Advances in Sciences and Engineering 2020;12(1):28-40
https://doi.org/10.32732/ase.2020.12.1.28

Dibenzoylhydrazines as Insect Growth Modulators: Topology-Based QSAR Modelling

J.P. Doucet, A. Doucet-Panaye

University Paris-Diderot,15 rue jean de Baïf ,75013 Paris, France
E-mail: doucet@paris7.jussieu.fr

Received: 10 May 2020; Accepted: 5 June 2020; Available online: 20 June 2020

Abstract: Dibenzoylhydrazines X_a-$(C_6H_5)_a$-CO-N-(*t*-Bu)-NH-CO-$(C_6H_5)_b$-Y_b are efficient insect growth regulators with high activity and selectivity toward lepidopteran and coleopteran pests. For 123 congeneric molecules, a quantitative structure activity relationship model was built in the framework of the QSARINS package using 2D, Topology-based, PaDEL descriptors. Variable selection by GA-MLR allows building an efficient multilinear regression linking pEC_{50} values to nine structural variables. Robustness and quality of the model were carefully examined at various levels: data-fitting (recall), leave-one (or some) - out, internal and external validation (including random splitting), points not in depth investigated in previous works. Various Machine Learning approaches (Partial Least Squares Regression, Projection Pursuit Regression, Linear Support Vector Machine or Three Layer Perceptron Artificial Neural Network) confirm the validity of the analysis, giving highly consistent results of comparable quality, with only a slight advantage for the three-layer perceptron.

Keywords: Dibenzoylhydrazines; Insect growth regulators; QSAR models; Topological descriptors.

1. Introduction

Insect growth regulators (IGR) stopping larvae development and inducing lethal processes during moulting are efficient tools in insect control for crop protection. As stressed by Nakagawa *et al.* [1] since their introduction in the mid 1980's, diacyhydrazines (DAH) , and among them dibenzoylhydrazines, of general formula X_a-$(C_6H_5)_a$-CO-N(*t*-Bu)-NH-CO-$(C_6H_5)_b$-Y_b, received an increasing interest as larvicides, owing to their easy synthesis at affordable cost, high efficiency and specificity against lepidopteran and coleopteran pests. These molecules act as moulting accelerating compounds, activating the ecdysone receptor, part of the steroidal 20-hydroxyecdysone moulting hormone receptor. This hormone regulates moulting and metamorphosis. An external high dose of moulting hormone agonists maintains a premature abnormal moulting process, rapidly leading larvae to death [1-3]. Additionally this hormone receptor is not present in mammals, making the ecdysone receptor an interesting target for larvicide development.

For limited homogeneous series of chemicals (with a fixed substitution pattern on one of the phenyl rings), several QSAR models, linking the larvicidal activity of acylhydrazines to structural characteristics, have been proposed from Hansch-Fujita [1,4,5] multilinear models or in the framework of Free Energy Relationships, involving "classical" electronic and steric substituent constants (Hammett's sigma and Verloop's constants) [5]. However, strong interactions in di- or tri-substituted phenyl ring A bearing an ortho group, required the presence of "ad hoc" indicator variables. On the other hand, Partial Least Squares analyses of electronic and steric field contributions on grid points surrounding the substrates (CoMFA treatment) were also carried out on diverse species of insects [1, 3, 6].

In this framework, extended analysis was carried out by Wheelock *et al.* [3] for a large population of 172 compounds where both phenyl rings may be simultaneously substituted, and the central moiety (CO)-N-(*t*-Bu)-NH-(CO) differently substituted. The used CoMFA analysis led to satisfactory performance in data fitting (recall), and gave some clues about the more important intervening interactions with the corresponding receptor. But the determination coefficient Q^2 in leave-one-out cross validation was poor ($Q^2 = 0.447$ on 158 compounds), and the approach involved a huge number of variables before reduction by Partial Least Squares. External validation of the results (prediction tests) was not considered. Furthermore calculated values for compounds experimentally observed as "inactive" were much higher than the activity threshold retained. In a more recent publication [2] Crisan *et al.* also examined a limited population of 33 compounds, in the framework of the MLR models from QSARINS [7], and pharmacophore search.

We present here various 2D topology-based QSAR models linking the pEC_{50} values (co-logarithm concentration of half maximum response) to molecular structure. Topology-based models have been widely used with satisfactory performances in developing SAR and QSAR models, and appeared very efficient for prediction of new or unknown chemicals [8-12]. Although physical interpretation of the selected 2D structural parameters is often more difficult than for quantum-derived descriptors, this approach gets rid of energy minimisation processes, structure alignment and (often heavy) MO calculations, and requires only swift evaluation of structure-invariant descriptors, available from the knowledge of only the structural formulae [12]. It may be hoped that working on few, easily evaluated descriptors rather than on a huge number of field values on nodes surrounding the aligned molecules, would lead to flexible models, more easily applicable to the study of new potentially active chemicals.

2. Materials and methods

2.1 Data set

In the present work, experimental data, for a population of 126 dibenzoylhydrazines of general formula

X_a-$(C_6H_5)_a$-CO-N-(*t*-Bu)-NH-CO-$(C_6H_5)_b$-X_b,

were retrieved from the previous CoMFA study carried out by Wheelock *et al.* [3]. pEC$_{50}$ values, correspond to the co-logarithm of the minimal concentration for obtaining an effect on 50% of the cells), determined in a *Bombyx mori* cell-based, high-throughput screening via a reporter gene assay. These pEC$_{50}$ values (expressed in M) cover a range 8.91-4.33.

From the initial population of 133 compounds, 7 chemicals from this original study were discarded, owing to observed inactivity (pEC$_{50}$ < 4.00). However *(vide infra)* they may give some clues about the quality of the models we built (by comparison with the calculated activity values in a somewhat “rough test set”).

In the present work, the 126 compounds with precise activity values (8.91-4.33) were first ordered by decreasing activity (regardless of structural similarity) and identified by an ID number (1…126). Original Wheelock “names” W… were also indicated for easier retrieval. Structural formulae and activities (pEC$_{50}$) are gathered in **Table 1**.

Table 1. Activity of the studied dibenzoylhydrazines (pEC$_{50}$ M): X_a-$(C_6H_5)_a$-CO-N (*t*-Bu)-NH-CO-$(C_6H_5)_b$-X_b

NAME	ID	Act-Exp	PHI-A	PHI-B	NAME	ID	Act-Exp	PHI-A	PHI-B
W001	75	6.37	H	H	W068	93	6.02	2-Cl	3-CN
W002	68	6.51	2-F	H	W069	82	6.22	2-Cl	3-Me
W003	39	7.12	2-Cl	H	W070	31	7.28	2-Cl	3-OMe
W004	55	6.77	2-Br	H	W071	73	6.41	2-Cl	4-F
W005	45	6.98	2-I	H	W072	48	6.94	2-Cl	4-Cl
W006	85	6.20	2-CF_3	H	W073	25	7.47	2-Cl	4-Br
W007	62	6.65	2-NO_2	H	W074	9	8.12	2-Cl	4-I
W008	63	6.65	2-Me	H	W075	14	8.00	2-Cl	4-CF_3
W009	81	6.23	2-Et	H	W076	107	5.46	2-Cl	4-NO_2
W010	127	< 4.00	2-Phi	H	W077	83	6.22	2-Cl	4-CN
W011	79	6.28	2-OMe	H	W078	18	7.73	2-Cl	4-Me
W012	128	< 4.00	2-O-*s*-Bu	H	W079	7	8.24	2-Cl	4-Et
W013	129	< 4.00	2-OCH_2Ph	H	W080	10	8.04	2-Cl	4-*n*-Pr
W014	91	6.07	2-SMe	H	W081	22	7.61	2-Cl	4-*i*-Pr
W015	60	6.68	3-F	H	W082	132	< 4.00	2-Cl	4-Ph
W016	47	6.94	3-Cl	H	W083	29	7.29	2-Cl	4-OMe
W017	78	6.29	3-Br	H	W084	97	5.89	2-Cl	4-SO_2Me
W018	88	6.15	3-I	H	W085	46	6.98	2-Cl	4-COMe
W019	102	5.74	3-CF_3	H	W086	12	8.03	2-Cl	2,3-Cl_2
W020	114	5.13	3-NO_2	H	W087	20	7.70	2-Cl	2,3-Me_2
W021	99	5.82	3-CN	H	W088	17	7.81	2-Cl	2-Me,3-OMe
W022	41	7.07	3-Me	H	W089	65	6.65	2-Cl	2,4-Cl_2
W023	64	6.65	4-F	H	W090	53	6.79	2-Cl	2,4-Me_2
W024	54	6.78	4-Cl	H	W091	104	5.69	2-Cl	2,5-Cl_2
W025	90	6.10	4-Br	H	W092	100	5.79	2-Cl	2,5-Me_2
W026	117	4.88	4-I	H	W093	24	7.53	2-Cl	2,6-F_2
W027	118	4.83	4-CF_3	H	W094	56	6.77	2-Cl	2-F,6-Cl
W028	120	4.69	4-NO_2	H	W095	113	5.17	2-Cl	2,6-Cl_2
W029	121	4.67	4-CN	H	W096	76	6.34	2-Cl	3,4-Cl_2
W030	108	5.45	4-Me	H	W097	42	7.06	2-Cl	3,4-Me_2
W031	130	< 4.00	4-t-Bu	H	W098	72	6.42	2-Cl	3,5-Cl_2
W032	131	< 4.00	4-Phi	H	W099	115	5.07	2-Cl	3,5-Me_2
W033	119	4.72	4-OMe	H	W100	133	< 4.00	2-Cl	3,5-(O-*n*-Bu)$_2$
W034	124	4.36	4-O-$(CH_2)_3$-Ph	H	W101	44	6.99	3,5-Me_2	2-Me

W035	116	5.04	2,3-Cl$_2$	H	W102	23	7.55	3,5-Me$_2$	3-Me
W036	98	5.84	2-Me,3-Cl	H	W103	71	6.47	3,5-Me$_2$	3-OH
W037	95	5.89	2,3-Me$_2$	H	W104	34	7.18	3,5-Me$_2$	3-OMe
W038	49	6.93	2,4-Cl$_2$	H	W105	77	6.32	3,5-Me$_2$	3-OEt
W039	80	6.24	2,4-Me$_2$	H	W106	15	7.92	3,5-Me$_2$	4-Me
W040	106	5.48	2,5-Me$_2$	H	W107	1	8.91	3,5-Me$_2$	4-Et
W041	94	6.00	2-OMe,5-*n*-Pr	H	W108	16	7.88	3,5-Me$_2$	4-*n*-Pr
W042	110	5.36	2,6-F$_2$	H	W109	2	8.87	3,5-Me$_2$	4-*i*-Pr
W043	105	5.59	2-F,6-Cl	H	W110	57	6.75	3,5-Me$_2$	4-*n*-Bu
W044	89	6.14	3,4-Me$_2$	H	W111	4	8.61	3,5-Me$_2$	4-*t*-Bu
W045	126	4.33	3,4-OMe$_2$	H	W112	30	7.29	3,5-Me$_2$	4-*n*-Pentyl
W046	122	4.64	2,3,4-Cl$_3$	H	W113	19	7.72	3,5-Me$_2$	4-Cl
W047	35	7.16	3,5-Me$_2$	H	W114	5	8.43	3,5-Me$_2$	4-CF$_3$
W048	109	5.40	2,5-Cl$_2$,3-CF$_3$	H	W115	3	8.62	3,5-Me$_2$	2,3-Me$_2$
W049	111	5.34	2-OMe,3,5-Me$_2$	H	W116	50	6.89	3,5-Me$_2$	2-Me,3-OH
W050	61	6.67	2,3,4,5,-F$_4$	H	W117	8	8.22	3,5-Me$_2$	2-Me,3-OMe
W051	125	4.34	2,3,4,5,6,-F$_5$	H	W118	32	7.28	3,5-Me$_2$	2-Me,3-OEt
W052	28	7.29	2-Cl	2-F	W119	33	7.26	3,5-Me$_2$	2,3,4-F$_3$
W053	70	6.48	2-Cl	2-Cl	W120	27	7.31	3,5-Me$_2$	2,4,5-F$_3$
W054	69	6.50	2-Cl	2-Br	W121	13	8.01	3,5-Cl$_2$	4-Me
W055	96	5.89	2-Cl	2-I	W122	37	7.16	3,5-Cl$_2$	4-C l
W056	101	5.76	2-Cl	2-CF$_3$	W123	51	6.88	3,5-Br$_2$	2-Me
W057	84	6.21	2-Cl	2-NO$_2$	W124	21	7.68	3,5-Br$_2$	4-Me
W058	38	7.14	2-Cl	2-Me	W125	11	8.04	3,5-Br$_2$	4-Et
W059	123	4.58	2-Cl	2-Phi	W126	59	6.72	3,5-Br$_2$	4-NO$_2$
W060	112	5.21	2-Cl	2-OMe	W127	67	6.55	2-Me	2-Me
W061	87	6.17	2-Cl	2-SMe	W128	74	6.39	4-Cl	4-Cl
W062	36	7.16	2-Cl	3-F	W129	26	7.38	4-Cl	4-Me
W063	52	6.81	2-Cl	3-Cl	W130	66	6.64	4-Et	4-Et
W064	58	6.72	2-Cl	3-Br	W131	40	7.08	2,6-F$_2$	4-Cl
W065	86	6.20	2-Cl	3-I	W132	6	8.31	H	4-Et
W066	92	6.03	2-Cl	3-CF$_3$	W133	43	7.06	H	4-Cl
W067	103	5.70	2-Cl	3-NO$_2$					

2.2 Descriptor generation and model selection

2D topological-type structural descriptors were generated from the software PaDEL [13] leading to an initial pool of about 1200 values for compound, to be introduced in the QSARINS software [7]. These descriptors encompass nature of the atoms, autocorrelation vectors and elements of adjacency or distance matrices, E-states, etc. Elimination of descriptors with (nearly) constant values and pruning pairs of highly inter-correlated values (correlation coefficient higher than 0.9) led to a reduced set of 168 (potentially significant) variables. Subsequent selection was carried out in QSARINS by using OLS-MLR coupled with a genetic algorithm (GA) based procedure [14].

This was carried out in an external validation step [15, 16]. For allowing, in parallel, complementary internal validations, five subsets (m = 0 to m = 4) where created with different compositions of training and prediction sets: In subset m, prediction set is composed of chemicals with (ID modulo 5) = m (that is rest of the division of ID by 5 equals m). In other words, one every five compounds is placed in prediction. For example for subset 2 prediction will be carried out on compounds ID = 2, 7, 12…122, (Note that, in parallel, the full data set is accessible with m = 5).This procedure we previously used in several applications [17-19], allows for a rather regular splitting all along the reactivity range, irrespective of structural similarities for both training and prediction samples.

From the 168 initially retained variables (structural descriptors), further selection was carried out on subset m = 1. This variable selection procedure was performed in a two steps process embedded in the software QSARINS. In a first step, all the possible triples of descriptors were explored using an exhaustive selection procedure (“All Subsets”). In a second step, the pool of the best 100 models generated by All Subsets was then extended via the GA to explore models with higher complexity, in order to find the model with the best Q^{2}loo (and RMSE), and satisfying the QUICK Rule [16, 20].

Set up values for GA selection were: number of generations per size = 1000, population size = 100 models, mutation rate = 50%, and 100 models were saved for each model dimension (i. e. number of variables included). In addition, to avoid overfitting and inclusion of inefficient variables, uselessly complicating the models, or leading to insignificant supplementary benefit, we limited the increase of the model complexity up to 9 variables. This led to a ratio (nb compounds/nb descriptors) largely higher than the currently admitted threshold of five according to the OECD guidance [21].

For subset m = 1, chosen for selection driving, the quality (and robustness) of the model was determined by the Q^2loo coefficient on the training part, and its predictive ability, examined on the prediction part. This corresponds to true "external validation" since the predicted chemicals were never involved in the development of the GA-based population of models. This allowed us to define a restricted pool of 9 descriptors (to be detailed below (**&2.3**) that will be the basis of the different proposed models (MLR and machine learning approaches). Beside the robustness and precision in data fitting, and preceding external validation process, quality of the MLR models built on these descriptors was also examined at several levels of internal validation:

1) on the remaining subsets (m = 0,2,3,4) in recall, and in leave-one (or some)-out on the training part,

2) in prediction on the corresponding part for each subset,

3) on recall, loo and prediction on randomized samples.

This was carried out first with OLS-MLR models and extended with the same descriptor set to various Machine Learning approaches (at least for the five created subsets).

2.3 Selected descriptors

The nine selected structural descriptors are respectively:

1) GATS1c, GATS5e, GATS8s

Geary autocorrelation terms (lag 1, 5, or 8) weighted by charge, Sanderson electronegativity or I-state. They are calculated on centered property values (w_i), but weighted by the square of the centered property value on all atoms minus one (So, mean and standard deviations are accounted for [22])

$$Ck = (1/2\Delta k)\sum_i^A \sum_j^A (w_i\text{-}w_j)^2\ \delta_{ij}\ /(1/A\text{-}1)\ \sum_i^A (w_i\text{-}\hat{w})^2.$$

with A number of atoms, Δk number of atom pairs at (topological) distance k. δ_{ij} =1 for a topological distance k between atoms i and j equal to the lag, zero otherwise.

2) SM1-Dzm

The Barysz distance matrix is defined as a weighted distance matrix (from the Hydrogen depleted graph) that simultaneously accounts the presence of multiple bonds and heteroatoms in the chemicals [23]. SM1-Dzm is the spectral moment of order 1 from Barysz matrix, weighted by mass.

3) VE3_Dzp

Logarithmic coefficient of Randic-like sum of the last eigenvector (absolute values) from Barysz matrix weighted by polarizabilities.

4) SpMax2_Bhp

Largest absolute eigenvalue of Burden modified matrix- n2, weighted by relative polarizabilities [24].

5) SHBint6

Sum of E-State products of strength for potential internal hydrogen bonds of path length 6. Electro-topological state (E-State) belongs on Roy topological indices based on the valence electron mobile (VEM) count [25-27].

6) MPC8

Molecular path count of order 8.

7) TopoShape

Petitjean topological shape index [28], relying on notion of generalized radius and diameter.

The first trials on the full set (126 chemicals) with the selected 9 descriptors lead to a rather "acceptable" MLR model: for subset m = 1, R^2 = 0.711, Q^2 = 0.657 and Q^2pred = 0.544, but strong residuals (about 1.1 to 2.1 log unit) were observed for compounds ID = 6, 12, 61. So, we decided to discard these compounds and work on 123 chemicals. We checked that the selected 9 descriptors can be satisfactorily applied to this (slightly), reduced data set.

3. Results and discussion

Selected descriptors relevance will be first examined in a Multilinear Regression (by ordinary least squares) approach. They will be then applied to various Machine Learning methods using different representations of the descriptor space.

3.1 OLS-MLR model

Multilinear regression performance is now characterized at different levels: data fitting ("recall"), cross validation, prediction. Particular attention will be devoted to robustness and accuracy.

In OLS-MLR model (ordinary least-squares multi-linear relationship), the relation between a (univariate) dependent variable y (activity here) and several independent variables xi (structural descriptors) is expressed as:

$$\mathbf{y} = \mathbf{X}\ \mathbf{b} + \mathbf{e}$$

where **X** is the matrix of the independent variables xi, **b** and **e** being the column vectors of coefficients and residuals respectively. Minimizing the residuals by OLS method, it comes:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$$

and the calculated response $\hat{\mathbf{y}}$ is:

$$\hat{\mathbf{y}} = \mathbf{X}\,\mathbf{b}$$

3.1.1 Data fitting and external validation (Subset m = 1)

For subset m = 1, after discarding deviating compounds 6, 12, 61, the selected 9 descriptors led to:

$$\begin{aligned} pEC_{50} = & - 169.8789 + 3.0705\ GATS1c - 1.6719\ GATS5e + 1.1188\ GATS8s - 0.4646\ SM1\_Dzm \\ & + 0.2347\ VE3\_Dzp + 46.9041\ SpMax2\_Bhp - 0.0742\ SHBint6 - 0.0939\ MPC8 \\ & + 2.3559\ topoShape \end{aligned} \tag{1}$$

with, in data fitting (recall on training set): N = 99, $R^2 = 0.749$, RMSE = 0.51, MAE = 0.41 and s = 0.53. In leave-one-out cross validation: $R^2 = 0.699$, RMSE = 0.56, MAE = 0.45 and $Q^2 = 0.697$. And, for the prediction set, $R^2 = 0.736$, RMSE = 0.59, MAE = 0.47 and Q^2-F2 = 0.706.

Graphs for recall and prediction (on subset 1) are reproduced in **Figure 1**.

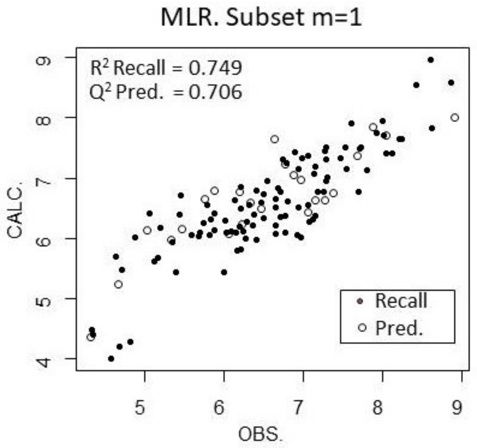

Figure 1. MLR correlations between observed and calculated (eq. 1) pEC_{50} values. * Circles correspond to test compounds (subset m = 1), disks to training compounds ("recall", aka "data fitting").

From the standardized coefficients, the decreasing sequence of descriptor relative importance is: SpMax2_Bhp, MPC8, GATS5e, VE3_Dzp, GATS1c, SM1_Dzm, GATS8s, SHBint6, TopoShape.

An important aspect of MLR treatment is the "applicability domain" [7]. It characterizes "influential" objects: those that in training have a heavy importance in the definition of the model, and in prediction, points falling outside this AD, that must be considered with caution. In the leverage approach, the influence of each object on the regression result (its "leverage") is given by the corresponding diagonal element h of the "Hat" matrix **H**:

$$\mathbf{H = (X^TX)^{-1}\,X^T}$$

where **X** represents the matrix of the descriptors characterizing the samples.

For a study involving n training samples and k variables, objects with h larger than the threshold value h* = 3(k+1)/n are considered outside the AD. Williams' plot (standardized residuals *vs* Hat diagonal values h) immediately highlights points outside the AD and outliers with residuals larger than 2.5 times the standard deviation (the common norm). Six points fall outside the Applicability Domain **(**See Williams'plot**, Figure 2)**. Compounds ID #85 and 92 (h about 0.46) as compounds 109, 125 (h about 0.31) are well calculated, whereas ID #123 and 118 (h about 0.40) show deviations about 1.5 std residual. Note that all these points belong to the training set.

To more firmly establish the robustness and quality of the proposed model (and the corresponding set of descriptors) several confirmations were examined.

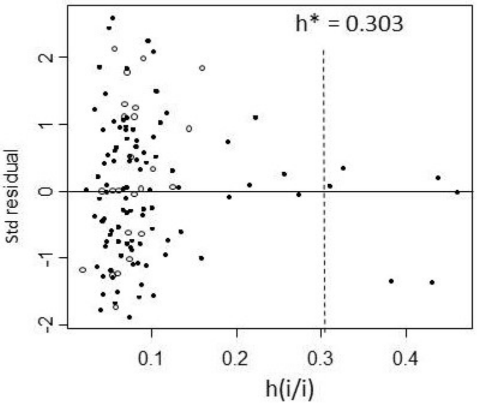

Figure 2. Williams ‘ plot for subset m=1. * Circles correspond to test compounds (subset m=1), disks to training compounds (“recall”, aka “data fitting”)

3.1.2 Pluri internal validation processes

First, MLRs were built with the same structural descriptors for the other subsets previously defined (m = 0, 2, 3, 4) and for the full set (corresponding to m = 5). Results, gathered in **Table 2** (training tr, loo cross validation lo and test te, respectively) show that satisfactory correlations of comparable quality are obtained in these various trials. Remark that in these processes, when elaborating the corresponding MLRs (with the relevant descriptors selected on subset m = 1), the current training set (used to build the corresponding MLR) encompasses only 66% of the chemicals involved for the selection of the retained 9 descriptors ((in subset m = 1). For easier comparison, the results obtained for m = 1 are repeated in the table.

On another hand, in these treatments, each compound is considered four times in training and once in prediction. To have a global estimate of the predictive ability, we also gathered, in a single file, the predicted values obtained in each subset (m = 0 - 4), and calculated the usual statistical parameters for the correlation of these grouped values with the experimental ones. Results are reported in the last line in **Table 2**.

Table 2. Statistical elements for MLRs built on the 6 subsets m = 0-5

m	R^2tr	RMSEtr	MAEtr	R^2lo	RMSElo	MAElo	Q^2lo	R^2te	RMSEte	MAEte	Q^2te
0	0.765	0.50	0.41	0.707	0.56	0.46	0.703	0.646	0.62	0.52	0.621
1	0.749	0.51	0.41	0.699	0.56	0.45	0.697	0.736	0.59	0.47	0.706
2	0.767	0.50	0.40	0.712	0.56	0.45	0.708	0.645	0.62	0.52	0.610
3	0.737	0.53	0.43	0.690	0.58	0.47	0.688	0.750	0.54	0.42	0.712
4	0.736	0.53	0.43	0.666	0.60	0.48	0.658	0.760	0.55	0.47	0.716
5	0.744	0.52	0.42	0.701	0.56	0.46	0.699	NA	NA	NA	NA
								0.681	0.58	0.48	0.675

3.1.3 Randomized leave-many-out cross validation

To confirm the choice of the selected descriptors, we carried out 2000 runs of cross validation with 20% data left out. For a more homogeneous sampling, we randomly selected, in the ID-ordered list of 123 compounds, 13 compounds in the more reactive half of the population and 13 in the less reactive part, to build the corresponding validation group. The histograms of R^2 (fitting), R^2 and Q^2 (validation aka prediction) are given in **Figure 3**. The obtained values (0.749, 0.717 and 0.689 respectively) confirm the validity of the selected descriptor set.

Similarly, we also verified that randomly shuffling activity values (2000 runs) led to very low correlation coefficients. ($Q^2 = 0.137$ for the m = 1 subset, for example). Consistency of these results prompts us to consider that the nine selected structural variables led to satisfactory fitting and prediction for the various populations studied. Presumably, this choice would not always be the optimal one when looking independently at each splitting. But we consider it gives a unique set of structural variables, actually applicable to the various subsets and that can also be used for the other correlation methods we proposed.

3.1.4 “Inactive compounds”

In the initial data set, seven compounds were discarded since “inactive” ($pEC_{50} < 4$). It may be interesting to examine at what level our MLR model calculates their pEC_{50}. **Table 3** gathers the evaluations from CoMFA values (from Wheelock *et al* study [3]) and our model, **eq. (1)**, established from subset m = 1, but nearly identical results would be obtained considering MLR from the whole set of 123 chemicals (m = 5).

Although these two series of values are not obtained in identical conditions (158 compounds in CoMFA, 123 in our work) it is noteworthy than CoMFA predictions are largely overestimated, whereas our proposed values are more in agreement with the observed inactivity, at least for six compounds out of seven. Of course, external factors (other than the balance of structural influences taken into account by the selected descriptors): experimental uncertainty, activity cliff, or change in mechanism, may intervene in the experimentally observed activity. However, our results might be considered as a good “rough external validation” of our model.

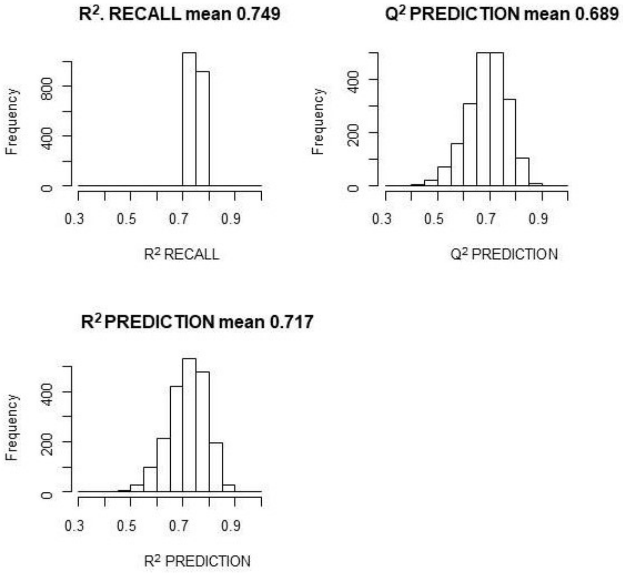

Figure 3. Histogram of R^2 and Q^2 results for 2000 random runs on the full set (123 comp.)

Table 3. Predictions (pEC_{50}) for “inactive” compounds by CoMFA [3] and our MLR model (this work)

COMPOUNDS		CoMFA	MLR
W10	ID127	5.12	3.30
W12	ID128	6.16	4.23
W13	ID129	6.62	2.56
W31	ID130	5.25	6.62
W32	ID131	5.23	3.63
W82	ID132	7.24	4.72
W100	ID133	6.04	3.19

3.2 Machine Learning approaches

Although MLR is the most widely used approach in QSAR/QSPR studies, due to its efficient and straightforward implementation, Machine Learning methods [10] are increasingly used in the field, particularly with recent application to property modelling of nanoparticles [17-19, 29-40]. The approaches here used: Support Vector Machine (SVM), Projection Pursuit Regression (PPR), Partial least Squares (PLS), Three-Layer perceptron (TLP-ANN), a Back Propagation Artificial Neural Network , have been previously largely presented in various publications.[29-40] More details . can be found in specific literature [10, 41-45]. Such approaches usually do not propose any explicit, directly usable, formula for property prediction. However they offer easy settings, rapid training and generally guarantee to find the global minimum on the error surface. Furthermore the introduction of non-linear methods may give more flexibility to adjust the model to experimental observations.

On the other hand, some of the methods here employed, work on projections of data into structural spaces of varied dimensionality: enlarged with SVM, or reduced with PPR, or on transformed data (PLS). So it may be interesting to see whether these transformations may affect the encoding of structural information embedded in the descriptors and emphasize or deaden some of their specific characteristics. It will be also interesting to examine whether they could be used as alternatives, equivalent to MLR, or overwhelm it. In several publications indeed, machine learning approaches have proved their ability to cope with nonlinear responses, and proposed improved results as compared to MLR correlation models [10, 29, 31, 32]. Suffices it here to recall some basic elements of these approaches.

3.2.1 Basics on used Machine Learning approaches

1) Partial Least Squares

PLS [43] operates not on original variables but on components (latent variables) built from them, so as to represent (at best) simultaneously the variability of both the structural descriptors and the response (pEC_{50}). This is an important difference with Principal Component Regression where the principal components optimize only the variable matrix. One advantage of PLS (not important here, however) is that it can be used when variables are numerous, highly collinear and even more numerous than samples (as for example in CoMFA analyses).

2) Projection Pursuit Regression

Developed by Friedman and Stueltzle [44] this non-parametric method relies on an (empirically determined) sum of nonlinear local smooth (univariate) ridge functions introduced iteratively. Schematically, given a trial direction vector **a**, the descriptor matrix **X** (k variables*n samples) is projected as:

$$\mathbf{Z}= \mathbf{a}^{T}\,\mathbf{X}$$

The model operates on these **Z** projections (linear combinations of the initial variables*)* and approximates the regression function (linking the property **y** to associated predictors **X**) by a finite sum (empirically determined) of smooth ridge functions of the new predictor variables **Z**. As there are infinitely many possible projections from higher dimension to lower dimension spaces, it is important to have an optimization technique to pursue a sequence of projections that can reveal the most interesting structure in the data set. Once the smoothing function selected, the (tuneable) number of projections is automatically determined by optimizing cross-validation results (Q^2loo).

3) Support Vector Machine

First proposed by Vapnik [41] for classifications, SVM was soon extended to correlations, and is now largely used in QSAR studies, including recent applications to nanoparticles [10, 30-39, 42] SVM is rooted to two key ideas. First, robustness of the model is privileged over good performance in data fitting (“recall”). Second, (using a kernel function) the data are projected in a higher dimensionality space where it may be hoped that a simpler, linear representation is possible. Parameters to be set are:

a) Regularisation constant C; that controls the balance between precision and complexity of the model (too small a value gives limited importance to data fitting. A too large value complicates the model and may cause overfitting).

b) Epsilon insensitive loss, that defines the diameter of the “error tube along the regression line” where deviations between observed and calculated values are ignored when building the model.

c) Parameters of the kernel function (if necessary). As to the two most largely used kernel function, the linear one (here used) k = xx’ (where x,x’ are independent variables) does not require any parameter, whereas the Gaussian kernel ($k = \exp(-(x-x')/\sigma^2)$) depends on the “width” σ.

4) Three Layer Perceptron (Back Propagation Artificial Neural Network)

The first layer is fed with the structural descriptors of the investigated pattern. This information (scaled by the weights of the connexions input-hidden layers) is send to the units of the “hidden layer”. On each of these units, these inputs are summed up, and transmitted, thanks to a transfer function (for example, a logistic one) to the output unit where their summation delivers the calculated property value. Connection weights (between units of the successive layers) are iteratively optimised from a training set using a “back-propagation algorithm”, operating from the output layer to the input layer [45]. In the present application, only one hidden unit was introduced.

3.2.2 Results from Machine Learning approaches

In the present study, we used for machine learning approaches, the nine descriptors selected by MLR. Diverse other **se**lection routines have been proposed, particularly in the framework of the caret package [46], but they often relied on classification problem and operate by backward selection. Clearly, using here the descriptors selected by MLR and in view of their good results, a drastic improvement of performance would not be expected. However, it was interesting to verify whether the changes in the structural space, induced by machine learning approaches, may be beneficial.

Calculations were carried out in the framework of the Cran-R project, using the caret package [46, 47] or home-made combinations of available R routines. Parameter adjustments involve the number of latent variables (PLS), number of projections (PPR), number of hidden units (TLP-ANN), regularisation parameter and noise (epsilon) in SVM. In this case, although some programs have been proposed for this setting [48, 49] we prefer using a grid-type method: for 7 possible values of epsilon (0.10-default to 0.50), several values of the regularisation parameter C (1,2,4… 32) are examined and our program automatically determines the best choice. For PPR one projection was selected for all subsets. Similarly for ANN the hidden layer encompasses only one neuron (corresponding to a 9-1-1 architecture of the network) and 100 iterations were carried out. For PLS, a high number of latent variables was necessary (7 to 9), in agreement with the observation that the selected variables are not deeply correlated, and (not surprisingly) results are nearly identical to those obtained in MLR.

For the diverse approaches here used, results are presented in **Table 4**.for the different subsets, in training, tr, loo cross validation lo and prediction te. Line m = 5 corresponds to the results obtained in training and loo for the full population. The last line indicates the correlation observed between experimental pEC50 and the prediction gathered in the five subsets m = 0- to m = 4. Results for MLR (**Table 2**) are repeated for easier comparison.

Table 4. Machine learning approaches.

m	R^2tr	RMSEtr	MAEtr	R^2lo	RMSElo	MAElo	Q^2lo	R^2te	RMSEte	MAEte	Q^2te
MLR											
0	0.765	0.5	0.41	0.707	0.56	0.46	0.703	0.646	0.62	0.52	0.621
1	0.749	0.51	0.41	0.699	0.56	0.45	0.697	0.736	0.59	0.47	0.706
2	0.767	0.5	0.4	0.712	0.56	0.45	0.708	0.645	0.62	0.52	0.61
3	0.737	0.53	0.43	0.69	0.58	0.47	0.688	0.75	0.54	0.42	0.712
4	0.736	0.53	0.43	0.666	0.6	0.48	0.658	0.76	0.55	0.47	0.716
5	0.744	0.52	0.42	0.701	0.56	0.46	0.699	NA	NA	NA	NA
								0.681	0.58	0.48	0.675
PLS											
0	0.765	0.5	0.41	0.693	0.57	0.47	0.688	0.646	0.62	0.52	0.622
1	0.749	0.51	0.41	0.698	0.56	0.45	0.695	0.737	0.58	0.47	0.706
2	0.767	0.5	0.4	0.71	0.56	0.45	0.705	0.645	0.62	0.52	0.610
3	0.737	0.53	0.43	0.689	0.58	0.47	0.688	0.751	0.54	0.42	0.714
4	0.736	0.53	0.43	0.661	0.61	0.49	0.652	0.761	0.55	0.47	0.717
5	0.744	0.52	0.42	0.701	0.56	0.46	0.699	NA	NA	NA	NA
								0.682	0.58	0.48	0.670
PPR											
0	0.802	0.46	0.37	0.719	0.55	0.46	0.715	0.661	0.61	0.5	0.639
1	0.763	0.49	0.39	0.705	0.55	0.45	0.703	0.764	0.56	0.45	0.733
2	0.805	0.46	0.35	0.71	0.56	0.45	0.705	0.643	0.61	0.5	0.632
3	0.841	0.41	0.33	0.732	0.54	0.43	0.73	0.648	0.63	0.53	0.604
4	0.827	0.43	0.34	0.673	0.59	0.48	0.671	0.649	0.65	0.53	0.600
5	0.789	0.49	0.4	0.7	0.56	0.46	0.697	NA	NA	NA	NA
								0.652	0.61	0.5	0.644
SVM-LIN											
0	0.76	0.51	0.43	0.704	0.56	0.47	0.702	0.653	0.6	0.49	0.65
1	0.744	0.51	0.42	0.697	0.56	0.46	0.696	0.728	0.59	0.49	0.698
2	0.762	0.5	0.41	0.709	0.56	0.45	0.706	0.607	0.65	0.54	0.572
3	0.732	0.54	0.44	0.696	0.57	0.46	0.693	0.753	0.54	0.42	0.715
4	0.726	0.54	0.45	0.648	0.61	0.5	0.645	0.743	0.54	0.47	0.720
5	0.739	0.53	0.44	0.709	0.55	0.46	0.708	NA	NA	NA	NA
								0.675	0.58	0.48	0.673
TLP-ANN											
0	0.789	0.47	0.38	0.741	0.52	0.42	0.74	0.639	0.62	0.51	0.625
1	0.765	0.49	0.39	0.716	0.54	0.43	0.716	0.766	0.56	0.46	0.731
2	0.79	0.47	0.37	0.724	0.54	0.43	0.722	0.619	0.62	0.5	0.611
3	0.751	0.51	0.41	0.699	0.57	0.45	0.699	0.797	0.45	0.36	0.795
4	0.761	0.5	0.4	0.705	0.56	0.45	0.704	0.787	0.54	0.42	0.727
5	0.764	0.5	0.4	0.725	0.54	0.43	0.724	NA	NA	NA	NA
								0.703	0.56	0.45	0.700

From **Table 4**, it appears that, in the diverse investigated approaches, results for the different subsets (m = 0,,…,4) and the full population (m = 5), with a same correlation method, are highly consistent and led to similar statistical criteria. For example, looking at recall results, R^2 varied from 0.767 to 0.736 in MLR, and 0.751 to 0.790 for TLP. In fact, this observation was not unexpected since a common set of descriptors was used, Comparing now the different approaches, although the descriptor space was differently treated, MLR, PLS and Linear SVM, in recall, leave-one-out cross validation or prediction, gave nearly identical statistical parameters, whereas PPR (in data fitting) and TLP (for fitting, loo and prediction) are slightly superior. As additional remarks, it may be seen that R^2 and Q^2-F2 in prediction values were close. This comes from the fact that, since the prediction compounds are regularly split along the reactivity scale, the mean values for training and prediction sets are close.

In **Figure 4** are illustrated some examples of these machine learning approaches at different levels: gathered predictions for TLP, LOO for SVM, recall for PPR, compared to MLR recall results, here repeated for easier comparisons.

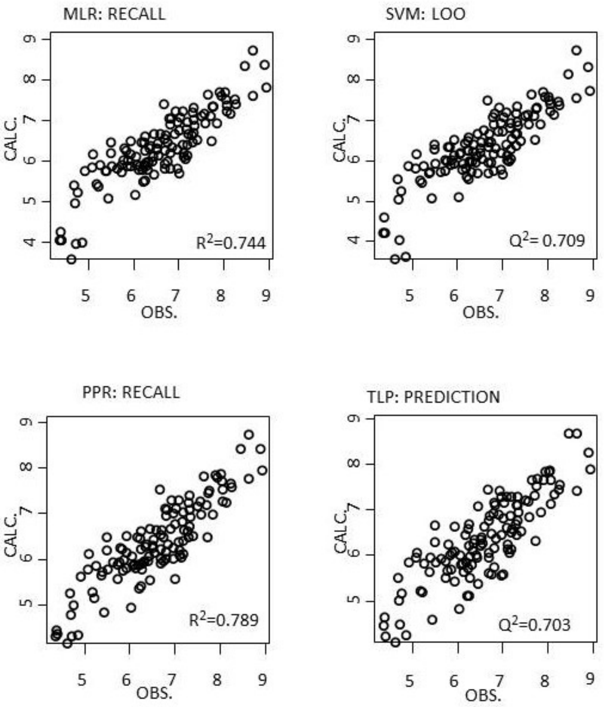

Figure 4. Correlations of observed pEC_{50} *vs.* calculated values by MLR and various machine learning approaches.

3.3 Structural information from selected descriptors

Importance of the various descriptors on the variations of calculated pEC_{50} values are exemplified in **Figure 5** where are indicated their contributions ((coefficients of MLR equation (1) times the descriptor value), according to ID ordering. Owing to the relatively large number of descriptors (9) intervening in the correlation models, and their topological character, it’s difficult to individually associate them to definite influences. Schematically, three variables are directly related to the “shape and volume”:

1) TopoShape: something like a generalized diameter /radius ratio
2) MPC8: count of atomic paths
3) SM1_Dzm: involving mass weighting

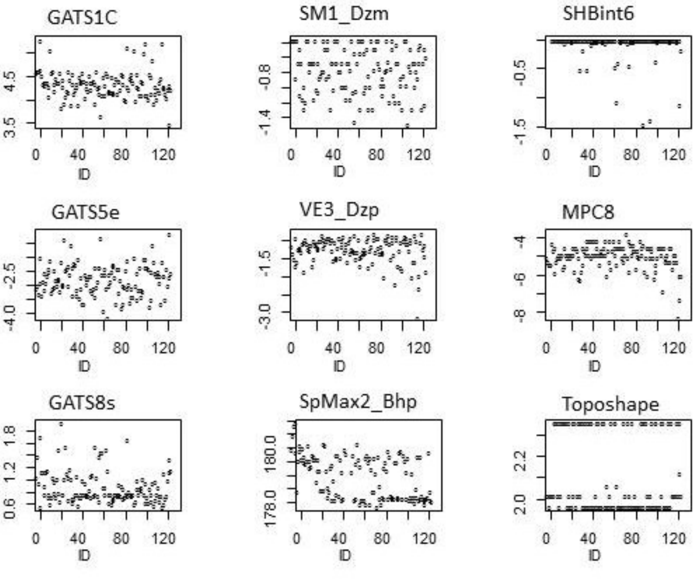

Figure 5. Individual contributions of the selected structural descriptor to pEC_{50}

But topological distance factors are also present in autocorrelation indices with lag 5 and 8, and in the constitution of Barysz and Burden matrices (SM1_Dzm, VE3_Dzp and SpMax2_Bhp). At last electronic characteristics (charge, electronegativity, polarizability, and E-states) also affect GATS1c, 5e and 8s, VE3_Dzp, SpMax2_Bhp and SHBint6. Note that the role of H-bond (at least between the carbonyls and NH group with the receptor) was cited by Nakagawa *et al*. [1] whereas consideration of logP did not improve the CoMFA model, as noted by Wheelock *et al.* [3]. Presumably the hydrophobicity influence of phenyl ring substituents (noted in previous publications) is more or less included (and borrowed) in unfavourable steric contributions.

An immediate remark is that for most of them, the variation ranges are of comparable extent, with a weight on the global pEC_{50} values about 1.5 to 4 log units, so that their introduction in the MLR model is mandatory. Only few variables are largely nearly invariant for most of compounds and take different values only for few isolated chemicals (SHBint6, and to a lesser degree GATS8s). At last TopoShape intervenes for only 0.4 log unit. However eliminating these (less efficient) variables is detrimental for the quality of the resulting model. So, for the whole set (123 compounds), eliminating one of these less important descriptors lowers R^2 and Q^2loo from 0.744, 0.699 to values about 0.705, 0.660 (elimination of SHBint6, TopoShape or GATS8s). Similarly elimination of two variables out of these three, led to values of R^2 and Q^2loo about 0.680 and 0.664.

4. Conclusion

In this paper, the activity of 123 dibenzoylhydrazines, acting as growth regulators was investigated through 2D, topology-based, QSAR models. In the framework of open source softwares R and PaDEL and the free application QSARINS, the approach relies on topology-based 2D characteristics. These descriptors, easily calculable, and structurally invariant, avoid the choice of an active conformation, subsequent energy minimization and (often heavy) quantum calculations, and are attainable with only the knowledge of the molecular graph.

After descriptor selection by MLR-Genetic Algorithm of QSARINS, MultiLinear Regression and several machine learning approaches (PLS, PPR, SVM, TLP-ANN) were investigated. They gave satisfactory results, highly consistent and robust, of comparable performance. Quality of the models was duly tested not only in data fitting, but also in cross validation, varied internal prediction steps or random splitting (two aspects not in depth investigated in a previous extended CoMFA study [3]).

5. References

[1] Nakagawa Y, Hormann RE, Smagghe G. SAR and QSAR studies for in vivo and in vitro activities of ecdysone agonists in Ecdysone: structure and functions. G. Smagghe Ed Springer Science and Business Media New York, 2009;Chap 20:475-509.

[2] Crisan L, Funar-Timofei S, Borota A. QSAR and ligand based pharmacophore models of dibenzoylhydrazines with insecticide activity against the silkworm *Bombyx Mori* L. Revue Roumaine de Chimie. 2007;62:(8-9):687-698.

[3] Wheelock CE, Nakagawa Y, Harada T, Oikawa N, Akamatsu M, Smagghe G, Stefanou D, Iatroue K, Swevers L. High-throughput screening of ecdysone agonists using a reporter gene assay followed by 3-D QSAR analysis of the molting hormonal activity. Bioorg Med Chem. 2006;14:1143-1159.

[4] Fujita T, Nakagawa Y. SAR and QSAR analyses of substituted dibenzoylhydrazines for their mode of action as ecdysone agonists in endocrine disruption modeling, Devillers J (Ed.), CRC Press, Boca Raton, FL. 2009;357-377.

[5] Fujita T, Nakagawa Y. QSAR and mode of action studies of insecticidal ecdysone agonists. SAR QSAR Environ Res. 2007;18:77-88.

[6] Dinan L, Hormann RE, Fujimoto T. An extensive ecdysteroid CoMFA. J Comput-Aided Mol Des. 1999; 13:185-207.

[7] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem. 2013;34:2121–2132.

[8] Dubois JE, Doucet JP, Panaye A. Behaviour similarity and local shapes. Topolgy reflects topography. Bull. Soc. Chim. Belges. 1989;98:31-44. (in French)

[9] Gozalbes R, Doucet JP, Derouin F. Application of topological descriptors in QSAR and drug-design: History and new trends. Current Drug Targets-Infectious Disorders. 2002;2:93-102.

[10] Doucet JP, Panaye A. Three-dimensional QSAR applications in pharmacology and toxicology. CRC Press, BocaRaton, FL. 2010.

[11] García-Domenech R, Aguliera J, El Moncef A, Pocovi S, Gálvez J. Application of molecular topology to the prediction of mosquito repellents of a group of terpenoid compounds. Mol Diversity. 2010;14:321–329.

[12] Dearden JC. Challenges and applications in computational chemistry and physics in Advances in QSAR Modelling. Ed. Roy K, Springer International Publishing.AG. 2017;24:57-88.

[13] Yap CW. PaDEL descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32:1466-1474.

[14] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemom. 1992;6:267-281.

[15] Gramatica P. Principle of QSAR models validation: Internal and external. QSAR Comb Sci. 2007;26:694-701.

[16] Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques. J Chemom. 2010;24:194-201.

[17] Doucet JP, Papa E, Doucet-Panaye A, Devillers J. QSAR models for predicting the toxicity of piperidine derivatives against *AEDES aegypti*. SAR QSAR Environ Res. 2017;28:451–470.

[18] Doucet JP, Doucet-Panaye A. Quantitative structure activity relationship for carboxamides and related compounds active on *Aedes aegypti* adult females. Vector Biology Journal. 2018;3:1.

[19] Doucet JP, Doucet-Panaye A, Papa E. Topological QSAR modelling of carboxamides repellent activity to *Aedes aegypti*. Mol Inf. 2019; DOI: 10.1002/minf.201900029.

[20] Todeschini R, Maiocchi A, Consonni V. The K correlation index: Theory development and its application in chemometrics. Chemom Intell Lab Sys. 1999;46:13–29.

[21] OECD. Guidance document on the validation of (Quantitative) Structure-Activity Relationship (Q)SAR Models. 2007;2:1-154.

[22] Puzyn T, Leszczynski J, Cronin MT. Recent advances in QSAR studies: methods and applications. Springer Science and Business media. 2010.

[23] Dehmer M, Varmuza K, Bonchev D. Statistical modelling of molecular descriptors in QSAR/QSPR. Wiley; Blackwell. 2012.

[24] Yali ZP, Fatemi MH. Novel consensus quantitative structure-retention relationship method in prediction of pesticides retention time in nano-LC. Nanochem Res. 2018;3:205-211.

[25] Hall LH, Kier LB. Electrotopological state indices for atom types: A novel combination of electronic topological and valence state information. J Chem Inf Comput Sci. 1995;35:1039-1045.

[26] Roy K, Ghosh G. QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. J Chem Inf Comput Sci. 2004;44:559-567.

[27] Roy K, Das RN. On some novel extended topochemical atom (ETA) parameter for effective encoding of chemical information and modelling of fundamental physicochemical properties. SAR QSAR Environ Res. 2011;22:451-472.

[28] Petitjean M. Applications of the radius diameter diagram to the classification of topological and geometrical shapes of chemical compounds. J Chem Inf Comput Sci. 1992;32:331-337.

[29] Ren S. Modeling the toxicity of aromatic compounds to Tetrahymena pyriformis. The response surface methodology with nonlinear methods. J Chem Inf Comput Sci. 2003;43:1679–1687.

[30] Golmohammadi H, Dashtbozorgi Z. QSPR studies for predicting polarity parameter of organic compounds in methanol using support vector machine and enhanced replacement method. SAR QSAR Environ Res. 2016;27:977-997.

[31] Yao XY, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis, function neural networks and multiple linear regression. J Chem Inf Comput Sci. 2004;44:257–266.

[32] Panaye A, Fan BT, Doucet JP, Yao XJ, Zhang RS, Liu MC, Hu ZD. Quantitative structure-toxicity relationships (QSTRs): A comparative study of various nonlinear methods: General regression neural network radial basis function neural network and support vector machine in predicting toxicity of nitro- and cyano-aromatics to *Tetrahymena pyriformis*. SAR QSAR Environ Res. 2006;17:75-91.

[33] Thissen U, Pepers M, Üstün B, Melssen WJ, Buydens LMC. Comparing support vector machine to PLS for spectral regression application. Chemom Intell Lab Syst. 2004;73:169–179.

[34] Tetko IV, Solovev VP, Antonov AV, Yao XJ, Doucet JP, Fan BT, Hoonakker F, Fourches D, Jost P, Lachiche N, Varnek A. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. J Chem Inf Model. 2006;46:808–819.

[35] Tanabe K, Kurita T, Nishida K, Lučić B, Amić D, Suzuki T. Improvement of carcinogenicity prediction performances based on sensitivity analysis in variable selection of SVM models. SAR QSAR Environ Res. 2013;24:565–580.

[36] Papa E, Doucet JP, Doucet-Panaye A. Linear and nonlinear modelling of the cyrotoxicity of TiO2 and ZnO nanoparticles by empirical descriptors. SAR QSAR Environ Res. 2015;26:647-665.

[37] Papa E, Doucet JP, Doucet-Panaye A. Computational approaches for the prediction of the selective uptake of magnetofluorescent nanoparticules into human cells. RSC Advances. 2016;6:68806-68818.

[38] Papa E, Doucet JP, Sangion A, Doucet-Panaye A. Investigation of the influence of protein corona composition on gold nanoparticle bioactivity using machine learning approaches. SAR QSAR Environ Res. 2016;27:521-538.

[39] Winkler DA, Burden FR, Weissleder BY, Tassa C, Shaw S, *et al.* Modelling and predicting the biological effects of nanomaterials. SAR QSAR Environ Res. 2014;25:161-172.
[40] Toropova AP, Toropov AA, Benfenati E, Puzyn T, Leszczynska D, *et al.* Optimal descriptor as a translator of eclectic information into the prediction of membrane damage: The case of a group of ZnO and TiO2 nanoparticles. Ecotoxicol Environ Safe. 2014;108:203-209.
[41] Cortes C, Vapnik V. Support vector networks. Mach Learn. 1995;20:273-297.
[42] Doucet JP, Barbault F, Xia HR, Panaye A, Fan BT. Nonlinear SVM approaches to QSPR/QSAR studies and Drug Design. Curr Comput-Aided Drug Des. 2007;3:263-289.
[43] Wold S, Sjöström M, Eriksson L. PLS regression: a basic tool of chemometrics. Chemom Intell. Lab. Systems. 2001;58:109-130.
[44] Friedman JH, Stuetzle W. Projection pursuit regression. J Am Stat Assoc. 1981;76:817–823.
[45] Devillers J. Neural networks in QSAR and drug design. Academic Press, London. 1996.
[46] Kuhn M. Building predictive models in R using the caret package. J Stat Soft. 2008;28:1-26.
[47] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.. Available at: http://www.R-project.org. 2014.
[48] Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab an S4 package for kernel methods in R. J Stats Soft. 2004;11:1-20.
[49] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Net. 2004;17:113-126.