RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2018, **8**, 4662

Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti*: application of ETA indices†

Priyanka De, Rahul B. Aher and Kunal Roy *

Dengue, zika and chikungunya have severe public health concerns in several countries. Human modification of the natural environment continues to create habitats in which mosquitoes, vectors of a wide variety of human and animal pathogens, thrive, which can bring about an enormous negative impact on public health if not controlled properly. Quantitative structure–activity relationship (QSAR) modeling has been applied in this work with the aim of exploring features contributing to promising larvicidal properties against the vector *Aedes aegypti* (Diptera: Culicidae). A dataset of 61 plant derived compounds reported in previous literature was used in this present study. A genetic algorithm (GA) was used for QSAR model development employing the "Double Cross Validation" (DCV) tool available at http://teqip.jdvu.ac.in/QSAR_Tools/. The DCV tool removes any bias in descriptor selection from a fixed composition of a training set and often provides an optimum solution in terms of predictivity. Simple topological descriptors, the "Extended Topochemical Atom" (ETA) indices developed by the present authors' group, were used for model development. These descriptors do not require pretreatment of molecular structures by conformational analysis or energy minimization before model development, thus saving computational time and resources. They also avoid ambiguities with respect to the existence of compounds in various conformational states leading to the loss of predictive capability in QSAR models. A number of models were generated from GA, and further, the descriptors appearing in the best model obtained from GA were subjected to partial least squares (PLS) regression to obtain the final robust model. The developed model was validated extensively using different validation metrics to check the reliability and predictivity of the model for enhancing confidence in QSAR predictions. Based on the insights obtained from the PLS model, we can conclude that the presence of hydrogen bond acceptor atoms, the presence of multiple bonds as well as sufficient lipophilicity and a limited polar surface area play crucial roles in regulating the activity of the compounds.

Received 8th December 2017
Accepted 17th January 2018

DOI: 10.1039/c7ra13159c

rsc.li/rsc-advances

1. Introduction

During the past 20 years, there has been a spectacular reappearance or emergence of epidemic arboviral diseases transmitted by mosquitoes affecting both human and domestic animal health.[1] Human modification of the natural environment continues to create habitats in which mosquitoes, vectors of a wide variety of human and animal pathogens, thrive, which can bring about an enormous negative impact on public health if not controlled properly. Morbidity and mortality have been reported to be increasing at an alarming rate[2] while a large number of lives are under threat due to mosquito borne diseases, like zika, malaria, chikungunya, dengue and yellow fever. The outbreak of these diseases has been observed mostly in countries like Brazil, Colombia, Mexico, Argentina and India.

Recent reports of local transmission of chikungunya have been made in south-eastern France, with 13 cases (four confirmed, one probable and eight suspected) of people aged between 3 and 77 years.[3] Also 183 cases have been notified in the Lazio Region of Italy, with 109 confirmed and 74 additional cases.[4] During the last few years an outbreak of the zika virus transmitted by mosquitoes has been observed in West Africa and in America (in Brazil and Colombia) due to weak health infrastructures and the decline in programmes for mosquito control.[5] The species responsible for these transmissions were

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. E-mail: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in; Web: https://sites.google.com/site/kunalroyindia/; Fax: +91-33-2837-1078; Tel: +91 98315 94140

† Electronic supplementary information (ESI) available: Table S1 in supporting materials lists molecular structures of the compounds used for modeling with their larvicidal activity data against *Aedes aegypti*. Table S2 in supplementary materials show the values of the descriptors appearing in eqn (1) for both training and test set compounds along with model derived (computed) larvicidal activity values. Fig. S1 and S2 in supplementary materials show the analysis of applicability domain and randomization test for the developed model, respectively. See DOI: 10.1039/c7ra13159c

found to be *Aedes aegypti*, *Aedes leucocaelenus*, *Aedes albopictus* and *Aedes sabethes* which proliferate rapidly due to continuous change in the environment leading to invasion of new territories.

The use of safe and efficacious insecticides against the adult and larval populations of mosquito vectors can be an effective way to control the transmission of zika virus and other viruses transmitted by *Aedes* mosquitoes, such as chikungunya and dengue. Pesticides play an effective role in the development of public health by working as a sustainable form of mosquito management.[6] Synthetic insect repellents like dichloro-diphenyl-trichloroethane (DDT) and *N*,*N*-diethylmetatoluamide (DEET) are used.[6,7] However, over the time, the vector mosquito has become highly resistant to DDT, which also creates a nuisance by becoming highly accumulated in the environment and producing toxic effects to humans, birds, fish and other animals.[7]

Over the last 50 years, the use of synthetic repellents has been one method of personal protection against mosquito bites. For example, compounds such as dimethyl phthalate (DMP), ethyl hexanediol (EHD) and diethylmetatoluamide (DEET) have been developed for this purpose. DEET, which is still being used worldwide, has some problems with efficacy, irritating effects on the skin, low retention and anaphylactic reactions.[8]

Botanical compounds known as essential oils (EOs) can be an alternative method to conventional pesticides where the former act as repellents, ovicides, adulticides, feeding inhibitors, or attractants for various insect species.[9,10] A large number of secondary metabolites like alkaloids, terpenoids and phenylpropanoids are found in considerable amounts in various parts of a plant. Compounds like menthol, citronellal, pulegone, linalool and other terpenes have shown insecticidal, fungicidal and larvicidal activities.[11,12] In a study, oils of 41 plants were evaluated for their effects against *Aedes*, *Anopheles*, and *Culex* larvae, among which 13 oils were found to induce 100% mortality after 24 hours or less in *Aedes aegypti*.[13]

The general aims of developing an ideal repellent are: it should be potent enough to repel a diverse class of vectors, should be effective for about eight to twelve hours, should not cause toxicity to the host, should be non-irritant to the skin and should not bring about systemic toxicity. However, no such compounds could be found with all these properties, and moreover the exploration of new insecticides needs time, a budget and several analytical set-ups.[7]

Quantitative structure–activity relationship (QSAR) modeling is an approach for determining the chemical features contributing to a target activity. This approach can be used for the compounds from plant essential oils with larvicidal activities in order to find a congener with optimum activity.[14] In the current study, we have utilized a dataset of 61 natural or semi-synthetic compounds with larvicidal activity for QSAR model development, using simple Extended Topochemical Atom (ETA) descriptors developed by the present authors' group.[15,16] The developed models are aimed at providing statistically robust predictions for the larvicidal activity of the compounds, expressed as the median lethal concentration (LC_{50}).

2. Materials and methods

2.1. The dataset

The experimental larvicidal lethal concentration (LC_{50}) values for 61 plant derived compounds were collected from the literature.[17–20] The concentrations of the chemical in air that kills 50% of the test population during the observation period is the LC_{50} value. The lethal concentration is usually applied for chemicals that are breathed into the body. In all the above-mentioned pieces of research, third instar larvae were used to determine the LC_{50} values of the compounds. The LC_{50} values were converted into their logarithmic scale equivalents (pLC_{50}) for the purpose of modeling. The structures of 61 compounds were drawn in the MarvinSketch (version 14.10.27)[21] application with proper aromatisation and explicit hydrogen addition. In Table S1 in ESI,† various classes of heterogeneous molecular structures involving terpenes, phenylpropanoids, ketones and oxygenated compounds along with their LC_{50} values are given.

2.2. Molecular descriptors

In the present work, there is only a single class of descriptors (Extended Topochemical Atom or ETA indices).[22] The descriptors were calculated using the PaDel-Descriptor software tool.[23] Variables with constant or near constant values (standard deviation less than 0.0001), descriptors with at least one missing value, descriptors with all values missing and descriptors with (absolute) pair correlation larger than or equal to 0.95 were excluded from the initial pool of descriptors. In the end, a set of 42 ETA descriptors were obtained which were used for model development. Since we have used only 2D descriptors in the present research, the model development does not require any conformational analysis or energy minimization of molecular structures. In addition to 2D descriptors not requiring molecular structure optimization, this approach involves some additional advantages; for instance, topological descriptors are simpler to interpret than geometrical descriptors. In fact, 2D descriptors avoid ambiguities with respect to the existence of compounds in various conformational states, which can lead to the loss of predictive capability in QSAR models.

2.3. Dataset division

The whole dataset was divided into training (66% of the all available data points) and test (34%) sets based on a simple and fast algorithm for *k*-Medoids clustering. For this, we employed a software tool "Modified *k*-Medoids" (version 1.2) developed in our laboratory.[24] The process categorizes a set of objects into clusters, so that the objects within a cluster are similar to each other but are dissimilar to objects present in other clusters.[25] The indicative objects within a cluster are called medoids. After arranging the whole dataset according to the cluster number with the corresponding activity values, we selected approximately 34% of compounds from each cluster as test set compounds ($n_{test} = 20$) and the remaining 66% as a training set ($n_{train} = 41$). The training set was used for model development and the test set was applied for the purpose of model validation.

2.4. Model development

In this study, we have developed a QSAR model using LC_{50} values of the plant derived compounds as the response variable for model development. Initially various statistical tools, such as multiple linear regression (MLR), stepwise regression[26] and double cross-validation (DCV),[27,28] were applied to develop the models; finally the most statistically significant and robust model was obtained by a genetic algorithm (GA)[29] within the DCV tool, followed by partial least squares (PLS) regression analysis.

Double cross-validation[27,28] is a statistical technique used for the generation and selection of models to produce a better predictive model. The fixed composition of a training set can often influence the selection of descriptors and can lead to a bias in descriptor selection. A double cross-validation method, in which the training set is further divided into '*n*' calibration and validation sets, can result in diverse compositions of the modeling set, thus removing any bias in descriptor selection. In addition, a model with the lowest prediction errors in the validation set is chosen; thus, this procedure is expected to provide an optimum solution in terms of predictivity in most cases. The tool comprises two nested cross-validation loops recognized as internal cross-validation and external cross-validation loops. In the external loop, the compounds in the dataset are divided into training set compounds and test set compounds. The training set compounds are involved in the internal loop for the purpose of model development and model selection, and the test set is used solely for the intention of checking model predictivity. In the internal loop, the training set is further repetitively split into calibration and validation sets by employing the *k*-fold cross-validation technique (in this study, $k = 10$)[27] and producing *k* iterations to construct calibration and validation sets. In the end, the best models are selected based on various validation metrics.

The double cross-validation technique in MLR model building and selection is a better choice compared to the conventional hold-out method. In the hold-out method, the composition of the training set remains the same, so there is a chance of bias in the descriptor selection. On the other hand, in the DCV method, the training set is further divided into '*n*' calibration and validation sets resulting in diverse compositions. So, there are more chances for optimal selection of descriptors for model development.

PLS regression is a generalization of multiple linear regression (MLR).[30] PLS provides an approach to the quantitative modeling of the often complex relationships between predictors, *X*, and responses, *Y*, and it is more general and robust than MLR. We performed PLS regression for development of the final model. We used the set of 5 descriptors from the previous step (GA-MLR) and ran PLS, which can handle overfitting and extensive noise during predictive model development. Information about the original variables is stored in latent variables (LV) generated by PLS. Although we used 5 descriptors in our model, it should be noted that PLS modeling is more robust than multiple linear regression, and it uses a reduced number of regression variables (latent variables, which are functions of the original variables). In our case, we used only 3 latent variables. This means that the actual number of regression variables is only 3 (and not 5) allowing an acceptable number of degrees of freedom.

2.5. Statistical validation metrics

In this present study, we employed multiple approaches for the evaluation of model quality, for measurement of the fitness, stability, robustness and predictivity of the developed model. The determination coefficient (R^2)[28] is a measure of goodness-of-fit whereas internal validation (which deals with the predictive ability of the model based on training set compounds) is usually determined by a cross-validated correlation coefficient, ${Q_{LOO}}^2$ (leave-one-out). Q^2 provides a measure of model robustness, but is not sufficient to determine the performance of the model when new sets of compounds are employed. The external validation of the model was estimated using various parameters, ${Q_{F_1}}^2$ and ${Q_{F_2}}^2$.[31] The external validation deals with the predictive ability of the model for the test set compounds. Additionally, the root mean square error (RMSE)[32] was estimated, which summarizes the overall error. We also included the values of standard error of estimate (*s*) and variance ratio (*F*) at the specified degrees of freedom (df) for the training set, to indicate the quality of fit and robustness of the regression coefficients of the developed model, respectively.

3. Results and discussion

In the current study, we have developed a PLS regression model using descriptors selected in GA-MLR employed in the DCV tool, as described in the Materials and methods section. The statistical quality of the model developed was sound. The final PLS model developed with five descriptors using three LVs is depicted below:

$$\begin{aligned} \mathrm{pLC}_{50} &= 5.998 + 1.241\mathrm{ETA\_EtaP\_F} - 5.338\mathrm{ETA\_dEpsilon\_D} - 13.458\mathrm{ETA\_dAlpha\_B} - 6.170\mathrm{ETA\_BetaP\_s} \\ &\quad - 2.129\mathrm{ETA\_dEpsilon\_C} \\ n_{\mathrm{training}} &= 41,\ R^2 = 0.726,\ Q^2 = 0.635,\ \overline{r_{m\mathrm{LOO}}{}^2} = 0.518,\ \mathrm{RMSE(train)} = 0.256,\ S_{\mathrm{train}} = 0.269\ F = 32.7(\mathrm{df}3, 37) \\ n_{\mathrm{test}} &= 20,\ {Q_{F_1}}^2 = 0.672,\ {Q_{F_2}}^2 = 0.650,\ \overline{r_{m(\mathrm{test})}{}^2} = 0.535,\ \mathrm{RMSE(test)} = 0.299,\ S_{\mathrm{test}} = 0.333 \end{aligned} \quad (1)$$

The model showed acceptable values of the coefficient of determination R^2 (0.726) and cross-validated correlation coefficient (LOO–$Q^2 = 0.635$), and a low standard error of estimate (S), signifying the statistical reliability of the model. The significant F value (at $p < 0.05$) suggests the robustness of the regression coefficients. The predictivity of the model was judged by means of predictive R^2 ($R_{pred}{}^2$) or Q^2F_1 ($Q^2F_1 = 0.672$), which shows a moderate predictive ability for the model. The values of the descriptors appearing in eqn (1) for both training and test set compounds along with model derived (computed) response values are provided in Table S2 in ESI.†

The regression coefficient plot[30] (Fig. 1) gives knowledge about the positive or negative contribution of descriptors towards the activity of the compounds. A descriptor with a positive correlation coefficient (*i.e.*, ETA_EtaP_F) signifies that as the descriptor value increases, the larvicidal activity value also increases, whereas a descriptor with a negative coefficient (*i.e.*, ETA_dEpsilon_D, ETA_dAlpha_B, ETA_BetaP_s, ETA_dEpsilon_C) indicates that as its value increases, the larvicidal activity decreases.

From the variable importance plot (VIP) (Fig. 2), the significance of each of the descriptors obtained in the final PLS model can be described for their importance to the larvicidal activity of the compounds. The most and the least important descriptors contributing to the larvicidal activity of the used compounds can be identified with the help of this plot (Fig. 2). A variable with VIP score >1 shows higher statistical significance as compared with one with a low VIP value.[33] The descriptors are arranged in the plot according to their importance (maximum contribution to minimum contribution) and their significance level is found to be in the following order: ETA_dEpsilon_D, ETA_EtaP_F, ETA_dAlpha_B, ETA_BetaP_s and ETA_dEpsilon_C.

The descriptor contributing most to the response is ETA_dEpsilon_D, which is a measure of contribution of hydrogen bond donor atoms, *i.e.*, the presence of groups such as –OH, –NH_2, –SH *etc.* The negative coefficient of the descriptor shows that there will be a decrease in the desired activity of the compound with an increase in descriptor value, *i.e.*, an increase in the number of hydrogen bond donor atoms. Compounds like

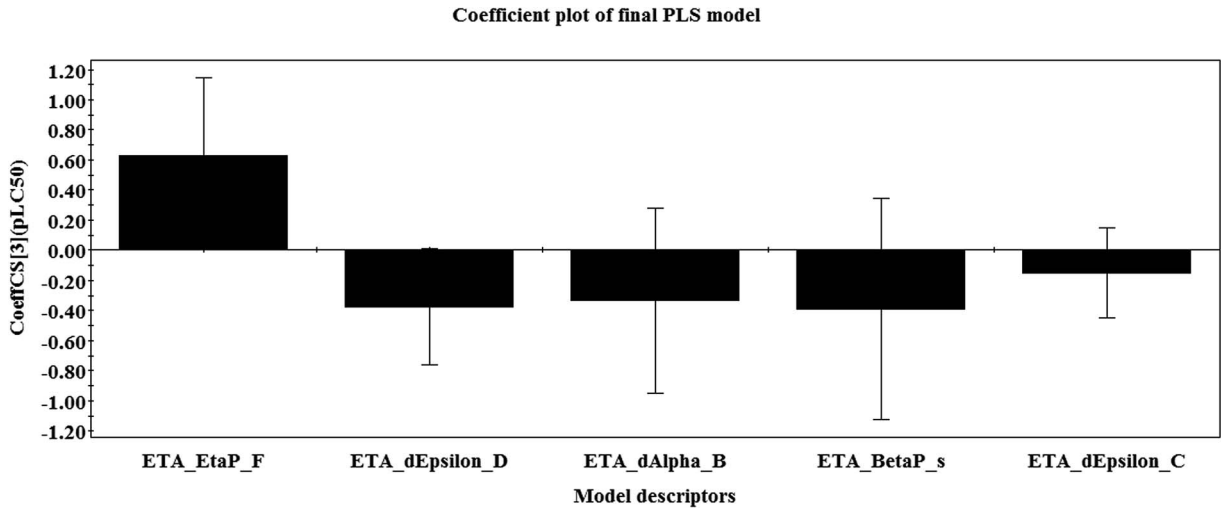

Fig. 1 Regression coefficient plot of the final PLS model.

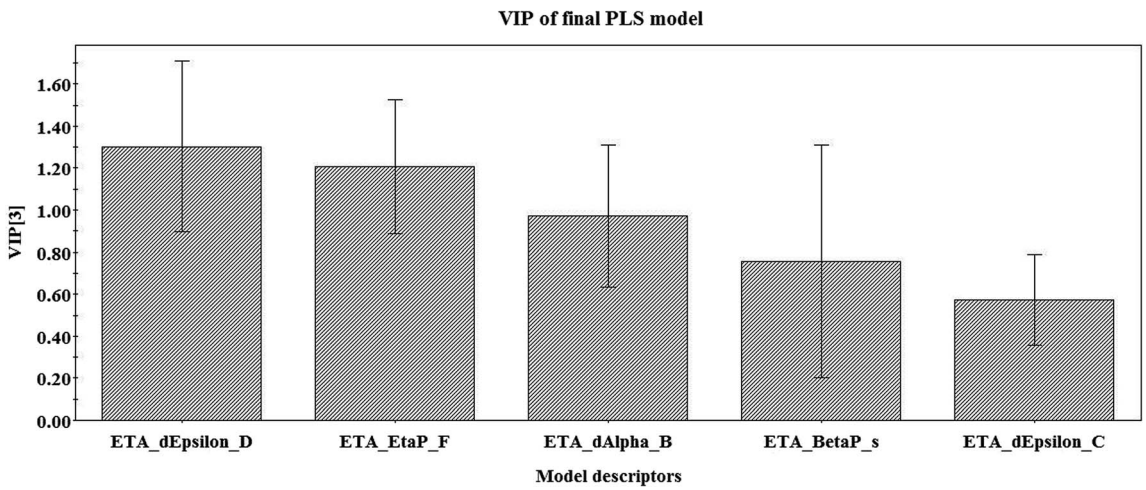

Fig. 2 Variable importance plot of the final PLS model.

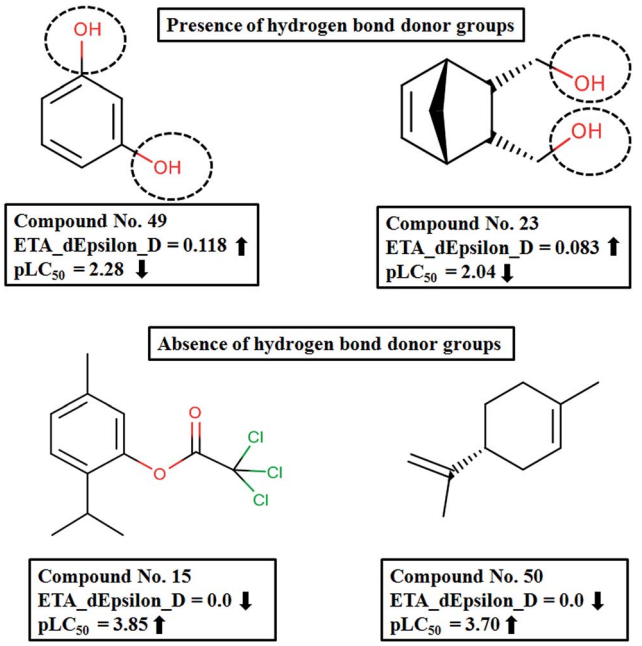

Fig. 3 Contribution of ETA_dEpsilon_D to pLC_{50} of the compounds.

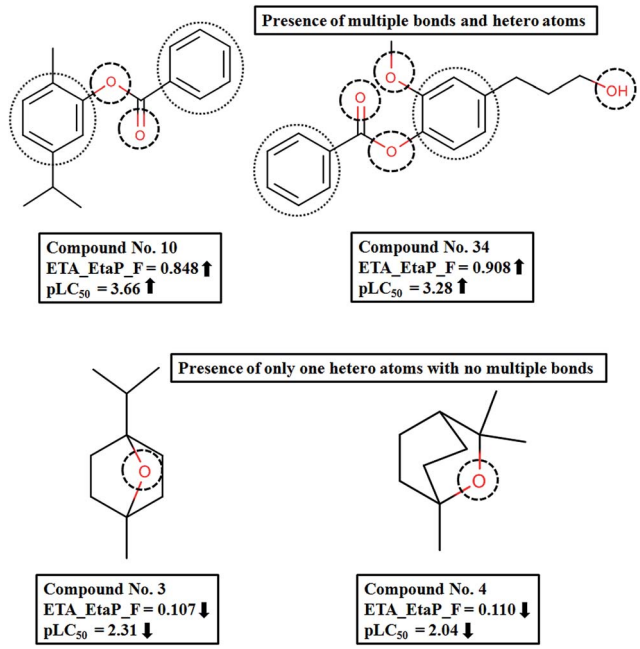

Fig. 4 Effect of ETA_EtaP_F on pLC_{50} of the compounds.

49 (resorcinol), **23** (5-norbornene-2-*endo*-3-*endo*-dimethanol) and **35** (4-hydroxy-3-methoxy-benzenepropanol) have a higher number of hydrogen bond donor atoms contributing to lower activity values, whereas compounds like **15** (thymyl trichloroacetate), **50** (*R*-limonene) and **10** (carvacryl benzoate) have low ETA_dEpsilon_D values leading to higher activity. The effect of the ETA_dEpsilon_D descriptor on the activity of the compounds is depicted in Fig. 3.

The next most important descriptor is ETA_EtaP_F, which is a functionality index relative to molecular size. It gives a measure of the presence of heteroatoms and multiple bonds. The positive regression coefficient of the descriptor denotes that an increased number of heteroatoms and multiple bonds will increase the larvicidal activity against the *Aedes* mosquito. In compounds like **10** (carvacryl benzoate), **17** (thymyl benzoate) and **34** (1-benzoate-2-methoxy-4-(3-hydroxypropyl)-phenol), the number of heteroatoms (like oxygen) and multiple bonds (as in benzene rings) are higher; accordingly the descriptor values are also higher, contributing to increased activity. On the other hand, compounds like **3**(1,4-cineole) and **4**(1,8-cineole) have low descriptor values, thus leading to lower activity (Fig. 4). From these observations, we can conclude that hydrophobicity is important for larvicidal activity.

The descriptor ETA_dAlpha_B($\Delta\alpha_{\downarrow}B$) is the next most important descriptor, which is a measure of polar surface area. The negative contribution of this descriptor indicates that the presence of polar groups is detrimental to the activity, as shown in compounds like **35** (4-hydroxy-3-methoxy-benzenepropanol) and **49** (resorcinol). In contrast, compounds like **15** (thymyl trichloroacetate), **10** (carvacryl benzoate) and **14** (thymyl chloroacetate) which have hydrogen bond acceptor atoms have higher activity (Fig. 5).

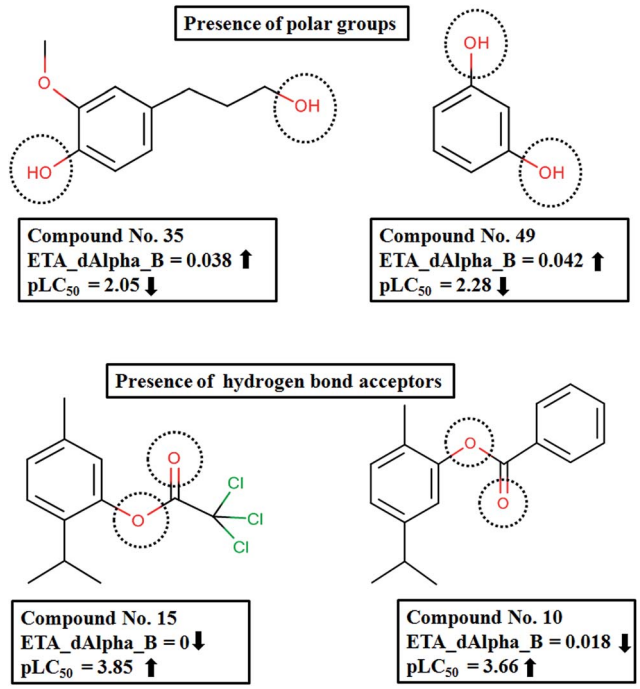

Fig. 5 Contribution of ETA_dAlpha_B to pLC_{50} of the compounds.

The fourth important descriptor is ETA_BetaP_s ($\Sigma\beta$), which is the sum of the β values for all the sigma bonds (VEM sigma contribution)[15,34] relative to the number of vertices.

$$\Sigma\beta'_s = \Sigma\beta_s/N_v$$

The descriptor ETA_BetaP_s gives a measure of the electronegative atom count of the molecule relative to the molecular

size. The negative contribution suggests that with an increase in the descriptor value the activity will decrease. From the above equation, the descriptor values obtained can be justified.[35] According to the ETA scheme, the sigma contribution of two bonded atoms with similar electronegativity is 0.5 and that for ions with different electronegativity is 0.75. Therefore, considering the relative values (relative to the number of vertices), we can see that in compounds with a higher number of heteroatoms like **26** (2-[2-methoxy-4-(2-propen-1-yl)phenoxy] acetic acid) and **44** (1,2-carvone oxide), the descriptor values are higher (higher sigma contribution). Also the contribution of the descriptor to the activity is also well explained by these compounds, since their activity values are low. Next, if we consider compounds like **50** (*R*-limonene) and **54** (*S*-limonene), which have a nonfunctional carbocyclic skeleton, they have lower descriptor values and consequently their activity values are higher (Fig. 6).

The descriptor with the least importance is ETA_dEpsilon_C($\Delta\varepsilon_C$), which is a measure of electronegativity. The descriptor can be expressed as $\Delta\varepsilon_C = \varepsilon_3 - \varepsilon_4$, where the terms ε_3 and ε_4 can be defined as:

(i) ε_3: sum of epsilon (ε) values relative to the total number of atoms (N_R) including hydrogens in the connected molecular graph of the reference alkane. A reference alkane of a molecule corresponds to a structure where all heteroatoms are replaced with carbon atoms and multiple bonds (covalent) with single bonds.[34]

$$\varepsilon_3 = \frac{\left[\sum \varepsilon\right]_R}{N_R}$$

(ii) ε_4: Sum of epsilon (ε) values relative to the total number of atoms (N_{ss}) including hydrogen for a saturated carbon skeleton moiety of the normal molecule, *i.e.*, with carbon–carbon multiple bonds considered as single bonds.[35]

$$\varepsilon_4 = \frac{\left[\sum \varepsilon\right]_{ss}}{N_{ss}}$$

This descriptor shows a negative influence on the pLC_{50} values; thus an increase in the ETA_dEpsilon_C value will result in a decrease in the response and *vice versa*. In compounds **1** ((−)-Camphene) and **59** (3-Carene), there is an absence of any electronegative atoms and the reference alkane and the saturated carbon skeleton for these two compounds will be the same. Therefore, the values for ε_3 and ε_4 will be the same and hence their difference, $\Delta\varepsilon_C$, is zero for both compounds. On the other hand, compounds like **8** (carvacryl trichloroacetate) and **15** (thymyl trichloroacetate), which possess a considerable number of electronegative atoms (five electronegative atoms in both cases), will have higher ε_4 values than ε_3 making $\Delta\varepsilon_C$ negative (Fig. 7).

3.1. Score plot of the PLS model

The distribution of the compounds in the latent variable space as defined by the scores is expressed in a score plot, as given in Fig. 8. Here, we have plotted the scores of the first two components t1 and t2. The ellipse indicates the applicability domain of the model, as defined by Hotelling's t2. Hotelling's t2 is a multivariate generalization of Student's *t*-test. It provides a check for compounds adhering to multivariate normality.[36] In this plot, compounds which are situated near each other have similar characteristics or properties, whereas compounds

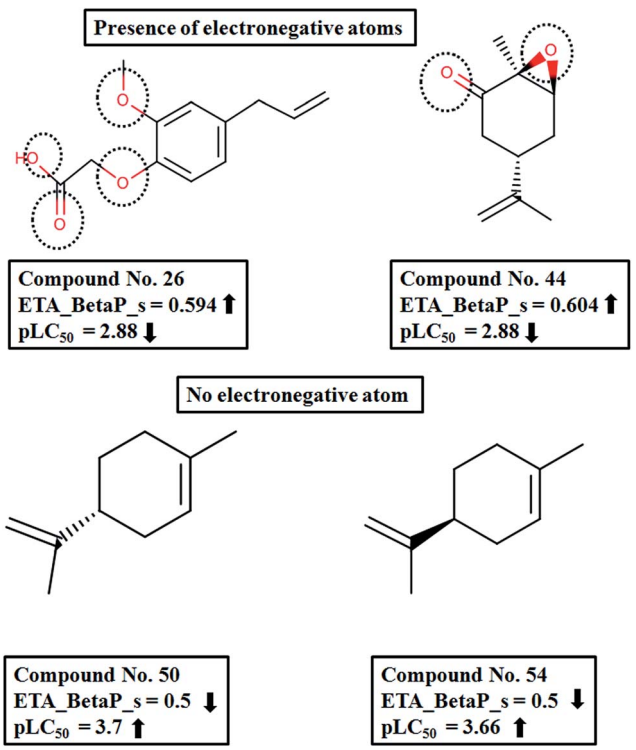

Fig. 6 Contribution of ETA_BetaP_s on pLC_{50} of the compounds.

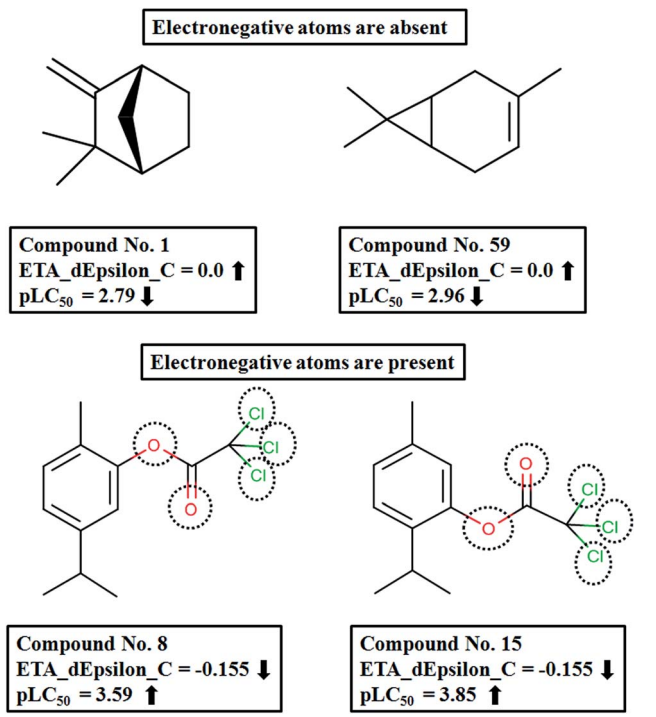

Fig. 7 Effect of ETA_dEpsilon_C on pLC_{50} of the compounds.

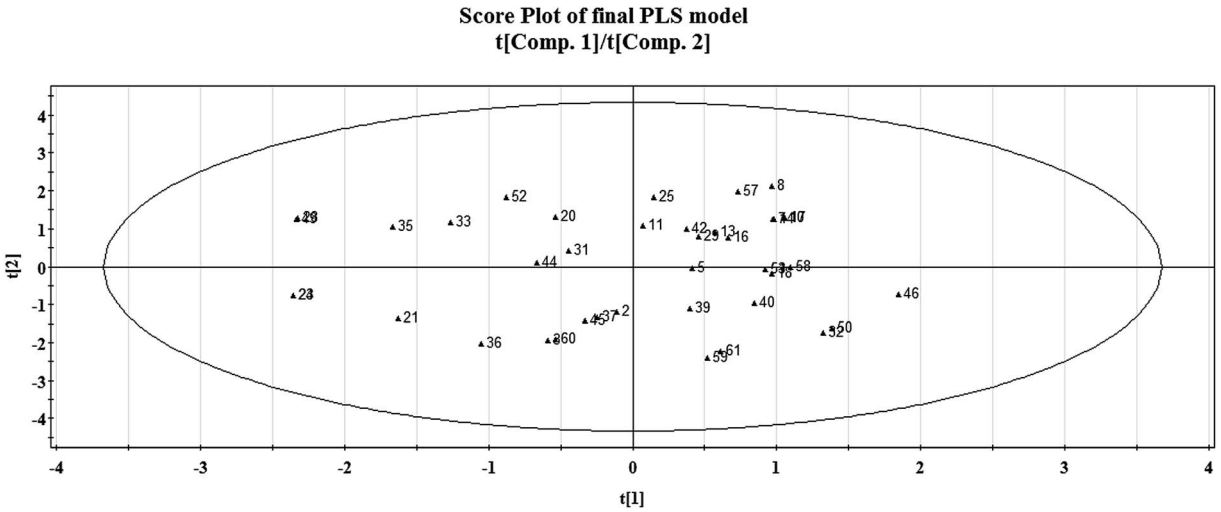

Fig. 8 Score plot of the final PLS model.

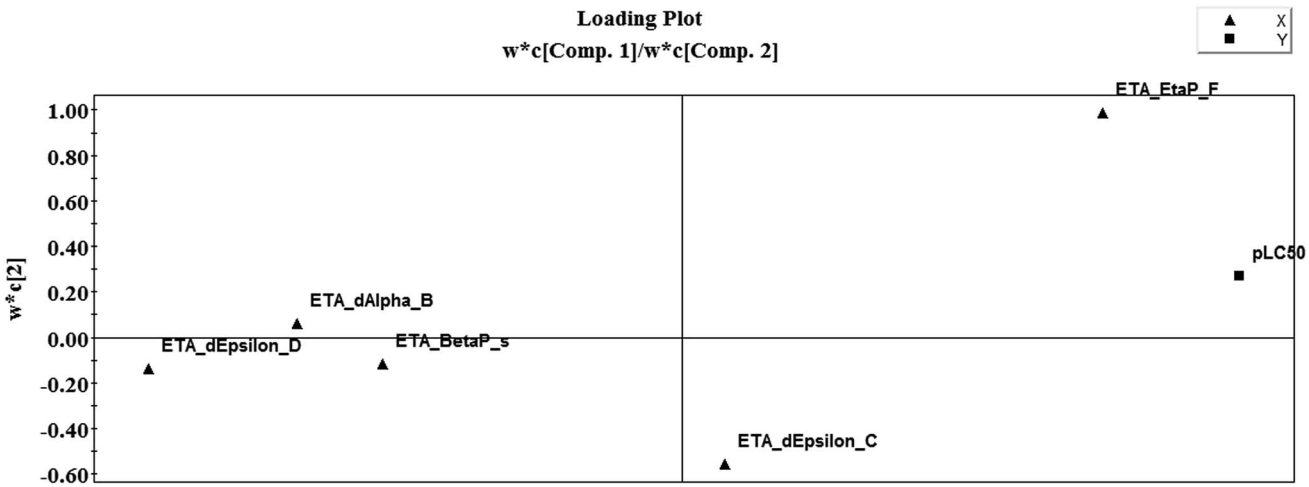

Fig. 9 Loading plot of the final PLS model.

which are far from each other have dissimilar properties with respect to their larvicidal activity against the zika vector. For example, compounds which are located in the upper right hand corner like **42** (1,2-dimethoxy-4-(2-propen-1-yl)-benzene) and **29** (1-ethoxy-2-methoxy-4-(2-propen-1-yl)-benzene) have some similarity in properties whereas compounds which are far from each other like those in the lower left hand corner (for example compound number **21** or 5-norbornene-2-ol) and upper right hand corner (for example compound number **8** or carvacryl trichloroacetate) represent heterogeneity in the property space. The compounds which are close to the centre of the plane have average properties. Since there are no compounds present outside the ellipse, we can conclude that there are no outliers according to this method.

3.2. Loading plot of the PLS model

A loading plot of a PLS model (Fig. 9) gives the relationship between *X*-variables and *Y*-variables, as shown in Fig. 9, where five *X*-variables and one *Y*-variable (pLC_{50}) are shown. The loading plot was developed using the first two components. The loading plot gives us an insight into how the different variables produce an impact on the model and which variable produces the maximum footprint. For interpretation of the PLS model, we should consider the distance from the plot origin. Similar types

Table 1 Weightage of descriptors for first two PLS components

	Weightage based on the first two components	
Descriptors	Component 1	Component 2
ETA_EtaP_F	0.102729	0.709627
ETA_dEpsilon_D	−0.765005	0.100524
ETA_dAlpha_B	−0.644193	0.384778
ETA_BetaP_s	−0.405975	0.241983
ETA_dEpsilon_C	0.385227	−0.568717

Table 2 Comparison of current model with previously developed models

Models	Total no. of compounds used	No. of compounds in the training set	No. of compounds in the test set	Descriptor type	Number of descriptors used in the initial pool	R^2	No. of descriptors in the final model	Q^2 (LOO)	$Q_{F_1}^2$	S (train)	S (test)
Current study	61	41	20	2D	42	0.726	5 (3 LVs)	0.635	0.672	0.269	0.333
Saavedra *et al.*, 2018 (ref. 7)	62	52	10	2D + 3D	4885	0.690	5	0.600	—	0.28	0.39
Scotti *et al.*, 2014 (ref. 20)	55	41	14	3D	128	0.714	6	0.679	0.623	—	—

of variables contributing similar information are grouped together and are correlated. The variables which are situated far away from the plot origin are considered to have a stronger impact on the model for that particular variable. The algebraic sign of the PLS loading is also taken into account, which gives important information about correlation among the variables. The *X*-variable ETA_EtaP_F is influential for the *Y*-variable pLC_{50} because of its closeness to the *Y*-variable. Hence, if the numerical value of this descriptor increases, the larvicidal activity against the *Aedes* mosquito will also increase. In the case of the descriptor ETA_dEpsilon_D, which is present on the opposite side of the plot origin with respect to pLC_{50}, this suggests that an increase in ETA_dEpsilon_D value will result in a decrease in activity. From the loading plot, we can also identify the weighting of the *X*-variables based on the first component and second component. From weighting (Table 1) analysis, we can conclude that component 1 considers the hydrogen bonding property of compounds and component 2 considers the electron richness of the compounds.

3.3. Applicability domain of PLS model

The applicability domain (AD) gives a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable.[37] The AD assessment of the proposed model for PCPs was performed according to the DModX (distance to model) in the *X*-space approach using SIMCA-P[38] software. From Fig. S1 (in ESI†) we can see that there is only one outlier to be found in the training set, *i.e.*, compound **8** (or carvacryl trichloroacetate) and one compound outside the AD, *i.e.*, compound **15** (or thymyl trichloroacetate) at a 99% confidence level (D-critical = 0.00999898).

3.4. Randomization model of PLS model

The statistical significance of the model is analyzed by randomization plot (Fig. S2 in ESI†). The randomization plot has been developed in order to confirm that the model is not the result of any chance correlation.[39] In randomization, a number of models are generated by permuting different combinations of *X* or *Y* variables based on the fit of the reordered model. In our study, for the training set, the *X* data remained intact and the *Y* data were shuffled randomly (*Y*-randomization), and the model was fitted to the permuted data and compared with the best fit. The number of permutations can vary; here we used 100 permutations. The basic statistics of randomization models (Q^2 and R^2) should be poor and not within the range of those for acceptable regression models. Otherwise, each resulting model may be considered as a chance correlation.[40] The value of the R_Y^2 intercept should not exceed 0.3 and the value of the Q_Y^2 intercept should not exceed 0.05. The obtained model in our study shows the intercept at $R_Y^2 = 0.0487$, $Q_Y^2 = -0.355$ (in Fig. S2 in ESI†), signifying the validity of the model. This shows that the developed model is non-random and robust, and is suitable for prediction of the larvicidal activity of compounds within the AD of the model.

3.5. Comparison with previously published models

We compared the currently developed model with previously developed models[7,20] for larvicidal activity against *Aedes aegypti* in terms of quality measures (Table 2). However, due to the different compositions of the training and test sets in these studies, a critical comparison of the models is not possible. The advantage of the current model is that it has been developed by using simple 2D ETA descriptors which do not require conformation analysis or energy minimization prior to their calculation. Also, these descriptors have been calculated using freely available software (PaDel-Descriptor).[23] The model we developed using a single class of descriptors (ETA) is comparable to or of better quality than those developed previously[7,20] using computationally more expensive 3D descriptors.

4. Conclusion

The present research used chemometric tools for investigating a set of 61 compounds of natural origin showing larvicidal activity against the zika vector *Aedes aegypti*. Based on the information obtained from the final PLS model (as also illustrated in the regression coefficient plot, variable importance plot, loading plot and score plot, in Fig. 1, 2, 8 and 9), we can conclude that: (i) the presence of hydrogen bond donor groups like –OH, $–NH_2$, –SH *etc.* will attenuate the larvicidal activity against the zika vector; (ii) heteroatoms and multiple bonds are essential to increase the activity; (iii) the presence of electronegative hydrogen bond acceptor atoms helps to increase the larvicidal activity; (iv) a higher polar surface area is detrimental to the activity. The QSAR model developed here with simple and interpretable descriptors highlights the structural requirements and molecular properties needed to be present in the

compounds for them to show acceptable larvicidal properties. The topological descriptors used also do not require the application of time-consuming computational procedures like conformational analysis or energy minimization; thus the developed model may be suitable for the quick screening of database compounds. The developed model further helps in the prediction of the activity of new analogues even before their synthesis and/or evaluation.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

PD is thankful to All India Council for Technical Education, New Delhi for providing an MPharm scholarship. KR thanks the UGC, New Delhi providing financial assistance under UPE-II scheme.

References

1 D. J. Gubler, *Arch. Med. Res.*, 2002, **33**, 330–342.
2 A. R. Katritzky, Z. Wang, S. Slavov, M. Tsikolia, D. Dobchev, N. G. Akhmedov, C. D. Hall, U. R. Bernier, G. G. Clark and K. J. Linthicum, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 7359–7364.
3 http://www.who.int/csr/don/25-august-2017-chikungunya-france/en/, accessed on 3.11.2017.
4 http://www.who.int/csr/don/29-september-2017-chikungunya-italy/en/, accessed on 3.11.2017.
5 http://www.who.int/mediacentre/commentaries/yellow-fever/en/, accessed on 3.11.2017.
6 R. I. Rose, *Emerging Infect. Dis.*, 2001, **7**, 17.
7 L. M. Saavedra, G. P. Romanelli, C. E. Rozo and P. R. Duchowicz, *Sci. Total Environ.*, 2018, **610**, 937–943.
8 S. Licciardi, J. P. Hervé, F. Darriet, J. M. Hougard and V. Corbel, *Med. Vet. Entomol.*, 2006, **20**, 288–293.
9 A. L. Tapondjou, C. Adler, D. A. Fontem, H. Bouda and C. H. Reichmuth, *J. Stored Prod. Res.*, 2005, **41**, 91–102.
10 P. J. Rice and J. R. Coats, *J. Econ. Entomol.*, 1994, **87**, 1172–1179.
11 H. c. Carrasco, M. Raimondi, L. Svetaz, M. D. Liberto, M. V. Rodriguez, L. Espinoza, A. Madrid and S. Zacchino, *Molecules*, 2012, **17**, 1002–1024.
12 J. K. Kim, C. S. Kang, J. K. Lee, Y. R. Kim, H. Y. Han and H. K. Yun, *Entomol. Res.*, 2005, **35**, 117–120.
13 K. Murugan, P. Murugan and A. Noortheen, *Bioresour. Technol.*, 2007, **98**, 198–201.
14 A. Leo and D. H. Hoekman, *Exploring QSAR:. Fundamentals and applications in chemistry and biology*, An American Chemical Society Publication, 1995.
15 K. Roy and R. N. Das, in *Quantitative Structure–Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment*, 2015, p. 48.
16 K. Roy and R. N. Das, *SAR QSAR Environ. Res.*, 2011, **22**, 451–472.
17 S. R. L. Santos, V. B. Silva, M. A. Melo, J. D. F. Barbosa, R. L. C. Santos, D. o. P. de Sousa and S. c. C. H. Cavalcanti, *Vector Borne Zoonotic Dis.*, 2010, **10**, 1049–1054.
18 S. R. L. Santos, M. A. Melo, A. V. a. Cardoso, R. L. C. Santos, D. o. P. de Sousa and S. c. C. H. Cavalcanti, *Chemosphere*, 2011, **84**, 150–153.
19 J. D. F. Barbosa, V. B. Silva, P. B. Alves, G. Gumina, R. L. C. Santos, D. o. P. Sousa and S. c. C. H. Cavalcanti, *Pest Manage. Sci.*, 2012, **68**, 1478–1483.
20 L. Scotti, M. Tullius Scotti, V. Barros Silva, S. Regina Lima Santos, S. c. Ch Cavalcanti and F. Jb Mendonca Junior, *Med. Chem.*, 2014, **10**, 201–210.
21 https://www.chemaxon.com.
22 K. Roy and G. Ghosh, *Curr. Pharm. Des.*, 2010, **16**, 2625–2639.
23 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.
24 http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab.
25 H.-S. Park and C.-H. Jun, *Expert Syst. Appl.*, 2009, **36**, 3336–3341.
26 P. T. Pope and J. T. Webster, *Technometrics*, 1972, **14**, 327–340.
27 D. s. e. Baumann and K. Baumann, *J. Cheminf.*, 2014, **6**, 47.
28 K. Roy, R. N. Das, P. Ambure and R. B. Aher, *Chemom. Intell. Lab. Syst.*, 2016, **152**, 18–33.
29 R. Leardi, *J. Chemom.*, 2001, **15**, 559–569.
30 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
31 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52**, 2044–2058.
32 T. Chai and R. R. Draxler, *Geosci. Model Dev.*, 2014, **7**, 1247–1250.
33 N. Akarachantachote, S. Chadcham and K. Saithanu, *Int. J. Pure Appl. Math.*, 2014, **94**, 307–322.
34 K. Roy and G. Ghosh, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 559–567.
35 K. Roy and R. N. Das, in *Advanced methods and applications in chemoinformatics: Research progress and new applications*, IGI Global, 2012, pp. 380–411.
36 J. E. Jackson, *A user's guide to principal components*, John Wiley & Sons, 2005.
37 D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti and O. Nicolotti, *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, 2016, **1**, 45–63.
38 U. Simca-P, 10.0, info@umetrics.com, www.umetrics.com, Umea, Sweden, 2002.
39 J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, 1979, **22**, 1238–1244.
40 C. Rücker, G. Rücker and M. Meringer, *J. Chem. Inf. Model.*, 2007, **47**, 2345–2357.

