J Mol Model (2010) 16:137–153
DOI 10.1007/s00894-009-0528-8

ORIGINAL PAPER

Docking and 3D QSAR studies of protoporphyrinogen oxidase inhibitor 3H-pyrazolo[3,4-d][1,2,3]triazin-4-one derivatives

Kunal Roy · Somnath Paul

Received: 3 February 2009 / Accepted: 22 April 2009 / Published online: 19 June 2009

Abstract Docking and three dimensional quantitative structure - activity relationship (3D-QSAR) studies have been performed for protoporphyrinogen oxidase (PPO) inhibitor 3H-pyrazolo[3,4-d][1,2,3]triazin-4-one analogues which are potential herbicides to protect agricultural products from unwanted weeds. The 3D-QSAR studies have been carried out using shape, spatial, electronic and molecular field descriptors along with a few structural parameters. The chemometric tools used for the analyses are genetic function approximation (GFA), partial least squares (PLS) and genetic partial least squares (G/PLS). The whole data set (n=34) was divided into a training set (75% of the data set) and a test set (remaining 25%) on the basis of *K*-means clustering technique applied on topological, spatial and electronic descriptor matrix. Models developed from the training set were used to predict the activity of the test set compounds. All the models have been validated internally, externally and by Y-randomization technique. Docking studies suggest that the molecules bind with a hydrophobic pocket of the enzyme formed by some nonpolar amino acid (Ile168, Ile311, Ile412, Met365, Phe65 and Val164) residues. The QSAR studies suggest that for better activity the molecules should have symmetrical shape in the 3D space. For better PPO inhibitory activity, there should be a balance between the electrophilic and nucleophilic characters of the inhibitors. The charged surface area descriptors suggest that, the positive charge distributed over a large surface area may enhance the activity. Molecular field probes reflect that increase in steric volume and positively charged surface area may enhance the herbicidal activity.

Keywords GFA · PLS · G/PLS · K-means cluster · Docking · QSAR · Protoporphyrinogen oxidase

Introduction

Agriculture has played a very important role in the development of human civilization. It is widely believed that domestication of plants, allowed humans to settle in a place and give up their previous migrant life style. However, there are certain threats like weeds, insects, fungus, pests *etc*. which can cause reduction of the agricultural productivity. Among them, weeds play a major role to restrict the agricultural efficiency in modern civilization. Weeds are generally considered as unwanted plants in human made settings like gardens, agricultural areas *etc*. because (a) they might restrict light to the desirable plants, (b) they can take the nutrients from soil leaving the desired plant unfed and making them less productive, (c) they can spread plant pathogens that infect and diminish the quality of crop [1]. There should be a control measure to protect the agricultural products from the above mentioned harmful threats either chemically or biologically (genetically). Genetic control is very complicated and expensive, making chemical control the first choice. Herbicides have been broadly classified into two groups according to their activity. They are (1) contact herbicides (destroy only the plant tissue in contact with the chemical) and (2) systemic herbicides (chemicals translocated through the plant circula-

Electronic supplementary material The online version of this article (doi:10.1007/s00894-009-0528-8) contains supplementary material, which is available to authorized users.

K. Roy (✉) · S. Paul
Drug Theoretics and Cheminformatics Laboratory,
Department of Pharmaceutical Technology, Jadavpur University,
Kolkata 700032, India
e-mail: kunalroy_in@yahoo.com
URL: http://www.geocities.com/kunalroy_in

tion system, either from foliar application down to the roots, or from soil application up to the leaves) [2].

Computer-aided chemical design has been extensively applied in the area of modern drug discovery, ecotoxicological modelling and design of agrochemicals for its high efficiency in the design of new compounds and optimization of lead compounds, thus saving both time and economic costs in the large-scale experimental synthesis and biological tests [3]. Quantitative structure-activity relationship (QSAR) helps us to understand structure-activity relationship (SAR) in a quantitative manner. It is one of the most important applications of chemometrics, giving information useful for the design of new compounds acting on a specific target. QSAR attempts to find a consistent relationship between biological activity or toxicity and molecular properties. Thus, QSAR models can be used to predict the activity of new compounds. QSAR models have been reported by different groups of researchers for agrochemicals like herbicides, fungicides and insecticides. Xi *et al.* [4] have performed density functional theory based QSAR study for herbicidal sulfonylurea analogues using general quantum chemical descriptors. Hao Peng *et al.* [5] have performed molecular docking and three-dimensional QSAR studies on the herbicides 1-(substituted phenoxyacetoxy) alkylphosphonates which bind to the E1 component of pyruvate dehydrogenase. Li Zhang *et al.* [6] have developed a DFT-based QSARs study of protoporphyrinogen oxidase inhibitor phenyl triazolinones. Jian-Guo Wang *et al.* [7] have performed CoMFA and CoMSiA analyses of a new family of sulfonylurea herbicides. Yang *et al.* [8] have synthesized a series of 3H-[3,4-d][1,2,3] triazin-4-one derivatives which act as systemic herbicides by inhibiting protoporphyrinogen oxidase. The protoporphyrinogen oxidase (PPO, E.C. 1.3.3.4) is the last enzyme in the common tetrapyrrole biosynthesis pathway before the pathway branches towards chlorophyll (plant) and heme (animal) synthesis. These PPO inhibiting herbicides cause peroxidative destruction of cellular membrane and bleaching of plant tissues in the presence of light. In contrast to the other herbicides, PPO inhibitors have certain advantages like quick onset of action (necrosis within 24 hrs), long lasting effect and have a wide range of activity [9]. Previously it was assumed that either a heterocyclic structure with one or more nitrogen atoms or a polysubstituted benzene ring that links with the nitrogen atom of the heterocyclic ring [10–16] may provide good herbicidal activity. When the polysubstituted benzene ring was replaced by a benzo-heterocycle ring, the corresponding compounds (flumioxazin) also possessed excellent PPO inhibitory activity and herbicidal activities [10]. Yang *et al.* [17] have designed and synthesized a series of pyrazolo [5,1-d][1,2,3, 5] tetrazin-4(3H)-one derivatives and evaluated their herbicidal activities against *Brassica campestris* in a previous paper. For further improvement Yang *et al.* have synthesized a series of 3H-pyrazolo [3,4-d][1,2,3] triazin-4-one derivatives [8] which act as systemic herbicides.

In the present paper, we have docked 3H- pyrazolo [3,4-d][1,2,3] triazin-4-one derivatives into the enzyme PPO isolated from *Myxococcus xanthus* to explore important interactions between the ligands and the active site of the PPO enzyme. Further we have performed the 3D QSAR analysis to obtain a clear insight of the structure-activity relationship of this class of compounds.

Materials and methods

The protoporphyrinogen oxidase inhibitory data (IC_{50}) of 33 3H-pyrazolo [3,4-d] [1,2,3] triazin-4-one derivatives and one reference compound (flumioxazin) [8] against corn protoporphyrinogen oxidase (PPO) were converted to reciprocal logarithmic values [$pIC_{50}=-\log IC_{50}$] which have been used for the QSAR analysis. There are four regions of structural variations in the 3H-pyrazolo [3,4-d][1,2,3] triazin-4-one compounds (Table 1). Subclass **A** of these compounds differs from subclass **B** in the position of R_1 substituent. Flumioxazin is structurally different from the other compound and forms the subclass **C** (Table 1). The structures of the compounds of subclasses **A**, **B** and **C** are shown in **Schemes** 1, 2 and 3, respectively. The observed and calculated PPO inhibitory activities are listed in Table 1. The range of the PPO inhibitory activity values is quite wide (3.98 log units). The herbicidal activity of these 34 compounds were also tested against the various herbs like *Brassica campestrics* (Mustard), *Amaranthus retroflexus* (Tumble weed), *Digitaria sanguinalis* (Crabgrass) and *Echinochloa crus-galli* (Barn yard grass) by Yang *et al.* [8] and satisfactory results were obtained.

Docking

Molecular docking is an application, wherein molecular modelling techniques are used to predict how a protein (enzyme) interacts with small molecules (ligand) [18]. The ability of a protein (enzyme) to interact with small molecules plays a major role in the dynamics of protein which may enhance/inhibit its biological function. In our current paper, we have performed docking of 3H-pyrazolo [3,4-d][1,2,3] triazin-4-one derivatives into the active site of protoporphyrinogen oxidase enzyme. The crystal structure of the PPO enzyme (E.C. 1.3.3.4, 2IVD.pdb) has been obtained from RCSB protein data bank (http://www.pdb.org). We have worked on the PPO enzyme complexed with acifluorfen isolated from *Myxococcus xanthus*. The PPO enzyme is present as co-crystallised with flavin adenine dinucleotide (FAD). There are two subunits (A subunit and B subunit) of the enzyme. A subunit consists of 464 amino

Table 1 PPO inhibitory activity (observed and calculated) of 33 3H-pyrazolo [3,4-d][1,2,3] triazin-4-one derivatives and flumioxazin

Series	Compound Sl. Nos.	PPO inhibitory activity			
		Observed [8]	Calculated		
			Eq. (M1a)	Eq. (M2a)	Eq. (M3)
A	1	7.12	6.80	6.88	6.89
	2	6.33	6.61	6.89	6.34
	*3	6.31	6.61	7.16	6.24
	4	5.71	5.86	5.43	6.22
	5	6.49	6.52	6.60	6.20
	6	6.10	6.19	6.73	6.32
	7	7.85	6.96	6.98	7.59
	*8	7.34	7.52	6.92	7.16
	9	7.37	7.11	6.71	7.18
	10	5.74	5.87	5.51	6.37
	*11	5.41	6.22	5.31	6.51
	12	6.63	6.62	7.00	6.59
	13	6.28	6.42	6.64	6.58
	14	7.94	7.78	7.35	7.49
	15	7.19	7.87	7.27	7.32
	16	8.02	7.87	7.85	7.94
	*17	7.78	8.39	8.08	7.86
	18	7.87	7.14	7.44	7.55
	19	7.75	7.70	7.70	8.08
	20	6.23	7.19	6.26	6.00
	*21	6.12	6.50	6.58	6.27
	22	6.53	6.26	6.49	7.20
	23	6.61	6.03	6.58	6.35
	24	6.32	5.88	6.58	5.94
	25	6.55	5.96	6.39	6.19
	*26	6.58	6.05	6.72	6.32
	*27	6.64	7.85	6.98	7.72
	28	6.09	6.46	6.75	6.43
B	*29	4.51	5.31	5.46	4.55
	30	5.27	5.35	5.74	5.11
	*31	4.83	5.83	5.56	5.66
	32	5.00	5.44	5.05	5.07
	33	5.04	5.99	4.91	5.09
C	34	8.49	8.64	8.78	8.47

* Test set members

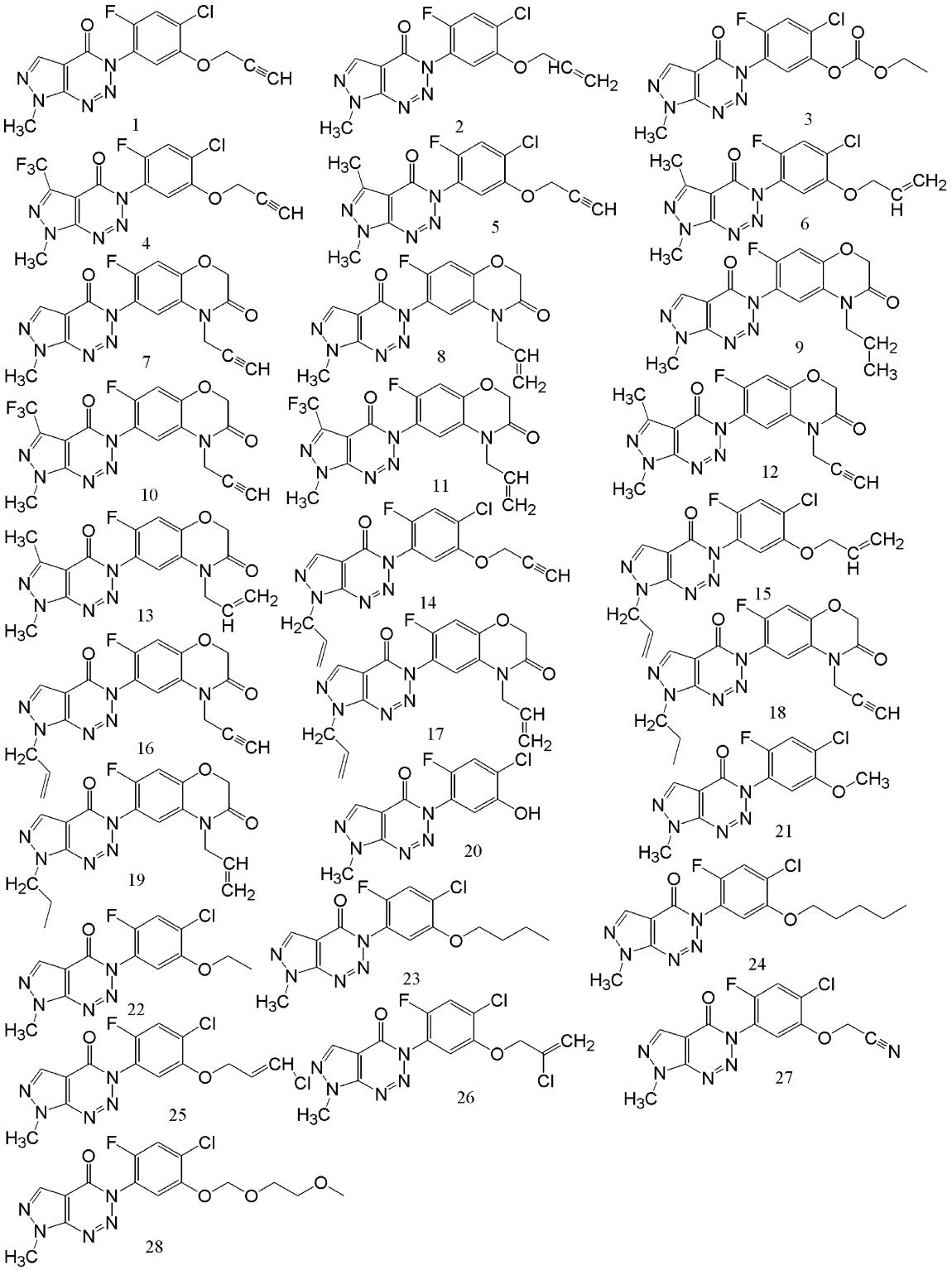
acid residues and B subunit contains 466 amino acid residues. Both the subunits are complexed with acifluorfen. We have docked the inhibitor molecules at the B subunit. We have performed the docking studies by using LigandFit of Receptor-ligand interactions protocol section of *Discovery Studio 2.1* [19]. Initially there was a pretreatment process for both the ligands and the enzyme (PPO). For ligand preparation, all the duplicate structures were removed and options for ionization change, tautomer generation, isomer generation, Lipinski filter and 3D generator have been set true. For enzyme preparation, the whole enzyme has been selected and hydrogen atoms were added to it. The pH of the protein has been set in the range of 6.5 to 8.5. Then we have defined the PPO enzyme as a total receptor and the active site was selected based on the ligand binding domain of acifluorfen. Then the preexisting ligand (acifluorfen) was removed and freshly prepared ligand (3H-pyrazolo [3,4-d] [1,2,3] triazin-4-one derivative) prepared by us was placed. Then from the receptor- ligand interaction section LigandFit was chosen. We have used the preprocessed receptor and ligand as inputs. Dreiding was selected as the energy grid. The conformational search of the ligand poses was performed by Monte Carlo trial method. Torsional step size for polar hydrogen was set at 10. The docking was performed with consideration of electrostatic energy. Maximum internal energy was set at 10000 Cal. Pose saving and interaction filters were set as default. Fifty poses were docked for each compound. During the procedure of docking, no attempt was made to minimize the ligand - enyme complex (rigid docking). After completion of docking, the docked enzyme (protein-ligand complex) was analyzed to investigate the type of interactions. The 50 docking poses saved for each compound were ranked according to their dock score function. The pose (conformation) having the highest dock score was selected for further analysis.

Descriptors

We have performed QSAR studies on the data set reported by Yang *et al*. [8] with three-dimensional (shape, spatial, electronic and molecular field) descriptors along with a few structural descriptors. The categorical list [20] of descriptors used in the development of QSAR models was reported in Table 2.

Cluster analysis

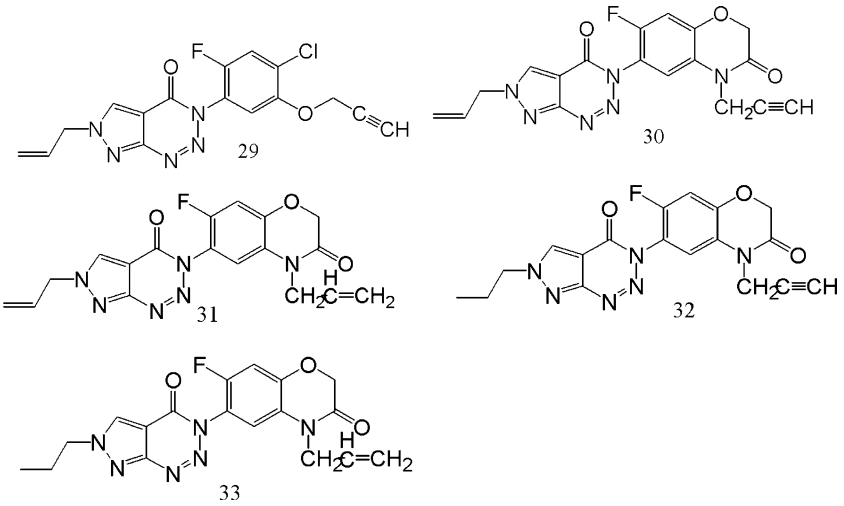
The ultimate target of any QSAR modelling is that the developed model should be strong enough to be capable of making accurate and reliable predictions of biological activities of new compounds. The models were cross validated using leave-one-out method. However, internal validation does not ascertain that the model will perform well on a new set of data. For maximum cases, appropriate external data set is not available for prediction purpose. Hence, the whole data set is divided into a training set and a test set or external evaluation set. In the present study the models developed from a training set (subset of the original set) were externally validated using a test set. Predictive capacity of a model for new chemical entities is influenced



Scheme 1

by chemical nature of the training set molecules used for development of the model [21–23]. In the actual case, the test set molecules will be predicted well when these molecules are structurally very similar to the training set molecules. The reason is that the model has considered all features common to the training set molecules. There are different techniques available for division of the data set into training and test sets like statistical molecular design,

Scheme 2



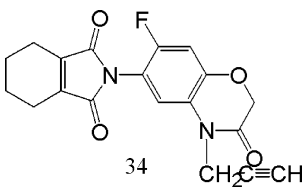
self-organizing map, clustering, Kennard - Stone selection, sphere exclusion, *etc*. [24]. In the present case we have used clustering technique as the method for training set selection. Cluster analysis [25] is a technique to arrange the objects into groups.

In our present work, the total data set (n=34) was divided into training set (n=25) and test (external evaluation) set (n=9) (75% and 25% respectively of the total number of compounds) based on clusters obtained from K-means clustering [26] applied on topological, spatial and electronic descriptor matrix. The whole data set was clustered into three subgroups from each of which 25% of compounds were selected as members of the test set. Serial numbers of compounds under different clusters are shown in Table 3.

Molecular shape analysis

Molecular shape analysis (MSA) was used as a 3D QSAR technique. In our study, the steps to perform MSA were [27] —

1) *Conformational analysis.* The first operation in MSA is the conformational analysis of the analogues. The conformers were generated with the "optimal search method" option followed by energy minimization.
2) *Hypothesizing an active conformer.* The aim of this step is to select a structure that is present in the rate-limiting step for the activity in a biological reaction. The minimum-energy conformer (global minima) of the most active compound **34** was taken as the active conformer.
3) *Selection of a candidate shape reference compound.* Shape reference compound is the molecule that is used when shape descriptors are calculated. MSA compares all other molecules to the shape reference compound (global minimum of compound **34**) and provides information about each comparison.
4) *Performing pair-wise molecular superposition.* Each study molecule was aligned to the shape reference compound using the maximum common sub graph (MCSG) method to calculate the shape descriptors.
5) *Measurement of molecular shape commonality.* After alignment, various shape descriptors, based on relative shape similarity with the shape reference compound, were calculated for each study molecule.
6) *Other molecular descriptors.* Determination of other molecular features by calculating spatial, structural and electronic parameters was done in addition to the shape descriptors.
7) *Construction of QSAR.* QSAR equations were generated using genetic function approximation (GFA) with linear option as the statistical tool.



Scheme 3

For comparison, we have performed another MSA using the docked conformations of the ligands instead of using global energy minimum conformers.

Molecular field analysis

Molecular field analysis (MFA) [28] attempts to postulate and represent the essential features of a receptor site from the aligned common features of the molecules in 3D space. MFA is a 3D-QSAR approach that computes the steric and electrostatic interactions of a given series of molecules, using probes within a regularly spaced grid. Molecular field analysis (MFA) is a method for quantifying the interaction

Table 2 Categorical list of descriptors used in the development of QSAR models

Category of descriptors	Name of the descriptors
Shape	DIFFV, COSV, F_o, NCOSV, ShapeRMS.
Electronic	Dipole-mag, HOMO, LUMO, Sr.
Spatial	RadOfGyration, Jurs_SASA, Jurs_PPSA_1, Jurs_PNSA_1, Jurs_DPSA_1, Jurs_PPSA_2, Jurs_PNSA_2, Jurs_DPSA_2, Jurs_PPSA_3, Jurs_PNSA_3, Jurs_DPSA_3, Jurs_FPSA_1, Jurs_FNSA_1, Jurs_FPSA_2, Jurs_FNSA_2, Jurs_FPSA_3, Jurs_FNSA_3, Jurs_WPSA_1, Jurs_WNSA_1, Jurs_WPSA_2, Jurs_WNSA_2, Jurs_WPSA_3, Jurs_WNSA_3, Jurs_RPCG, Jurs_RNCG, Jurs_RPCS, Jurs_RNCS, Jurs_TPSA, Jurs_TASA, Jurs_RPSA, Jurs_RASA, Shadow_XY, Shadow_XZ, Shadow_YZ, Shadow_XYfrac, Shadow_XZfrac, Shadow_YZfrac, Shadow_nu, Shadow_Xlength, Shadow_Ylength, Shadow_Zlength, Area, Vm, Density, PMI_mag.
Structural	MW, Rotlbonds, Hbond acceptor, Hbond donor.
Molecular field (Probe used)	H^+, CH_3, CH_3^-.

energy between a probe molecule and a set of aligned target molecules in QSAR. Interaction energies measured and analyzed for a set of three-dimensional structures can be useful in establishing structure-activity relationships. To generate an energy field (also known as a probe map), a probe molecule is placed at a random location and then moved about a target molecule within a defined three-dimensional grid. At each defined point in the grid, an energy calculation is performed, measuring the interaction energy between the probe and the target molecule. Atoms in the target molecule are fixed, so that intra-molecular energy in the target is ignored. When a complete probe map is calculated for each molecule in the target set, energy values for each point in the grid can be reported in columns added to the study table. For a set of structures for which energy fields are generated, some or all the grid data points can be used as descriptors in generating QSARs and analyzing structure—activity relationships. The selection of the independent variable columns was done based on variances of the columns. MFA was performed using the

Table 3 Serial numbers of compounds under different clusters

Cluster number	Serial number of compounds
1	3,4,5,6,7,8,9,12,13,14,15,23,24,25,26,29,34.
2	1,2,20,21,22,27.
3	10,11,16,17,18,19,28,30,31,32, 33.

QSAR module of Cerius2 4.10 [20] version. Docked conformations of the molecules were used for the analysis. A regression analysis was performed using the G/PLS method that combines the best features of genetic function approximation (GFA) and partial least squares (PLS). A rectangular field was generated using the probes H^+, CH_3, CH_3^-. A grid spacing of 2 Å was used, and fields with 720 points were generated. The energy cutoff was kept at −30 to +30 kcal. The charge calculation method was set to Gasteiger type. Alignment was done on the basis of the common substructure (CSS) method. The CSS method starts with defining a core model substructure to find a match in all of the molecules under the study. The mutation probabilities were kept at 5000 iterations. Smoothness (*d*) was kept at 1.00. Initial equation length value was selected as 4 and the length of the final equation was not fixed. All the variables were scaled.

Genetic function approximation-multiple linear regression

Genetic algorithms are derived from an analogy with the evolution of DNA [29]. The genetic function approximation algorithm was initially anticipated by 1) Holland's genetic algorithm and 2) Friedman's multivariate adaptive regression splines (MARS) algorithm. In this algorithm a model is represented as a one-dimensional string of bits. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as do most other statistical methods. Genetic algorithm makes superior models to those developed using stepwise regression techniques because it selects the basis functions genetically. Descriptors, which were selected by this algorithm, were subjected to multiple linear regression for generation of models. A "fitness function" or lack of fit (LOF) is used to estimate the quality of an individual or model, so that the best individual or model receives the best fitness score. The error measurement term LOF is determined by the following equation:

$$LOF = \frac{LSE}{\left(1 - \frac{c+d*p}{M}\right)^2}. \quad (1)$$

In Eq. (1), 'c' is the number of basis functions (other than constant term), 'd' is smoothing parameter (adjustable by the user), 'M' is number of samples in the training set, LSE is least squares error, and 'p' is total numbers of features contained in all basis functions.

Once models in the population have been rated using the LOF score, the genetic cross over operation is repeatedly performed. Initially two good models are probabilistically selected as parents and each parent is randomly cut into two pieces and a new model (child) is generated using a piece from each parent. After many mating steps, *i.e.* genetic crossover type operation, average fitness of models in the population

increases as good combinations of genes are discovered and spread through the population. It can build not only linear models but also higher-order polynomials, splines and Gaussians. In our present work, linear terms have been used. For the development of genetic function approximation (GFA) models, Cerius2 4.10 version [20] has been used. The mutation probabilities were kept at 5000 iterations. Smoothness (*d*) was kept at 1.00. Initial equation length value was selected as 4 and the length of the final equation was not fixed.

Partial least squares

PLS regression is a technique that generalizes and combines features from principle component analysis (PCA) and multiple regression. PLS is a useful technique when the number of factors is large and they are highly collinear. For PLS [30, 31], "leave-one-out" method was used for cross-validation to obtain the optimum number of components. In the case of PLS analysis, based on the standardized regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components. It gives a statistically more robust solution than MLR. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data. In the present paper, PLS analysis has been done to remove the intercorrelation problem with the descriptors selected by GFA-MLR technique.

G/PLS

Genetic partial least squares (G/PLS) [20, 30, 31] is a statistical method that combines the best features of genetic function approximation (GFA) and partial least squares (PLS). Both of these methods are valuable statistical techniques for QSAR modelling where the number of descriptors is more than the number of samples. Genetic function approximation is used to select the appropriate variables to be used in the development of a model. It is followed by PLS regression as a fitting technique to weigh the relative contribution of the selected variables in the final model. G/PLS retains the ease of interpretation of GFA by back transforming the PLS components to the original variables. There is no chance of over-fitting of the model.

Validation methods

The robustness of the models should be verified by using different types of validation criteria. For validation of QSAR models usually four strategies [32] are adopted: (1) internal validation or cross-validation, (2) validation by dividing the data set into training and test compounds, (3) data randomization or Y-scrambling, (4) true external validation by application of model on new external data. However, due to the lack of true external evaluation set, the total data set was divided into an internal evaluation (training) set and external evaluation (test) set. So, we have performed only the first three validation techniques. Most of the QSAR modelling methods implement the leave-one-out (LOO) or leave-many-out (LMO) cross-validation procedures, which are internal validation techniques. The outcome from the cross-validation procedure is cross-validated R^2 (LOO-Q^2 or LMO-Q^2) which is used as a criterion of both robustness and predictive ability of the model. In this paper, we have performed the leave-one-out validation method as the internal validation tool. Cross-validated squared correlation coefficient R^2 (LOO-Q^2) is calculated according to this equation.

$$Q^2 = 1 - \frac{\sum \left(Y_{obs(training)} - Y_{pred(training)}\right)^2}{\sum \left(Y_{obs(training)} - \overline{Y}_{training}\right)^2} \tag{2}$$

In Eq. (2), $\overline{Y}_{training}$ represents average activity value of the training set while $Y_{obs(training)}$ and $Y_{pred(training)}$ represent observed and predicted activity values of training set compounds, respectively. Often, a high Q^2 value ($Q^2>0.5$) is considered as a proof of high predictive ability of the model [33].

Models are generated based on training set compounds and predictive capacity of the models is judged based on the predictive R^2 (R_{pred}^2) values calculated according to the following equation [34]:

$$R_{pred}^2 = 1 - \frac{\sum \left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\sum \left(Y_{obs(test)} - \overline{Y}_{training}\right)^2}. \tag{3}$$

In Eq. (3), $Y_{pred(test)}$ and $Y_{obs(test)}$ indicate predicted and observed activity values respectively of the test set compounds and $\overline{Y}_{training}$ indicates the mean activity value of the training set compounds. The value of R_{pred}^2 for an acceptable model should be more than 0.5.

Further statistical significance of the relationship between the PPO inhibitory activity and chemical structure descriptors was obtained from randomization (Y-randomization) of the developed models. The test was done by repeatedly scrambling the activity values to generate QSAR models and then comparing the resulting scores with the score of the original QSAR model generated from non-randomized activity values. If the score of the non-random QSAR model is significantly better than that of the random

models then that model should be considered as a statistically robust model [35].

Software

MINITAB [36] was used for linear regression and partial least squares methods. Cerius2 version 4.10 [20] was used for GFA, G/PLS and MFA analyses. SPSS [37] were used for cluster analysis and intercorrelation matrix of the descriptors. LigandFit of the receptor-ligand interactions section available under Discovery Studio 2.1 [19] was used to dock the inhibitor molecules into the active site of the enzyme PPO.

Results and discussion

Docking

In the present study, to understand the interactions between the protoporphyrinogen oxidase (PPO) and its inhibitors and to explore their binding mode, docking study was performed using LigandFit of receptor-ligand interactions section available under *Discovery Studio 2.1*. Docking studies yielded crucial information concerning the orientation of the inhibitors in the binding pocket of the enzyme (PPO). The ligand-enzyme interaction analysis shows that Ile168, Ile311, Ile412, Met365, Ser64, Phe65, Val164 and Asn449 are the important residues present at the active site and are the main contributors to the receptor-ligand interaction. Non-polar amino acid residues (Ile168, Ile311, Ile412, Met365, Phe65 and Val164) form a hydrophobic pocket to which the inhibitors bind. It has been observed that, for better herbicidal activity, four amino acid (Ile168, Ile311, Ile412, and Met365) residues should optimally interact with the substituted pyrazolo-triazine-4-one ring system. In case of compound **10** (Fig. 1), the pyrazolo triazin-4-one ring structure is far away from Ile311 and Ile412 residues. Thus the pyrazolo triazin-4-one ring of that molecule does not fit properly in the hydrophobic pocket and there is a bump between the oxygen atom of benzo-morpholine ring system and the amino acid residue Met365. This bump may disturb the optimal position of the molecule in the pocket and thus it hinders the interaction with the amino acid residues and thus its herbicidal activity is poor. In case of compound **14** (Fig. 2), all the four important amino acids (Ile168, Ile311, Ile412, and Met365) are close to the substituted pyrazolo-triazin-4-one ring. For compound **14**, the above mentioned ring system fits well in the pocket of the enzyme. A hydrogen bond has been formed between the fluorine atom of substituted phenyl ring and the hydrogen atom of the flavin adenine dinucleotide (FAD) present as a co-enzyme in the PPO enzyme. The oxygen atom of the morpholine ring system of compound **11** (Fig. 3) forms a hydrogen bond with the hydrogen atom of the FAD molecule. This hydrogen bond can help the inhibitor to place itself within the hydrophobic pocket of the enzyme.

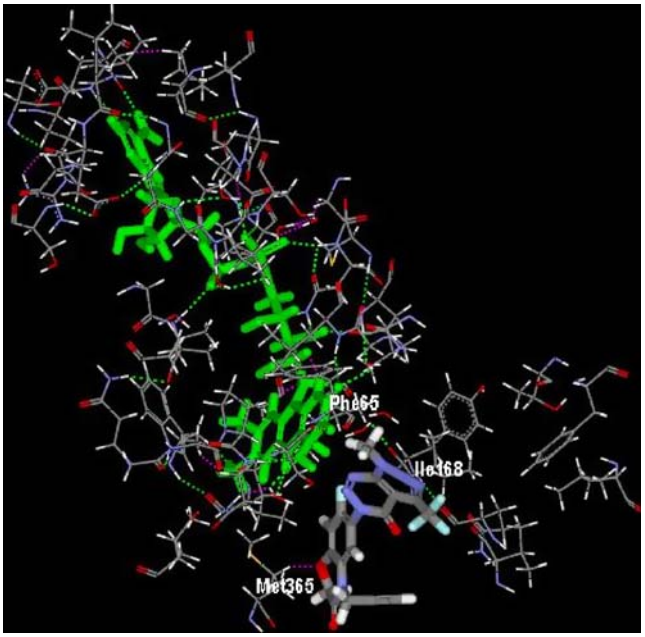


Fig. 1 Docked conformation of compound 10 along with the important amino acid residues

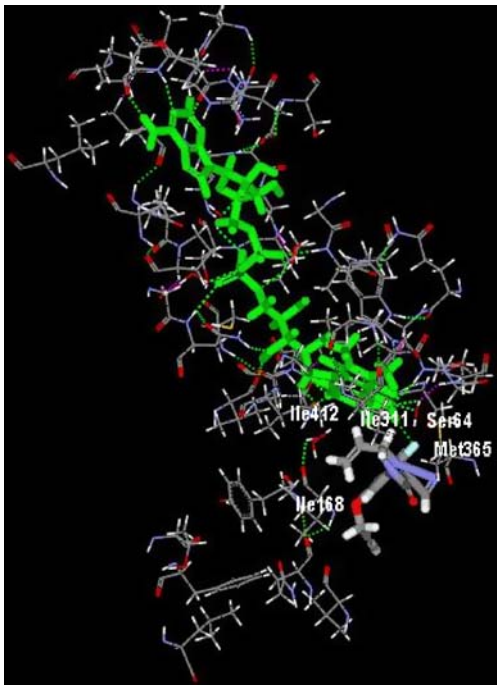


Fig. 2 Docked conformation of compound 14 along with the important amino acid residues

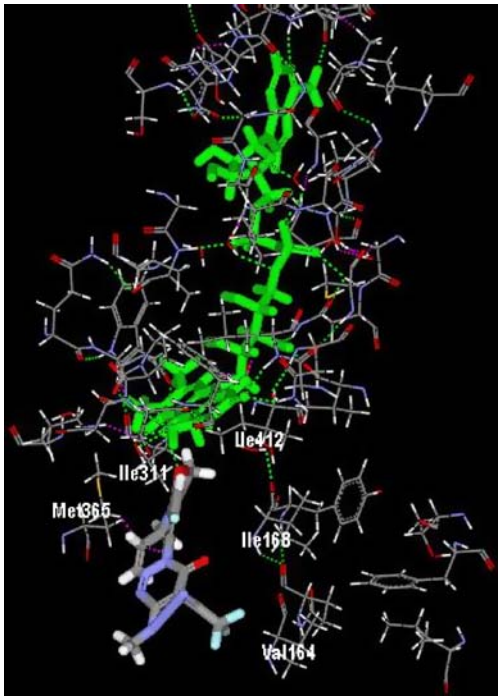


Fig. 3 Docked conformation of compound 11 along with the important amino acid residues

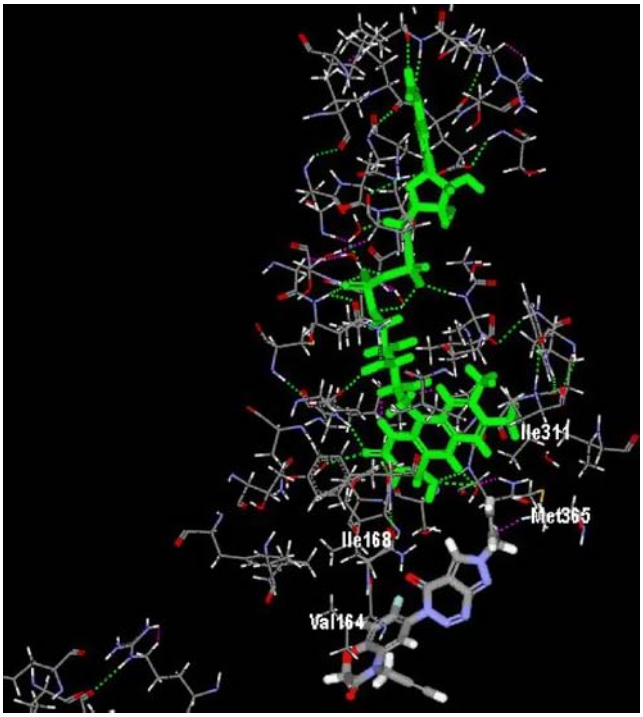


Fig. 4 Docked conformation of compound 30 along with the important amino acid residues

However, two bumps (intermolecular) have been formed which indicate the inhibitor molecule is not fitted well in the hydrophobic pocket. One bump is present between the allyl group of the benzomorpholine ring of compound **11** and the amino acid residue Met365 and another bump (intramolecular) is present between a hydrogen of the phenyl ring and the carbon of the allyl group. The activity of compound **11** is low due to the presence of these two bumps though the compound forms a hydrogen bond. In the case of compound **30** (Fig. 4), the position of the R_1 substitution has changed from that of the previous compounds. The allyl group present at the pyrazolo-triazin-4-one ring system has formed a bump with the important amino acid residue Met365. The position of the allyl substitution may hinder the placement of the inhibitor molecule in the hydrophobic cavity. Thus the activity of this molecule is low. In case of compound **34** (Fig. 5), the pyrazolo-triazin-4-one ring system has been replaced by tetrahydroisoindole-1,3-dione ring structure and the new ring system has optimally interacted with the four important non-polar amino acid (Ile168, Ile311, Ile412 and Met365) residues and this molecule posses very high activity. Thus we can say that, for better herbicidal activity pyrazolo-triazin-4-one ring can be replaced by a tetrahydroisoindole-1,3-dione ring system. In the case of the pyrazolo-triazin-4-one ring system, five nitrogen atoms are present while in the case of the tetrahydroisoindole-1,3-dione ring structure only one

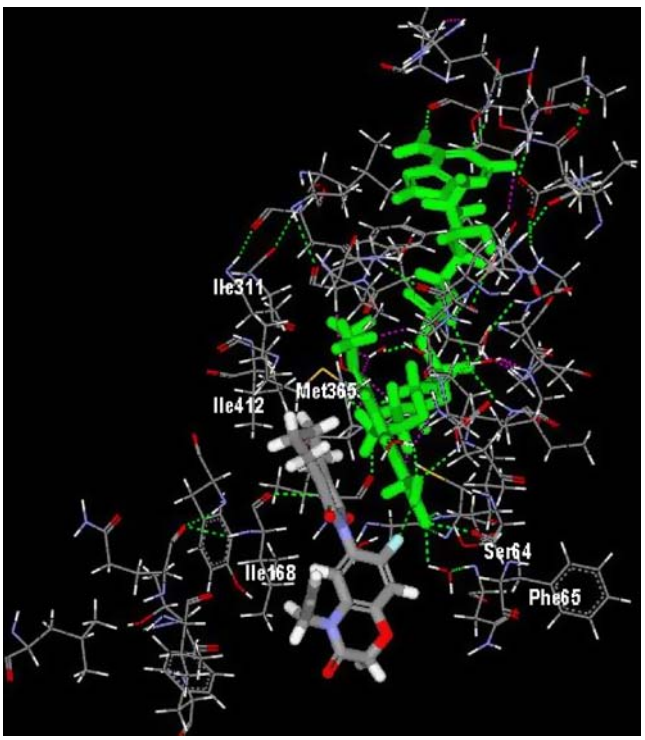


Fig. 5 Docked conformation of compound 34 along with the important amino acid residues

nitrogen atom is present. The presence of fewer number of nitrogen atoms in the case of the latter system increases the positively charged surface area which may increase the PPO inhibitory activity.

Molecular electrostatic potential surface (wire mesh) of the energy minimized geometry (calculated at AM1 level using Chem 3D) [38] of compound **29** shows (Fig. 6) that the pyrazolo-triazine-4-one ring system forms a negatively charged surface. This occurs due to the presence of more nitrogen atoms in the above mentioned ring system. Hence, positively charged surface is less, and the PPO inhibitory activity (4.51) of this compound is also less. However, in the case of compound **34** (Fig. 7) tetrahydroisoindole-1,3-dione ring structure forms a positively charged surface as fewer number of nitrogen atoms are present there. Here, positively charged surface area is large, and the PPO inhibitory activity (8.49) of this compound is also high. So, better PPO inhibitory activity, the pyrazolo triazine-4-one ring system can be replaced by tetrahydroisoindole-1,3-dione ring structure.

We have also validated the ligand binding process by docking the most active (compound **34**) inhibitor molecule with the enzyme PPO using other two docking tools LibDock and CDOCKER of receptor-ligand interactions section available under *Discovery Studio 2.1* [19]. The docked geometries of the most active compound obtained from LibDock and CDOCKER tools have been shown in Fig. S1 and Fig. S2 (Supplementary Materials), respectively. The docked geometries obtained from the latter two docking tools were very much close to that obtained from LigandFit tool of *Discovery Studio 2.1*. The important amino acids (Ile168, Ile311, Ile412, Met365, Ser64 and Phe65) in the docked geometries obtained from LibDock and CDOCKER tools interact with the ligand in a similar pattern as they do in the case of LigandFit. This supports that, our docking process is robust and reproducible. We have performed another type of validation study to determine whether our docking process is robust or not.

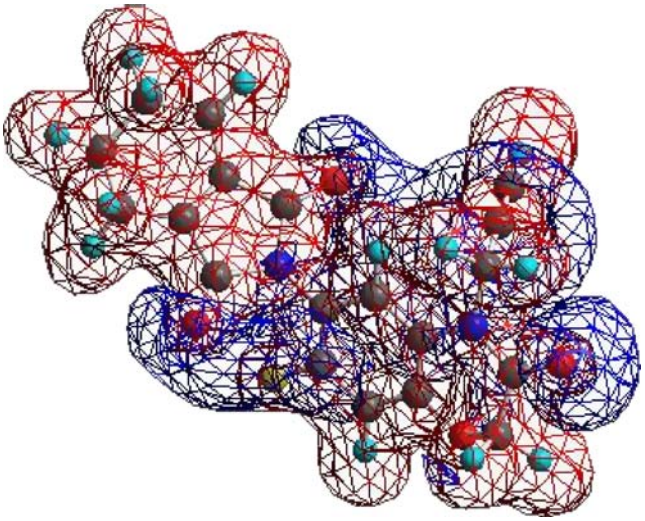


Fig. 7 Molecular electrostatic potential surface (wire mesh) of the energy minimized geometry of compound 34 (blue points in the surface indicate negatively charged areas, red points indicate positively charged areas)

We have docked acifluorfen (the original co-crystallized ligand) with the enzyme PPO and the analyzed the docked geometry with the original crystal structure (E. C. 1.3.3.4, 2IVD.pdb) obtained from protein data bank. This analysis shows that, the amino acid residues present close to the inhibitor molecule are the same as in the enzyme-inhibitor complex of the protein data bank. This suggests that, our docking process is reproducible.

Molecular shape analysis

We have performed a three dimensional quantitative structure-activity relationship to obtain the information about the effect of shape, spatial arrangement of atoms in three dimensional space and charge distribution of the substituents on the biological activity. This study was conducted using MSA descriptors along with additional

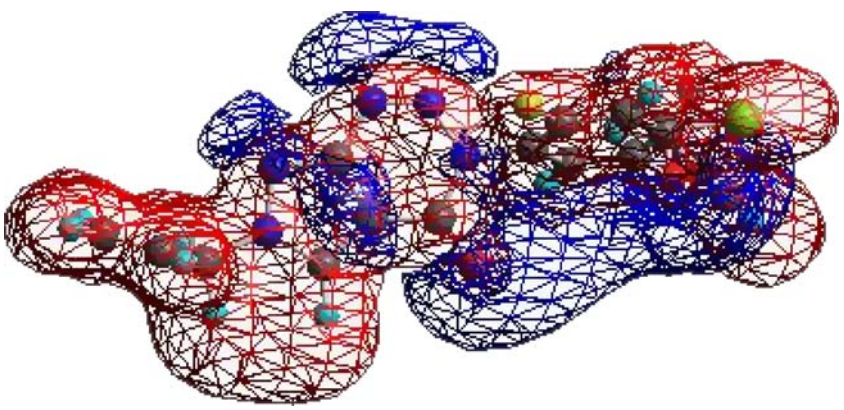


Fig. 6 Molecular electrostatic potential surface (wire mesh) of the energy minimized geometry of compound 29 (blue points in the surface indicate negatively charged areas, red points indicate positively charged areas)

descriptors like spatial and electronic parameters and a few structural descriptors. Figure 8 shows the aligned geometry of the training set compounds used in MSA.

Initially we have performed the MSA by generating conformers using Cerius 2 version 4.10 software as detailed in materials and methods section. Models were generated with shape, spatial and electronic descriptors using genetic function approximation with linear option as the statistical tool. The mutation probability was kept at 5000 iterations. In case of GFA linear technique, the following equation was obtained with acceptable leave-one-out (LOO) internal variance (Q^2) and external predicted variance (R_{pred}^2).

$$\begin{aligned} pI_{50} =& 100.940(\pm 19.960) - 5.844(\pm 1.296)LUMO \\ &+ 7.022(\pm 1.868)HOMO - 4.599(\pm 0.961)RadofGyration \\ &+ 0.918(\pm 0.190)Rotlbonds \\ &+ 2.581(\pm 0.617)Jurs\_FPSA\_2 \\ n_{Training} &= 25, R^2 = 0.801, R_a^2 = 0.749, \\ F &= 15.3(df\,5, 19), Q^2 = 0.606, \\ PRESS &= 8.757, n_{Test} = 9, R_{pred}^2 = 0.767, r^2 = 0.864, \\ r_0^2 &= 0.858, r_m^2 = 0.797. \end{aligned} \tag{M1}$$

The above model could explain 74.9% of the variance (adjusted coefficient of variation). The leave-one-out predicted variance was found to be 60.6%. The predictive potential of this model was determined by predicted R^2 of the test set compounds and it was found to be 0.767. The squared correlation coefficient between the observed and predicted activity of the test set compounds was 0.864. The squared correlation coefficient between the observed and predicted activity of the test set compounds, setting intercept to zero, was found to be 0.858.

Using the standardized variable matrix for regression, the significance level of the descriptors was found to be of the order: *LUMO*, *HOMO*, *RadofGyration*, *Rotlbonds* and *Jurs_FPSA_2*. However, in the above model, some variables are highly intercorrelated (*LUMO*, *HOMO* and *RadofGyration*, *Rotlbonds*) though the equation shows acceptable internal and external validation statistics. To eliminate the problem of intercorrelation, we have performed partial least squares (PLS) using descriptors selected by the GFA technique. The PLS model is free from intercorrelation problem. In case of PLS technique, the following equation was obtained with acceptable leave-one-out (LOO) internal variance (Q^2) and external predicted variance (R_{pred}^2).

$$\begin{aligned} pI_{50} =& 65.689 - 3.439LUMO + 3.541HOMO \\ &- 5.778RadofGyration + 1.119Rotlbonds \\ &+ 3.257Jurs\_FPSA\_2 \\ n_{Training} &= 25, R^2 = 0.761, R_a^2 = 0.713, F = 16(df\,4, 20), \\ Q^2 &= 0.545, \\ PRESS &= 10.120, n_{Test} = 9, R_{pred}^2 = 0.601, r^2 = 0.776, \\ r_0^2 &= 0.775, r_m^2 = 0.758. \end{aligned} \tag{M1a}$$

Though the external validation statistics of Eq. (M1a) is slightly inferior to Eq. (M1), the latter (M1a) does not suffer from the problem of high intercorrelation. *LUMO* is the energy of lowest unoccupied molecular orbital. This represents the electrophilicity of a molecule. Molecules with low *LUMO* energy can accept electrons more easily than those having high *LUMO* energy. *LUMO* has unfavourable contribution towards the herbicidal activity as evidenced by the negative regression coefficient. Another model not reported here, suggests that if the value of *LUMO* energy is greater than 1.812 eV then it shows negative contribution and vice versa. This can be explained by the compounds **7**, **15**, **16**, **30**, **32**–**33**. In case of compounds **15**, **7** and **16** herbicidal activity has increased as the value of *LUMO* is enhanced. However, when the value of *LUMO* crosses the threshold value (1.812 eV) then further increase in *LUMO* energy leads to reduction of activity as seen in case of compounds **30**, **32**–**33**. Thus, for better herbicidal activity, the molecule should be highly electrophilic.

HOMO is the energy of highest occupied molecular orbital. Molecules with high *HOMO* energy can easily donate electron and it represents nucleophilicity of a molecule. *HOMO* has favourable contribution towards the herbicidal activity as evidenced by the positive regression coefficient. So, for better herbicidal activity, the numerical values of *HOMO* should be high. This can be explained by the compounds **4**, **14** and **16**. In case of compound number **4** the *HOMO* energy (−10.517 eV) is less than that of

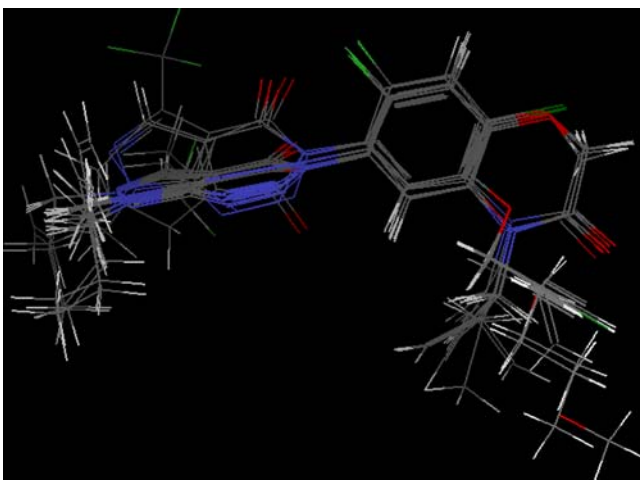


Fig. 8 Aligned geometry of the training set members used in MSA

compound **14** (−10.057 eV). Thus the activity of compound **4** (5.71) is quite less than that of the compound **14** (7.94). As the *HOMO* energy of compound **16** is very higher (−9.953 eV) than the above mentioned two compounds, so the herbicidal activity (8.02) of this compound is also higher than those. As the above mentioned two descriptors have regression coefficients of opposite sign, so there must be a balance between them. In Eqs. (M1) and (M1a), the *HOMO* term balances the negative coefficient of *LUMO*. For better PPO inhibition activity, there should be a balance between the electrophilic and nucleophilic character of the inhibitors.

Radius of gyration (Å) is a measure of the size of an object, a surface, or an ensemble of points. It is calculated as the root mean square distance of the objects' parts from either its centre of gravity or an axis. This can be calculated by the following equation:

$$RadofGyration = \sqrt{\left(\sum \frac{(x_i^2 + y_i^2 + z_i^2)}{N}\right)}. \quad (4)$$

Here, N is the number of atoms and x, y, z are the atomic coordinates relative to the centre of mass. It has unfavourable contribution towards the herbicidal activity as evidenced by the negative regression coefficient. This reflects that the shape of the molecules plays an important role for herbicidal activity. For better herbicidal activity, the molecules should have a symmetrical shape, *i.e.* spherical shape. This can be explained taking examples of the compounds **6** and **7**. As compound **6** is less symmetrical in shape than compound **7**, its radius of gyration possesses higher value than that of compound **7**. For this reason activity of compound **7** is quite higher than compound **6**. There is another example of compounds **15** and **17**. Compound **17** possesses more symmetrical shape than compound **15**, so its radius of gyration is less than compound **15** and herbicidal activity is higher than compound **15**.

Rotlbonds are the counts the number of bonds in the current molecule having rotations that are considered to be meaningful for molecular mechanics and it reflects the flexibility of a molecule. *Rotlbonds* has favourable contribution towards the herbicidal activity as evidenced by the positive regression coefficient. This can be explained by the following pairs of examples. In case of compound **13**, number of rotable bonds are less than that of compound **14**, so the activity of compound **14** is also high. This effect is also observed in the case of compounds **19** and **20** and compounds **17** and **21**.

Jurs_FPSA_2 is fractional charged partial positive surface area. It can be calculated by the total charge weighted positive surface area (PPSA-2) divided by the total molecular solvent accessible surface area (SASA).

$$Jurs\_FPSA\_2 = \frac{PPSA\_2}{SASA} \quad (5)$$

It has favourable contribution towards the herbicidal activity. This implies that increase in the total positive charge may enhance the activity. For example, in case of compound **15**, chloro substitution at R_3 position may reduce the total positive charge. Thus in compound **15** the value of *Jurs_FPSA_2* is less than that of the compound **16**, and accordingly the activity of compound **16** is higher than compound **15**.

Further we have performed molecular shape analysis using docked conformers obtained from *Discovery Studio 2.1*. Figure 9 shows the aligned geometry of docked conformers of the training set compounds used in MSA. Model has been generated with shape, spatial and electronic descriptors using genetic function approximation with linear option as the statistical tool. The mutation probability was kept at 5000 iterations. In case of GFA linear technique the following equation was obtained with acceptable leave-one-out (LOO) internal variance (Q^2) and external predicted variance (R_{pred}^2).

$$\begin{aligned} pI_{50} =& 15.887(\pm 1.655) - 0.005(\pm 0.001)PMI\_mag \\ &+ 0.440(\pm 0.070)Jurs\_WPSA\_3 - 1.673(\pm 0.334)LUMO \\ &- 2.706(\pm 0.839)Shadow\_nu \\ &+ 0.008(\pm 0.008)DIFFV \\ & n_{Training} = 25, R^2 = 0.829, R_a^2 = 0.784, F = 18.4(df\,5, 19), \\ & Q^2 = 0.628, \\ & PRESS = 8.259, n_{Test} = 9, R_{pred}^2 = 0.788, r^2 = 0.833, \\ & r_0^2 = 0.823, r_m^2 = 0.750. \end{aligned} \quad (M2)$$

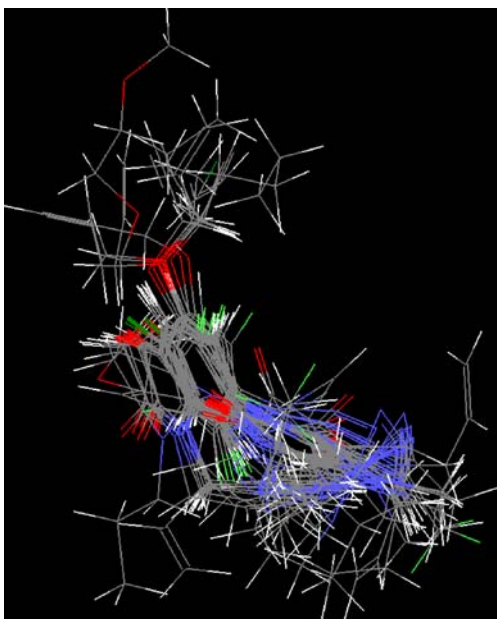


Fig. 9 Aligned geometry of docked conformers of the training set members used in MSA

Using the standardized variable matrix for regression, the significance level of the descriptors was found to be of the order: *PMI_mag, Jurs_WPSA_3, LUMO, Shadow_nu* and *DIFFV*. However, in the above model, the variables are highly intercorrelated (*PMI_mag, Jurs_WPSA_3* and *DIFFV*) though the equation shows acceptable internal and external validation statistics. To obviate this type of intercorrelation, we have performed partial least squares technique (PLS) using descriptors selected by GFA technique. The PLS model is free from intercorrelation problem. In case of PLS technique the following equation was obtained with acceptable leave-one-out (LOO) internal variance (Q^2) and external predicted variance (R_{pred}^2).

$$\begin{aligned} pI_{50} =& 15.875 - 0.005PMI\_mag + 0.438Jurs\_WPSA\_3 \\ & - 1.663LUMO - 2.675Shadow\_nu + 0.009DIFFV \\ & n_{Training} = 25, R^2 = 0.829, R_a^2 = 0.795, F = 24.3(df\,4, 20), \\ & Q^2 = 0.668, \\ & PRESS = 7.369, n_{Test} = 9, R_{pred}^2 = 0.762, r^2 = 0.836, \\ & r_0^2 = 0.822, r_m^2 = 0.737. \end{aligned} \tag{M2a}$$

PMI_mag is the moment of inertia, the resultant of the moment of inertia of three axes, which are calculated for a series of straight lines through the centre of mass. These are associated with the principal axes of the ellipsoid. *PMI_mag* has unfavourable contribution towards the herbicidal activity. In case of compounds **29**–**33,** the position of R_1 substituent is different from the compounds **1**–**28**. For this reason the value of *PMI_mag* is higher in the former compounds than that of the latter ones. Thus, the activity is less in case of compounds **29**–**33**.

Jurs_WPSA_3 is the surface weighted charged partial positive surface area. It can be calculated from the atomic charge weighted positive surface area (PPSA_3) multiplied by the total solvent-accessible surface area (SASA) and divided by 1000.

$$Jurs\_WPSA\_3 = \frac{PPSA\_3 * SASA}{1000} \tag{6}$$

It has favourable contribution towards the herbicidal activity. This implies that increase in the partial charges over all positively charged atoms and the total solvent accessible surface area enhance the activity. In the case of compound **16**, the value of *Jurs_WPSA_3* is high and the PPO inhibitory activity (8.02) is also high. For better activity, a compound should contain positive charge distributed over a large surface area. This has already been suggested by the molecular electrostatic potential surface of compounds **29** (Fig. 6) and **34** (Fig. 7).

Shadow_nu is the ratio of largest to smallest dimension. It has unfavourable contribution to the herbicidal activity as it has negative regression coefficient. If the compound has a symmetrical shape considering all three directions then the value of *Shadow_nu* will be smaller and herbicidal activity will rise. This can be explained taking the examples of compounds **19** and **28**. As compound **19** is more symmetrical in shape than compound **28**, the former has more herbicidal activity than the latter.

DIFFV is the difference between the volume of the individual molecule and the volume of the shape reference compound. It has favourable contribution towards the herbicidal activity. When the values of *NCOSV* of two compounds are similar, then an increase in the value of *DIFFV* may lead to an increase in the activity as in the case of compounds **2**, **16** and compounds **1**, **22**.

Molecular field analysis

Molecular field analysis (MFA) is a method for quantifying the interaction energy between a probe atom/molecule and a set of aligned target molecules in QSAR. Interaction energies measured and analyzed for a set of 3D structures can be useful in establishing structure-activity relationships. Figure 10 shows the aligned geometry of docked conformers of the

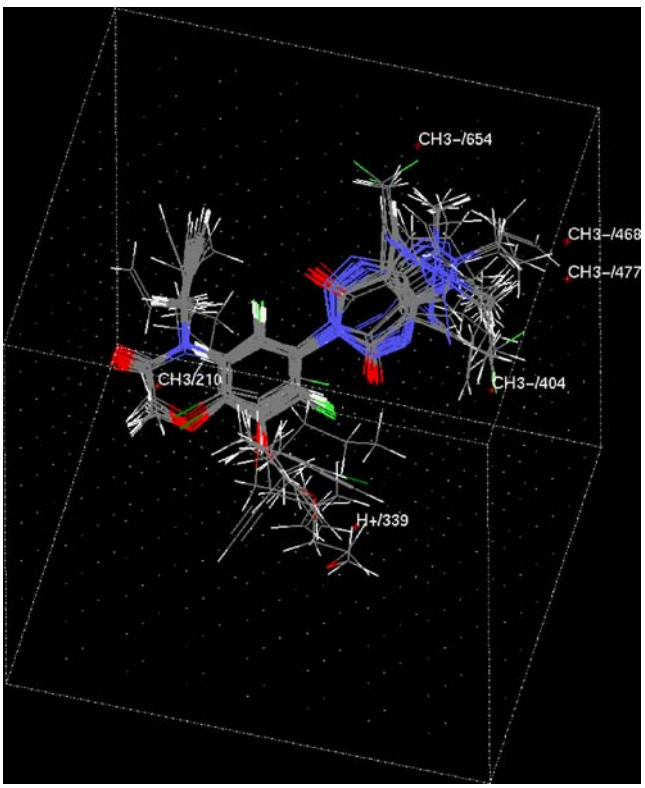


Fig. 10 Aligned geometry of docked conformers within the MFA grid showing important interaction points of the training set members used in MFA

training set compounds used in MFA. The mutation probability was kept at 5000 iterations. In the case of G/PLS linear technique the following equation was obtained with satisfactory leave-one-out (LOO) internal variance (Q^2) and predicted variance (R_{pred}^2).

$$\begin{aligned} pI_{50} =& 6.244 + 0.045(CH_3^-/404) + 0.020(CH_3/210) \\ & - 0.044(CH_3^-/468) - 0.021(CH_3^-/654) \\ & - 0.027(H^+/339) - 0.008(CH_3^-/477) \\ & n_{Training} = 25, R^2 = 0.890, R_a^2 = 0.874, F = 56.4(df\,3, 21), \\ & Q^2 = 0.813, \\ & PRESS = 4.152, n_{Test} = 9, R_{pred}^2 = 0.726, r^2 = 0.758, \\ & r_0^2 = 0.757, r_m^2 = 0.734. \end{aligned} \quad \text{(M3)}$$

$H^+/339$ indicates the interaction of the electrostatic probe at grid point number 339 with the molecules. It has unfavourable contribution to the activity. Interaction with H^+ at the grid point 339 is possible when a long chain substituent especially with electronegative atom is present at the R_4 position. This can be encountered in case of compounds **25** and **26**. They have long chain substituent with electronegative (chlorine) atom at the R_4 position, so their activity is less. The term $CH_3/210$ represents the interaction of the steric probe at grid point 210 with the molecules. It has a positive effect on the herbicidal activity. Significant interaction at this grid point occurs only in case of high steric volume of the substituents at the R_3 and R_4 substitution positions. This implies that, for better herbicidal activity, surface area should be greater, *i.e.* bulky substituents at R_3 and R_4 position may increase the activity. The term $CH_3^-/404$ indicates the interaction of the electronegative probe at grid point 404 with the molecules. It has favourable contribution to the herbicidal activity. For better herbicidal activity, the triazine-4-one ring system should be situated far from this grid point. However, the position of the triazin-4-one ring system in 3D space is dependent on the position of the substituent at R_1 of the pyrazole ring fused with triazine ring system. The term $CH_3^-/468$ represents the interaction of the electronegative probe at grid point 468 with the molecules. It has detrimental effect on the activity. This is dependent on the position of the substituent at R_1 of the pyrazole ring fused with triazine ring system. Due to this interaction, the activity of compounds **29**–**33** was lower than other compounds. The terms $CH_3^-/477$ and $CH_3^-/654$ represent the interactions of the CH_3^- probe at the grid points 477 and 654 respectively with the molecules. Both terms have unfavourable contribution towards the herbicidal activity. To avoid these interactions, either the pyrazolo-triazin-4-one ring system should be far away from these points or the triazin-4-one ring system can be replaced by tetrahydroisoindole-1,3-dione ring system for better activity (compound **34**). Figure 11 shows the most active compound (compound **34**) within the MFA grid showing the important interaction points.

Additional test on external validation

The models were also subjected to the test for criteria of external validation as suggested by Golbraikh and Tropsha [39]. To know predictive potential of the models, squared correlation coefficient values between the observed and predicted values of the test set compounds with intercept (r^2) and without intercept (r_0^2) were calculated. Interchange of the axes gives the value of r'^2_0. According to Golbraikh and Tropsha [39], models are considered acceptable, if they satisfy all of the following conditions:

(1). $Q^2 > 0.5$
(2). $r^2 > 0.6$
(3). $(r^2 - r_0^2)/r^2 < 0.1$ or $(r^2 - r'^2_0)/r^2 < 0.1$
(4). $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$.

When the observed values of the test set compounds (Y-axis) are plotted against the predicted values of the compounds (X-axis) setting intercept to zero, the slope of the fitted line gives the value of k. Interchange of the axes gives the value of k′. A list of values of different validation parameters defined above for different models have given in Table 5.

It has been previously shown [40] the R^2_{pred} may not be sufficient to indicate external predictivity of a model. The value of R^2_{pred} is mainly controlled by $\sum (Y_{obs(test)} - \overline{Y}_{training})^2$, *i.e.* the difference between observed values of test set compounds and mean observed activity values of training data set. Thus, it may not truly reflect the predictive capability on a new dataset. Besides this, the squared regression coefficient (r^2) between observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to observed activity (there may be considerable numerical difference between the values though maintaining an overall good intercorrelation). So, for better external predictive potential of the model, a modified r^2 ($r^2_{m(test)}$) was introduced by the following equation [40]:

$$r^2_{m(test)} = r^2 * \left(1 - \sqrt{r^2 - r_0^2}\right). \quad (7)$$

In Eq. (7), r_0^2 is squared correlation coefficient between the observed and predicted values of the test set compounds with intercept set to zero. The value of $r^2_{m(test)}$ should be greater than 0.5 for an acceptable model. The values of $r^2_{m(test)}$ for the different models have been reported in Table 4.

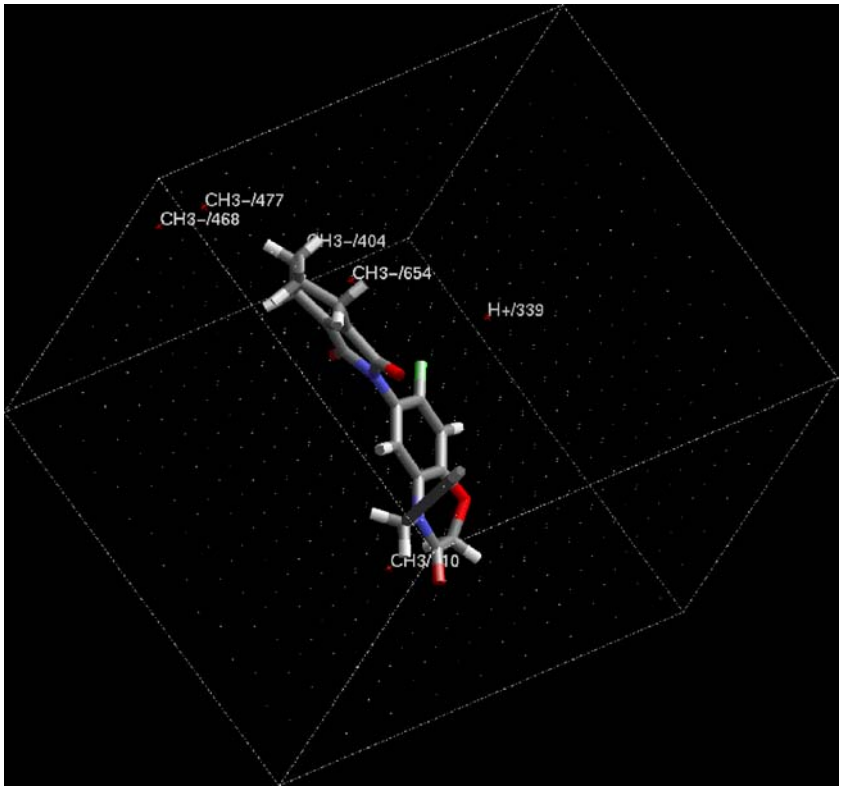


Fig. 11 The most active compound (compound 34) within the MFA grid showing its important interaction points

Initially the concept r^2_m was applied only to the test set prediction [40], but it can also be applied for the training set if one considers the correlation between observed and leave-one-out (LOO) predicted values of the training set compounds [41]. More interestingly, this can be used for the whole set considering LOO-predicted values for the training set and predicted values of the test set compounds. The advantages of such consideration are: (1) Unlike external validation parameters (R^2_{pred} *etc.*), the $r_m{}^2_{(overall)}$ statistic is not based only on a limited number of test set compounds. It includes prediction for both test set and training set (using LOO predictions) compounds. Thus, this statistic is based on prediction of comparably large number of compounds. In many cases, test set size is considerably small and regression based external validation parameter may be less reliable and highly dependent on individual test set observations. In such cases, the $r_m{}^2_{(overall)}$ statistic may be advantageous. (2) In many cases, comparable models are obtained where some models show comparatively better internal validation parameters and some other models show relatively superior external validation parameters. This may create a problem in selecting the final model. The $r_m{}^2_{(overall)}$ statistic may be used for selection of the best predictive models from among comparable models. For the present QSAR study, we have determined $r_m{}^2$ values for both training (based on LOO predicted values) and test sets and also for the whole set for the reported models and the results are shown in Table 4.

Model randomization

Further statistical significance of the relationship between the PPO inhibitory activity and descriptors were checked by randomization test (Y-randomization) of the models. This technique ensures the robustness of the model. The values of dependent variable were randomly scrambled and new QSAR models were developed keeping the independent variable matrix unchanged. The randomization tests for the models have been performed at 99% confidence level. The test has been done by shuffling the PPO inhibitory activity values and the average value of the correlation coefficient (R_r) of random models was calculated. For an acceptable QSAR model, the average correlation coefficient (R_r) of randomized models should be less than the correlation coefficient (R) of non-randomized model. No clear-cut recommendation was found in the literature for the desired difference between the average correlation coefficient (R_r) of randomized models and the correlation coefficient (R) of non-randomized model. We have used a parameter $R_p{}^2$ [42], which penalizes the model R^2 for the difference between squared mean correlation coefficient ($R_r{}^2$) of randomized models and squared correlation coefficient (R^2) of non-randomized model. The above mentioned novel parameter can be calculated by the following equation:

$$R_p^2 = R^2 * \sqrt{R^2 - R_r^2}. \quad (8)$$

Table 4 Comparison of statistical quality parameters and validation parameters of the models

Eq. no.	Type of descriptors	Model Type	Model quality				Internal validation parameter			External validation parameters		Overall validation parameter	Model randomization	
			R^2	R_a^2	F	s	Q^2	PRESS	$r^2_{m(LOO)}$	R^2_{pred}	$r^2_{m(test)}$	$r^2_{m(overall)}$	R^2_r*	R^2_p
M1	Shape + Spatial + Electronic + Structural	GFA-linear[#]	0.801	0.749	15.3	0.482	0.606	8.757	0.450	0.767	0.797	0.533	0.187	0.628
M1a	Shape + Spatial + Electronic + Structural	PLS-linear[#]	0.761	0.713	16	0.515	0.545	10.120	0.473	0.601	0.758	0.537	–	–
M2	Shape + Spatial + Electronic + Structural	GFA-linear[a]	0.829	0.784	18.4	0.447	0.628	8.259	0.477	0.788	0.750	0.562	0.182	0.667
M2a	Shape + Spatial + Electronic + Structural	PLS-linear[a]	0.829	0.795	24.3	0.436	0.668	7.369	0.574	0.762	0.737	0.656	–	–
M3	Molecular field descriptors	G/PLS-linear[a]	0.890	0.874	56.4	0.342	0.813	4.152	0.788	0.726	0.734	0.771	0.012	0.834

* Squared mean R for random models

[#] Using conformers generated by conformational analysis

[a] Using docked conformers

This novel parameter R_p^2 ensures that the models developed are not obtained by chance. For an acceptable QSAR model, the value of R_p^2 should be greater than 0.5. The values of R^2, R_r^2 and R_p^2 for different models have been reported in the Table 4.

Overview and conclusions

In our present paper, we have performed docking of 34 PPO inhibitors reported by Yang *et al*. [8] into the active site of PPO enzyme. Also we have done the QSAR studies with three dimensional (shape, spatial, electronic and molecular field) descriptors along with a few structural descriptors. The whole dataset (n=34) was divided into a training set (75% of the dataset) and a test set (remaining 25%) on the basis of K-means clustering technique. Models developed from training set compounds were used to predict the activity of the test set compounds. A comparison of statistical quality of different models was given in Table 4. The docking study suggests that in the active site of the PPO enzyme, important amino acid residues present are Ile168, Ile311, Ile412, Met365, Ser64, Phe65, Val164 and Asn449. The non-polar amino acid residues (Ile168, Ile311, Ile412, Met365, Phe65 and Val164) form a hydrophobic pocket to which the PPO inhibitors bind. Models generated from molecular shape analysis (MSA) reflects the importance of structural (*Rotlbonds*), electronic (*HOMO*, *LUMO*), shape (*DIFFV*) and spatial (*RadofGyration*, *Jurs_FPSA_2*, *Jurs_WPSA_3*, *Shadow_nu* and *PMI_mag*) descriptors. Molecular field analysis (MFA) suggests the importance of probes (H^+, CH_3, CH_3^-) at definite locations. Among the models generated by MSA, the model derived from the docked conformers is better than that developed by conformational analysis because the former is statistically more robust. According to the internal variance (Q^2=0.813), equation M3 is the best one. However, when we consider the external predictive variance, equation M2a (R^2_{pred}=0.762) is the best one. To avoid this type of contradiction, we have developed a novel parameter ($r^2_{m(overall)}$) [41]. Again, according to the $r^2_{m(overall)}$ Eq. M3 (0.771) is the best one. According to the newly introduced parameter (R_p^2), also Eq. M3 (0.834) is the best one.

Docking studies suggest that the molecules bind with a hydrophobic pocket of the enzyme formed by some nonpolar amino acid (Ile168, Ile311, Ile412, Met365, Phe65 and Val164) residues. The co-enzyme FAD plays a major role in the receptor binding of the inhibitors. The inhibitors form hydrogen bonds to bind properly with the enzyme. However, steric bumps have a detrimental effect on the PPO inhibition activity. As compounds **10**, **30**, **32** have formed bumps either intermolecular or intramolecular, their PPO inhibitory activity is lower. The quantum chemical descriptor *LUMO* suggests that, for better herbi-

Table 5 External validation criteria according to Golbraikh and Tropsha [39] of all the models

Equation no.	Model type	r^2	Q^2	$(r^2-r_0^2)/r^2$	k
M1	GFA-undocked	0.864	0.606	0.007	0.941
M1a	PLS-undocked	0.776	0.545	0.0007	0.922
M2	GFA-docked	0.833	0.628	0.012	0.955
M2a	PLS-docked	0.836	0.668	0.017	0.947
M3	G/PLS-docked (MFA)	0.758	0.813	0.001	0.952

cidal activity the molecules should be highly electrophilic. However, another electronic descriptor (*HOMO*) also shows positive contribution. So, there must be a balance between *HOMO* and *LUMO* energies, *i.e.* electrophilic and nucleophilic characters of the inhibitors. The charged surface area descriptors suggest that, the positive charge distributed over a large surface area may enhance the activity. The spatial descriptors show that, for better activity the molecules should have symmetrical shape in all directions in a 3D space. Molecular field probes suggest that an increase in steric volume may enhance the herbicidal activity. The position of the R_1 substituent may affect the PPO inhibition activity. Due to this reason the activity of the **B** series compounds is quite lower than the others. Instead of triazin-4-one ring system, tetrahydroisoindole-1,3-dione ring structure may enhance (compound **34**) the PPO inhibition activity. The results of our present study may be useful for the design and development of novel compounds having better PPO inhibitory activity against the unwanted herbs and weeds which reduce agricultural productivity.

Acknowledgments Financial assistance from the Ministry of Human Resource Development, Govt. of India, New Delhi in the form of a scholarship to SP is thankfully acknowledged.

References

1. http://en.wikipedia.org/wiki/Weed
2. http://en.wikipedia.org/wiki/Herbicide
3. Yang GF, Huang X (2006) Curr Pharm Design 12:4601–4611
4. Xi Z, Yu Z, Niu C, Ban S, Yang G (2006) J Comput Chem 27 (13):1571–1576
5. Peng H, Wang T, Xie P, Chen T, He H, Wan J (2007) J Agric Food Chem 55:1871–1880
6. Zhang L, Wan J, Yang G (2004) Bioorg Med Chem 12:6183–6191
7. Wang JG, Li ZM, Ma N, Wang BL, Jiang L, Pang SS, Lee YT, Guddat LW, Duggleby RG (2005) J Comput Aided Mol Des 19:801–820
8. Li HB, Zhu YQ, Song XW, Hu FZ, Liu B, Li YH, Niu ZX, Liu P, Wang ZH, Song HB, Zou XM, Yang HZ (2008) J Agric Food Chem 56:9535–9542
9. Hirai K, Uchida A, Ohno R (2002) In: Boger P, Wakabayashi K, Hirai K (eds) Herbicide Classes in DeVelopment. Springer-Verlag, Berlin, Heidelberg, pp 255–274
10. Yoshida R, Sakaki M, Sato R, Nagano E, Oshio H, Kamoshita H (1991) S-53482 A new phthalimide herbicide. Proc Brighton Crop Protection ConferencesWeeds. BCPC, Farnham, Surrey, UK, pp 69–75
11. Nagano E, Hashimoto S, Yoshida R, Matsumoto H, Kamoshita K (1987) (Sumitomo Chemical Company) U.S. Patent 4670046
12. Grossmann K, Schiffer H (1999) Pestic Sci 55:687–695
13. Dickmann R, Melgarejo J, Loubiere P, Montagnon M (1997) Oxadiargyl: A novel herbicide for rice and sugarcane. Proc Brighton Crop Protection ConferencesWeeds. BCPC, Farnham: Surrey, UK, pp 51–57
14. Auti K, Trombini A, Giammarusti L, Sbriscia C, Harder H, Gabard J (1997) Azafenidin: A new low use rate herbicide for weed control in perennial crops, industrial weed control and forestry. Proc Brighton Crop Protection ConferencesWeeds. BCPC, Farnham, Surrey, UK, pp 59–66
15. Van Saun WA, Bahr JT, Crosby GA, Fore ZQ, Guscar HL, Harnish WN, Hooten RS, Marques MS, Parrish DS, Theodoridis G, Tymonko JM, Wilson KR, Wyle MJ (1991) F6285-A new herbicide for the post-emergence selective control of broad-leaved weeds soybeans. Proc Brighton Crop Protection Conferences-Weeds. BCPC, Farnham, Surrey, UK , pp 77–82
16. Van Saun WA, Bahr JT, Bordouxhe LJ, Gargantiel FJ, Hotzman FW, Shires SW, Sladen NA, Tutt FS, Wilson KR (1993) F8426-A new rapidly acting, low-rate herbicide for the postemergence selective control of broad-leaved weeds in cereals. Proc Brighton Crop Protection ConferencesWeeds. BCPC, Farnham, Surrey, UK, pp 19–22
17. Zhu YQ, Wu C, Li HB, Zou XM, Si XK, Hu FZ, Yang HZ (2007) J Agric Food Chem 55:1364–1369
18. Kirkpatrick P (2004) Nature Rev Drug Disc 3:299
19. Discovery Studio 2.1 is a product of Accelrys Inc, San Diego, CA, USA
20. Cerius2 Version 4.10 is a product of Accelrys Inc, San Diego, CA, USA
21. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM (2003) Env Health Perspect 111:1361–1375
22. Guha R, Jurs PC (2005) J Chem Inf Model 45:65–73
23. Leonard JT, Roy K (2006) QSAR Comb Sci 25:235–251
24. Roy K (2007) Expert Opin Drug Discov 2:1567–1577
25. Everitt B, Landau S, Leese M (2001) Cluster analysis. Arnold, London
26. Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, Bittner M, Trent JM (2002) J Comput Biol 9:105–126
27. Hopfinger AJ, Tokarsi JS (1997) Three-dimensional Quantitative structure activity relationship analysis. In: Charifson PS (ed) Practical Applications of Computer-Aided Drug Design. Dekker, New York, pp 105–164
28. Hirashima A, Eiraku T, Kuwano E, Eto M (2003) Internet Electron J Mol Des 2:511–526
29. Rogers D, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854–866
30. Wold S (1995) PLS for Multivariate Linear Modeling. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 195–218
31. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN (2001) J Med Chem 44:3254–3263
32. Roy PP, Leonard JT, Roy K (2008) Chemom Intell Lab Sys 90:31–42
33. Kubinyi H, Hamprecht FA, Mietzner T (1998) J Med Chem 41:2553–2564
34. Marshall GR (1994) Binding-Site nmodeling of unknown receptors. In: Kubinyi H (ed) 3D QSAR in Drug Design—Theory, Methods and Applications. ESCOM, Leiden, pp 117–133
35. Deswal S, Roy N (2006) Eur J Med Chem 11:1339–1346
36. MINITAB is a statistical software of Minitab Inc, USA
37. SPSS is a statistical software of SPSS Inc, Chicago, IL
38. Chem 3D Pro version 5.0 program of CambridgeSoft Inc, Cambridge, USA
39. Golbraikh A, Tropsha A (2002) J Mol Graphics Mod 20:269–276
40. Roy PP, Roy K (2008) QSAR Comb Sci 27:302–313
41. Roy PP, Roy K (2008) Chem Biol Drug Des 72:370–382
42. Roy K, Paul S (2009) QSAR Comb Sci 28:406–425