Environmental Science and Pollution Research (2020) 27:6205–6214
https://doi.org/10.1007/s11356-019-06630-9

RESEARCH ARTICLE

A non-conformational QSAR study for plant-derived larvicides against Zika *Aedes aegypti* L. vector

Laura M. Saavedra[1] · **Gustavo P. Romanelli**[2,3] · **Pablo R. Duchowicz**[1]

Received: 23 April 2019 / Accepted: 25 September 2019 / Published online: 21 December 2019

Abstract

A set of 263 plant-derived compounds with larvicidal activity against *Aedes aegypti* L. (Diptera: Culicidae) vector is collected from the literature, and is studied by means of a non-conformational quantitative structure-activity relationships (QSAR) approach. The balanced subsets method (BSM) is employed to split the complete dataset into training, validation and test sets. From 26,775 freely available molecular descriptors, the most relevant structural features of compounds affecting the bioactivity are taken. The molecular descriptors are calculated through four different freewares, such as PaDEL, Mold^2, EPI Suite and QuBiLs-MAS. The replacement method (RM) variable subset selection technique leads to the best linear regression models. A successful QSAR equation involves 7-conformation-independent molecular descriptors, fulfiling the evaluated internal (*loo*, *l30%o*, *VIF* and Y-randomization) and external (test set with $N_{\text{test}} = 65$ compounds) validation criteria. The practical application of this QSAR model reveals promising predicted values for some natural compounds with unknown experimental larvicidal activity. Therefore, the present model constitutes the first one based on a large molecular set, being a useful computational tool for identifying and guiding the synthesis of new active molecules inspired by natural products.

Keywords QSAR analysis · MLR method · Larvicidal activity · *Aedes aegypti* vector · Freeware

Introduction

Mosquitoes are one of the deadliest arthropods ever known worldwide; their ability to carry and spread out infectious diseases on the human race leads annually to more than one million deaths in risk zones (tropical and subtropical countries). Several vector-borne diseases such as dengue, yellow fever and Chikungunya fever (*flaviviruses*) are transmitted by the genus *Aedes*, mainly the *Aedes aegypti* (Diptera: Culicidae) mosquito, which is also responsible for transmitting Zika virus (ZIKV) (Department of control of neglected tropical diseases/WHO et al. 2017).

The predominant mode of transmission for ZIKV is through the infected mosquito's bite. However, this virus may also be propagated by sexual contact, blood transfusions and perinatal transmission (ECDC 2017). The ZIKV has become increasingly notorious due to the fact that it coincides with increased microcephaly cases, the Guillain-Barré syndrome (GBS) and other neurological disorders associated with intrauterine central nervous system (CNS) infection (Srinivasan et al. 2015). Hence, the *A. aegypti* vector is considered a potential threat to the world public health.

There exist three main strategies for controlling and preventing the transmission of vector-borne infections: vaccines, antivirus and mosquito control programs. Unfortunately, to date, these methods have not proven to be successful. Within the mosquito control programs, larvicidal products are found to be effective, although their repeated and indiscriminate use has triggered

Responsible Editor: Giovanni Benelli

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-019-06630-9) contains supplementary material, which is available to authorized users.

✉ Laura M. Saavedra
laurasaa0913@gmail.com

✉ Pablo R. Duchowicz
pabloducho@gmail.com

[1] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina

[2] Departamento de Química, Facultad de Ciencias Exactas, CONICET, UNLP, Centro de Investigación y Desarrollo en Ciencias Aplicadas "Dr. J.J. Ronco" (CINDECA), Calle 47 No. 257, B1900AJK La Plata, Argentina

[3] Cátedra de Química Orgánica, Centro de Investigación en Sanidad Vegetal (CISaV), Facultad de Ciencias Agrarias y Forestales, Universidad Nacional de La Plata, Calles 60 y 119 s/n, B1904AAN La Plata, Argentina

negative environmental impacts, the development of resistance in mosquito populations and bioaccumulation in non-target organisms (Lima et al. 2015; World Health Organization (WHO)1992). Consequently, many studies have focused on developing new strategies based upon plant-derived larvicides and their analogues, which allow performing a selective, environmentally friendly and effective larval control, being renewable and biodegradable feedstocks with low mammalian toxicity (Kim et al. 2013; Yu et al. 2015).

It is well-known that the experimental design of new active natural larvicides is limited, as the biological and clinical assays require time and economical resources. Therefore, the search, identification and development of new selective and potent biomolecules against vectors have usually been assisted by *in silico* techniques.

The main hypothesis behind the quantitative structure-activity relationships (QSAR) theory (Hansch et al. 1995) relies on the fact that the molecular structure of a chemical compound determines its observed properties. The QSAR formalism has proved to be a successful computational tool for studying biological, organoleptic and physicochemical properties of interest. The essence of any QSAR study is not to predict the involved mechanism of action but the property, which is a final result of this mechanism.

A QSAR model allows finding a logical mathematical relationship between the biological response (bioactivity) and a set of representative molecular descriptors capturing specific structural information of the constitutional, topological, geometrical or electronic type. Such correlation between the molecular descriptors and a biological activity may be established through linear or non-linear techniques, leading to the best possible structure-activity parallelisms (Hansch and Verma, 2009; Katritzky and Goordeva 1993; Roy et al. 2015; Devillers et al. 2014).

Nowadays, few QSAR studies have been focused on natural or semi-synthetic classes of molecules with larvicidal activity against *A. aegypti*. Besides, all the published models are based on small-size datasets (Devillers et al. 2014; Da Silva et al. 2015). In 2014, Scotti and co-workers develop a chemometric study with 55 larvicides. Principal component analysis (PCA), consensus PCA (CPCA) and partial least squares regression (PLS) methods are employed for analysing 128 3D-molecular interaction fields (MIFs) with GRID force field descriptors obtained from VolSurf+ program, establishing a suitable model with correlation coefficients: $R^2_{\text{train}} = 0.71, R^2_{\text{test}} = 0.68$, (with 14 compounds) and $R^2_{LOO} = 0.67$ through the PCA technique (Scotti et al. 2014). Subsequently, a set of 31 monoterpenes with acute toxicity against the *A. aegypti* larvae is studied by Alencar Filho and co-workers in 2016 (Alencar Filho et al. 2016), by means of multivariable linear regression (MLR) technique. They find a QSAR model with 3-molecular descriptors from E-Dragon that has good predictive ability $(R^2_{\text{train}} = 0.83, S_{\text{train}} = 0.19, R^2_{\text{test}} = 0.83$ (with 7 molecules) and $R^2_{LOO} = 0.77$. Recently, we have performed a QSAR study for 62 plant-derived compounds against Zika *A. aegypti* vector (Saavedra et al. 2018a). The replacement method (RM) variable subset selection technique coupled with MLR (Duchowicz et al. 2006) proves to be useful for exploring 4885 Dragon 6 descriptors. A suitable QSAR model involving five descriptors with acceptable predictive capability for both the training set $(N_{\text{train}} = 52\ , R^2_{\text{train}} = 0.69,\ S_{\text{train}} = 0.28)$ and the test $(N_{\text{test}} = 10, R^2_{\text{test}} = 0.78, S_{\text{test}} = 0.39)$ is established.

In a next study, we have proposed an alternative QSAR model for the same molecular set (excluding two molecules with high experimental measurement error), with the purpose of applying non-conformational descriptors calculated with freely available softwares. The RM technique is applied on 18,326 conformation-independent descriptors (Saavedra et al. 2018b). Thus, a robust and reliable 5-descriptors model $\left(N_{\text{train}} = 50, R^2_{\text{train}} = 0.84, S_{\text{train}} = 0.20\right)$ is achieved with a high predictive capability $\left(N_{\text{test}} = 10, R^2_{\text{test}} = 0.92, S_{\text{test}} = 0.23\right)$ that surpasses previously published ones.

The main objective of the present work is based on our continuous efforts for developing predictive QSAR models that contribute to the *A. aegypti* vector control. It is our purpose to establish a mathematical model that predicts the larvicidal activity of a molecular structure set based on 263 plant-derived compounds, which are extracted from the literature, thus demonstrating the structure-activity hypothesis of QSAR. In this way, the so-developed QSAR model may serve as a useful computational tool for identifying and guiding the synthesis of new active molecules inspired by natural products. Simple and interpretable models solely based on 1D and 2D structural information are established, employing several freely available programs, such as PaDEL (Yap, 2011; PaDEL, 2018), Mold^2 (Hong, et al. 2008), EPI Suite (US EPA, 2016) and QuBiLs-MAS software (Valdes-Martini, 2012).

Materials and methods

Experimental dataset

The QSAR study is performed on 263 plant-derived molecules with larvicidal activity against *A. aegypti*. This molecular set comprises structurally diverse larvicidal compounds, including quinones, polyketides, phenylpropanoids, coumarins, flavonoids, terpenoids, alkaloids and their analogues. The larvicidal activity is defined as the median lethal concentration LC_{50} (μg mL^{-1}), which represents the concentration at which 50% of third or early fourth instar larvae show lethal effects after 24 to 48 h of treatment with the testing solution. The dataset is collected from the literature (Geris et al. 2012;

Dias and Fernandes 2014; Kishore et al. 2014) and for modelling purposes, each LC_{50} value is converted into the logarithmic scale ($\log_{10}LC_{50}$) The complete list of molecules studied here is provided in Table 1S as Supplementary material.

Molecular modelling and molecular descriptors calculation

The analysed 263 chemical structures are generated in both canonical SMILES notation and 2D structures are drawn with the ACDLabs ChemSketch open-source software (Weininger 1988; ACD/ChemSketch program 2016), without performing geometrical optimization, and saved in MDL mol (V2000) format. All the compounds studied here are listed in Table 1S of the Supplementary material section. The file format conversions are performed with Open Babel for Windows (O'Boyle et al. 2011).

An advantage of not analysing molecular conformations is that the only experimental data required for developing the QSAR models is the studied experimental larvicidal activity. In addition, it is known that the exclusion of geometrical (3D) descriptors, e.g. the charge distribution descriptor which requires optimizing the molecular structure, avoids ambiguities arising from the existence of a molecule in various conformational states (Doucet et al. 2017).

The set of non-conformational descriptors is calculated using PaDEL version 2.20 freely-available software (Yap 2011; PaDEL 2018). PaDEL currently calculates 1444 0D-2D descriptors and 12 types of fingerprints (16,092 bits). Also, Mold^2 version 2.0 freeware is used to compute a set of 777 descriptors, by encoding the 2D chemical structure information from molecules in MDL sdf format (Hong et al. 2008). Furthermore, 14 semiempirical descriptors from EPI Suite are added (US EPA 2016); these variables are based on physicochemical properties and environmental fate estimations, such as the Henry's law constant ($\log K_H EPI$), the sorption coefficient for soil and sediment ($\log K_{oc} EPI$) and the logarithm of the octanol/water partition coefficient ($\log K_{ow} EPI$). Finally, two-dimensional descriptors are calculated through quadratic bilinear and *N*-Linear maps (QuBiLs) (Valdes-Martini et al. 2012), employing the graph-theoretic electronic-density matrices and atomic weightings (MAS) module from the ToMoCoMD-CARDD free multi-platform software. The QuBiLs-MAS algebraic module has the capability of calculating 8448 tensor-based indices belonging to 176 types of bilinear, quadratic and linear algebraic maps, which are based on *N*-tuple spatial metric (dis-similarity matrices and atomic weightings) indices. The *N*-tuple matrices are used to represent the relationships among two, three and four atoms, and can also be used to codify information related to groups or atom-types belonging to a specific molecular fragment (Valdes-Martini et al. 2017).

The QuBiLs-MAS freeware is used by selecting the following options: 'bilinear (B)', 'quadratic (Q)' and 'linear (F)' algebraic forms; 'atom-based (AB)', 'non-chiral (nCi)', 'duplex' constraints; 'non-stochastic (NS)', 'simple stochastic (SS)', 'double stochastic (DS)', 'mutual probability (MP)' matrix forms with 15 of maximum order; 'keep all (KA)' cut off; 'total' groups. Moreover, the included atomic properties are 'Ghose-Crippen LogP (A)', 'charge (C)', 'electronegativity (E)', 'mass (M)', 'polarizability (P)', 'polar surface area (PSA)', 'refractivity (R)', 'mass and van der Waals volume (V)'; 'Euclidean distance (N2)', 'arithmetic mean (alpha = 1) (AM)' and 'standard deviation (SD)', invariants with non-standardized option.

A great number of 26,775 non-conformational descriptors are calculated in this work, in order to explore the most relevant structural characteristics affecting the analysed larvicidal activity. Afterwards, linearly dependent descriptor pairs are identified, and one variable is removed from each pair. Also, non-informative descriptors (i.e. variables with constant or near-constant values and variables with at least one missing value) are excluded from the original matrix of variables in order to remove redundant information. Thus, a matrix with 10,604 linearly independent non-conformational descriptors is achieved.

Molecular descriptors selection based on MLR

The MLR technique has proven to be of great utility in several disciplines for establishing predictive QSAR models. MLR models clearly show the effect of including/excluding descriptors in the linear equation; then, it is possible to suggest cause/effect relationships through such simple parallelisms. Another advantage of MLR based models is that they do not require too many optimized parameters during the model design (just a regression coefficient per descriptor) (Duchowicz et al. 2017; Duchowicz 2018). Thus, we employ the replacement method (RM) variable subset selection technique in order to generate MLR models on the training set (train), by searching in a pool having $D = 10{,}604$ descriptors for an optimal subset containing *d* descriptor ($D \gg d$),with smallest standard deviation (S_{train}) or smallest root mean square error (RMS_{train}) (Duchowicz et al. 2005; Duchowicz et al. 2006).

The RM technique has been successfully applied in different QSAR studies (Aranda et al. 2016; Duchowicz et al. 2017; Duchowicz 2018), and the quality of the results achieved with this technique is quite similar to that obtained by performing an exact (combinatorial) full search (FS) of molecular descriptors, although, of course, it requires much less computational work. The RM provides models with better statistical parameters than the ones obtained with the forward stepwise regression (FSR) procedure, and quite similar to the results found by the genetic algorithms (GA) approach (Morales et al. 2006). Table 2S includes a list of mathematical equations used in the present study. The Matlab-programmed algorithms involved in our calculations are available upon request (Matlab n.d.).

Model validation

The analysis of an external test set of molecules, never seen by the model during the calibration of its parameters, is known as the most reliable validation criterion. In this sense, the complete molecular set of 263 plant-derived larvicides is split into three subsets: training (train), validation (val) and test sets. The training set is used for calibrating and obtaining the parameters of the QSAR model, employing the RM technique, while the validation set helps to calibrate and partially validate the model by predicting the bioactivity of molecules not included in the training set. Finally, the test set contains compounds "never seen" during the model calibration with the training and the validation sets, and thus, it can demonstrate the real predictive power of the QSAR.

A rational partition into training, validation and test sets should lead to similar structure-activity relationships in each set, due to the fact that a random splitting of compounds does not lead to suitable prediction results. Hence, we perform a rational splitting of the molecular set through the balanced subsets method (BSM) (Rojas et al. 2015; Aranda et al. 2017), which has been developed by our group, and is based upon the k-means cluster analysis (k-MCA) method executed in Matlab. The main idea behind the BSM technique is to create k-clusters or groups of compounds in such a way that compounds in the same cluster are very similar in terms of a distance metrics (i.e. Euclidean distance), and compounds in different clusters are very distinct. Thus, the BMS procedure ensures that the training set is representative of both the validation and test sets.

Afterwards, the QSAR model is theoretically validated through the leave-one-out (*loo*) cross-validation technique (Wold et al. 1995; Hawkins et al. 2003; Gramatica 2007). The statistical parameters R^2_{loo} and (RMS_{loo}) (square correlation coefficient and root mean square error of(*loo*)) measure the stability of the QSAR model upon inclusion/exclusion of molecules. Moreover, the more rigorous leave-30%-out (*l30%o*) cross-validation technique is employed; it uses 40 molecules of the training set. Throughout 80,000 cases of random data removal, the results are expressed by the $R^2_{l30\%}$ and RMS$_{l\ 30\%}$statistical parameters (Rücker et al. 2007).

On the other hand, we verify the model's robustness through the Y-randomization procedure (Stanton et al. 1993), which scrambles the experimental property values in such a way that they do not correspond to the respective compounds. After calculating 150,000 cases, the obtained root mean square error (RMS^{rand}) has to be a poorer value than the one found by considering the true calibration(RMS_{train}). Thus, if, $RMS^{\mathrm{rand}} > RMS_{\mathrm{train}}$ it is assumed that the QSAR is not fortuitous and does not result from happenstance, confirming a genuine structure-activity relationship.

Owing to the need of checking the inter-correlation effect among molecular descriptors, we also analyse the variance inflation factor (*VIF*). The *VIF* values indicate how much the variance of the descriptor coefficient is inflated as compared to the case where the descriptors are completely orthogonal to each other. Ideally, the *VIF* value of each descriptor should be lower than 10 (Mullen et al. 2011; Rafiei et al. 2016).

Additionally, some important validation criteria proposed by Golbraikh et al. (2003) are used here, wherein some model's parameters should fulfil specific requirements for ensuring the predictive capability: $0.85 \leq k \leq 1.15$ *or* $0.85 \leq k' \leq 1.15$, as well as $1 - R^2_0/R^2_{\mathrm{test}} < 0.1$ *or* $1 - R'^2_0/R^2_{\mathrm{test}} < 0.1$, and $R^2_m > 0.5$

Applicability domain analysis

A predictive model is only able to predict molecules falling within its applicability domain (AD), so that the predicted activity is not a result of substantial extrapolation, considering that not even a predictive model has the capability to reliably predict the modelled activity for the whole universe of molecules (Roy et al. 2015). The AD is a theoretically defined area that depends on the model's descriptors and the experimental activity.

In this study, we determine the AD through the well-known leverage approach (Gramatica 2007), where a test set compound *i* must have a calculated leverage (h_i) smaller than the warning leverage (h^*). Table 2S includes the definitions for (h_i) and h^*. Then, when $h_i > h^*$or h_i is quite similar to h^* for a test set compound, a warning should be given: this means that the prediction for this test set compound is a result of substantial extrapolation of the model and cannot be considered as reliable (Eriksson et al. 2003). In order to represent the AD from the selected model, the Williams plot is drawn.

The degree of importance of selected descriptors

The relative importance of the *jth* descriptor of the linear model is determined by standardizing its regression coefficient (b^s_j, see Table 2S). Thus, the larger is the absolute value of b^s_j, the greater is the importance of such descriptor (Draper and Smith 1998; Cañizares-Carmenate et al. 2017).

Results and discussion

We begin our QSAR analysis with a large set of 263 molecular structures built upon secondary metabolites and their analogues, with larvicidal activity against *A. aegypti* mosquito. In order to split the dataset into training ($N_{\mathrm{train}} = 133$), validation ($N_{\mathrm{val}} = 65$) ans test ($N_{\mathrm{test}} = 65$) sets, the BSM is employed to ensure that representative sets are obtained. Thus, the calibration compounds in the training and validation sets constitute 75% of the whole dataset. Table 1S from the Supplementary material denotes the members of val (*) and test (^) sets.

The most relevant structural characteristics of the training set are searched by means of the RM technique, providing a way to explore 10,604 linearly independent descriptors. The model selection criteria are based upon the minimum *RMS*, the maximum coefficient of determination (R^2) and the minimum $R^2_{ij\max}$ value between descriptor pairs in the model. Thus, the best MLR models involving the most representative 1-8 molecular descriptors are detailed in Table 1. A brief description of such descriptors is also given in Table 3S. From the results of Table 1, it is clearly appreciated that both the $R^2_{\text{train}}\,and\,RMS_{\text{train}}$ parameters improve with the addition of molecular descriptors into the linear equation until. $d = 8$ We use the criterion of keeping the model's dimension as small as possible, thereby allowing us to select the 7 descriptors model whose statistical quality for both the training and validation sets is acceptable. Besides, it is the only model from Table 1 capable of accomplishing with the internal and external validation criteria employed in this work.

$$\log_{10}LC_{50} = 5.6 + 08(\pm 0.3)M16 - 1.8(\pm 0.3)PC34 + 0.6(\pm 0.2)PC199 - 0.7(\pm 0.2)KR1592 + 1.1(\pm 0.3)AP653 - 1.2(\pm 0.2)Sub282 - 1.2(\pm 0.1)D589 \tag{1}$$

$$N_{\text{train}} = 133, d = 7, R^2_{\text{train}} = 0.68, RMS_{\text{train}} = 0.43, N_{\text{train}}/d = 19,$$
$$R^2_{(ijmax)} = 0.19, VIF^{\max} = 1.1, o3 = 0$$
$$R^2_{loo} = 0.64, RMS_{loo} = 0.46, R^2_{130\%o} = 0.55, RMS_{130\text{‰}} = 0.50, RMS^{rand} = 0.64$$
$$N_{val} = 65, R^2_{val} = 0.72, RMS_{val} = 0.41$$
$$N_{test} = 65, R^2_{test} = 0.58, RMS_{test} = 0.37$$

Here, *N* denotes the number of compounds in each set, the *o3* indicates the number of outlier compounds having a residual (difference between experimental and predicted activity value) greater than three-times $RMS_{\ train}$, and the N_{train}/d ratio indicates that the model satisfies the rule of thumb.

A plot for the $\log_{10}LC_{50}$ prediction given by Eq. 1 as a function of the experimental value in each molecular set is provided by Fig. 1 (numerical data provided in Table 4S). Figure 2 draws the dispersion plot of residuals (residual as a function of the predicted $\log_{10}$ LC_{50} value), which tends to obey a random pattern around the zero line, suggesting that Eq. 1 predicts the whole dataset without systematic errors.

Our proposed 7-descriptor model approves the internal validation process of *loo* and 130%o (80,000 cases) cross-validation procedures through the exclusion of 1 or 40 molecules at a time from the training set, respectively. The obtained results for both the *loo* and 130%o techniques indicate that Eq. 1 does not deteriorate so much with the removal of one (R^2_{loo}=0.64, RMS_{loo}=0.46) or more compounds ($R^2_{130\%o}$=0.55, $RMS_{130\%o}$=0.50). According to the specialized literature (Hawkins et al. 2003), the cross-validation R^2_{loo} and $R^2_{130\%o}$ explained variances should be greater than 0.5, although this is a necessary but not sufficient condition for determining the real predictive power of the model.

In order to demonstrate that our proposed QSAR model is not a result of happenstance correlation, the Y-randomization technique is employed, showing that $RMS_{train} < RMS^{rand}$ (0.64) after analysing 150,000 randomization cases; in this way, a valid structure-activity relationship is achieved. Furthermore, the external validation criteria suggested by Golbraikh et al. (2003) are checked in order to ensure the predictive power of Eq. 1: $1\text{-}R^2_0/R^2_{test}(0.02) < 0.1$ or $1\text{-}R'^2_0/R^2_{test}(0.2) < 0.1$; $0.85 \leq k(0.97) \leq 1.15$ or $0.85 \leq k'(0.97) \leq 1.15$; $R^2_m(0.52)>0.5$. These parameters are calculated with the equations from Table 2S of the Supplementary material.

We also analyse the $R^2_{ij\max}$ parameter for Eq. 1, which is the maximum squared correlation coefficient between descriptor pairs: $R^2_{ij\max}$=0.19 indicates that there is no serious problem about structural information overlapping. Likewise, the model's correlation matrix provided in Table 2 shows that the 7-descriptors from the QSAR model have very low intercorrelations; and the VIF^{max} value for all the descriptors are lower (close to 1) than 10 (Gramatica 2007), indicating the absence of multicollinearity.

It is well known that a successful QSAR model is established whenever it surpasses the external validation process. For that reason, we check the model's ability to predict the experimental activity of plant-derived compounds that are not considered during the model's calibration. Equation 1 presents an acceptable predictive power for the external test set with 65 'never seen' experimental $\log_{10}$ LC_{50} values, according to R^2_{test}=0.58 and RMS_{test}=0.37, and Figs. 1 and 2. In this way, we prove that the QSAR model given by Eq. 1 can be very useful for predicting new larvicidal compounds with unknown LC_{50}.

Now, it is possible to provide a brief description for each non-conformational descriptor included in Eq. 1. The seven molecular indices are of two different types: six indicator descriptors (fingerprints) and one Mold^2 topological descriptor, which are detailed below:

- a MACCS fingerprint descriptor: *M*16, which refers to structures that have cycloalkanes (with 5 or 6 members) bonded to hydroxyl (*-OH*), acetyl (*-Ac*) or acetoxy (*-OAc*) groups (Durant et al. 2002).
- two PubChem fingerprint descriptors: *PC*34, which indicates the presence or count of individual chemical atoms represented by the atomic symbol *2S*; *PC*199, which denotes any ring that does not share three consecutive atoms with any other ring in the chemical structure (ESSSR). In this case, the *PC*199 indicator descriptor details the

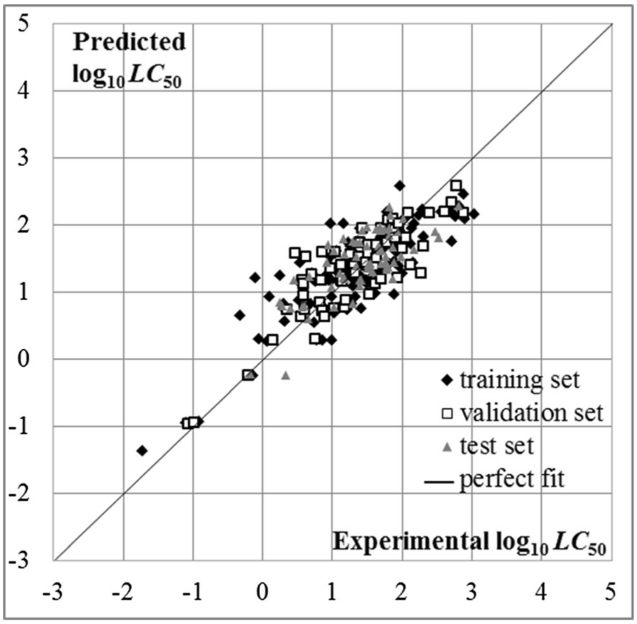

Fig. 1 Experimental and predicted log$_{10}$ LC_{50} values for the 263 plant-derived compounds according to the QSAR of Eq. 1

presence of any ring of size six, which has one ESSSR ring (Bolton et al. 2008).

- a Klekota-Roth fingerprint descriptor: *KR*1592, which indicates the presence of a SMART substructure representing 3,3'-Dimethyl-1,1'-biphenyl.
- a 2D-Atom-pairs fingerprint descriptor: *AP*653, which denotes the presence of atom pairs (O-Br) at topological distance 9.
- a Substructure fingerprint descriptor: *Sub*282, which symbolises the presence of a chemical substructure with 5 or 6-membered ring containing one O and an acetal-like bond at position 2.
- a Mold^2 topological descriptor: *D*589, which denotes the highest eigenvalue from Burden matrix weighted by polarizabilities order-2.

The contribution degree for each descriptor ($b^2{}_j$) is supplied, showing that *D*589 has the greatest contribution in Eq. 1, followed by the *PC*34 and *Sub*282 fingerprints: *D*589 (0.65) > *PC*34 (0.34) > *Sub*282 (0.32) > *PC*199 (0.24) > *KR*1592 (0.20) > *AP*653 (0.17) > *M*16 (0.12). The numerical values of these descriptors are provided in Table 5S; it can be appreciated that all of them are indicator descriptors (fingerprints) that are represented by means of the binary code with the exception of *D*589, which has positive numerical values. Besides, the sign of the regression coefficients in the linear Eq. 1 indicates when the descriptor contribution increases or decreases the predicted log$_{10}$$LC_{50}$ values.

Hence, it is possible to suggest the following useful QSAR guide for the chemical synthesis of new plant-derived larvicides: if *PC*34, *Sub*282 and *KRI*592 indicate their presence (positive numerical values higher than 0) in the chemical structure, while *PC*199, *AP*653 and *M*16 denote their absence (0 values), and simultaneously the *D*589 topological descriptor decreases, then more potent larvicidal compounds could be achieved, exhibiting lower predicted log$_{10}$ LC_{50} values, as happens with molecules **15**, **27-30** from the training set, **26** and **31** from the validation set and **24** of the test set, which are predicted at concentrations LC_{50} < 4µg mL^{-1} against *A. aegypti* larvae.

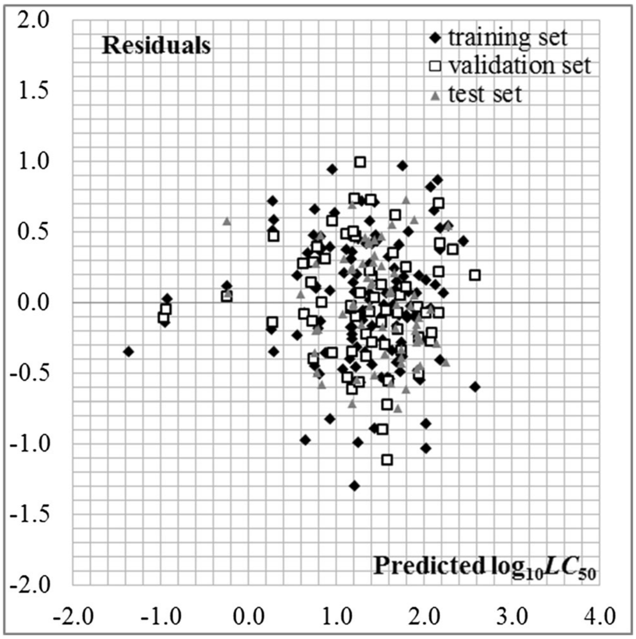

Fig. 2 Dispersion plot of residuals obtained for each analysed subset by Eq. 1

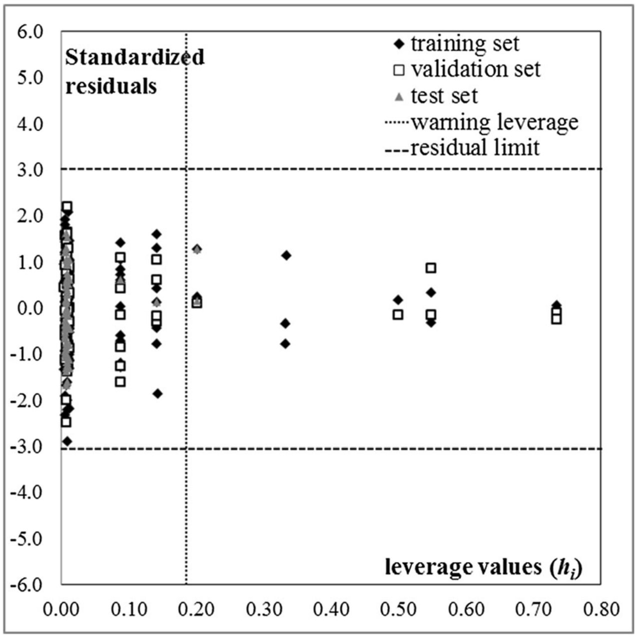

Fig. 3 Williams plot for Eq. 1

Table 1 The best QSAR models of different size established on 263 plant-derived compounds. The selected model appears in bold

d	R^2_{train}	RMS_{train}	R^2_{val}	RMS_{val}	$R^2_{ij\max}$	Molecular descriptors
1	0.31	0.63	0.42	0.59	0	*D*590
2	0.42	0.58	0.64	0.46	0.01	*AATS5v*, *Sub*282
3	0.50	0.54	0.65	0.46	0.01	*AATS5v*, *KR*1592, *Sub*282
4	0.56	0.50	0.67	0.44	0.20	*PC*34, *PC777*, *Sub*282, *D*589
5	0.62	0.47	0.69	0.43	0.06	*VE*2_*Dze*, *PC*34, *KRI*592, *Sub*282, *D*589
6	0.65	0.45	0.70	0.42	0.47	*nRing*, *PC*34, *KRI*592, *KR*3584, *Sub*282, *D*589
7	**0.68**	**0.43**	**0.72**	**0.41**	**0.19**	***M*16, *PC*34, *PC*199, *KR1*592, *AP*653, *Sub*282, *D*589**
8	0.71	0.43	0.69	0.46	0.25	*Ve*1_*Dze*, *PC*34, *KR*1592, *AP*653, *Sub*282, *KRC*4736, *D*178, *D*589

In order to analyse the applicability domain (AD) of the proposed QSAR model, the standardized residual is plotted as a function of the leverage (h_i) value, employing the Williams plot in Fig. 3. Within the leverage approach, a compound with high leverage would reinforce the model when the compound take part of the training or validation set (good leverage); but when such compound belongs to the test set, it would have an unreliable predicted data, as the result of substantial extrapolation of the model (poor leverage) (Eriksson et al. 2003). The obtained leverage values are included in Table S4, revealing that the compounds **115** and **246** included in the test set do not fulfil the AD of Eq. 1, with $h_i > h^*= 0.1805$. It is noteworthy that this particular behaviour can be attributed to the complexity of the whole dataset, i.e. the large structural diversity of the molecules considered in this work. However, Fig. 3 reveals that the predicted $\log_{10}LC_{50}$ values for most of the test set compounds have h_i values falling under the warning h^* value, and thus may be considered as reliable.

The regression model of Eq. 1 has proven to successfully quantify the larvicidal activity (LC_{50}) of 263 plant-derived molecules. Now, we demonstrate that Eq. 1 can be converted into a classification model by classifying compounds with experimental $LC_{50} \leq 40\mu g\ mL^{-1}$ as highly active larvicidals, and compounds with experimental $LC_{50} \leq 40\mu g\ mL^{-1}$ as poorly active larvicidals. As the WHO has not established a standard criterion for determining the larvicidal activity of natural products, in this study, the classification proposed by Cheng et al. (2003) is used. Then, through the Cooper statistics (Cooper et al. 1979), three main statistical parameters are calculated: accuracy (*A*%) referred to concordance; sensitivity (*SE*) that measures correctly predicted toxic compounds, and specificity (*SP*), which calculates rightly predicted non-toxic compounds (Benfenati, 2012). These parameters are defined as shown in Table 2S from the Supplementary material.

The obtained classification parameters show acceptable values in the test set: *A*%= 72, *SE*= 0.75 and *SP*= 0.69, indicating that Eq. 1 achieves classifying 65 experimental $\log_{10}$ LC_{50} values 'never considered' during the model's calibration. Moreover, in order to evaluate the estimated acute toxicity (LC_{50}) of the test set compounds, we have grouped them into four categories: 27 toxic compounds predicted as toxic (*TP*) and 9 as non- toxic (*FN*), as well as 20 non-toxic compounds predicted as non-toxic (*TN*) and 9 as toxic (*FP*). Hence, 73% of the N_{test} compounds are correctly predicted.

In order to perform a practical application of the established QSAR model of Eq. 1, an unknown set of 237 plant-derived larvicidals against *A. aegypti* is collected from the literature. Such compounds have measured activities obtained through different bioassays not considered in the present study. The first step to predict the $\log_{10}$ LC_{50} value consists on calculating the leverage parameter for each compound from the unknown set. The calculated leverage values are recorded in Table 6S, showing that 27 compounds have leverages higher than h^*; then, their predictions cannot be considered reliable. However, the 210 remaining molecules present leverage values under the warning leverage, indicating that the

Table 2 Correlation coefficients matrix of the selected descriptors with their *VIF* values

Model descriptors	*M16*	*PC34*	*PC199*	*KR1592*	*AP653*	*Sub282*	*D589*	*VIF*
M16	1	0.0004	0.0015	0.0008	0.0002	0.0763	0.0049	1.0
PC34		1	0.0023	0.0013	0.0004	0.0005	0.0062	1.0
PC199			1	0.0055	0.0015	0.0419	0.1852	1.1
KR1592				1	0.0008	0.0011	0.0072	1.0
AP653					1	0.0003	0.0157	1.0
Sub282						1	0.0148	1.0
D589							1	1.1

established QSAR model (Eq. 1) is able to reliably calculate the LC_{50} activity against *A. aegypti* vector.

Then, from Table 6S, it is observed that the most prominent compounds exhibit acute toxicity against *A. aegypti* L in a range of 1.8-2.5μg mL^{-1}, such as happens with the natural molecules **11**, **43**, **50**, **91** and **92**, which belong to the naphthoquinones, terpenes, terpenoids, chromones and alcohols chemical groups. In the same manner, Table 6S reveals those compounds with reduced larvicidal activity at concentrations $LC_{50} > 250\ \mu g\ mL^{-1}$, as occurs with molecules **53**, **217** and **218**, which are based on triterpenes groups, a member of furan, an epoxide, an organic heterohexacyclic compound and a lactone.

Finally, it is known that in silico approaches play key roles to find out chemical strategies in vector control. In this sense, a successful QSAR model employing free available programs is built here; this constitutes the first one based on a large and heterogeneous molecular set inspired by phytochemicals. Through Eq. 1, 78% of the whole molecular set with larvicidal activity against *A. aegypti* at concentrations between 0.02 and 790 μg mL^{-1} is correctly predicted, as well as 89% of the unknown set is quantified within the AD of the regression model. For these reasons, we highlight that our linear QSAR model represents a useful computational tool to guide the synthesis or discovery of new active molecules with plausible larvicidal activity.

Conclusions

The *Aedes aegypti* vector is responsible for transmitting several arboviral diseases, such as dengue, chikungunya and zika worldwide, so the identification of larvicidal compounds inspired by natural products has been of great interest during the last years. In this framework, a linear regression QSAR model based upon a large molecular set of 263 plant-derived larvicides is proposed, which involves seven non-conformational descriptors and has an acceptable predictive capability in the external test set.

The established QSAR model is able to fulfil other necessary mathematical conditions, such as *loo* and *l30%o* cross-validation, Y-randomization and *VIF*. For chemical structures falling within the applicability domain of this model, a QSAR guide for the synthesis or identification of new plant-derived larvicides is provided as follows: the molecular structures of active larvicides should have the *PC*34, *Sub*282 and *KR*1592 fragments present, as well as the *PC*199, *AP*653 and *M*16 absent, and simultaneously lower values of *D*589 in order to achieve the best larvicidal effects.

Thus, the simplicity of the linear QSAR model, the easy interpretation, the availability of the involved descriptors and the proper predictive power make the proposed approach attractive for guiding the design of new potentially active molecules based upon renewable sources.

Acknowledgements PRD acknowledges the financial support from the National Scientific and Technical Research Council of Argentina (CONICET) PIP11220130100311 project and to the Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities. PRD and GPR are members of the scientific researcher career of CONICET.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

ACD/ChemSketch program. https://www.acdlabs.com., 2016.

Alencar Filho EB, Castro Silva JW, Cavalcanti SCH (2016) Quantitative structure-toxicity relationships and molecular highlights about Aedes aegypti larvicidal activity of monoterpenes and related compounds. Med Chem Res 25:2171–2178. https://doi.org/10.1007/s00044-016-1650-7

Aranda JF, Bacelo DE, Leguizamón Aparicio MS, Ocsachoque MA, Castro EA, Duchowicz PR (2017) Predicting the bioconcentration factor through a conformation-independent QSPR study. SAR QSAR Environ Res 28:749–763. https://doi.org/10.1080/1062936X.2017.1377765

Aranda JF, Garro Martinez JC, Castro EA, Duchowicz PR (2016) Conformation-independent QSPR approach for the soil sorption coefficient of heterogeneous compounds. Int J Mol Sci 17:1247–1255. https://doi.org/10.3390/ijms17081247

Benfenati E. Theory, guidance and applications on QSAR and REACH. Orchestra: Milan, Italy 2012. Available online: http://ebook.insilico.eu/insilico-ebook-orchestra-benfenati-ed1_rev-June2013.pdf ().

Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. Annual Rep Comput Chem 4:217–241. https://doi.org/10.1016/S1574-1400(08)00012-1

Cañizares-Carmenate Y, Hernander-Morfa N, Torrens F, Castellano G, Castillo-Garit JA (2017) Larvicidal activity prediction against Aedes aegypti mosquito using computational tools. J Vector Born Dis 54:164–171

Cheng SS, Chang HT, Chang ST, Tsai KH, Chen WJ (2003) Bioactivity of selected plant essential oils against the yellow fever mosquito Aedes aegypti larvae. Bioresour Technol 89:99–102. https://doi.org/10.1016/S0960-8524(03)00008-7

Cooper J, Saracci R, Cole P (1979) Describing the validity of carcinogen screening tests. Br J Cancer 39:8–89

Da Silva JBP, Navarro DMAF, da Silva AG, Santos GKN, Dutra KA, Moreira DR, Ramos MN, Espíndola JWP, de Oliveira ADT, Brondani DJ, Leite ACL, Hernandes MZ, Pereira VRA, da Rocha LF, de Castro MCAB, de Oliveira BC, Lan Q, Merz KM Jr (2015) Thiosemicarbazones as Aedes aegypti Larvicidal. Eur J Med Chem 100:162–175. https://doi.org/10.1016/j.ejmech.2015.04.061

Department of control of neglected tropical diseases/WHO. Sixth meeting of the vector control advisory group. World Health Organization, Geneva 2017. Publishing PhysicsWeb.http://www.who.int/neglected_diseases/vector_ecology/resources/WHO_HTM_NTD_VEM_2017.05/en/ [].

Devillers J, Lagneau C, Lattes A, Garrigues JC, Clémenté MM, Yébakima A (2014) In silico models for predicting vector control chemicals targeting Aedes aegypti. SAR QSAR Environ Res 25: 805–835. https://doi.org/10.1080/1062936X.2014.958291

Dias C, Fernandes D (2014) Essential oils and their compounds as Aedes aegypti L. (Diptera: Culicidae) larvicides: review. Parasitol Res 113: 565–592. https://doi.org/10.1007/s00436-013-3687-6

Doucet JP, Papa E, Doucet-Panaye A, Devillers J (2017) QSAR models for predicting the toxicity of piperidine derivatives against Aedes aegypti. SAR QSAR Environ Res 28:451–470. https://doi.org/10.1080/1062936X.2017.1328855

Draper NR, Smith H (1998) Applied Regression Analysis, Third edn. Wiley, New York

Duchowicz PR, Castro EA, Fernández FM (2006) Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem 55:179–192

Duchowicz PR, Castro EA, Fernández FM, González MP (2005) A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules. Chem Phys Lett 412:376–380. https://doi.org/10.1016/j.cplett.2005.07.016

Duchowicz PR, Fioressi SE, Castro EA, Wróbel K, Ibezim NE, Bacelo DE. Conformation-independent QSAR Study on human epidermal growth factor receptor-2 (HER2) inhibitors. Chemistryselect 2017; 2: 3725-3731, doi:https://doi.org/10.1002/slct.201700436.

Duchowicz PR (2018) Linear regression QSAR models for Polo-Like Kinase-1 inhibitors. Cells **7**:13–24. https://doi.org/10.3390/cells7020013

Durant JL, Leland BA, Henry DR, Nourse J (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42: 1273–1280. https://doi.org/10.1021/ci010132r

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARS. Environ Health Perspect 111:1361–1375

European Centre for Disease Prevention and Control. Rapid risk assessment. Zika virus disease epidemic, Tenth update, 4 April 2017. Stockholm: ECDC.

Geris R, Ribeiro PR, Da Silva M, Garcia HH, Garcia I (2012) Bioactive natural products as potential candidates to control Aedes aegypti, the vector of dengue: Atta-ur-Rahman (ed) Studies in Natural Products Chemistry, vol 37. Academic Press, Elsevier, London

Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253. https://doi.org/10.1023/A:1025386326946

Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701. https://doi.org/10.1002/qsar.200610151

Hansch C, Leo A, Exploring QSAR (1995) fundamentals and applications in chemistry and biology. by American Chemical Society, Washington, pp 139–205

Hansch C, Verma RP (2009) Larvicidal activities of some organotin compounds on mosquito larvae: A QSAR study. Eur J Med Chem 44: 260–273. https://doi.org/10.1016/j.ejmech.2008.02.040

Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross validation. J Chem Inf Model 43:579–586. https://doi.org/10.1021/ci025626i

Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W (2008) Mold2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 48: 1337–1344. https://doi.org/10.1021/ci800038f

Katritzky AR, Goordeva EV (1993) Traditional topological indices vs. electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J Chem Inf Comput Sci 33:835–857. https://doi.org/10.1021/ci00016a005

Kim MG, Jeon JH, Lee HS (2013) Larvicidal activity of the active constituent isolated from tabebuia avellanedae bark and structurally related derivatives against three mosquito species. J Agric Food Chem 61:10741–10745. https://doi.org/10.1021/jf403679h

Kishore N, Mishra MM, Tiwari VK, Tripathi V, Lall N (2014) Natural products as leads to potential mosquitocides. Phytochem Rev 13: 587–627. https://doi.org/10.1007/s11101-013-9316-2

Lima TC, Santos SR, Uliana MP, Santos RC, Brocksom TJ, Cavalcanti SCH, de Sousa DP (2015) Oxime derivatives with larvicidal activity against Aedes aegypti L. Parasitol Res 114:2883–2891. https://doi.org/10.1007/s00436-015-4489-9

Matlab 7.0. Masachussetts, USA: The MathWorks, Inc., http://www.mathworks.com.

Morales AH, Duchowicz PR, Cabrera Pérez MA, Castro EA, Cordeiro MNDS, González MP (2006) Application of the replacement method as a novel variable selection strategy in QSAR Carcinogenic potential. Chemom Intell Lab Syst 81:180–187. https://doi.org/10.1016/j.chemolab.2005.12.002

Mullen LMA, Duchowicz PR, Castro EA (2011) QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. Chemom Intell Lab Syst 17:269–275. https://doi.org/10.1016/j.chemolab.2011.04.011

O'Boyle N, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. Aust J Chem 3:33–47. https://doi.org/10.1186/1758-2946-3-33

PaDEL, http, //wwwyapcwsoftcom [accessed 10 July 2018].

Rafiei H, Khanzadeh M, Mozaffari S, Bostanifar MH, Avval ZM, Aalizadeh R, Pourbasheer E (2016) QSAR study of HCV NS5B polymerase inhibitors using the genetic algorithm-multiple linear regression (GA-MLR). EXCLI J 15:38–53. https://doi.org/10.17179/excli2015-731

Rojas C, Duchowicz PR, Tripaldi P, Diez RP (2015) QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemom Intell Lab Syst 140:126–132. https://doi.org/10.1016/j.chemolab.2014.09.020

Roy K, Supratik K, Rudra ND. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment; Academic press: Elsevier: USA, 2015.

Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357. https://doi.org/10.1021/ci700157b

Saavedra LM, Romanelli GP, Duchowicz PR (2018a) QSAR analysis of plant-derived compounds with larvicidal activity against Zika Aedes aegypti (Diptera: Culicidae) vector using freely available descriptors. Pest Manag Sci 74:1608–1615. https://doi.org/10.1002/ps.4850

Saavedra LM, Romanelli GP, Rozo CE, Duchowicz PR (2018b) The quantitative structure–insecticidal activity relationships from plant derived compounds against chikungunya and zika Aedes aegypti (Diptera:Culicidae) vector. Sci Total Environ 611:937–943. https://doi.org/10.1016/j.scitotenv.2017.08.119

Scotti L, Scotti MT, Silva VB, Santos SRL, Cavalcanti SCH, Mendonça FJB Jr (2014) Chemometric studies on potential larvicidal compounds against Aedes Aegypti. Med Chem 10:201–210. https://doi.org/10.2174/15734064113099990005

Srinivasan R, Natarajan D, Shivakumar MS, Vinuchakkaravarthy T, Velmurugan D (2015) Bioassay guided isolation of mosquito larvicidal compound from acetone leaf extract of Elaeagnus indica Servett Bull and its in-silico study. Ind Crop Prod 76:394–401. https://doi.org/10.1016/j.indcrop.2015.07.032

Stanton DT, Murray WJ, Jurs PC (1993) Comparison of QSAR and molecular similarity approaches for a structure-activity relationship study of DHFR inhibitors. Quant Struct -Act Relat 12:239–245. https://doi.org/10.1002/qsar.19930120304

US EPA (2016) Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, DC, USA1

Valdes-Martini JR, García Jacas CR, Marrero-Ponce Y, Silveira Vaz d'Almeida Y, Morrel C (2012) QuBiLS-MAS: free software for molecular descriptors calculator from quadratic, bilinear and linear maps based on graph–theoretic electronic-density matrices and atomic weightings; Version 1.0; CAMD-BIR Unit, CENDA Number of Register: 2373-2012. Central University of Las Villas, Villa Clara, Cuba

Valdes-Martini JR, Marrero-Ponce Y, Garcia-Jacas CR, Martinez-Mayorga K, Barigye SJ, Silveira Vaz d'Almeida Y, Pham-The H, Perez-Gimenez F, Morell CA (2017) QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. Aust J Chem 9:35–61. https://doi.org/10.1186/s13321-017-0211-5

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36

Wold S, Eriksson L, Clementi S (1995) Statistical validation of QSAR results. Chemometrics. methods in molecular design. Van de Waterbeemd. H. Eds. Weinheim, Wiley VCH Verlag GmbH

World Health Organization (WHO).Vector resistance to pesticides: fifteenth report of the WHO expert committee on vector biology and control. WHO Technical Report Series 818, Geneva, 1992.

Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/jcc

Yu KX, Wong CL, Ahmad R, Jantan I (2015) Larvicidal activity, inhibition effect on development, histopathological alteration and morphological aberration induced by seaweed extracts in Aedes aegypti (Diptera: Culicidae). Asian Pac J Trop Med 8:1006–1012. https://doi.org/10.1016/j.apjtm.2015.11.011

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.