ANALYSIS ARTICLE | OPEN ACCESS

DRUG DISCOVERY

Molecular descriptors in quantitative relationship with herbicidal activity of novel derivatives of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid

Nidhi Shekhawat[1*], Prithvi Singh[2]

ABSTRACT

The performed research focused on analyzing the herbicidal properties of several new analogs of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid using three different methods, namely partial least squares, PLS analysis, the combinatorial protocol in multiple linear regression, CP-MLR and the non-parametric Fujita-Ban, FB technique. The FB method employs the principle of additivity, which suggests that substituents contribute to the activity of the parent molecule. The study investigated the impact of different substituents, positioned in various ways, on the activity of the analogs. The most active compound in the series exhibited a pattern of substituents exhibiting positive contribution relative to the parent molecule. The CP-MLR technique employs statistically derived quantitative structure-activity relationship, QSAR models to elucidate the herbicidal action of congeners. In this procedure, three-dimensional descriptors such as asphericity, ASP, the second component accessibility directional WHIM index weighted by atomic masses, E2m, and the maximum autocorrelation of lag 2 weighted by atomic masses, R2m+ were identified as the most important. Furthermore, external validation of the models was performed using data generated from the test-set, and the models demonstrated their capacity for forecasting through applicability domain, AD study. These findings can be helpful in identifying potential analogs of the series. The dominance of the descriptors discovered by the CP-MLR investigation was corroborated by PLS analysis, and the computed activity values were shown to agree with observed ones using the FB, CP-MLR, and the PLS analyses.

Keywords: QSAR study, Molecular 3D-descriptors, Herbicidal activity, Derivatives of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid

To Cite:
Shekhawat N, Singh P. Molecular descriptors in quantitative relationship with herbicidal activity of novel derivatives of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid. *Drug Discovery* 2023; 17: e34dd1959
doi: https://doi.org/10.54905/disssi.v17i40.e34dd1959

Author Affiliation:
[1]Department of Zoology, Vedanta Post-Graduate Girls' College, Reengus-332 404, Rajasthan, India
[2]Ex-Senior Lecturer, SK Government College, Sikar-332 001 & Ex-Emeritus Research Fellow, UGC, New Delhi

[*]Corresponding author
Department of Zoology, Vedanta Post-Graduate Girls' College,
Reengus-332 404, Rajasthan
India
Email: nidhi14sikar@gmail.com

Contact List

Nidhi Shekhawat	nidhi14sikar@gmail.com
Prithvi Singh	psingh_sikar@rediffmail.com

Peer-Review History
Received: 08 September 2023
Reviewed & Revised: 12/September/2023 to 11/November/2023
Accepted: 15 November 2023
Published: 18 November 2023

Peer-Review Model
External peer-review was done through double-blind method.

Drug Discovery
pISSN 2278–540X; eISSN 2278–5396

DISCOVERY
SCIENTIFIC SOCIETY

1. INTRODUCTION

Herbicides are crucial in controlling unwanted weeds that grow with crops, thereby enhancing agricultural yield. However, the extensive and regular use of certain herbicides can lead to the development of weed resistance, necessitating the development of new herbicides that are less toxic, have minimal resistance, and have higher efficacy (Qu et al., 2021). Synthetic auxin herbicides, which have different structural moieties, are significant compounds with diverse modes of action and explicit binding sites in target proteins (Busi et al., 2018). Thus, these chemicals hold great potential for the development of new herbicides. Recently, herbicides containing the pyrazole moiety have demonstrated substantial herbicidal efficacy (Havrylyuk et al., 2016; Huang et al., 2017; Wu et al., 2012; Zhang et al., 2022).

Researchers have been exploring new chemicals, such as 4-amino-3,5-dichloro-6-pyrazolyl-2-picolinic acids with a phenyl-substituted pyrazole replacing the chlorine atom at position six of the picloram herbicide, following the discovery of 6-aryl-2-picolinate herbicides (Feng et al., 2023). The study involves synthesizing 33 new herbicidal compounds and evaluating their inhibitory activity, IC50, against *Arabidopsis thaliana* root growth. The objective of present communication is to establish the correlations between IC50 and chemometric molecular descriptors for a series of 33 compounds. The approach is called as the quantitative structure-activity relationship, QSAR study. The activity being the dependent variable is quantitatively expressed in terms of independent variables, or descriptors, Xis, using the multiple linear regression analysis, MLR.

The study utilizes an interpolative approach to calculate the activity values of all compounds in the series and compare them to their observed values. The observed and computed activities must match. The extrapolative nature of a QSAR study is beneficial in predicting the activities of new potential compounds beyond the synthesized compounds in the series, provided that the mode of action of new analogs remains similar to that of compounds in the parent series. The prediction, therefore, helps to reduce both the time and cost of exploring additional compounds in the series. The independent descriptors participated in a valid correlation Equation have the ability to reflect the nature of the forces that are operative during interaction with the receptor site(s). As a result, such information can assist in formulating a plausible molecular mechanism of action.

2. MATERIAL AND METHODOLOGY

The present study provides a report on the herbicidal activity values of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid derivatives, expressed as IC50 (μmol L-1), determined by measuring the concentration required to elicit 50% of the desired effect against the growth of *A. thaliana* roots. The estimates of herbicidal activity values for the compounds were obtained from published works (Feng et al., 2023). However, for present study, the activity estimate of each compound is expressed on a molar basis as -logIC50 (M) or as pIC50 (M). Table 1 presents the compounds and their corresponding pIC50 (M) values, while Figure 1 illustrates the general molecular structure of these congeners.

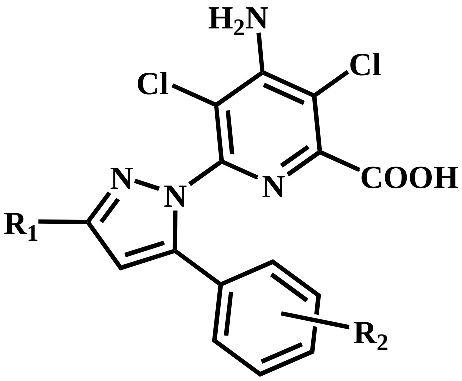

Figure 1 The general structure of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid derivatives.

Table 1 Observed and modeled herbicidal activities of novel derivatives of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid (Figure 1)

Compd	R1	R2	Obsd pIC50 (M)	Calcd pIC50 (M)		
				FB	Eq (2)	PLS
1	Me	4-Me	5.67	6.38	5.89	6.16
2	CHF2	4-Me	7.16	6.23	6.35	6.48
3	CF3	4-Me	5.94	6.17	5.74	6.01
4	Me	4-F	5.84	5.86	5.58	5.75
5	CHF2	4-F	5.66	5.71	5.95	5.89
6	CF3	4-F	5.73	5.66	5.47	5.46
7	Me	4-Cl	7.41	6.56	6.05	6.04
8	CHF2	4-Cl	5.85	6.41	5.65	5.78
9	CF3	4-Cl	6.07	6.36	5.30	5.62
10	Me	3-Cl	4.60	5.07	4.95	5.03
11	CHF2	3-Cl	5.15	4.92	5.03	5.02
12	CF3	3-Cl	5.12	4.87	5.17	5.10
13	Me	4-Br	6.05	6.03	6.04	5.94
14	CHF2	4-Br	5.91	5.88	5.89	5.94
15	CF3	4-Br	5.77	5.82	5.57	5.74
16	Me	2-Br	6.73	6.47	6.67	6.69
17	CHF2	2-Br	6.18	6.32	6.74	6.50
18	CF3	2-Br	6.14	6.27	6.30	5.94
19	Me	4-*i*-Pr	4.69	4.62	4.70	4.77
20	CHF2	4-*i*-Pr	4.44	4.47	4.43	4.51
21	CF3	4-*i*-Pr	4.36	4.41	4.71	4.58
22	Me	3,4-Cl2	5.17	5.29	5.28	5.56
23	CHF2	3,4-Cl2	5.19	5.14	5.53	5.39
24	CF3	3,4-Cl2	5.15	5.09	4.95	5.03
25	Me	4-Et	5.87	5.75	5.57	5.90
26	CHF2	4-Et	5.38	5.60	5.23	5.29
27	CF3	4-Et	5.65	5.55	5.86	5.56
28	Me	4-*n*-Pr	4.47	4.60	4.61	4.31
29	CHF2	4-*n*-Pr	4.35	4.45	4.38	4.37
30	CF3	4-*n*-Pr	4.64	4.40	4.28	4.19
31	Me	4-*t*-Bu	4.56	4.43	4.53	4.48
32	CHF2	4-*t*-Bu	4.14	4.28	4.03	4.24
33	CF3	4-*t*-Bu	4.24	4.22	4.08	4.08

Fujita-Ban analysis

The Fujita-Ban analysis is a non-parametric method Fujita and Ban, (1971) and is based on the additive principle of substituent's contribution of activity to the parent moiety. In this method, pIC50 is considered a free energy-related parameter that is additive in nature. The analysis is limited to the parent data-set, but it identifies the substituents that have a positive impact on activity relative to the parent compound. To expand the scope of Fujita-Ban analysis, a multiple linear regression, MLR analysis was performed to establish important correlations between the molecular descriptors and activity profiles of the compounds under investigation. The models generated from the MLR analysis have the potential to reveal new candidate compounds outside of the named series, as well as offer insights into the potential modes of action of these compounds at different receptor sites.

The MLR analysis is a multi-step process that involves computing molecular descriptors and filtering out only the important ones. These important descriptors are then correlated with activity values to develop statistically significant models. The models are validated both internally and externally. Finally, the molecular descriptors that are shared in the developed models may be interpreted in terms of various binding forces, including covalent bonding, ionic (electrostatic) interactions, ion-dipole and dipole-dipole interactions, hydrogen bonding, charge-transfer interactions, hydrophobic interactions, halogen bonding, van der Waals interactions, etc. The following sections provide a detailed discussion of each step involved in the MLR analysis.

Calculations of molecular descriptors

First, the structures of all 33 compounds were drawn in 2D ChemDraw then converted into 3D modules. These modules underwent an energy minimization process in MOPAC, using the AM1 procedure for closed-shell systems, to ensure a well-defined conformer relationship among the compounds. The DRAGON software was used to compute the molecular descriptors of these compounds. The software computes hundreds of molecular descriptors pertaining to 0D, 1D, 2D, and 3D classes. After elimination of inter-correlated descriptors, a total number of 434 descriptors for 0D to 2D classes and 648 descriptors for the 3D class were saved in two separate files.

Development of regression models

As the descriptors have varying magnitudes, the regression coefficients and intercept would reflect this imbalance. The descriptors of the data-set were scaled between 0 and 1 to assign equal weights in a model, preventing bias against unscaled descriptors with higher or lower values (Golbraikh and Tropsha, 2002). The combinatorial protocol in a multiple linear regression, CP-MLR computational process was used to develop QSAR models with scaled descriptors (Prabhakar, 2003). When performing a QSAR analysis, it is crucial to select the most significant descriptors from the multivariate space to obtain meaningful models. The CP-MLR is one of the many techniques available that uses a filter-based variable selection method to simplify the selection process and generate unique and statistically significant models. Our previous publications (Sharma et al., 2010; Sharma et al., 2011; Sharma et al., 2013; Singh, 2013) provide detailed information on the strategy and its applications.

The CP-MLR analysis computation software has four filters implanted in it. The first filter allows only those descriptors with inter-descriptor correlations equal to or greater than 0.79 to be entered. The second filter regulates the entry of descriptors into a regression model by setting a threshold for their coefficient's t-values at 2.0. The third filter makes it possible to compare models with different descriptor counts using r-bar, which is the square root of the adjusted multiple correlation coefficient of the model Equation. The fourth filter measures the internal robustness of the model using the leave-one-out index Q2LOO, where 0.3 ≤ Q2LOO ≤ 1.0. The r-bar value of the prior optimum model (third filter) was enhanced by increasing the number of useful descriptors, which became the new upper limit for subsequent model creation. Repeated randomization of the activity profile was done to test for chance correlations in each cross-validated model (So and Karplus, 1997; Prabhakar et al., 2004).

Every model was put through 100 simulation runs with random activity for this. To describe the percent chance correlation of the model under discussion, the scrambled activity models with regression statistics better than or equal to those of the original activity model were counted. To evaluate the statistical significance of a model, the multiple correlation coefficient, r, standard deviation, s and F-ratio between the variances of calculated to observed activities were used. The F-ratio is represented as Fn, n-k-1, where n is the number of compounds and k is the number of independent descriptors, and is compared with critical F-values. The leave-one-out and leave-five-out procedures were used to determine the internal robustness of the model. The resulting statistical indices Q2LOO and Q2L5O, greater than 0.5, indicate a reliable model. Additionally, the Kubinyi function, FIT Kubinyi, (1994), Friedman's lack of fit, LOF Friedman, (1990), and Akaike's information criteria, AIC Akaike, (1973), Akaike, (1974) were calculated to evaluate the best model.

For external validation of the developed models, a test-set was chosen, containing almost 24% of the total population, while the remaining compounds were included in the training-set. The r2Test indices were determined accordingly. The selection of the compounds for the test-set was made through SYSTAT using the single linkage hierarchical cluster procedure, which involved the Euclidean distances of the activity values. A cluster tree was generated, and the test-set compounds were selected in such a way as to keep them at the maximum possible distance from each other. The normalized Euclidean distances were computed to join the objects of the cluster, and the single linkage clustering procedure was followed since it generates long clusters and provides different object intervals to choose from.

Setting the applicability domain

To ensure that all compounds of a given series fall within the applicability domain, AD, a study was performed to identify any “outlier” or structurally influential compounds. This is crucial in accurately predicting new analogs of the series. The AD is determined using the Williams plot, which plots standardized residuals against the leverage values of all compounds in the training-domain (Gramatica, 2007; Eriksson et al., 2003). The domain is established within a plot by considering a measure (±β×s.d.) and the leverage threshold value, h*. The value of h* is specified at 3(k + 1)/n, where k is the number of independent variables of the model under consideration and n is the number of compounds in the training-set. The figure may then be used to visually identify the Y-outlier or response outlier and the X-outlier or structurally influential compound. When a compound's leverage value is less than the threshold value h*, the prediction becomes trustworthy, and the calculated and observed activity values of the training-set chemicals agree. However, when the leverage value exceeds h*, the forecast becomes unreliable.

The PLS regression analysis

The partial least squares, PLS linear regression is a powerful modeling technique that addresses the challenges of multiple linear regression, MLR when working with incomplete, noisy, and collinear variables in both the descriptor matrix, X and the activity profile matrix, Y. This technique utilizes a small number of latent variables, LVs known as PLS components, which are linear combinations of the original variables. Additionally, the Y matrix is used to identify the most appropriate LVs in X for predicting Y variables. Before applying the PLS approach, descriptors are auto-scaled, and data is mean-centered to ensure that the scaling of X is consistent despite varying magnitudes. This critical step is necessary since preprocessed descriptors may have different orders of magnitude. To determine the optimal number of LVs, cross-validation was carried out by setting aside a portion of the calibration data for prediction. The process was repeated for each sample, with the predicted values of the excluded data compared to the observed values using the predicted residual sum of squares, PRESS. Every time a new LV was added to the model, PRESS was calculated.

3. RESULTS AND DISCUSSION

In the Fujita-Ban analysis, compound 1 from Table 1 was considered as the parent compound. To explore the impact of substituents at R1 and three different positions (2-, 3-, and 4-) of R2, a matrix was formulated consisting of 33 compounds in the series. The matrix was analyzed using the MRA, and the contributions of the parent compound, μ0 and different substituents are listed in (Table 2). The data within parentheses are the 90% confidence intervals.

Table 2 The herbicidal activity contributions of substituents and parent moiety (Figure 1 for general structure)

Substituent		Contribution
R1	CF3	-0.205 (±0.303)
	CHF2	-0.151 (±0.303)
2- R2	Br	0.125 (±0.821)
3- R2	Cl	-1.275 (±0.580)
4- R2	Br	-0.350 (±0.580)
	Cl	0.186 (±0.580)
	Et	-0.628 (±0.580)
	F	-0.516 (±0.580)
	H	-0.033 (±0.821)
	i-Pr	-1.762 (±0.580)
	n-Pr	-1.775 (±0.580)
	t-Bu	-1.949 (±0.580)
Parent contribution, μ0		6.378 (±0.446)

Additionally, the statistical parameters calculated from the study are given below:

n = 33, r = 0.925, s = 0.411, F(13, 19) = 9.370

The r2 value accounted for 86% of variance in the observed activity profiles while F-value remained significant at 99% level (F13, 19(0.01) = 3.249). The calculated activities closely matching the observed ones are shown in (Table 1). For convenience, a plot comparing the calculated and observed pIC50s is included in (Figure 2A).

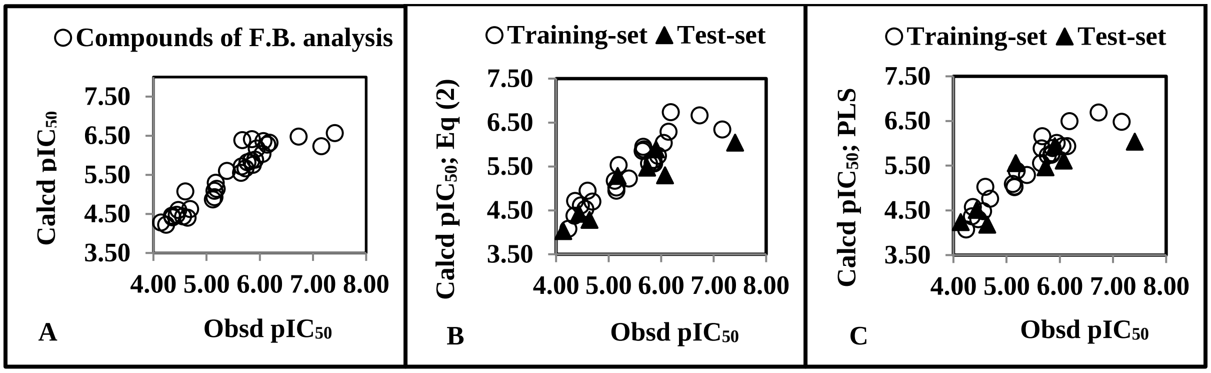

Figure 2 The plot between observed and calculated pIC50 values, Fujita-Ban analysis; A, regression Equation (2); B and PLS analysis; C.

According to Table 1, the parent compound 1 has a substituent pattern of Me at R1, H at 2-R2, H at 3-R2, and Me at 4-R2. The substituents CF3 and CHF2 at R1 and Cl at 3-R2 have negative contribution to activity, which means that Me and H, in that order, are preferred. On the other hand, the substituents Br and Cl at positions 2-R2 and 4-R2, respectively, are desirable since they have positive activity contribution. Such features are observed in compound 7 (Table 1), which is the highest active analog of the series. The CP-MLR method used next to correlate the herbicidal capabilities of compounds in Table 1 with 3D molecular descriptors. Initially, 648 scaled 3D descriptors were used for this purpose. Attempts were made to associate the activity profiles of the substances using 0D to 2D class descriptors. However, such descriptors were unacceptable to address the herbicidal actions of compounds as they resulted in poor models when compared to the 3D descriptors.

As a result, 3D descriptors were employed to obtain models in one descriptor, two descriptors, and three descriptor increments. A test-set comprising of eight compounds (6, 7, 9, 14, 20, 22, 30 and 32; Table 1) was considered to validate the developed models. The models involving three descriptors remained statistically significant in explaining the variance in observed activities of the congeners. To examine prediction models, compounds from Table 1 were used to correlate their herbicidal activity in terms of scaled 3D descriptors. A total of nine models were obtained, considering r-bar value of 0.915. These models come with the necessary statistical parameters. They are composed of 16 shared descriptors, which are listed in Table 3, with their name, class, physical meaning, average regression coefficient and incidence.

Table 3 Identified descriptors along with their class, physical meaning, average regression coefficient and incidence, in modeling of herbicidal activity

No	Name	Class	Physical meaning	Avg reg coefficient (incidence)
1	ASP	Geometrical	Represents asphericity	-2.041 (1)
2	RDF020m	RDF	Radial distribution function-2.0/weighted by atomic masses	0.925 (1)
3	Mor22u	3D-MoRSE	3D-MoRSE-signal 22/unweighted	-1.617 (1)
4	Mor23u	3D-MoRSE	3D-MoRSE-signal 23/unweighted	2.244 (1)
5	Mor12m	3D-MoRSE	3D-MoRSE-signal 23/weighted by atomic masses	1.233 (1)
6	Mor15m	3D-MoRSE	3D-MoRSE-signal 15/weighted by atomic masses	1.821 (2)
7	Mor24m	3D-MoRSE	3D-MoRSE-signal 24/weighted by atomic masses	1.198 (1)
8	Mor27m	3D-MoRSE	3D-MoRSE-signal 27/ weighted by atomic masses	-1.553 (1)
9	Mor22e	3D-MoRSE	3D-MoRSE-signal 22/weighted by atomic Sanderson electronegativities	-1.515 (4)
10	Mor23e	3D-MoRSE	3D-MoRSE-signal 23/weighted by atomic Sanderson electronegativities	2.189 (1)
11	E2m	WHIM	2nd component accessibility directional WHIM index/weighted by atomic masses	-1.511 (1)
12	P2v	WHIM	2nd component shape directional WHIM index/weighted by atomic van der Waals volumes	1.534 (1)
13	Km	WHIM	K global shape index/weighted by atomic masses	-2.007 (1)
14	R6u	GETAWAY	R autocorrelation of lag 6/unweighted	-1.726 (7)
15	R2m+	GETAWAY	R maximal autocorrelation of lag 2/weighted by atomic masses	1.734 (1)

16	R7m+	GETAWAY	R maximal autocorrelation of lag 7/weighted by atomic masses	1.006 (2)

However, out of these nine models, only two were able to meet the requirement of having the r2Test value greater than 0.5. These two models are presented in increasing order of statistical significance as Equation (1) and Equation (2).

$$\begin{aligned}&pIC50 = 1.198(\pm 0.331)Mor24m + 1.533(\pm 0.213)P2v - 2.007(\pm 0.246)Km + 5.256\\&n = 25,\ r = 0.929,\ s = 0.305,\ F(3,21) = 43.855,\ AIC = 0.129,\ FIT = 3.870,\\&LOF = 0.135,\ Q2LOO = 0.787,\ Q2L5O = 0.799,\ r2Test = 0.660 \qquad (1)\end{aligned}$$

$$\begin{aligned}&pIC50 = -2.041(\pm 0.245)ASP - 1.511(\pm 0.198)E2m + 1.734(\pm 0.253)R2m+ + 6.569\\&n = 25,\ r = 0.929,\ s = 0.304,\ F(3,21) = 44.150,\ AIC = 0.128,\ FIT = 3.895,\\&LOF = 0.135,\ Q2LOO = 0.811,\ Q2L5O = 0.833,\ r2Test = 0.666 \qquad (2)\end{aligned}$$

The F-values, in the Equations above, are significant at 99% level (F3, 21(0.01) = 4.874). Moreover, the enclosed standard errors associated with regression coefficients are significant at a level exceeding 95%. The indices Q2LOO and Q2L5O (> 0.5) indicate the internal robustness of the models derived, while the index r2Test greater than 0.5 suggests that the selected test-set can be used for external validation of the models. The signs of the regression coefficients indicate the direction of influence of explanatory variables. A positive coefficient associated with a descriptor indicates that it will improve the activity of a compound. In contrast, a negative coefficient will lead to a detrimental effect.

Equation (1) incorporates three descriptors: Mor24m, P2v and Km. Mor24m denotes the 3D-MoRSE-signal 24 weighted by atomic masses, whereas, P2v represents the 2nd component shape directional WHIM index weighted by atomic van der Waals volumes. Lastly, Km represents the K global shape index weighted by atomic masses. In Equation (2), the descriptors are ASP, E2m, and R2m+. ASP denotes the asphericity representation, E2m represents the 2nd component accessibility directional WHIM index weighted by atomic masses and R2m+ represents the maximal autocorrelation of lag 2 weighted by atomic masses. Positive regression coefficients for Mor24m, P2v, and R2m+ indicate that a higher value of these descriptors will increase the activity, while negative regression coefficients for Km, ASP, and E2m indicate that a lower value of these descriptors will be beneficial to improve the activity of a compound.

Based on the results of Equation (2), it may be deduced that it explains 86% of the variance in the observed activity profiles, as demonstrated by the squared correlation coefficient, r2 of 0.863. The statistical parameters of this particular Equation have been fine-tuned to the most significant model, making it the best choice for calculating the herbicidal activities of all 33 congeners in the series. These calculated values are listed in Table 1 for easy comparison with observed values. Furthermore, a graphical representation (labeled as B in Figure 2) indicates a close match between the observed and calculated pIC50s for both the training-set and test-set compounds.

The study mentioned above was supported further by the implementation of PLS analysis (Wold, 1978; Kettaneh et al., 2005; Stahle and Wold, 1988). The most influential six descriptors were subjected to the PLS analysis relating to the herbicidal activity of the compounds. The outcomes are given in Table 4 wherein the symbols, SE and FC are the standard error associated to PLS coefficient and fraction contribution of regression coefficient respectively.

Table 4 PLS and MLR-like PLS Equations from the descriptors of CP-MLR identified models for herbicidal Activity

A PLS Equation				B PLS Regression statistics			
PLS components		PLS coefficient (SE)		Symbol		Estimate	
Component-1		0.481 (0.035)		n		25	
Component-2		0.206 (0.042)		r		0.953	
Constant		5.431		s		0.244	
				F		108.151	
				Q2LOO		0.878	
				Q2L5O		0.876	
				r2Test		0.685	
C MLR-Like PLS Equation							
S. No.	Descriptor	MLR-like Coefficient	FC (order)	S. No.	Descriptor	MLR-like Coefficient	FC (order)
1	ASP	-0.765	0.139 (5)	5	Km	-1.178	-0.217 (1)
2	Mor24m	0.771	-0.106 (6)	6	R2m+	0.815	0.145 (4)

3	E2m	-0.803	-0.187 (3)	-	Constant	5.713	-
4	P2v	0.984	0.205 (2)	-	-	-	-

The descriptors were uniformly weighted as they were auto-scaled to attain a zero mean and unit standard deviation. Two components were found to be optimal for cross-validation, explaining 91% (r2 = 0.908) of the variances in the observed activity values. Provided in Table 4 are the PLS Equations for two components, along with MLR-like PLS coefficients of descriptors for herbicidal activities. As demonstrated in Table 1, the calculated pIC50 values of both the training-set and test-compounds closely align with the observed values. Additionally, Figure 2C plots the calculated *vs.* observed pIC50 values for the training- and test-compounds. Table 4 shows the different orders of identified descriptors and their levels of significance with the biological activity. Figure 3 provides a graphical representation of the fraction contribution of normalized regression coefficients of these descriptors to the activity.

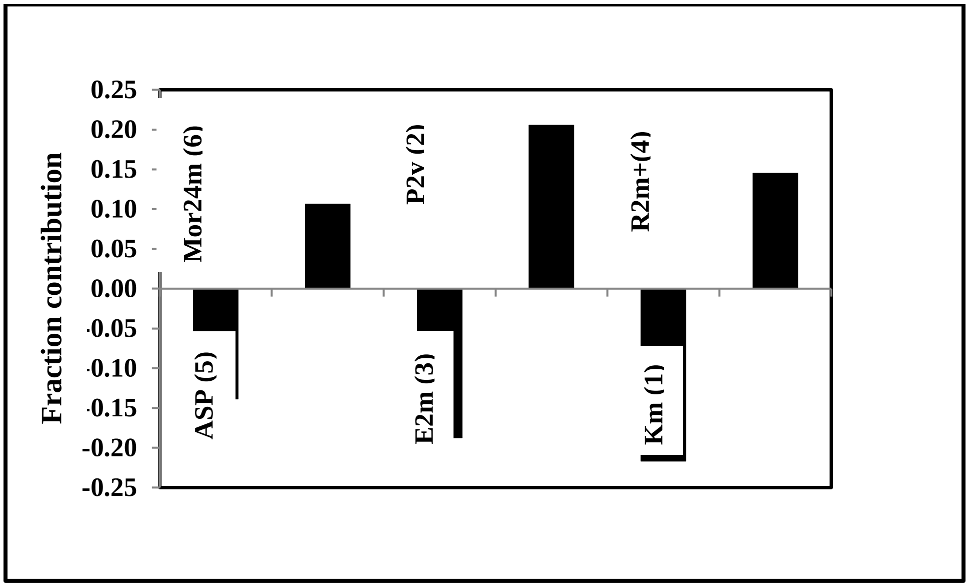

Figure 3 The plot of fraction contribution of MLR-like PLS coefficients (normalized) against six identified descriptors (Table 4) associated with herbicidal activities of the compounds.

A lower order indicates a higher level of significance for a given descriptor in addressing the biological activity. Descriptors with positive contributions enhance the activity, and higher values are desirable for further improvement. However, descriptors with negative contributions reduce the activity, and lower or more negative values of such descriptors may enhance the activity of a compound. Following these guidelines, an investigation was carried out on various analogs to ascertain if they demonstrate superior activity profiles compared to the compounds listed in (Table 1). Through this investigation, three congeners were discovered, featuring structural modifications at R1 and R2 (as visualized in Figure 1), and predicted pIC50 values; obtained via Equation (2) and PLS analysis, which yielded the most significant results. Below are the predicted activity values for compounds I, II, and III using Equation (2) and PLS analysis:

Compound	R1	R2	Predicted pIC50(M)	
			Eq (2)	PLS
I	Me	2-Br, 4-Cl	8.52	8.45
II	Me	2-Br, 4-Et	8.08	7.88
III	CHF2	2-Br, 4-Cl	7.61	7.88

Many attempts were made to vary -Me, -CHF2, and -CF3 at R1, and di-substitution from -Br, -Cl, -Me, -Et, CH2Br, -CH2Cl in various permutations at 2- and 4-positions pertaining to R2. However, the predicted herbicidal activity profiles obtained for the above three compounds were found superior compared to the titled compounds that have been listed in (Table 1). Therefore, these compounds are suggested for further exploration. The applicability domain, AD of models derived from the complete data-set was assessed through the Williams plot. It involves the plot of standardized residuals against leverage (hi) values. The most significant descriptors (ASP, E2m, and R2m+) were considered to develop a model for the complete data-set. The resulting model is depicted through regression Equation (3).

$$pIC50 = -2.223(\pm0.284)ASP\ -1.525(\pm0.207)E2m +1.656(\pm0.277)R2m+ +6.788$$
$$n = 33,\ r = 0.898,\ s = 0.386,\ F(3,29) = 40.126,\ AIC = 0.190,\ FIT = 2.866,$$
$$LOF = 0.195,\ Q2LOO = 0.755,\ Q2L5O = 0.694 \quad (3)$$

The assessment of the AD involves utilizing standardized residuals and leverage values. By considering the limits of standardized residuals as ±β ×s.d., Y-outliers are identified, while the leverage threshold is taken as h* (= 3(k + 1)/n). A visual representation of the influential descriptors is presented in Figure 4, displaying the training-set and the test-set compounds.

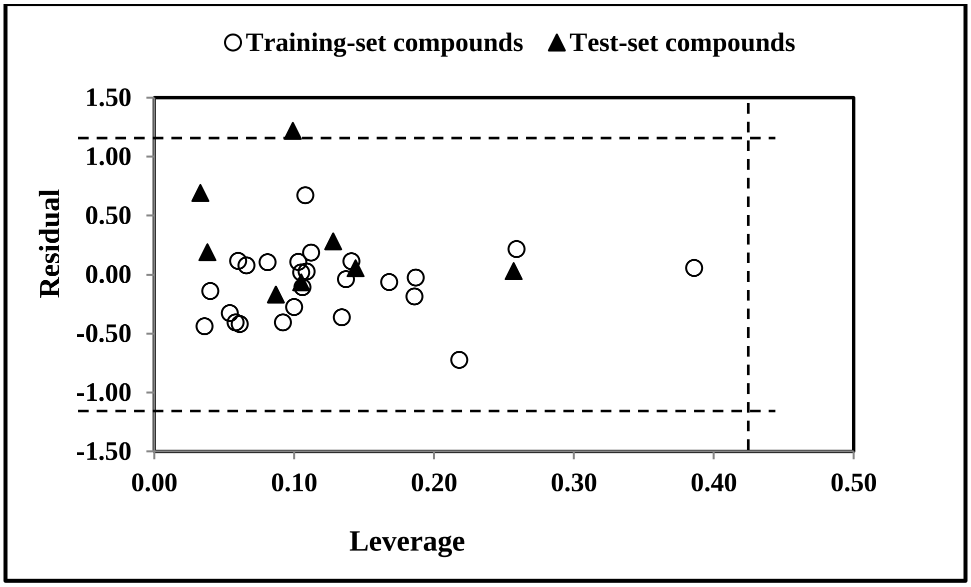

Figure 4 The Williams plot for whole data-set for DPP-4 inhibition activities of titled compounds, listed in Table 1 (leverage threshold h* = 0.480 and residual limits = ±1.158).

The suggested model effectively matches the most significant parameters, with strong fitting power for both the training-set and test-set compounds, and is capable of accurately evaluating external data. All compounds, except compound 7, were within the AD indicating that the model could assess both the training-set and test-set compounds correctly. However, compound 7 appeared to be an outlier but could not be removed from the study as it was the most potent congener in the series. Removal of the lowest or the highest active compound could mislead the results in a QSAR study.

4. CONCLUSION

The effectiveness of 33 novel derivatives of 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid in killing weeds was analyzed quantitatively using three methods. These methods are the Fujita-Ban, FB approach, the combinatorial protocol in multiple linear regression, CP-MLR computational process and the partial least squares, PLS analysis. The FB methodology is a non-parametric approach, based on the additive principle of the substituent's activity contribution to the parent compound. This approach is limited to the parent data-set, but it identifies the activity contribution of the substituents relative to the parent compound. The analysis focused on the activity contributions of different substituents at R1 and R2 positions (Figure 1) that led to the most active analog. Compound 7 (Table 1) had the highest observed activity profile, with patterns of substituents that were effective. The calculated activity values were in close agreement with the observed ones.

The CP-MLR procedure uses quantitative structure-activity relationship, QSAR models to explain the herbicidal activity of a group of congeners. Among the most significant 3D-descriptors were the representation of asphericity, ASP, the 2nd component accessibility directional WHIM index/weighted by atomic masses, E2m, and the R maximal autocorrelation of lag 2/weighted by atomic masses, R2m+. The identified models were validated through statistics from the test-set. The AD analysis confirmed that the model has adequate predictability, as both the training-set and test-set analogs were present within the domain.

Therefore, the model correctly predicted the herbicidal activities of all the compounds in the series. The discussion highlighted some guidelines that were helpful in exploring new potential analogs of the series. The effectiveness of the MLR-like PLS coefficient (normalized) was calculated concerning the activities of the compounds in conjunction with six identified descriptors. The PLS and MLR-like model Equations were utilized to confirm the superiority of the descriptors obtained from the CP-MLR study. Moreover, the computed pIC50s were consistent with the observed ones.

Acknowledgment

We appreciate the help from our colleagues and one of the authors' institutions.

Author contribution

The survey of literature, work on Chemdraw software and computation of molecular descriptors were carried out by NS; handling CP-MLR and PLS computational software, establishment of AD, prediction of new compounds, writing work of paper, and monitoring the entire presentation of the MS is performed by PS.

Informed consent

Not applicable.

Ethical approval

Not applicable.

Conflicts of interests

The authors declare that there are no conflicts of interests.

Funding

No internal or external funding was available with authors to support this work.

Data and materials availability

All data associated with this study are mentioned in appropriate sections of this MS.

REFERENCES AND NOTES

1. Akaike H. A new look at the statistical identification model. IEEE Trans Automat Control 1974; 19(6):716-23. doi: 10.1109/TAC.1974.1100705
2. Akaike H. Information theory and an extension of the minimum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 1973; 267-281.
3. Busi R, Goggin DE, Heap IM, Horak MJ, Jugulam HM, Masters RA, Napier RM, Riar DS, Satchivi NM, Torra J, Westra P, Wrught TR. Weed resistance to synthetic auxin herbicides. Pest Manag Sci 2018; 74(10):2265-2276. doi: 10.1002/ps.4823
4. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003; 111(10):1361-75. doi: 10.1289/ehp.5758
5. Feng T, Liu Q, Xu ZY, Li HT, Wei W, Shi RC, Zhang L, Cao YM, Liu SZ. Design, synthesis, herbicidal activity, and structure-activity relationship study of novel 6-(5-aryl-substituted-1-pyrazolyl)-2-picolinic acid as potential herbicides. Molecules 2023; 28(3):1431. doi: 10.3390/molecules28031431
6. Friedman J. Laboratory for computational statistics. In: Technical report number 102. Stanford, Stanford University, 1990.
7. Fujita T, Ban T. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. J Med Chem 1971; 14(2):148-52. doi: 10.1021/jm00284a016
8. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model 2002; 20(4):269-76. doi: 10.1016/s1093-3263(01)00123-1
9. Gramatica P. Principles of QSAR models validation: Internal and external. QSAR Comb Sci 2007; 26(5):694-701. doi: 10.1002/qsar.200610151
10. Havrylyuk D, Roman O, Lesyk R. Synthetic approaches, structure activity relationship and biological applications for pharmacologically attractive pyrazole/pyrazoline-thiazolidine-based hybrids. Eur J Med Chem 2016; 113:145-66. doi: 10.1016/j.ejmech.2016.02.030

11. Huang D, Liu A, Liu W, Liu X, Ren Y, Zheng X, Pei H, Xiang J, Hang M, Wang X. Synthesis and insecticidal activities of novel 1*H*-pyrazole-5-carboxylic acid derivatives. Heterocycl Commun 2017; 23(6):455-60. doi: 10.1515/hc-2017-0110
12. Kettaneh N, Berglund A, Wold S. PCA and PLS with very large data-sets. Comput Stat Data Anal 2005; 48(1):69-85. doi: 10.1016/j.csda.2003.11.027
13. Kubinyi H. Variable selection in QSAR studies. I. An evolutionary algorithm. Quant Struct-Act Relat 1994; 13:285-94. doi: 10.1002/qsar.19940130306
14. Kubinyi H. Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. Quant Struct-Act Relat 1994; 13:393-401. doi: 10.1002/qsar.19940130403
15. Prabhakar YS, Solomon VR, Rawal R, Gupta MK. CP-MLR/PLS Directed structure-activity modeling of the HIV-1 RT inhibitory activity of 2,3-diaryl-1,3-thiazolidin-4-ones. QSAR Combn Sci 2004; 23(4):234-244. doi: 10.1002/qsar.200330854
16. Prabhakar YS. A combinatorial approach to the variable selection in multiple linear regression: Analysis of Selwood et al. data-set - A case study. QSAR Comb Sci 2003; 22(6):583-595. doi: 10.1002/qsar.200330814
17. Qu RY, He B, Yang JF, Lin HY, Yang WC, Wu QY, Li QX, Yang GF. Where are new herbicides? Pest Manag Sci 2021; 77 (6):2620-2625. doi: 10.1002/ps.6285
18. Sharma BK, Singh P, Pilania P, Sarbhai K, Prabhakar YS. CP-MLR/ PLS directed study on apical sodium-codependent bile acid Transporter inhibition activity of benzothiepines. Mol Divers 2011; 15(1):135-47. doi: 10.1007/s11030-009-9220-2
19. Sharma BK, Singh P, Prabhakar YS. QSAR rationale of matrix metalloproteinase inhibition activity in a class of carboxylic acid-based compounds. J Pharm Res Int 2013; 3(4):697-721. doi: 10.9734/BJPR/2013/3903
20. Sharma BK, Singh P, Shekhawat M, A ratioale for the activity profile of benzenesulfonamide derivatives as cyclooxygenase (COX) inhibitors. Eur J Med Chem 2010; 45(6):2389-95. doi: 10.1016/j.ejmech.2010.02.019
21. Singh P. Molecular descriptor in modeling the tumour necrosis factor-α converting enzyme inhibition activity of novel tartrate-based analogues. Ind J Pharm Sci 2013; 75(1):36-44. doi: 10.4103/0250-474X.113539
22. So SS, Karplus M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. J Med Chem 1997; 40(26):4347-59. doi: 10.1021/jm970487v
23. Stahle L, Wold S. Multivariate data analysis and experimental design in biomedical research. Prog Med Chem 1988; 25:291-338. doi: 10.1016/S0079-6468(08)70281-9
24. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. Technomet 1978; 20:397-405. doi: 10.1080/00401706.1978.10489693
25. Wu J, Song BA, Hu D-Y, Yue M, Yang S. Design, Synthesis and insecticidal activities of novel pyrazoleamides containing hydrazone substructures. Pest Manag Sci 2012; 68(5):801-10. doi: 10.1002/ps.2329
26. Zhang X, Wei Z, Wang Y, Yang L, Yuan H, Feng J, Gao Y, Lei P, Ma Z. Synthesis and antifungal activity of 3-(difluoromethyl)-1-methyl pyrazole-4-carboxylic oxime esters. Chin J Pestic Sci 2022; 24(1):59-65. doi: 10.16801/j.issn.1008-7303.2021.0140