Heliyon 5 (2019) e02859

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Research article

Theoretical QSAR modelling and molecular docking studies of some 4-hydroxyphenylpyruvate dioxygenase (HPPD) enzyme inhibitors potentially used as herbicides

Saidu Tukur *, Gideon Adamu Shallangwa, Abdulkadir Ibrahim

Faculty of Physical Sciences, Chemistry Department, Ahmadu Bello University, P.M.B. 1044, Zaria, Kaduna State, Nigeria

ARTICLE INFO

Keywords:
Inorganic chemistry
Herbicide
Quantitative structure-activity relationship (QSAR)
Multiple linear regression (MLR)
Genetic function algorithm (GFA)
Applicability domain
Molecular docking
Binding affinity(BA)

ABSTRACT

Computational QSAR studies together with molecular docking calculations have been performed on 118 different derivatives of organic molecules potentially used as herbicides. The Becke's three parameter exchange functional (B3) hybrid with Lee, Yang and Parr correlation functional (LYP), termed as B3LYP hybrid function and 6-31G* basis set (B3LYP/6-31G*) were used to develop five models of QSAR using the GFA technique. Models 1, was preferred as the best model because it possesses certain statistical implications (Friedman LOF = 0.52567, R^2 = 0.9034, R^2_{adjst}= 0.8943, Q^2_{CV}= 0.87 98 and $R^2_{pred.}$= 0.8403)." The prepared model was validated internally and externally using training and test inhibitors. The molecular docking studies conducted in this study has actually outline the binding affinities of the 10 selected compounds (5, 25, 26, 27, 29, 35, 52, 55, 98 and 114) which were all in good correlation with their pIC_{50} values. The binding affinities of the 10 selected compounds range between -5.9 kcal/mol to -10.1 kcal/mol. The compounds 25 and 27 with binding affinities of -10.1 kcal/mol and -9.7 kcal/mol formed the most stable complexes with the receptor (HPPD) as compared to other inhibitors. The complexes of these inhibitors show two most important types of bonding; Hydrogen bonding and hydrophobic bond interaction with the target amino acid residues. The computational QSAR study together with the molecular docking has actually provided a valuable approach for agrochemical researchers in synthesizing and developing new herbicides with high potency against the target enzyme.

1. Introduction

The chemicals describe as herbicides are substances in a chemical form that tend to impede or inhibit different metabolic processes responsible for growth in plants [1]. The substances are broadly used for cultivated tenacities as a scheme used to compensate for undesirable crops (weeds) or to clear off uncultivated farmland. However, weeds habitually have the aptitude to strive with other variability of crops for sunshine, and nutrients. "In addition, weeds have the budding to slow down plant growth, causes deterioration in production and yield quality [2]. However, weeds are typically the main origin of disease outbursts in the plant. The first viable herbicides synthesized at the commencement of 1948 was 2.4-dichlorophenoxyacetic acid [3]. In addition, the structural characteristics and the chemical bustle of the 2.4-D were used in the description of numerous other inhibitors [4].

4-Hydroxyphenylpyruvate dioxygenase (HPPD) is an enzyme which is proficient to break tyrosine down into homogentisate [5] and with the help of an enzyme Homogentisic acid oxidase, the homogentisate serve as a precursor for the biosynthesis of plastoquinones and Tocopherols [6]. The Tocopherols is use in the formation of carotenoids and the carotenoids is the pigment that is responsible for protecting the chlorophyll from being destroyed by the sunlight [7]. Hence, when this enzyme is inhibited, the green plant can no longer synthesized its food (sugar) which result to wilting and drying off of the green plant due to lack of chlorophyll [8]. The trivial flowering plant *Arabidopsis thaliana* is intrinsic to Eurasia and Africa that fits the mustard family [9] which has become the distinguishing plant of special interest for study today in Plant biology [10]. The plant *Arabidopsis Thaliana* is viewed as a weed when it seems on farmland or in a fallow land. However, with the noteworthy understanding of plant evolution and development, it originated a good research area to the agricultural scientist to focus on the molecular heredities of the studied flowering plant and all other chlorophyll dependent plant.

The quantifiable structure-activity relationship (QSAR) is a practise

* Corresponding author.
E-mail address: saidutukur@gmail.com (S. Tukur).

https://doi.org/10.1016/j.heliyon.2019.e02859
Received 9 September 2019; Received in revised form 29 October 2019; Accepted 13 November 2019

Table 1
General minimum recommended value for an acceptable QSAR model.

Symbol	Name	Acceptable Value
R^2	Coefficient of determination	≥ 0.6
$P_{95\%}$	Confidence interval at 95% confidence level	< 0.05
Q^2_{cv}	Cross-validation Coefficient	≥ 0.5
$R^2 - Q^2_{cv}$	Difference between R^2 and Q^2_{cv}	< 0.3
$N_{ext.\ test\ set}$	Minimum number of external test set	≥ 5.0
cR^2_p	The coefficient of determination for Y-randomization	> 0.5

Table 2
QSAR models for the Herbicide derivatives (Sulfonyl urea, Pyridines, Pyrimidines, Triazines etc.).

S/N	Models	R^2	Q^2_{CV}	R^2_{ext}
	Standard	≥0.5	≥0.5	≥0.6
1	pIC50 = 0.256966263 *nCl +0.9491052000 *BCUTp-1l - 5.963111430 *SCH-5 + 0.126975986 *maxsOm+ 0.245629890 *LipoaffinityIndex+ 0.816750061 *MDEC-24 - 0.013154034 * ATS5s + 0.8714865420 * WNSA-1I + 1.272338365. $N_{training} = 94$, $N_{test} = 24$	0.9034	0.8798	0.8402
2	pIC50 = 0.231648312 *nCl +0.452867262 *BCUTp-1l -8.913193523*SCH-5 + 3.153579106 *VCH-7 + 0.125692436 *maxsOm +0.266741571 *LipoaffinityIndex - 0.013943208 *ATS5s - 0.013344655 *WNSA-1I + 1.031108425. $N_{training} = 94$, $N_{test} = 24$	0.9033	0.8709	0.8006
3	pIC50 = 0.232826016 *nCl +0.451981644 *BCUTp-1l- 8.915474315 *SCH-5 + 0.125528607 *maxsOm+ 0.266522498 *LipoaffinityIndex- 0.013796916 *ATS5s - 3.151712115 *VCH -7 - 0.013319734 * WNSA-1I + 3.390377329. $N_{training} = 94$, $N_{test} = 24$	0.9032	0.8708	0.7610
4	pIC50 = 0.397919950 *nCl +0.451626613 *BCUTp-1l - 6.043846561 * SCH-5 + 0.121564676 *maxsOm +0.231791378 * LipoaffinityIndex +0.834455606 *MDEC-24 - 0.545070632 *SpMax8_Bhs - 0.011385837 *WNSA-1I + 1.444156973. $N_{training} = 94$, $N_{test} = 24$	0.9027	0.8788	0.7214
5	pIC50 = 0.232484324 * nCl +0.451682402 * BCUTp-1l - 8.903303507 *SCH-5 + 0.125834826 *maxsOm +0.267695294 *LipoaffinityIndex- 0.013778622 * ATS5s - 3.206311385 * VCH-7 - 0.013245963 * WNSA-1I + 3.29483375. $N_{training} = 94$, $N_{test} = 24$	0.9020	0.8682	0.6818

that generates a good scientific or computational correlation among molecular descriptors and observed property (IC_{50}) by means of the MLR-GFA [12]. The molecular docking study helps us to understand the interaction mode between the inhibiting substances and the receptor when link together purposely to attain a stable conformation. "The core of this research was to develop a validated QSAR model that will well predict the IC_{50} values of the studied compounds." The docking process is supported out between studied molecules with the highest pIC50 values and the prepared crystalline structure of the HPPD receptor that was obtained from *Arabidopsis Thaliana* flowering plant.

Table 4
Y-randomization test parameters.

Model	R	R^2	Q^2
Original	0.911225	0.830332	0.765508
Random 1	0.307614	0.094626	-0.12441
Random 2	0.275489	0.075894	-0.09142
Random 3	0.273353	0.074722	-0.15235
Random 4	0.370389	0.137188	-0.04283
Random 5	0.320099	0.102463	-0.11135
Random 6	0.339911	0.115539	-0.11133
Random 7	0.254737	0.064891	-0.16892
Random 8	0.262175	0.068736	-0.12241
Random 9	0.255073	0.065062	-0.14825
Random 10	0.185379	0.034365	-0.17398
Random Models Parameters			
Average r:	0.284422		
Average r^2:	0.083349		
Average Q^2:	-0.12472		
cRp^2:	0.788848		

Table 5
List of the descriptors, their description, classes, and their statistical parameters.

S/N	Descriptors	Description	Descriptor Class	VIF	ME
1	nCl	Number of chlorine atoms	2D	1.616	0.048
2	BCUTp-1l	nhigh lowest polarizability weighted BCUTS	2D	2.035	0.344
3	SCH-5	Simple chain, order 5	2D	1.145	-0.016
4	maxsOm	Maximum atom-type E-State: -O-	2D	1.325	0.032
5	LipoaffinityIndex	Lipoaffinity index	2D	1.585	0.283
6	MDEC-24	Molecular distance edge between all secondary and quaternary carbons	2D	1.319	0.010
7	ATS5s	Broto-Moreau autocorrelation - lag 5/weighted by I-state	2D	2.195	0.045
8	WNSA-1	PNSA-1 (Partial negative surface area – the sum of surface area on negative parts of a molecule) * total molecular surface area/1000	3D	2.413	0.254

Table 3
Pearson's association matrix of the descriptors used in the model.

	nCl	BCUTp-1l	SCH-5	maxsOm	Lipoaffinity Index	MDEC-24	ATS5s	WNSA-1
nCl	1							
BCUTp-1l	-0.00042	1						
SCH-5	-0.02429	0.177746	1					
maxsOm	-0.09986	-0.37896	-0.22546	1				
LipoaffinityIndex	-0.10145	0.033268	0.064203	-0.22306	1			
MDEC-24	-0.24002	-0.00519	0.141807	0.024764	0.408618	1		
ATS5s	-0.33785	-0.42588	0.018333	0.04734	0.392048	0.21808	1	
WNSA-1	0.25611	-0.55883	0.067396	0.043026	0.241172	0.005096	0.49961	1

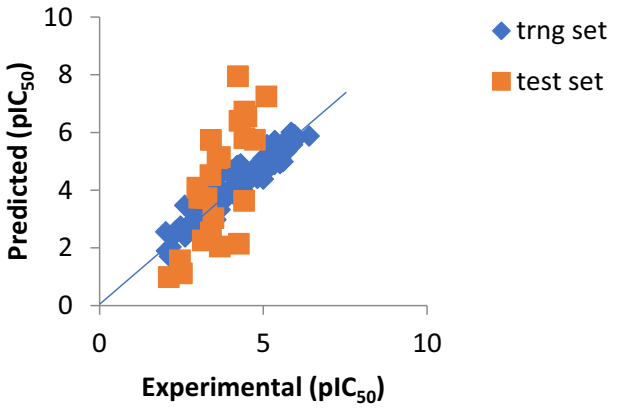


Fig. 1. Showing the plot of experimental pIC_{50} and predicted pIC_{50} values of training and test set compounds of model 1.

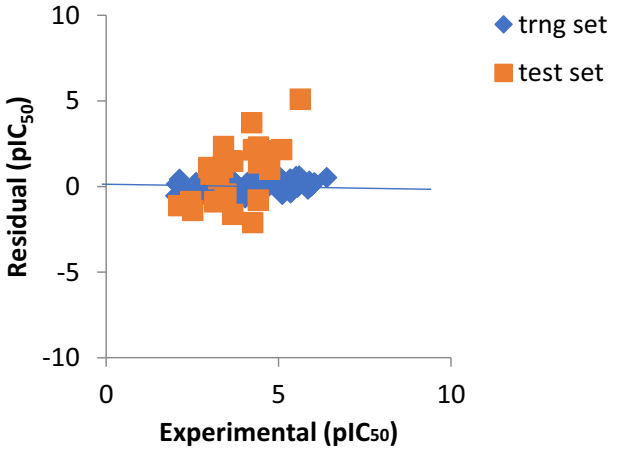


Fig. 2. A plot of Experimental pIC_{50} versus Residual values of the training and test set compounds of model 1.

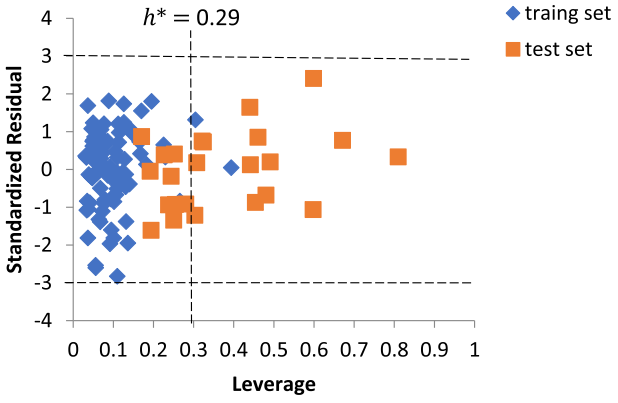


Fig. 3. Williams Plot, A plot of standardized residual versus Leverage of model 1.

2. Materials and methods

2.1. Experimental dataset

The large data set encompassing 334 different derivatives of compounds used as herbicides was gotten from the literature researched [11]. However, out of the 334 compounds, only 118 chemical compounds were designated for this study. This dataset includes the following classes of inhibitors; Sulfonylurea, triazines, pyrimidine, benzoic acid, ester, pyridine, etc. The chemical Abstract Service record numbers (CASRN), IUPAC names, generic names, and various physicochemical properties of the selected 118 compounds were found from a simplified molecular input line file entry system (SMILES), while the inhibitory activities (IC_{50}) against HPPD receptor of the selected ligands were obtained together with the SMILE document file in an Excel sheath.

2.2. Geometry energy minimization

The determination of the equilibrium geometry or the minimum energy conformation of a molecule is called Energy minimization [13]. The ChemDraw software version 12.0 was one among the widely known software used to draw the 2D structures of the compounds and the drawn 2D molecular structures will be saved in CDX file format [14]. The 3D conformations of the molecules were optimized using the wave function version I. 1.4 of the software Spartan [15]. The 3D molecular structures were pre-optimized using the semi-empirical method primary to lessen the tension absolutely. The Spartan file optimized for all molecules was then converted to "SDF " file format [16].

2.3. Determination of molecular descriptor

The common term molecular descriptor can be defined as a scientific or mathematical figure that designates the properties of the characteristics of a molecule attained from a precise algorithm or trial process [17]. The quantum chemical descriptors remained calculated using a software called Spartan 14 version 1.1.4 [18]. The molecular descriptors stayed calculated using the PaDEL descriptor software version 2.1.8. The middling of 1875 molecular descriptors remained generated by the software PaDEL and Spartan 14 software [19].

2.4. Pre-treatments of dataset

The molecular descriptor resulting from PaDEL-software was beforehand preserved with a drug theoretical Cheminformatics laboratory (DTC Lab.) software. The pre-processing process vicissitudes the correlation among the less dismissed descriptions [20]. The data dispensation procedure abolished the molecular characteristics viewing almost equal values and very low variance descriptions, with the complete dataset to reduce the Loch descriptor. The molecular characteristics with a lower level of correlation are maintained [21].

2.5. Test and Training set development

In QSAR model Expansion, datasets of 118 compounds were divided into 80% training and 20% test set rendering to the Kennard and Stone Algorithm using DTC Lab. Software [22]. However, 80% of the entire dataset was used for setting the model up while, the remaining 20% of the data collection was used for external validation of the former model.

2.6. Comparative implication of individually descriptor to the model

"The mean effect (*ME*) of each descriptor parameter which appraises the absolute distinction of each descriptor and its effect on the model, can be calculated using Formula (1) below:

$$ME = \frac{\beta_j \sum_i^n D_j}{\sum_J^m (\beta_j \sum_i^n D_j)} \quad (1)$$

ME, defines the average effect of descriptor ***J***, while β_j is a descriptor co-efficient, D_j is the value of each study material descriptor, ***m*** represents the number of characteristics in the model and the number of compounds representing ***N*** in the training set [23].

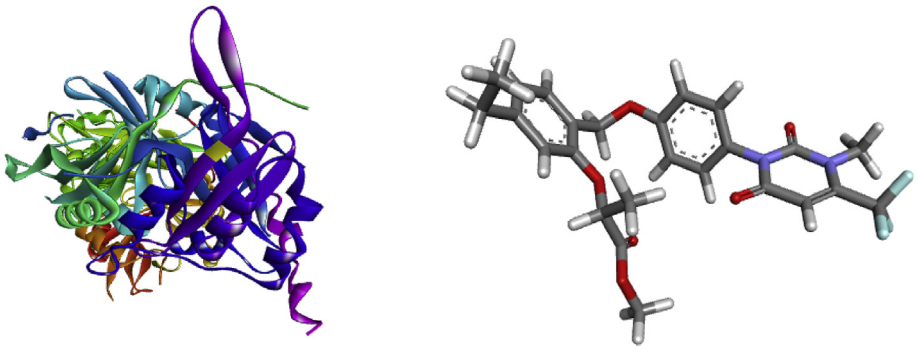


Fig 4. Prepared structure of 4-Hydroxyphenylpyruvate dioxygenase (HPPD) Receptor and 3D structure of the prepared Ligand (25).

2.7. Model development

The MLR/GFA techniques remained of used in developing models [24]. The technique of GFA definitely selects an outstanding number of descriptors in regression model and by so doing, it creates the MLR equation in a linear form and in higher polynomial [25]. The general MLR expression is given below in Eq. (2):

$$\boldsymbol{P} = \boldsymbol{p}_0 + \boldsymbol{p}_1\boldsymbol{X}_1 + \boldsymbol{p}_2\boldsymbol{X}_2 + \boldsymbol{p}_3\boldsymbol{X}_3 + \ldots \boldsymbol{p}_n\boldsymbol{X}_n \quad (2)$$

The symbol $\boldsymbol{P}$ denotes the IC_{50} value, $\boldsymbol{X_1, X_2,\ldots, X_n}$ are the value of descriptors prevailing in the model with their individual coefficients $\boldsymbol{p_1, p_2,\ldots, p_n}$. The term p_0 characterise the constant value in the regression equation. The meaning of the discrete descriptors $\boldsymbol{X_1, X_2,\ldots, Xn}$ is subject to their greatness (significant value) and their algebraic sign (±) [20]. The numeral of compounds and the number of descriptors exist in a ratio of 5:1 for an upright model.

2.8. Assessment of the QSAR models

The QSAR model gotten from the MLR model was weighed according to the following statistical limitations: $\boldsymbol{N}$ (number of compounds), $\boldsymbol{P}$ (descriptor number), $\boldsymbol{R}^2$ (square correlation coefficient), F-Test (Fischer value), $\boldsymbol{Q}^2_{CV}$ (cross-validation coefficient), $\boldsymbol{R}^2_{pred.}$ (external test set Correlation coefficient). In the assortment of the QSAR Model, justification was taken looking at the R^2 and Q^2_{CV} values [26]. The model is only pertinent if the following conditions are factual: $\boldsymbol{R}^2 > 0.6$, $\boldsymbol{Q}^2_{CV} > 0.6$ and $\boldsymbol{R}^2_{pred.} > 0.5$. Given the earlier criteria, the higher the $\boldsymbol{Q}^2_{CV}$ and. $\boldsymbol{R}^2$Values, the more robust and accurate is the prescribe model.

2.9. Calculation of internal authentication of the QSAR model

Internal validation is the foremost step to take into justification in validation of any QSAR model [27]. A square correlation coefficient $\boldsymbol{R}^2$characterizes all the models for the dispersal of the predicted variables. The nearer the $\boldsymbol{R}^2$value to 1.0, the better the model formed. $\boldsymbol{R}^2$ Can be calculated using Eq. (3) as shown below:

$$\boldsymbol{R}^2 = 1 - \frac{\sum \left(\boldsymbol{P}_{obs.} - \boldsymbol{P}_{pred.}\right)^2}{\sum \left(\boldsymbol{P}_{obs.} - \overline{\boldsymbol{P}}_{training}\right)^2} \quad (3)$$

where, $\boldsymbol{P}_{obs.}$, $\boldsymbol{P}_{pred.}$and $\overline{\boldsymbol{P}}_{training}$ constitute the experimental, projected and cruel activity value of the training set [28]. The value of $\boldsymbol{R}^2$ was found to be largely dependent on the number of model descriptors.

Therefore, the value of the field model $\boldsymbol{R}^2$ must be adjusted. The adjusted $\boldsymbol{R}^2$ can be calculated using Eq. (4) below:

$$\boldsymbol{R}^2_{adj.} = 1 - \left(1 - \boldsymbol{R}^2\right)\frac{\boldsymbol{n} - 1}{\boldsymbol{n} - \boldsymbol{m} - 1} = \frac{(\boldsymbol{n} - 1)\boldsymbol{R}^2 - \boldsymbol{m}}{\boldsymbol{n} - \boldsymbol{m} + 1} \quad (4)$$

Here, $\boldsymbol{n}$ stand for the number of training set and $\boldsymbol{m}$ stand for the number of descriptors used in the model.

However, the extrapolative strength of the QSAR model was also measured via Friedman's deficiency (LOF) [29], which was also among the key criteria for the assessment of other internal validation. LOF was calculated using Eq. (5), expressed below:

$$\boldsymbol{LOF} = \frac{\boldsymbol{SEE}}{\left(1 - \frac{c + dp}{N}\right)^2} \quad (5)$$

The acronym ***SEE***, describe the standard error of estimate or standard deviation (SD), $\boldsymbol{p}$ is the number of attributes in the model, $\boldsymbol{d}$ is defined by the debug parameter, $\boldsymbol{c}$ is the relentless term used in the model, $\boldsymbol{n}$ is the training set molecules. However, if the model *SEE* value is small, it is alleged that, the model established MLR model is robust. The *SEE* can be calculated using Formula (6), expressed below;

$$\boldsymbol{SEE} = \sqrt{\frac{\left(\boldsymbol{Y}_{exp.} - \boldsymbol{Y}_{pred.}\right)^2}{\boldsymbol{N} - \boldsymbol{P} - 1}} \quad (6)$$

Where $Y_{exp.}$and $Y_{pred.}$ characterise trial and foretold pIC_{50} values of training set. $\boldsymbol{N}$ is the number of compounds in the training set, $\boldsymbol{P}$ is the number of descriptors [30]. In addition, the second parameter, which should be regarded as very important when entering the internal validation of the QSAR model, is the leave-one-out cross validation coefficients. The Cross-validation regression coefficients (Q^2_{CV}) can be calculated using Eq. (7) below.

$$\boldsymbol{Q}^2_{CV} = 1 - \frac{\sum_{i=1}^{n}\left(\boldsymbol{Y}_{exp.} - \boldsymbol{Y}_{pred.}\right)^2}{\sum_{i=1}^{n}\left(\boldsymbol{Y}_{exp.} - \overline{\boldsymbol{Y}}\right)^2} \quad (7)$$

Where $\boldsymbol{Y}_{exp.}, \boldsymbol{Y}_{pred.}, and \overline{\boldsymbol{Y}}$are trial, foretold and the mean inhibition activity values of the training set compounds [14].

2.10. Assessment of External authentication of the QSAR model

The created QSAR model was validated externally to authorise its strength. Therefore, an external validation of the model was evaluated using the R^2_{test} expression for test compounds. Therefore, the external predictive force of the model was calculated using the regression coefficient expression as shown below in Eq. (8):

$$\boldsymbol{R}^2_{test} = 1 - \frac{\sum \left(\boldsymbol{P}_{pred.test} - \boldsymbol{P}_{exp.test}\right)^2}{\sum \left(\boldsymbol{P}_{pred.test} - \overline{\boldsymbol{P}}_{Training}\right)^2} \quad (8)$$

Where, $\boldsymbol{P}_{pred.test}$, $\boldsymbol{P}_{exp.test}$, are the forecast and tentative activity of the test set and $\overline{\boldsymbol{P}}_{Training}$ is the average experimental activity of the training set compounds [31].

In addition, other Computable relations between the activities and

Table 6
Binding Affinity, Hydrogen Bond and Hydrophobic Bond Interaction formed Between the Ligands with the highest pIC_{50} Values and the Active Site of the Hydroxyphenylpyruvate dioxygenase (HPPD) Receptor.

Ligand ID	Binding Energy [Kcal/mol]	Hydrophobic bond	Hydrogen bond	Hydrogen bond length [A^0]
5	-8.4	PHE392, PHE381, PHE392	GLN379	2.93328
25	-10.1	PHE381, PHE392, LEU265 PRO384, LEU265, VAL269 RO280, PHE419, PHE424	SER267, GLN307, HIS308	2.32652, 2.38559, 2.45207
26	-8.5	MET335, LEU265, ILE294, PHE381	GLN293, HIS308	2.41845
27	-9.7	HIS308, SER267, PHE392 PHE381, LEU265, LYS421	GLN379, PHE419, ASN282	2.84333, 2.43391, 2.48039
29	-7.6	LEU427, PRO280, HIS226 HIS308, PHE419, PHE424	SER267, ASN282	2.4535, 2.8447
35	-8.1	PHE381, PRO280, LEU427 LEU265, ILE294, PHE419 PHE424	SER267, HIS308, GLU252 ASN282	2.52058, 2.76242, 3.0342, 3.48822
52	-6.1	VAL228, PRO280, VAL269 HIS226, HIS308, PHE381 PHE419	SER267	2.0812
55	-8.0	PHE424, VAL228, LEU265 LYS421, LEU265, HIS308 PHE381	-	-
98	-8.7	VAL269, PRO280, LEU265 HIS226, HIS308, PHE381 PHE419, PHE424	SER267, ASN282	2.23157, 2.73159
114	-5.9	PHE72, ALA61, ARG62 LEU90, LEU217, VAL209 PHE72	ARG62, ARG62, SER65 PHE72, PRO216, PHE72 GLU210	2.5645, 2.99795, 2.18236, 2.15507, 3.04285, 3.54983, 3.39371

the descriptors were examined by means of randomisation testing. The range of the Y- segment is speckled and the new QSAR models were developed using the same variable priority as the variables present in the non-random model. However, we use the parameter that is represented by R_p^2, which is the model difference between the average square correlation coefficient $R_{and.}^2$ of the random model and square correlation coefficient R^2 of the non-random model. The R_p^2 parameter was calculated using Eq. (9) given below as:

$$\boldsymbol{R_p^2 = R^2 \times \sqrt{(R^2 - R_{and.}^2)}} \tag{9}$$

"The levitating R_p^2 ensures that the generated models were not acquired by fortuitous. Meanwhile, we have projected that the estimation of R_p^2 protrudes more at $p < 0.05$ (95%) confidence limit for a steadfast model.

2.11. External Y-randomization test

Y- Randomization test is used to remotely validate the established QSAR model. The Y-randomisation test was accomplished using the training set, as recommended by Tropsha [31]. However, for the model to fractious the Y-randomization test, the cR_p^2must be superior than 0.5 (cR_p^2>0.5). The cR_p^2can be calculated using Eq. (10) as shown below:

$$\boldsymbol{cR_p^2 = R\left[R^2 - (R_r)^2\right]^2} \tag{10}$$

Where, cR_p^2 is the Strength coefficient of Y-randomization test, R is the correlation coefficient of Y-randomization test, and R_r is the normal R of the random model [32]. The multivariate linear regression model in the randomization test, was prevented by mixing the pIC_{50} values while retaining the descriptor value unbothered. The models after the initial model usually have a lower R^2 and Q_{CV}^2values. Ten randomization tests were performed and the results showed that approximately 9 models have R^2 and Q_{CV}^2 of estimate <0.5. This evaluation confirms that the resulting model is strong enough and capable of making good predictions [26].

2.12. Domain of applicability evaluation of the generated QSAR model

"The domain of applicability (Williams plot) is the plot of standardized residual versus a distance (leverage). The applicability domain of the QSAR model was intentionally used to check for outliers and influential compounds [14]. Leverage is used to sketch the applicability domain of the model. Leverage can be calculated using Eq. (11), expressed below as:

$$\boldsymbol{h_i = Q_i(Q^T Q)^{-1} Q_i^T, \qquad (i = m, \ldots p,)} \tag{11}$$

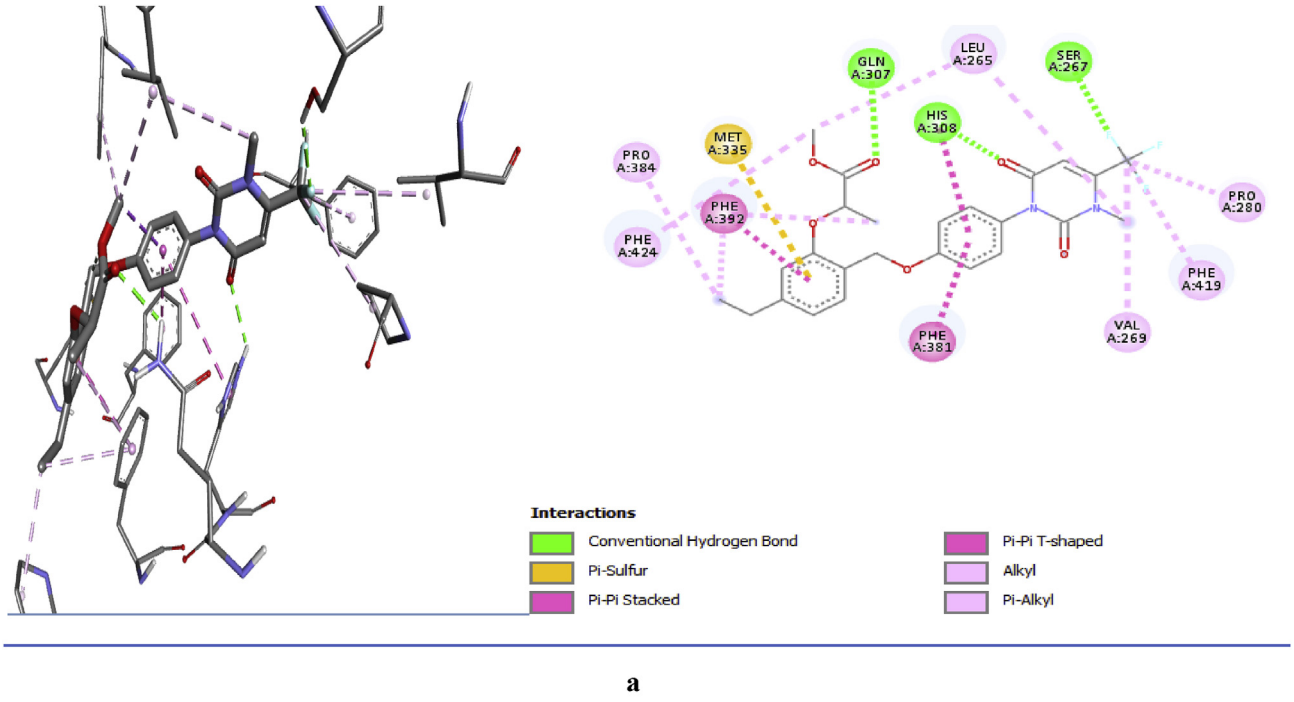
Where $\boldsymbol{Q_i}$ represent the training set matrix $\boldsymbol{I}$, $\boldsymbol{Q}$ is a $\boldsymbol{n \times m}$descriptor matrix of the training set, and $\boldsymbol{Q^T}$ is the transpose matrix $\boldsymbol{Q}$ used to develop the model. However, training or test set compounds with leverage values $\boldsymbol{hi}$ lower than the warning leverage $\boldsymbol{h^*}$ are capable of making predictions. The domain of applicability with the reliable predictive aptitude is that which has compounds (entire dataset) with leverage values within the threshold ($\boldsymbol{hi < h^*}$) and a standardized outstanding value of ±3 [16]. The applicability domain helps in choosing compounds with the best structural features [25]. The warning leverage ($\boldsymbol{h^*}$) can be calculated using Eq. (12), given below by the expression:

$$\boldsymbol{h^* = \frac{3(g+1)}{n}} \tag{12}$$

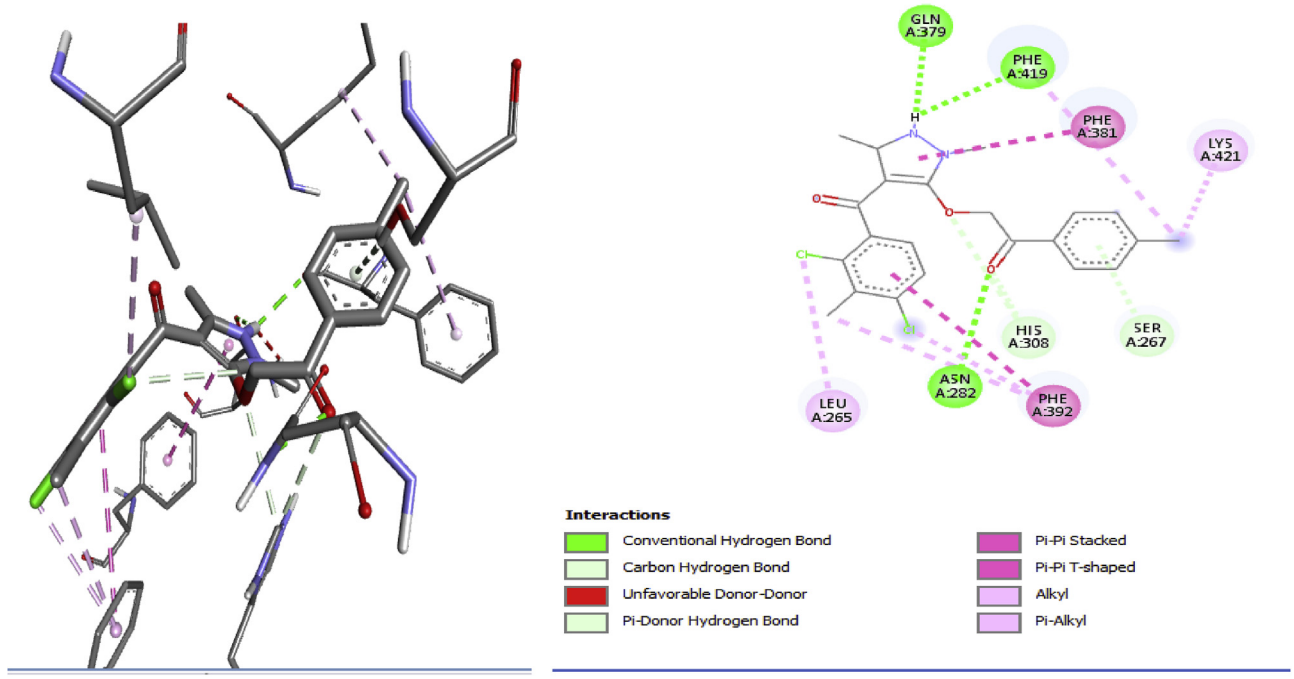
The symbol $\boldsymbol{n}$ represents the number of training set, and $\boldsymbol{g}$ is the number of descriptors used in developing the model.

2.13. Superiority declaration of the model

The internal and external QSAR authentication methods were the two main practices for appraising model strength. The validation constraints were related with the recommendation standard [32]. Table 1.0 below provides an overview of the standard recommended values for internal



a



b

Fig. 5. a) 3D and 2D molecular interaction for complex 2 (-10.1 kcal/mol). b) 3D and 2D molecular interaction for complex 4 (-9.7 kcal/mol).

and external endorsement parameters that pledge the acceptance of the model.

2.14. Molecular docking

The concept of Molecular docking refers to modeling method that predicts the direction or orientation of two molecules when they bind together in order to form a stable complex. A docking knowledge was steered between 10 selected compounds (5, 25, 26, 27, 29, 35, 52, 55, 98 and 114). All of the Spartan files for the nominated compound were transformed to a protein data bank (PDB). The crystal structure of the target receptor (HPPD) with a PDB code of 1sp9 was attained from the PDB website. Fig. 4 shows the receptor and a prepared ligand in 3D structure (compound 25). The prepared ligands were equipped with a prepared receptor, using the PyRx software and the Autodock Vina in the combined PyRx software [16].

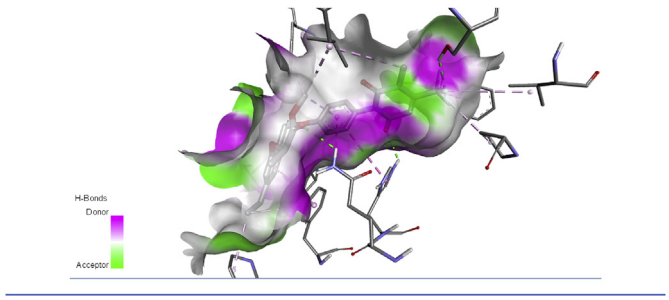


Fig. 6. H-bond molecular interaction between Ligand 25 and the target.

3. Result and discussion

The Multiple Linear Regression coupled with Genetic Function Algorithm (MLR- GFA) was used to produce five models as shown in Table 2 below. Model 1, was chosen because of its statistical effect as the best model (Friedman LOF = 0.52567, R^2 = 0.9034, R^2_{adjst}= 0.8943, Q^2_{CV}= 0.87 98 and $R^2_{pred.}$= 0.8403). However, the standard corroboration parameters for the acceptable model were given in Table 1. However, based on the model parameters above, model 1 has completed all the prerequisites for an acceptable QSAR model. The smaller the residual values indicate that the generated model has good forecasting capability. Eight descriptors remained designated to construct the linear model, which was able to predict the corresponding IC_{50} values of all the selected compounds using MLR-GFA practices.

Table 3 presents the Pearson correlation matrix for the eight descriptors. The correlation coefficient flanked by each descriptor of the model is significantly low, meaning that no much correlations exist. The Y-randomization test was shown in Table 4. However, the higher numerical values of the $R^2$and Q^2_{CV} indicate that the model was significant statistically. It was evident in the result that the cR^2_p value is greater than 0.5, therefore, the resulting model was powerful and did not receive a prepotential inference.

Table 5 make available the list of descriptions, descriptions, classes and other related statistical parameters (*VIF* and *ME*) that may take a superior influence on nominated descriptors. For all eight descriptors, the numerical values of the variance inflation factor (*VIF*) were all less than 3, indicating that the specifications of the model were coronal, and the model's consistency is of great significant [19]. Table 7 shows the chemical names of the compounds, the chemical Abstract Service registration number (CAS-RN), the trial pIC_{50}, the foreseen pIC_{50} and outstanding values.

Fig. 1 shows the trial pIC_{50} graph compared to the estimated pIC_{50} of the compounds in the training and test molecules. A graph of the residuals and trial pIC_{50} values of the entire dataset was shown in Fig. 2. The plot shows, in fact, there exist no sensible slip-up in the generated model because the standardized residual values were practically at zero line on both sides [28].

3.1. Interpretation of the molecular descriptors in model 1

The descriptor nCl, which stand for the Number of chlorine atoms, is a descriptor that belongs to the 2D class, while, the 2D descriptor BCUTp-1l, which is defined as the nhigh lowest polarizability weighted. Moreover, the 2D descriptor SCH-5, stands for Simple chain, order 5, the 2D descriptor maxsOm, refers to maximum atom-type.

E-State: -O-. Furthermore, the 2D descriptor LipoaffinityIndex, stand for the measure of the lipophilic property of a molecule and the 2D descriptor MDEC-24, which means the Molecular distance edge between all secondary and quaternary carbons. Moreover, ATS5s, a 2D descriptor which stand for Broto-Moreau autocorrelation - lag 5/weighted by I-state. Lastly, the 3D descriptor, WNSA-1 stand for the PNSA-1 (Partial negative surface area –

The sum of surface area on negative parts of a molecule) * total molecular surface area/1000. However, the eight descriptors nCl, BCUTp-1l, SCH-5, maxsOm, LipoaffinityIndex, MDEC-24, ATS5s, WNSA-1 has a positive mean effect value of 0.048, 0.344, -0.016, 0.032, 0.283, 0.010, 0.045, 0.254 which means that, increase in the concentration of these descriptors will lead to increase the inhibition concentrations of the inhibitor molecules which results to increase in the activity of the molecules against the target. While, SCH-5, descriptor with negative mean effect value of -0.016, reveal that, a decrease in the concentration of this descriptor will lead to increase in the activity of the inhibitor molecules.

The leverage values for the entire dataset were presented together with standardized residual values. The Williams plot is depicted in Fig. 3. However, based on our results, it is clear that all compounds (training and test molecules) were all within the square demarcation, in which all of the compound has a standardized residual value within ±3. Therefore, no extraneous compound was found in this study. In addition, it was found that approximately 40% of the test compounds were influential. This comportment was instigated by significant structural modifications between the compounds found in the domain and those found outside the domain.

3.2. discussion of docking results

Molecular docking studies were carried out on 10 selected compounds (5, 25, 26, 27, 29, 35, 52, 55, 98, 114) having the highest pIC_{50} values with the aim of studying their mode of interactions with the HPPD receptor. The docking results reported in Table 6 shows that the binding affinity of the selected ligands ranges between -5.9 kcal/mol to -10.1 kcal/mol. The docking results reveal that the ligand with the highest pIC_{50} value has the highest binding affinity, which indicates that the binding affinities of these ligands were proportional to their pIC_{50} values. The results show that Ligand 25 with the highest pIC_{50} value have the highest binding affinity of -10.1 kcal/mol, followed by Ligand 27, with the binding affinity of -9.7 kcal/mol. Moreover, using Discovery Studio Visualizer software, two complexes of Ligand 25, and 27 were selected for visualization in 3D and 2D, as shown in Figs. 5a and 5b respectively. Ligand 25 form three hydrogen bonds with SER267, GLN307 and HIS308 at 2.32652 Å, 2.38559 Å and 2.45207 Å of the target. The carbonyl C=O attached to the ester functional group and the C=O of the pyrazide ring both serves as hydrogen acceptors and forms one hydrogen bond each at GLN307 and HIS308 of the target. The fluorine atom attached to the pyrazide ring of ligand 25 also act as hydrogen acceptor and form one hydrogen bond with SER267 of the target. In addition, nine hydrophobic bond interactions were observed in ligand 25 at PHE381, PHE392 LEU265, PRO384, LEU265, VAL269, RO280, PHE419 and PHE424 of the target. Moreover, Ligand 27 also forms three hydrogen bonds with GLN379, PHE419 and ASN282 at 2.84333 Å, 2.43391 Å and 2.48039 Å of the target. The nitrogen atom attached to the diazole ring (-N-H) act as hydrogen acceptor and form two hydrogen bond with GLN379 and PHE419 at different distance of 2.84333 Å and 2.43391 Å of the target. Another hydrogen bond was observed with ligand 27 in which the carbonyl C=O of the keto functional group act as hydrogen acceptor and form one hydrogen bond with ASN282 of the target. Furthermore, six hydrophobic bond interactions were observed in ligand 27 at HIS308, SER267, PHE392, PHE381, LEU265 and LYS421of the receptor. Fig. 6, below shows the hydrogen bond interaction formed between the ligands 25 and the target. The hydrogen bond and the hydrophobic bond interaction formed between ligand 25 with the highest binding affinity of -10.1 kcal/mol suggest that ligand 25 forms the most stable complex and can be regarded as a template molecule in designing new potent

compounds against the HPPD receptor.

4. Conclusion

In summary, the eight molecular descriptors **nCl, BCUTp-1l, SCH-5, maxsOm, LipoaffinityIndex, MDEC-24, ATS5s** and **WNSA-1** used in the selected model (Model 1) were sufficiently potent and capable of predicting the herbicidal property of the selected compounds. A molecular docking study shows that the best two among the set of 10 compounds (**25** and **27**) with the lowest Gibbs free energies binding of **-10.1 kcal/mol** and **-9.7 kcal/mol**, formed the two most stable complexes after binding to the receptor. Furthermore, the two best compounds show two important type of interactions; Hydrogen and hydrophobic interactions with amino acid residues of the target receptor. Probably, these two compounds could be potentials inhibitors agents for better activity against the target enzyme.

Declarations

Author contribution statement

Gideon Adamu Shallangwa: Conceived and designed the experiments.

Saidu Tukur: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Abdulkadir Ibrahim: Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The corresponding author sincerely acknowledges the effort of Dr. Sani Uba, Mr. Muhammad Tukur Ibrahim, Mr. Abdullahi Bello Umar and Dr. David Ebuka from Department of chemistry, Faculty of Physical Science, Ahmadu Bello University Zaria, for their guidance toward the success of this study.

References

[1] S. Tukur, G.A. Shallangwa, A. Uzairu, A. Ibrahim, Insilico study for predicting the inhibition concentration of some heterocyclic and Phenylic compounds used as potent herbicides using the MLR-GFA approach, J. Eng. Exact Sci. 5 (2019) 4–7.

[2] R.K. Prasad, R. Sharma, 2D - QSAR analysis of pyrazine carboxamide derivatives as an herbicidal agent, J. Comput. Methods Mol. Des. 1 (2011) 7–13.

[3] P.W. Zimmerman, A.E. Hitchcock, Plant hormones, Annu. Rev. Biochem. 17 (1948) 601–626.

[4] Y. Liu, H. Zhao, Z. Wang, Y. Li, H. Song, H. Riches, D. Beattie, Y. Gu, Q. Wang, The Discovery of 3-(1-aminoethylidene) quinoline-2, 4 (1H, 3H)-Dione derivatives as novel PSII electron transport inhibitors, Mol. Divers. 17 (2013) 701–710.

[5] H.Y. Lin, J.F. Yang, D.W. Wang, G.F. Hao, J.Q. Dong, Y.X. Wang, W.C. Yang, J.W. Wu, C.G. Zhan, G.F. Yang, Molecular insights into the mechanism of 4-hydroxyphenylpyruvate dioxygenase inhibition: enzyme kinetics, J X-ray crystallog. comput. simulat. 286 (5) (2019) 975–990.

[6] M. Huang, D.Y. Yang, Z. Shang, 3D-QSAR studies on 4-hydroxyphenylpyruvate dioxygenase inhibitors by comparative molecular field analysis (CoMFA), Bioorg. Med. Chem. Lett 12 (17) (2002) 2271–2275.

[7] G.R. Moran, 4-Hydroxyphenylpyruvate dioxygenase, Arch. Biochem. Biophys. 433 (2005) 117–128.

[8] E. Haspolat, B. Huard, M. Angelova, Deterministic and stochastic models of Arabidopsis thaliana flowering, Bull. Math. Biol. 81 (1) (2019) 277–311.

[9] W. David, J. Meinke, C. Michael, D. Caroline, D. Steven, M.K. Rounsley, Arabidopsis thaliana: a model plant for genome analysis, Science 282 (5389) (1998) 662–682.

[10] V.L. Cruz, S. Martinez, J. Ramos, S.J. Martinez, 3d-QSAR as a tool for understanding and improving single-site polymerization catalysts. A review, Organometallics 33 (2014) 2944–2959.

[11] M.N. Gandy, M.G. Corral, J.S. Mylne, K.A. Stubbs, An interactive database to explore herbicide physicochemical properties, Org. Biomol. Chem. 13 (2015) 5586–5590.

[12] U. Abdulfatai, A. Uzairu, S. Uba, Investigation of some acetamido-N-benzylacetamide derivatives as potent anti-convulsant inhibitors, J. Comput. Methods Mol. Des. 5 (2015) 77–83.

[13] M.T. Ibrahim, A. Uzairu, G.A. Shallangwa, A. Ibrahim, In-silico studies of some oxadiazoles derivatives as anti-diabetic compounds, J. King Saud Univ. Sci. 1 (2018) 2.

[14] D.E. Arthur, A. Uzairu, P. Mamza, E. Abechi, G.A. `Shallangwa, Silico modelling of cytotoxic behaviour of anti-leukemic compounds on hl-60 cell line, Section A: Chemistry, J. Turk. Chem. Soc. 3 (2016) 147–158.

[15] S.E. Adeniji, S. Uba, A. Uzairu, A novel QSAR model for the evaluation and prediction of (E)-N'-Benzylideneisonicotinohydrazide derivatives as the potent anti-Mycobacterium tuberculosis antibodies using genetic function approach, Phys. Chem. Res. 6 (2018) 479–492.

[16] S.B. Olasupo, A. Uzairu, B.S. Sagagis, Density functional theory (B3lyp/6-31g*) study of toxicity of polychlorinated dibenzofurans, Int. J. Comput. Theor. Chem. 5 (2017) 4–24.

[17] U. Abdulfatai, A. Uzairu, S. Uba, Quantitative structure-activity relationship and molecular docking studies of a series of quinazolinonyl analogues as inhibitors of gamma-aminobutyric acid aminotransferase, J. Adv. Res. 8 (2017) 33–43.

[18] J.H. Panchal, S.R. Kalidindi, D.L. Mcdowell, Key computational modelling issues in integrated computational materials engineering, Comput. Aided Des. 45 (2013) 4–25.

[19] K. Roy, S. Kar, R.N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press 34 (12) (2013) 1071–1082.

[20] R. Kennard, W. Stone, A. Larry, Computer-aided design of experiments, Technometrics 11 (1969) 137–148.

[21] N. Minovski, Š. Župerl, V. Dragan, M. Novič, Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: *a case study*, Anal. Chim. Acta 759 (2013) 28–42.

[22] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships, J. Chem. Inf. Model. 34 (4) (1994) 854–866.

[23] D.H. Cho, S.K. Lee, B.T. Kim, K.T. Noa, Quantitative structure-activity relationship (QSAR) study ff new fluorovinyloxyacetamides, Bull. Korean Chem. Soc. 22 (2001) 388–394.

[24] D.E. Arthur, A. Uzairu, P. Mamza, S. Abechi, Quantitative structure-activity relationship study on potent anticancer compounds against molt-4 and P388 leukemia cell lines, J. Adv. Res. 7 (2016) 823–837.

[25] D.E. Arthur, A. Uzairu, P. Mamza, E. Abechi, G. Shallangwa, Insilco study on the toxicity of anti-cancer compounds tested against molt-4 and P388 cell lines using GFA-MLR technique, Beni-Suef University, J. Basic Appl. Sci. 5 (2016) 320–333.

[26] J. Friedman, Multivariate Adaptive Regression Splines, Technical Report No. 102, Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, Nov 1988 revised Aug 1990.

[27] H.M. Jalali, A. Kyani, Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a pca-mlr-ann approach, J. Chem. Inf. Comput. Sci. 44 (2004) 1328–1335.

[28] A. Tropsha, P. Gramatica, V.K. Gombar, Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.

[29] R. Christoph, R. Gerta, M. Markus Y, -Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (6) (2007) 2345–2357.

[30] J.R. Troyer, The Multiple discoveries of the first hormone herbicides, Weed Sci. 49 (2001) 290–297.

[31] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese, R.K. Agrawal, Validation of QSAR models-strategies and importance, Int. J. Drug Des. Discov. 3 (2011) 511–519.

[32] D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York, 1980.