	0.65	-0.16	0.58	-0.09
4i	0.50	0.52	-0.02	0.58	-0.08	0.56	-0.06
4g	0.54	0.49	0.05	0.54	0	0.58	-0.04
8	0.26	0.34	-0.08	0.4	-0.14	0.29	-0.03
4h	0.90	0.91	-0.01	0.7	0.2	0.8	0.1
4j	0.71	0.69	0.02	0.73	-0.02	0.6	0.11
4e	0.46	0.43	0.03	0.51	-0.05	0.4	0.06
Test set							
6c	0.52	0.427	0.12	0.57	-0.05	0.6	-0.08
4f	0.72	0.739	-0.019	0.69	0.03	0.6	0.12
4c	0.37	0.38	-0.01	0.49	-0.12	0.41	-0.04

Table 5: observed and predicted values of pIC_{50} and residuals

Compound	*Observed pIC_{50}	Predicted pIC_{50}					
		MLR	Res.	**PCR**	Res.	**PC-ANN**	Res.
4a	9.24	8.97	0.27	9	0.24	9	0.24
6a	9.17	9.17	0	8.93	0.24	9.14	0.03
6c	9.12	8.98	0.14	9.02	0.1	9	0.12
4b	8.97	9	-0.03	9	-0.03	9.03	-0.06
4d	8.93	8.7	0.23	9	-0.07	8.83	0.1
4i	8.93	8.74	0.19	8.88	0.05	8.7	0.23
4g	8.89	9	-0.11	8.9	-0.01	8.87	0.02
8	8.79	8.73	0.06	9	-0.21	8.9	-0.11
4f	8.72	8.85	-0.13	8.83	-0.11	8.75	-0.03
4h	8.25	8	0.25	8.5	-0.25	8.37	-0.12
4j	8.05	8.16	-0.11	8.3	-0.25	8.3	-0.25
4c	8.64	8.9	-0.26	9	-0.36	8.94	-0.3
Test set							
6b	9	8.89	0.11	9.02	-0.02	9.09	-0.09
ROF.	8.9	8.744	0.156	8.97	-0.07	8.89	0.01
4e	8.53	8.55	-0.02	8.7	-0.17	8.64	-0.11

*Original IC_{50}[nM] values can be found in the previous study (reference 8)

Table 6: Statistical summary to evaluate QSRR models performance

QSRR model				
Statistical parameter	**MLR**	**PCR**	**PC-ANN**	**Accepted range**
$R^2_{CV}(Q^2_{LMO}) = 1 - \frac{\sum(y - \hat{y}_{LMO})^2}{\sum(y - \bar{y})^2}$	0.78	0.66	0.78	$Q^2_{LMO} > 0.6$ (internal validation)
$R^2_{pred}(Q^2_{(F1)}) = 1 - \frac{\sum(y_{test} - \hat{y}_{test})^2}{\sum(y_{test} - \bar{y}_{training})^2}$	0.83	0.7	0.64	$R^2_{pred} > 0.6$ (External validation)
$RMSE_{CV} = \sqrt{\frac{\sum(\hat{y}_{LMO} - y)^2}{N_{training}}}$	0.082	0.092	0.075	Lower values preferable
$RMSE_{pred} = \sqrt{\frac{\sum(\hat{y}_{test} - y_{test})^2}{N_{test}}}$	0.07	0.077	0.086	Lower values preferable
$\frac{R^2_{test} - R^2_0}{R^2_{test}}$	-0.099	0.012	-0.4	<0.1
$\frac{R^2_{test} - \acute{R}^2_0}{R^2_{test}}$	-0.099	0.013	-0.4	<0.1
k	1.025	0.93	1.008	$0.85 \leq k \leq 1.15$
$\acute{k}$	0.966	1.052	0.97	$0.85 \leq \acute{k} \leq 1.15$

y : experimental value for training set, $\bar{y}$: the average of the experimental value for training set, $\hat{y}_{LMO}$ leave many out cross-validation predicted value for training set, y_{test}:experimental value of the external test set, $\hat{y}_{test}$predicted value for the external test set. N: number of samples, R^2_0 obtained from the plot of predicted values vs actual with zero intercept, $\acute{R}^2_0$ obtained from the plot of actual values vs predicted with zero intercept, k the slope of predicted values vs actual and $\acute{k}$ the slope of actual values vs predicted with zero intercept.

Table 7: Statistical parameters to evaluate QSAR model performance

QSAR model				
Statistical parameter	**MLR**	**PCR**	**PC-ANN**	**Accepted range**
$R^2_{CV}(Q^2_{LMO}) = 1 - \frac{\sum(y - \hat{y}_{LMO})^2}{\sum(y - \bar{y})^2}$	0.745	0.68	0.77	$Q^2_{LMO} > 0.6$ (internal validation)
$R^2_{pred}(Q^2_{(F1)}) = 1 - \frac{\sum(y_{test} - \hat{y}_{test})^2}{\sum(y_{test} - \bar{y}_{training})^2}$	0.7	0.72	0.835	$R^2_{pred} > 0.6$ (External validation)
$RMSE_{CV} = \sqrt{\frac{\sum(\hat{y}_{LMO} - y)^2}{N_{training}}}$	0.172	0.192	0.163	Lower values preferable

$RMSE_{pred} = \sqrt{\frac{\sum(\hat{y}_{test} - y_{test})^2}{N_{test}}}$	0.11	0.106	0.082	Lower values preferable
$\frac{R^2_{test} - R^2_0}{R^2_{test}}$	-0.063	-0.004	-0.068	<0.1
$\frac{R^2_{test} - \acute{R}^2_0}{R^2_{test}}$	-0.059	-0.01	-0.075	<0.1
k	1	0.99	0.99	$0.85 \leq k \leq 1.15$
$\acute{k}$	0.99	1.009	1.007	$0.85 \leq \acute{k} \leq 1.15$

y : experimental value for training set, $\bar{y}$: the average of the experimental value for training set, $\hat{y}_{LMO}$ leave many out cross-validation predicted value for training set, y_{test}:experimental value of the external test set, $\hat{y}_{test}$predicted value for the external test set. N: number of samples, R^2_0 obtained from the plot of predicted values vs actual with zero intercept, $\acute{R}^2_0$ obtained from the plot of actual values vs predicted with zero intercept, k the slope of predicted values vs actual and $\acute{k}$ the slope of actual values vs predicted with zero intercept.

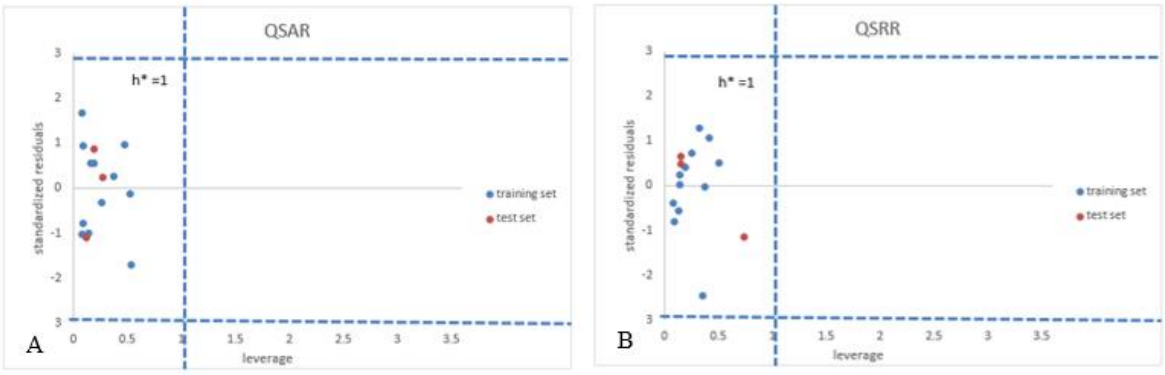

Figure 7. Applicability domain of (A) QS**A**R& (B) QS**R**R generated models

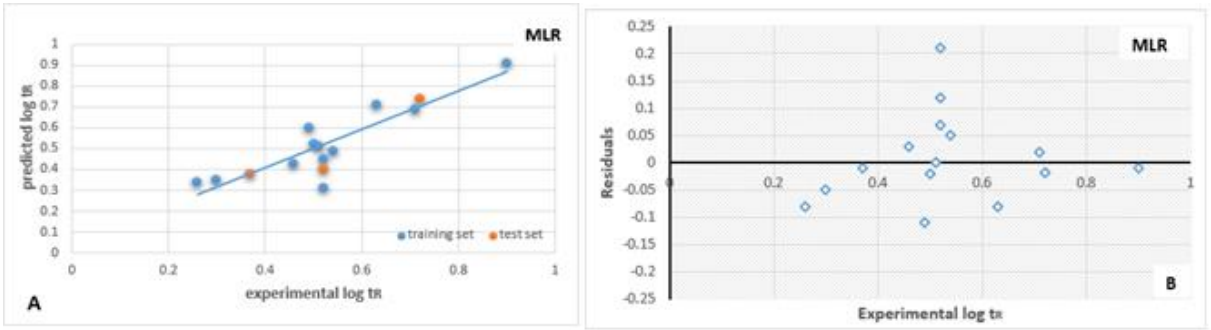

Figure 8. A: Plot of experimental vs predicted log t_R by MLR QSRR model, B: Plot of calculated residuals versus experimental values of log t_R

Applicability domain with leverage approach was shown by Williams plot (standardized residuals vs leverage), the squared area was between ±3 standardized residuals and the leverage threshold h*=1 calculated by equation 1. All the studied compounds fell in this square. However, if two-thirds of the test set fell into the AD area of the generated model, this model could be considered to have a good reliability for the predictions [31].

3.2.1.2 Principal component regression

QSRR study: PCA was performed on a data set consisting of 282 previously calculated molecular descriptors after excluding descriptors with constant values and descriptors with standard deviation equals 0. Descriptor standardization was performed to ensure that all the descriptors have the same weight in the analysis process. PCA was a powerful technique to summarize and compress the large pool of descriptors into only 14 informative PCs also they are orthogonal so the collinearity problem between the variables is solved [32,33]. These 14 PCs have eigen values more

than 1 and explain around 100% of variance in the original data matrix (

Table 3). To build relationship between these acquired PCs subset and the target property forward selection-linear regression was carried out. Herein, PC2 and PC3 showed significant linear correlation with the chromatographic retention according to the following equation:

$$log\ t_R = 0.533 - 0.089PC2 + 0.105PC3 \qquad equation\ 7$$

(R^2=0.7, R^2_{adj}=0.63, Q^2_{LMO}=0.66, R^2_{pred}= 0.709)

QSAR study: the same steps mentioned above were also applied here and PC1, PC2 and PC7 showed significant linear correlation with pIC_{50} and the PCR equation is:

$$pIC50 = 8.832 - 0.145PC1 + 0.207PC2 - 0.096PC7 \qquad equation\ 8$$

(R^2=0.69, R^2_{adj}=0.614, Q^2_{LMO}=0.68, R^2_{pred}= 0.72)

Table 3: PCs and initial Eigen values

Component	Initial Eigenvalues			Extraction Sums of Squared loading		
	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative%
1	95.993	34.04	34.040	95.993	34.04	34.040
2	46.47	16.482	50.522	46.47	16.482	50.522
3	33.082	11.731	62.253	33.082	11.731	62.253
4	22.965	8.144	70.397	22.965	8.144	70.397
5	18.274	6.480	76.877	18.274	6.480	76.877
6	14.134	5.012	81.889	14.134	5.012	81.889
7	11.801	4.185	86.074	11.801	4.185	86.074
8	9.992	3.54	89.614	9.992	3.54	89.614
9	9.475	3.360	92.973	9.475	3.360	92.973
10	7.143	2.533	95.506	7.143	2.533	95.506
11	4.599	1.631	97.137	4.599	1.631	97.137
12	3.705	1.314	98.451	3.705	1.314	98.451
13	3.078	1.091	99.542	3.078	1.091	99.542
14	1.29	0.458	100.000	1.29	0.458	100.000

Extraction Method: Principal Component Analysis

.

3.2.2 Nonlinear models PC-ANN:

QSRR study: ANN architecture: the input layer the first **6** PCs (ranked in descending order according to their eigenvalues) were utilized, this number of PCs gives the statistically accepted model. One hidden layer contains 3 neurons also this number of neurons was optimized, learning rate= 0.01, momentum constant α=0.5, maximum number of epochs (learning circles) = 1000, weight decay parameter = 10^{-2}, back-propagation training algorithm was used for training the neural network, the activation function was $tanh^{-1}$, the output layer was the predicted values.

Weight adjustment is an important issue during ANN training, as the ANN is trained more, weights will largely grow, and this could lead to an unstable network with poor performance to new unseen data. So, it is important to keep these weights small, for this purpose weight decay regularization parameter was used in this study as a driving force to reduce the weights and prevent them to be larger[34]. During ANN training process, the data set was divided into training set, validation set (20% of the data), the error of validation set was monitored to observe the decrease, when this error begins to increase this mean ANN is going to overfit and the training process should be stopped (this is called early stopping criterion). The obtained model showed good performance depending on Q^2_{LMO} and R^2_{pred} values but not the best comparing with SS-MLR method (Table 6)**.**

QSAR studies: ANN architecture: the input layer the first **7** PCs (ranked according to their eigenvalues) was used. Only one hidden layer with 3 neurons, learning rate= 0.01, momentum constant α=0.5, maximum number of epochs (learning circles) = 1000, weight decay parameter = 10^{-3}, back-propagation training algorithm was used for training the neural network, the predicted values showed in (

Table 5). This model was the best depending on external and internal validation results (Q^2_{LMO}=0.77, R^2_{pred}=0.835). Additional external validation metrics of this generated QSAR model showed in

Table 7.

In Figure 9A, the plot of experimental values of pIC_{50} versus predicted values obtained by the PC-ANN model, showed the good quality of the model. In Figure 9B, the plot of calculated residuals versus experimental values of pIC_{50} was distributed around zero, which reflect the good predictivity of the model.

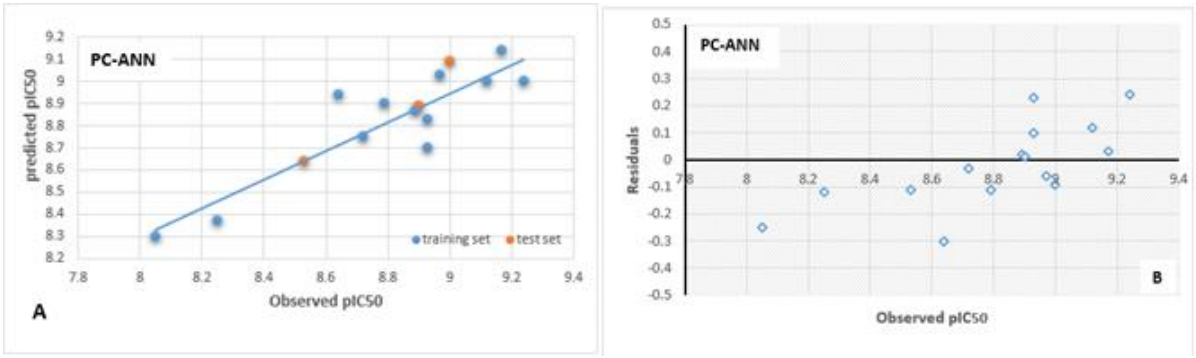

Figure 9. A: Plot of experimental vs predicted pIC_{50} by PC-ANN QSRR model, B: Plot of calculated residuals versus experimental values of pIC_{50}

4. Conclusion:

Fourteen reported newly synthesized PDE-4B inhibitors and roflumilast were highlighted in this research. Central composite design was used to select the optimal RP-HPLC conditions (C8, pH= 3 and 50:50 % organic ratio), retention times obtained under these optimal conditions were exploited in QSRR studies. Three different linear and nonlinear techniques (SS-MLR, PCR and PC-ANN) were utilized to build robust QSAR/QSRR models. For QSRR modelling, SS-MLR showed best results depending on internal and external validation statistical parameters, lipophilicity related descriptors have the most contribution in this model. PCR didn't give any significant improvement in comparison with MLR. QSAR studies, PC-ANN showed the best predictive results comparing with SS-MLR and PCR. The application of regularization (weight decay type) was important to improve the performance of ANN in our case study.

Conflict of interest:

The authors declare that they have no conflict of interest.

5. References

1. W. W. Labaki and S. R. Rosenberg, *Ann. Intern. Med*. (2020).
2. M. A. Makary and M. Daniel, *BMJ*, i2139 (2016).
3. N. A. Pinner, L. A. Hamilton, and A. Hughes, *Clin. Ther*. **34**, 56 (2012).
4. K. F. Rabe, E. D. Bateman, D. O'Donnell, S. Witte, D. Bredenbröker, and T. D. Bethke, *Lancet* **366**, 563 (2005).
5. T. J. Torphy, Am. J. Respir. Crit. *Care Med*. **157**, 351 (1998).
6. S.-L. C. Jin and M. Conti, *Proc. Natl. Acad. Sci*. **99**, 7628 (2002).
7. M. Ariga, B. Neitzert, S. Nakae, G. Mottin, C. Bertrand, M. P. Pruniaux, S.-L. C. Jin, and M. Conti, *J. Immunol*. **173**, 7531 (2004).
8. B. A. Moussa, A. A. El-Zaher, M. K. El-Ashrey, and M. A. Fouad, *Eur. J. Med. Chem*. **148**, 477 (2018).
9. B. A. Moussa, A. A. El-Zaher, M. K. El-Ashrey, and M. A. Fouad, *Drug Test. Anal.* **11**, (2019).
10. S. C. Peter, J. K. Dhanjal, V. Malik, N. Radhakrishnan, M. Jayakanthan, and D. Sundar, *in Encycl. Bioinforma. Comput. Biol*. (Elsevier, 2019), pp. 661–676.
11. R. Kaliszan and T. Bączek, in *Recent Adv. QSAR Stud.* (Springer Netherlands, Dordrecht, 2010), pp. 223–259.
12. M. A. Fouad, E. H. Tolba, M. A. El-Shal, and A. M. El Kerdawy, *J. Chromatogr. A* (2018).
13. S. Cateni, M. Vannucci, M. Vannocci, and V. Coll, in *Multivar. Anal. Manag. Eng. Sci.* (InTech, 2013).
14. B. Hemmateenejad, *Chemom. Intell. Lab. Syst*. (2005).
15. T. ArchanaH. and D. Sachin, *Int. J. Comput. Appl.* **122**, 4 (2015).
16. S. Yousefinejad and B. Hemmateenejad, *Chemom. Intell. Lab. Syst.* **149**, 177 (2015).
17. S. Yagiz, E. A. Sezer, and C. Gokceoglu, *Int. J. Numer. Anal. Methods Geomech.* (2012).
18.http://www.chemcomp.com/MOE-Molecular-Operating-Enviroment.htm
19. P. Gramatica, QSAR Comb. Sci. **26**, 694 (2007).
20.Roy et al. chemom Intell Lab Syst.145(2015)
21.http://www.knime.org
22. P. Gramatica, Methods Mol. Biol. (2013).
23. R. Kumar, R. Singh, N. Kumar, K. Bishnoi, and N. R. Bishnoi, *Chem. Eng. J.* **146**, 401 (2009).
24. B. Sadhukhan, N. K. Mondal, and S. Chattoraj, *Karbala Int. J. Mod. Sci.* **2**, 145 (2016).
25.D. Montgomery, Design and Analysis of Eperiments,5th edition, 2001, New York.
26. A. Asghar, A. A. Abdul Raman, and W. M. A. W. Daud, *Sci. World J.* ***2014***, 1 (2014).
27. S. Tong, N. Chen, H. Wang, H. Liu, C. Tao, C. Feng, B. Zhang, C. Hao, J. Pu, and J. Zhao, *Bioresour. Technol.* **171**, 389 (2014).
28. A. Goudzal, E. Hadaji, M. Bouachrine, H. El

Hamdani, and A. Ouammou, *Mater. Today Proc.* (2020).
29. H. Labjar, M. Kissi, R. Mouhibi, O. Khadir, H. Chaair, and M. Zahouily, *Int. J. Bioinform. Res. Appl.* **12**, 116 (2016).
30. A. Golbraikh and A. Tropsha, in *J. Mol. Graph. Model.* (Elsevier, 2002), pp. 269–276
31. M. Wen, Z. Deng, S. Jiang, Y. Guan, H. Wu, X. Wang, S. Xiao, Y. Zhang, J. Yang, D. Cao, and Y. Cheng, *Front. Pharmacol.* **0**, 391 (2019).
32. M. Shahlaei, A. Fassihi, and L. Saghaie, *Eur. J. Med. Chem.* **45**, 1572 (2010).
33. C. Yoo and M. Shahlaei, *Chem. Biol. Drug Des*. **91**, 137(2018).
34. M. Sammany, E. Pelican, and T. A. Harak, *Computing* **91**, 353 (2011).