

Quantile regression model for a diverse set of chemicals: application to acute toxicity for green algae

Jonathan Villain · Sylvain Lozano · Marie-Pierre Halm-Lemeille · Gilles Durrieu · Ronan Bureau

Received: 25 July 2014 / Accepted: 20 October 2014 / Published online: 29 November 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract The potential of quantile regression (QR) and quantile support vector machine regression (QSVMR) was analyzed for the definitions of quantitative structure-activity relationship (QSAR) models associated with a diverse set of chemicals toward a particular endpoint. This study focused on a specific sensitive endpoint (acute toxicity to algae) for which even a narcosis QSAR model is not actually clear. An initial dataset including more than 401 ecotoxicological data for one species of algae (*Selenastrum capricornutum*) was defined. This set corresponds to a large sample of chemicals ranging from classical organic chemicals to pesticides. From this original data set, the selection of the different subsets was made in terms of the notion of toxic ratio (TR), a parameter based on the ratio between predicted and experimental values. The robustness of QR and QSVMR to outliers was clearly observed, thus demonstrating that this approach represents a major interest for QSAR associated with a diverse set of chemicals. We focused particularly on descriptors related to molecular surface properties.

Keywords Algae species · Ecotoxicology · Molecular surface · Outliers · Quantile regression · Support vector machine

J. Villain · S. Lozano · M.-P. Halm-Lemeille · R. Bureau (✉)
Normandie University, Caen, France
e-mail: ronan.bureau@unicaen.fr

J. Villain · S. Lozano · M.-P. Halm-Lemeille · R. Bureau
UNICAEN, CERMN (Centre d'Etudes et de Recherche sur le
Médicament de Normandie) UPRES EA 4258-FR CNRS 3038
INC3M, Bd Becquerel, 14032 Caen, France

J. Villain · G. Durrieu
Laboratoire de Mathématiques de Bretagne Atlantique, Université de
Bretagne Sud et UMR CNRS 6205, Campus de Tohannic,
56017 Vannes, France

Introduction

Under REACH legislation [1], quantitative structure-activity relationship (QSAR) models are expected to be used as an alternative to save resources and to accelerate hazard and risk assessments. For algae, one of the three major endpoints in ecotoxicology, even QSAR models [2, 3] associated with a non-specific mode of action (MOA) are not clearly defined. The reason is explained by Aruoja et al. [4] and Netzeva et al. [5]. The issue comes from the lack of a consistent dataset with more than 100 values and the variability of algal test results due to the different methods and algae species used. Therefore, few non-polar narcotic QSAR models were defined for algae: one for *Selenastrum capricornutum* with only ten chemicals, one for *Chlorella vulgaris* with 34 chemicals, and one for green algae with 51 chemicals [6, 7]. For a global model, to our knowledge, only one study has been published (45 chemicals [8]) regarding the prediction of acute toxicity of chemicals to *Selenastrum capricornutum*. From these first QSAR studies, it appears that this endpoint is characterized by a particular sensitivity toward chemicals, and the presence of outliers affects the estimated models [9, 10]. An important characteristic of quantile regression (QR), compared to classical least squares regression, is its robustness to distribution assumptions and to outlying observations [11]. Experimental conditions associated with our study have integrated some measurement errors and systematic biases difficult to control (particularly true as soon as several MOA are related to our set). The use of QR should make the inference less biased and less sensitive to outliers. A major aspect associated with QSAR is the relationship between MOA and chemical derivatives [12, 13]. This assumption is relied upon when applying read-across analysis to data from REACH. Read-across information considers that the toxicity of a derivative could be estimated from the real toxicological data of a second derivative based on the chemical similarity between the two

structures and by assuming that they interact through the same MOA. In aquatic ecotoxicology, four MOAs are classically differentiated. Two are directly related to the relationship between one major descriptor ($\log K_{OW}$) and the biological activities, *i.e.*, baseline and polar narcosis mechanisms for the MOA. The other two correspond to chemical reactions with macromolecules (reactive functions related to native structures or metabolites) and to specific intermolecular interactions with macromolecules (modulation of biological pathways). The notion of toxic ratio (TR) represents one parameter to differentiate a non-specific (narcosis) from a specific MOA [13]. In this study, a large and diverse set of derivatives is considered with the integration of several major sources of ecotoxicological data. One was provided by the Japanese Ministry of Environment [14]. In this case, the biological tests were carried out according to the OECD test guidelines performed under Good Laboratory Practices (GLP). The others correspond to data extracted from the registration files accessible from the ECB website [15], the AQUIRE database [16], and an internal database called MATE [17].

The objective of our study is to define quantile QSAR models for this endpoint, models integrating a large number of chemicals. The sensitivity of the regressions to outliers is analyzed, and QR was used in combination with support vector machine (QSVMR [18]). A comparison of QSVMR with the classical SVM regression (SVMR) is also provided.

Methods

Training set

The ecotoxicological data were downloaded from the OECD website [14]. The biological tests on a specific algal species named *Pseudokirchneriella subcapitata* (*Selenastrum capricornutum*) followed the OECD-GLP standard and OECD test guidelines. Two types of 72 h-EC₅₀ values (OECD TG 201) were recorded in this set corresponding to the growth rate (EC_{50r}) and/or to the area under growth curve (AUG / EC_{50b}). A high correlation between the two types of EC₅₀ values ($n=249$, $r=0.967$) was observed. For the most recent data, only the values associated with the growth rate method are displayed. When examining the overall data and particularly the correlation between the two values, we chose to consider the lowest acute toxicity values recorded for the 72 h EC₅₀ regardless of the methods. With the cut-off values associated with hydrophobic and hydrophilic properties (*vide infra*), 277 chemical derivatives were part of the training set.

Testing set

A first dataset of biological data relating to algal growth effects of 2782 high-production volume chemicals was taken

from the ECB website [15]. Among these 2782 chemicals, 1749 structures can be downloaded, and only 47 have 72-h EC_{50r} value(s) for the same species as the training set: *Selenastrum capricornutum* (the lowest EC_{50r} value is taken in the case of several values displayed). A second data set was obtained from AQUatic Information RETrieval (AQUIRE) [16]. An advanced query was carried out on AQUIRE, and a set of 60 chemical structures with 72-h EC_{50r} value for the *Selenastrum capricornutum* was extracted. A third data set of 94 chemicals was retrieved from our internal database [17]. These three data sets were gathered in a testing set. By considering the cut-off values associated with hydrophobic and hydrophilic properties (*vide infra*) and the suppression of duplicates (comparison with the training set), 124 derivatives composed the testing set.

Descriptors

The octanol-water partition coefficients ($\log K_{OW}$) were determined by two *in silico* methods leading to a descriptor named logP (open-source software KOWWIN, [19]) and a second one named ALogP, an atom-based method [20]. Molecular solubility, expressed as logS with S in M, was estimated from multiple linear regression models defined by Tetko et al. [21]. Only derivatives with ALogP ≥ 0 and logS values ≥ -6 were retained. The 3D atomic coordinates were generated for each structure and a first energy minimization was performed with Pipeline Pilot [22] using a clean force field [23]. Then, another optimization was carried out with DMol3 [24] by considering PWC [25, 26] for the functional (DFT exchange correlation potential) and medium for the convergence. For the descriptors, special attention was given to molecular surface properties. The first ones correspond to molecular surface areas and their associated descriptors. In this case, total, polar, and solvent accessible surface areas were computed for each molecule using a 2D approximation. The fractional polar surface areas were also determined using the ratio between polar and total surface areas (the same process was applied to solvent accessible surface areas). For topological descriptors associated with molecular shapes, shadow indices and Jurs descriptors were calculated. Shadow indices [27] project the molecular shapes on three mutually perpendicular axes: XY, XZ, and YZ. The associated lengths (shadow Xlength, Ylength, Zlength) correspond to the maximum dimensions of the molecular surface projections. The ratio between the largest and the smallest dimension corresponds to the last descriptor (shadow_nu). The 30 Jurs descriptors [28] combine shape and electronic information. It is impossible to detail all these descriptors, but we can mention a descriptor named Jurs_PPSA_1 corresponding to the sum of the solvent accessible surface areas of all positive atomic charges or Jurs_PNSA_1, calculated in the same way but for negative atomic charges. Ehresmann et al. [29] have described new

molecular descriptors based on local properties at the molecular surface. This surface corresponds to a shrink-wrapped isodensity surface [30], with $10^{-4} \cdot e^{-\rho} \cdot \text{\AA}^{-3}$ for the electronic density, generated from semi-empirical molecular orbital calculations (VAMP [31] in this case). Four local properties, the molecular electrostatic potential (MEP), the local ionization energy (IE_L), the local electron affinity (EA_L), and the local polarizability (α_L , polarizability for the name of the descriptor) were calculated at the points on the surface. Two properties, the local hardness (η_L) and the local electronegativity (χ_L), were derived from IE_L and EA_L . Starting from these local properties, 81 descriptors were determined. Other descriptors were associated with steric and electronic properties (Pipeline pilot and Dmol3) like dipole moment descriptors [32, 33], the radius of gyration, the sum of atomic polarizabilities (Apol), principal moments of inertia (PMI), molecular weight, globularity, count descriptors (H bond acceptor and donor), pKa_acide (most acidic site), pKa_basic (most basic site), HOMO, LUMO, band gap energy (LUMO–HOMO), dielectric energy, solvation energy, molecular volume, and cavity volume.

Quantile regression (QR)

We consider the linear regression model:

$$Y = X\beta + \varepsilon, \quad (1)$$

where $Y = (Y_1, \dots, Y_n)$ is the vector of observations, X is the design matrix of dimension $n \times p$ where, for $i = 1, \dots, n$, $x_i^t \in \mathbb{R}^p$ is the i th line of the matrix X with t denoting the transpose of the vector, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ is a vector of independent errors with an unknown distribution function f and $\beta = (\beta_1, \dots, \beta_p)^t$ denotes the vector of unknown regression parameters to be estimated. The classical least squares linear regression estimator is ineffective if the errors are non-normal. To overcome this problem, in 1978, Koenker and Basset [11] proposed a quantile-based approach for linear regression models. The quantile regression estimator is more robust to non-normal errors and outlier observations. In this case, instead of focusing on the changes in the mean of Y , the QR approach tests whether there is a change in the θ th-quantile of Y for any given $\theta \in (0, 1)$. So, QR gives better characterization of the data, since it enables estimating the impact of a covariate on the entire distribution of the response variable rather than on its conditional mean. The least squares estimators in regression are designed to estimate the mean of the response variable Y conditional on X whereas in QR, the estimators are designed to estimate the relation of X with Y , conditional on quantiles of Y . QR can be viewed as an extension of the classical least squares estimation of conditional mean models for the estimation of models associated with several conditional quantile functions. Therefore, QR estimates the

conditional median or other quantiles of the response variable, unlike the ordinary least squares regression, which estimates the conditional mean. We also note that QR is invariant to monotonic transformations, such as logarithmic transformation, and QR algorithms are now available in most statistical packages.

The quantile regression estimators:

$$\hat{\beta}(\theta) = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\theta}(Y_i - x_i^t \beta), \quad (2)$$

are defined as a solution of the minimization problem where $\rho_{\theta}(z) = |z| \rho_{\theta}(z) = z(\theta - I(z < 0))$ and $I(P)$ takes the value 1 or 0 depending on whether the condition P is satisfied or not.

The QR loss function denoted by the function ρ_{θ} is an absolute loss function, that is a weighted sum of absolute deviations where the $(1 - \theta)$ weight is assigned to the negative deviations and the θ weight is used for the positive absolute deviations. More specifically, it can be shown that this loss function makes it possible to determine the θ -quantile, $\theta \in (0, 1)$. A special case of this class of estimator (obtained for $\theta = 1/2$) is the least absolute deviation (LAD) estimator or the median regression, which is obtained by resolution of the minimization problem (2). LAD is often chosen as an alternative to least squares estimators. It performs better in the presence of heavy tail distributions. Under the regularity conditions given in [34] (page 120), the asymptotic normality of the quantile regression estimator $\hat{\beta}(\theta)$ was proven by [11] and [34] under the assumption of independent and identically distributed (iid) errors, that is, ε_i are iid variables in the model. The asymptotic representation was given in [35] for independently, but not necessarily, identically distributed errors. The asymptotic variance of the estimator $\hat{\beta}(\theta)$ can be obtained by direct estimation using the non-parametric estimation of the sparsity function [36, 37]. When the observations are independent but not identically distributed, as often experienced in practical chemical applications, it is possible to extend the iid theory to produce a version of the Huber-Eicker-White sandwich formula for the limiting covariance matrix of $\hat{\beta}(\theta)$. Several estimators have been proposed for this problem, including a rank test as described in [38, 39] and following the work of [40], and bootstrap methods [41–43]. Statistical tests in quantile regression models need an estimator of the unknown nuisance sparsity function. A Wald test for the null hypothesis can be applied using the consistency of the sparsity function estimator and the asymptotic normality of the quantile regression estimator. The regression rank score [40] also provides an interesting approach to many inference problems while avoiding the density function estimation. The inference on quantile regression can also be considered using Khmaladze's extension [44, 45] of the Doob-Meyer construction. For more information on quantile regression methods, see Briollais and Durrieu publications [46, 36].

Segmented linear regression model

This approach led to the definition of the optimum number of descriptors for the equations. We consider the simple linear regression model with only one change point given by:

$$Y_i = \begin{cases} a_1 + b_1 X_i + \varepsilon_i & \text{if } X_i \leq \tau \\ a_2 + b_2 X_i + \varepsilon_i & \text{if } X_i > \tau \end{cases}$$

where ε , a_1 , a_2 , b_1 , b_2 and τ denote respectively the random error term, the unknown intercepts, slopes and change-point in the coefficient of the two linear regression models. The objective is to test, using the likelihood ratio test [47], the “no change in the regression coefficient” null hypothesis against the “one change in the regression coefficient” alternative hypothesis. For general information on the bilinear model applied to biological systems, see the initial work of Kubinyi [48].

Support vector machine regression (SVMR)

For the linear regression (in feature space) defined by $f(x, w) = \langle w, \phi(x) \rangle + b$ with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ and $\langle w, \phi(x) \rangle$ is the dot product in the feature space, the objective is to minimize:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-), \tag{3}$$

subject to:

$$y_i - f(x_i, w) \leq \delta + \xi_i^-, f(x_i, w) - y_i \leq \delta + \xi_i^+, \xi_i^+, \xi_i^- \geq 0, i \in \{1, \dots, n\},$$

where ξ_i^+ and ξ_i^- are respectively the slack variables associated with an overestimate and an underestimate of the calculated response for the input vector x_i , δ determines the limits of the approximation, and C is a positive constant that controls the penalty associated with deviation larger than δ .

The minimization problem can be formulated in its dual quadratic optimization form, which involves maximizing

$$-\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\lambda_i^- - \lambda_i^+) (\lambda_j^- - \lambda_j^+) \langle x_i, x_j \rangle - \delta \sum_{i=1}^n (\lambda_i^- + \lambda_i^+) + \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) y_i, \tag{4}$$

under the constraint

$$\sum_{i=1}^n (\lambda_i^- - \lambda_i^+) = 0 \quad \text{and} \quad \lambda_i^-, \lambda_i^+ \in [0, C],$$

where λ_i^-, λ_i^+ denote the Lagrange multipliers. Once the dual problem is solved for λ_i^- and λ_i^+ , the solution for a given x is obtained by:

$$w = \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) x_i$$

and therefore

$$f(x) = \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) \langle x_i, x \rangle + b, \tag{5}$$

where λ_i^- and $\lambda_i^+ \in [0, C]$.

For non-linear regression, the support vector machine algorithm can be performed by simply transforming the x_i by a non-linear mapping ϕ from the input space to some high-dimensional feature space (sometimes even infinite-dimensional). The optimization problem involves finding the flattest function in feature space, not in input space. The solution for x^* is obtained by:

$$f(x^*) = \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) K(x_i, x^*) + b.$$

SVMR performance depends on a correct setting of the hyper-parameters C , δ and the kernel function.

Quantile support vector regression (QSVMR)

The quantile function Y_i conditionally to $X=x_i$ is given for $i=1, \dots, n$ by:

$$Q(\theta | x_i) = \omega'_\theta \phi(x_i) \quad \text{for} \quad \theta \in (0, 1), \tag{6}$$

where ω_θ denotes the θ -quantile regression. QSVMR can be defined by minimizing for $\theta \in (0, 1)$

$$\frac{1}{2} \|\omega_\theta\|^2 + C \sum_{i=1}^n \rho_\theta(y_i - \omega'_\theta \phi(x_i)), \tag{7}$$

where C denotes the degree of penalization controlling the trade-off between the flatness of the quantile function estimate and the amount up to which deviations larger than zero are tolerated. A solution to the minimization problem (3) for $\theta \in (0, 1)$ is obtained by optimizing its quadratic dual version. The θ -quantile regression for x^* can be written:

$$\omega_\theta = \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) \phi(x_i) \quad \text{and} \quad Q(\theta | x^*) = \sum_{i=1}^n (\lambda_i^- - \lambda_i^+) K(x_i, x^*), \tag{8}$$

where λ_i^-, λ_i^+ are Lagrange multipliers and $K(x_i, x_j)$ denotes a kernel function.

Parameters and function

We consider the kernel Gaussian radial basis function (RBF) given by the equation:

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2} \|x_i - x_j\|^2\right), \quad (9)$$

where σ corresponds to the bandwidth parameter. The bandwidth parameter is estimated using the procedure developed in [49]. We also used the cross-validation method to determine the value of the bandwidth. We denote in the sequel by $\hat{\sigma}$ the bandwidth estimator of σ .

The parameter C determines the trade-off between the model complexity and the degree to which deviations larger than δ are tolerated in the optimization phase. The parameter δ controls (only SVMR) the width of the δ -insensitive zone used to fit the data. Its values affect the number of support vectors. Larger values result in fewer support vectors and more flat regression estimates.

To estimate C , we considered the approach of Cherkassy and Ma [50] given by:

$$\hat{C} = \max\left(\left|\bar{Y} + 3S\right|, \left|\bar{Y} - 3S\right|\right), \quad (10)$$

where \bar{Y} and S are respectively the empirical estimators of the mean and the standard deviation of the biological activities. This choice of C is more robust than the approach of Mattera and Haykin [51] when the data contains outliers.

The choice of δ in SVMR should be proportional to the variability of Y . Cherkassy and Ma [50] propose:

$$\hat{\delta} = 3S\sqrt{\frac{\log(n)}{n}}, \quad (11)$$

to determine δ where S corresponds to the empirical estimation of the standard deviation associated with a biological data error of 5 %, 10 %, and 15 %.

The regression was performed on the training set (277 derivatives) with a threefold cross-validation process to determine the optimum number of variables. The equation selected from SVMR or QSVMR was applied to the testing set (134 derivatives) leading to R^2_{test} values. Afterward, the two sets were reunified ($n=401$).

Statistical computation

The R statistical environment was used for the overall calculations. Principal component analysis was carried out with the ade4 package [52]. Stepwise regressions were carried out by

examining the best descriptors step by step. Within each fold of the cross-validation experiment, an arbitrary division of the dataset into training set (70 %) and testing set (30 %) was fixed. QR, SVMR, and QSVMR were applied using the kernlab package [53]. The coefficients of determination R^2_{cross} , R^2_{test} , R^2_{train} denote the cross-validation and the testing and training coefficients respectively. To select the optimum number of variables, we applied a segmented regression model using one unknown change point to be estimated. For all the statistical results, after checking the condition of application of statistical tests (normality, independence, homogeneity, etc.), a probability of $p < 0.05$ was considered significant.

Results

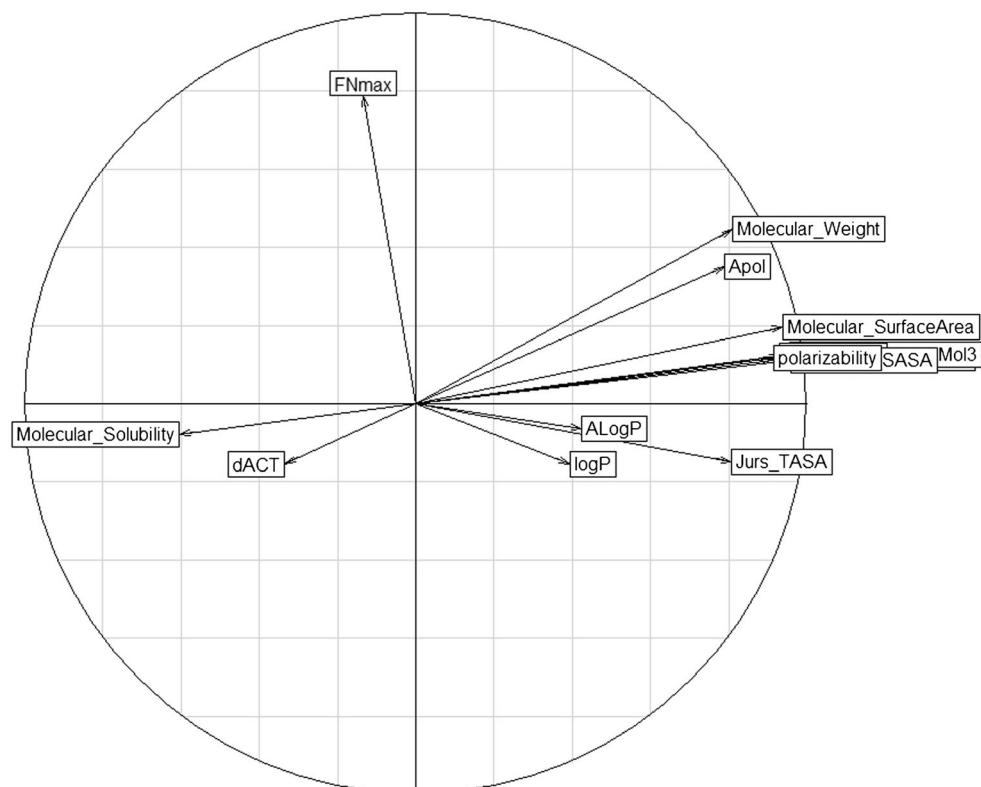
Comparison of the biological activities (training and testing set)

There is no significant difference in distribution between the training and testing sets when considering the biological activities ($p=0.46$ with the Kolmogorov-Smirnov test). Therefore, principal component analysis was performed on the joined training and testing sets. The first two principal component axes explain 55 % of the total variability. The following variables have a significant correlation coefficient ($p < 0.05$ with Spearman and Kendall tests) greater than 0.4 in relation to the biological activities: Surface_Area_DMol3, Cavity_Volume_DMol3, Apol, Jurs_TASA, Shadow_XY, ALogP, logP, Molecular_Solubility, Molecular_Weight, Molecular_SurfaceArea, Molecular_SASA, polarizability, and POLint. The representation of the projection of these variables into the correlation circle associated with the first two component axes (see Fig. 1) summarizes the correlations between the variables.

SVMR and QSVMR

SVMR ($\hat{C}=8.03$, $\hat{\delta}=0.2$, $\hat{\sigma}=0.11$) and QSVMR ($\hat{C}=8.03$, $\hat{\sigma}=0.11$) were carried out on the training set, and the models were computed on the testing set (see Fig. 2 and Table 1). Figure 2 shows the variations of R^2_{cross} , R^2_{test} , R^2_{cross} , and SE_{test} as a function of the number of components (descriptors). SVMR ($\hat{C}=8.24$, $\hat{\delta}=0.18$, $\hat{\sigma}=0.14$) and QSVMR ($\hat{C}=8.24$, $\hat{\sigma}=0.14$) were also carried out on the whole dataset (see Table 1). With QSVMR, when considering the addition of new descriptors in the regression, a stability of statistical values (R^2_{cross} , SE_{cross}) was observed, unlike SVMR for which a sensibility of R^2_{cross} and R^2_{test} values to this number was recorded. For QSVMR (277/124, Table 1), the three descriptors correspond to logP, solubility, and Apol.

Fig. 1 Projection of the variables into the plane spanned by the first two principal component axes (dACT for biological activities)



TR from external QSAR and MOA

To analyze the weakness of the initial global model and a potential relation with the association, in the same set of

compounds acting with different MOA, the TR was calculated [13]. TR represents the ratio between the predicted (QSAR approach) and the experimental values. The predicted values are determined from a QSAR model associated with a baseline

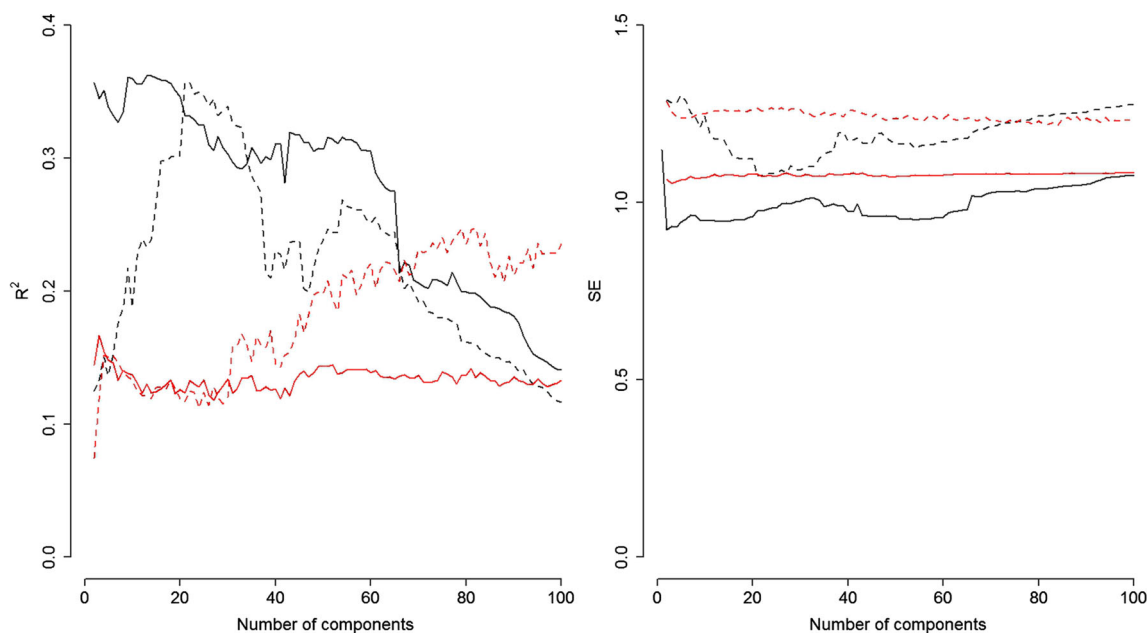


Fig. 2 Variation in function of the number of components of the R^2_{cross} , SE_{cross} (in solid lines) and R^2_{test} , SE_{test} (in dotted lines). The QSVMR results are in red and the SVMR results are in black

Table 1 Statistical results from SVMR and QSVMR for the training and testing datasets. In the first row, the sample size of the training and the testing sets is 277 and 124 respectively. In the second row, we consider the joined training and testing sets (n=277+124=401)

n	$\hat{\sigma}$	\hat{C}	$\hat{\delta}$	θ	Variables	$R^2_{\text{train}} / SE_{\text{train}}$	$R^2_{\text{cross}} / SE_{\text{cross}}$	$R^2_{\text{test}} / SE_{\text{test}}$
SVMR								
(277/124)	0.11	8.03	0.2		22	0.77 / 0.55	0.33 / 0.98	0.36 / 1.08
(401)	0.14	8.24	0.18		19	0.78 / 0.95	0.35 / 0.95	–
QSVMR								
(277/124)	0.11	8.03		0.5	3	0.61 / 0.95	0.16 / 1.03	0.15 / 1.28
(401)	0.14	8.24		0.5	29	0.60 / 1.02	0.20 / 1.09	–

narcosis for the MOA. A cutoff was fixed at ten for differentiating the two MOA (non-specific vs. specific). The model

$$\log \frac{1}{EC_{50}[M]} = 0.95 \log(D_{lipw})(pH7) + 1.16 \quad (12)$$

of Escher et al. [54] was chosen for this definition, keeping in mind the various remarks concerning the lack of a precise equation for this MOA.

The relationship [55] between $\log(D_{lipw})$ and $\log(K_{ow})$ was fixed by considering the model

$$\log(D_{lipw}) = 0.997 \log(K_{ow}) + 0.0851. \quad (13)$$

Our values correspond to a 72-h growth rate as opposed to a 24-h growth rate for model (12); hence, accounting for the expected time dependence of effect, EC_{50} are considered to be three times lower. The integration of the different points and our descriptor $\log P$ for $\log(K_{ow})$ led to the final model given by

$$\log \frac{1}{EC_{50}[M]} = 0.947 \log P + 0.77. \quad (14)$$

By Eq. (14), 294 derivatives out of 401 had a $TR < 10$. Interestingly, 64 % of the derivatives (n=254) were found with a TR between 0.1 and 10. Accordingly, a significant regression model ($R^2_{\text{train}}=0.77$, $R^2_{\text{cross}}=0.76$, $SE_{\text{cross}}=0.48$, n=254) given by

$$\log \frac{1}{EC_{50}[M]} = 0.76 \log P + 2.22, \quad (15)$$

was obtained on this set with $\log P$ as a descriptor. This equation represents our first general narcotic equation (associated with baseline and other narcosis like polar narcosis) for the algae endpoint.

A SVM classification [56, 57] was carried out to differentiate the chemical characteristics of the two groups ($TR \geq 10$ vs $TR < 10$). The optimum separation was obtained with $A \log P$, $\log P$, molecular solubility. This classification led to a group called group A (n=69) for which 75 % of chemicals have a

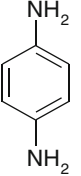
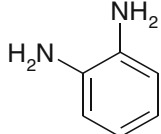
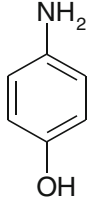
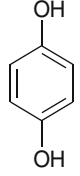
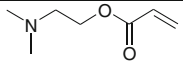
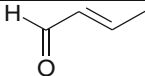
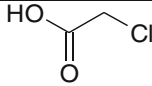
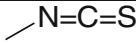
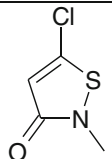
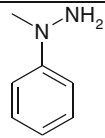

$TR > 10$ and a group called group B (n=332) for which 17 % (55 derivatives) of chemicals have a $TR > 10$. In fact, and logically when considering the intercept in regression model (14), most of the derivatives in group A have a low value of $\log P$ ($\log P < 1$). Our initial validity domain based on $A \log P$ and $\log S$ ($A \log P \geq 0$ and $\log S$ values ≥ -6) is justified and amplified, starting from this differentiation based on the agreement with the baseline narcosis model given in (14). In group A, the highly toxic derivatives ($-\log(EC_{50}) > 5$) with low values of $\log P$ correspond mostly to reactive derivatives (see Table 2). For phenylenediamine (ortho and para), aminophenol (ortho and para), and hydroquinone, the toxicities are related directly to their redox equilibrium with quinones. Alpha beta unsaturated carbonyl, activated halides, and isothiocyanate can also react with macromolecules. In fact, for reactive chemicals, the interval between predicted and experimental values based on $\log K_{OW}$ was described previously to be higher for hydrophilic derivatives than for hydrophobic derivatives.

TR from external QSAR and QSVMR

A QSVMR analysis ($\theta=0.5$) was carried out on group B after the suppression of the 13 derivatives with $\log P < 1$. The optimum relationship was obtained for 21 variables ($\hat{C}=8.27$; $R^2_{\text{cross}}=0.6$; $SE_{\text{cross}}=0.79$, n=319, see Fig. 3).

A TR determination was carried out by considering the equation associated with QSVMR as the basis for the predicted values. A total of 18 derivatives were found with a $TR > 10$. By discarding these 18 derivatives, no real increase in the statistical quality of the equation was observed, thus showing the stability of QSVMR ($\theta=0.5$) toward potential outliers ($R^2_{\text{cross}}=0.64$, $SE_{\text{cross}}=0.85$, n=301, $\hat{C}=7.97$, see Fig. 3). However, a decrease in the optimum number of descriptors was recorded with three descriptors ($A \log P$, molecular solubility, polarizability) instead of the previous 21 descriptors (for n=319). The ecotoxicity of most of these outliers is clearly associated with a specific MOA [58] well recorded in the literature (see Table 3). Phenylurea, triazinone, and bipyridylum derivatives are inhibitors of photosynthesis (photosystem I for bipyridylum and photosystem II

Table 2 Structural analysis of some derivatives (group A) with high toxicities and low logP values

 5.78	 5.12	 6.03	 6.31
 5.85	 5.17	 6.15	 5.71
 6.06	 7.15	 6.28	

otherwise). Chloracetamides are long-chain fatty acid inhibitors and inhibitors of mitosis and cell division. Diphenylethers are inhibitors of protoporphyrinogen oxidase or potentially uncouplers by considering the phenol function. Quinoline derivatives represent a specific class of antifungal drugs. For reactive chemicals, two unsaturated derivatives, one peroxide and one phenyl nitro derivative appear in this set. As always with reactive chemicals, the highest toxicities of alkyl thiols toward the corresponding nearest alcohol are clearly observed with the formation of free radical species for the explanation [59]. Two polyaromatic derivatives with aniline and nitro functions have a very high toxicity, which must be associated with a specific MOA, but in these cases, no explanation is provided in the literature.

TR/QR/QSVMR

Starting from the initial dataset (401 derivatives), a linear QR was done with logP as descriptor for the definition of the TR. The median quantile regression ($\theta=0.5$) is given by

$$\log \frac{1}{EC_{50}[M]} = 0.43 \log P + 3.35 \quad (16)$$

with $R^2=0.48$.

Depending on the quantiles, an evolution of the intercepts (2.41 ($\theta=0.1$) to 5.25 ($\theta=0.9$)) and an evolution of the slopes were observed (0.43 ($\theta=0.1$) to 0.33 ($\theta=0.9$)). So, logP exerts a change in the conditional distribution of the biological activities. The variance of activity decreases with an increase of logP.

From Eq. (16), a TR value for each derivative was determined, leading to 336 derivatives with a $TR < 10$ (as compared to 294 in the previous case). Starting from this set, we obtained the median quantile regression

$$\log \frac{1}{EC_{50}[M]} = 0.45 \log P + 3.08 \quad (17)$$

with a $R^2=0.64$.

Depending on the quantile, the slope is nearly stable (0.41 for $\theta=0.1$ and 0.44 for $\theta=0.9$) with an evolution of the intercept between 2.35 ($\theta=0.1$) and 3.90 ($\theta=0.9$). A QSVMR was applied to this set ($\hat{C}=7.04$, $\theta=0.5$) leading to interesting results, particularly for the value associated with the standard error of the estimates SE_{cross} ($R^2_{\text{cross}}=0.66$, $SE_{\text{cross}}=0.4$, see Fig. 3). This relationship was obtained with three descriptors (AlogP, molecular solubility, Apol). When examining the 95 % confidence intervals, we obtained less than one logarithmic unit (plus or minus) for the interval associated with predictions.

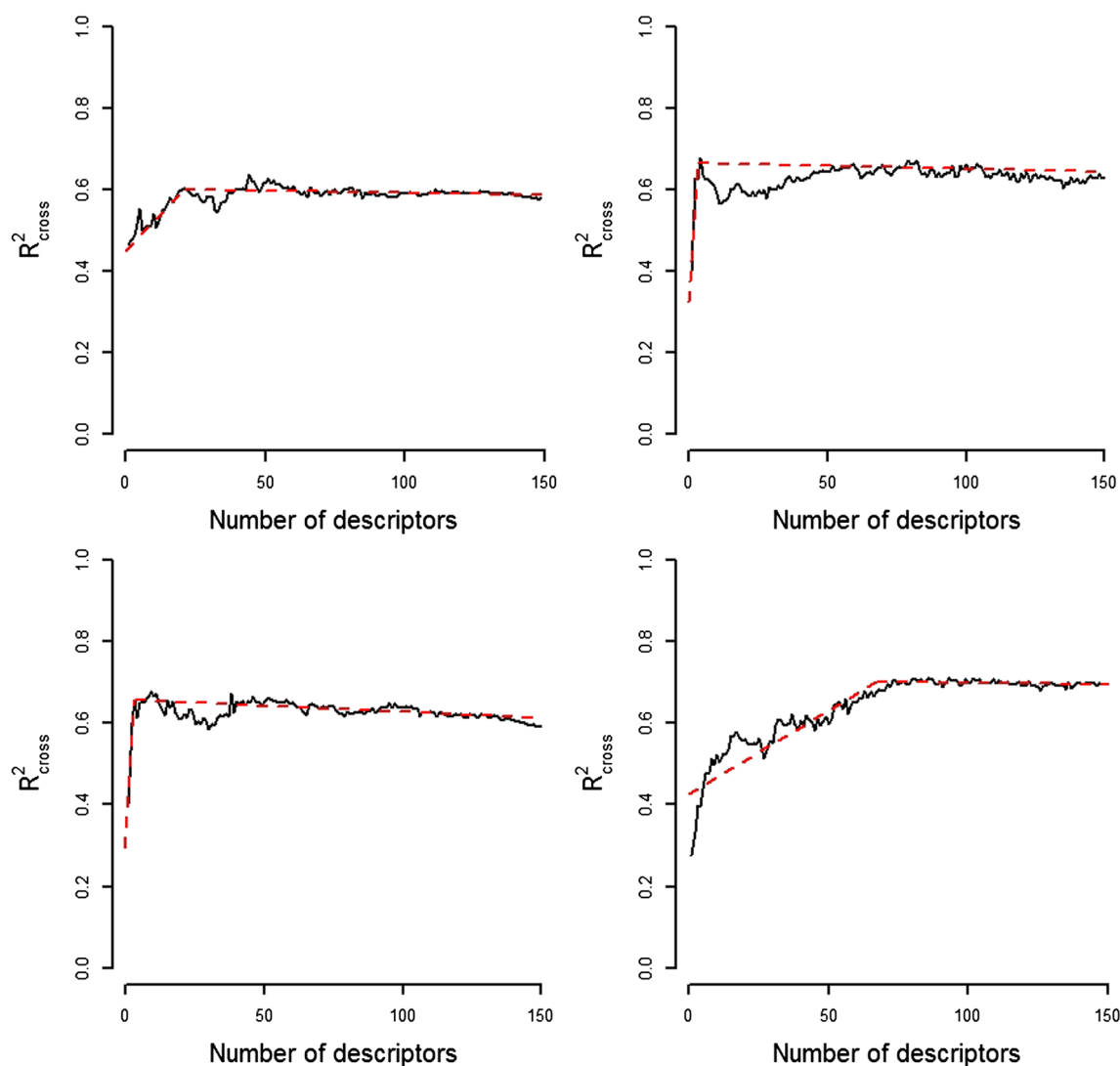


Fig. 3 Variation of R^2_{cross} as function of the number of components. QSVMR with $n=319$ (up left), $n=301$ (up right), $n=336$ (down left), $n=368$ (down right)

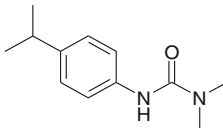
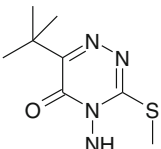
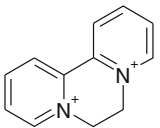
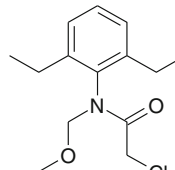
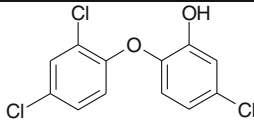
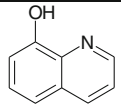
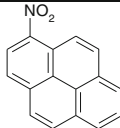
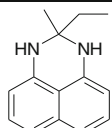

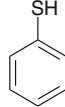
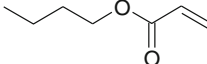
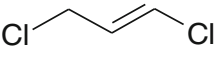
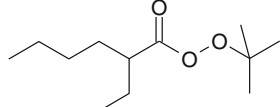
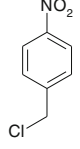
A SVM classification was carried out to understand the differences between the two groups of derivatives (336 with $TR < 10$ and 65 with $TR > 10$). No interesting result was obtained from our descriptors. However, for the isodensity surfaces and the properties (molecular shapes, electrostatics, donor and acceptor properties, polarizability) associated with these surfaces, a particular type of fingerprint named the rotationally invariant fingerprint (RIF) can be calculated [60]. Using SVM classification on the overall fingerprint (RIF), an error of classification of 1.75 % was observed on the training and 14.95 % on the cross-validation processes. After the cross-validation process, the dataset was separated into two groups with 368 derivatives and 33 derivatives ($TR > 10$) respectively. QSVMR was carried out ($\hat{C}=7.18$, $\theta=0.5$) on the set of 368 derivatives, leading to correct statistical results with 67 descriptors ($R^2_{\text{cross}}=0.67$, $SE_{\text{cross}}=0.41$, see

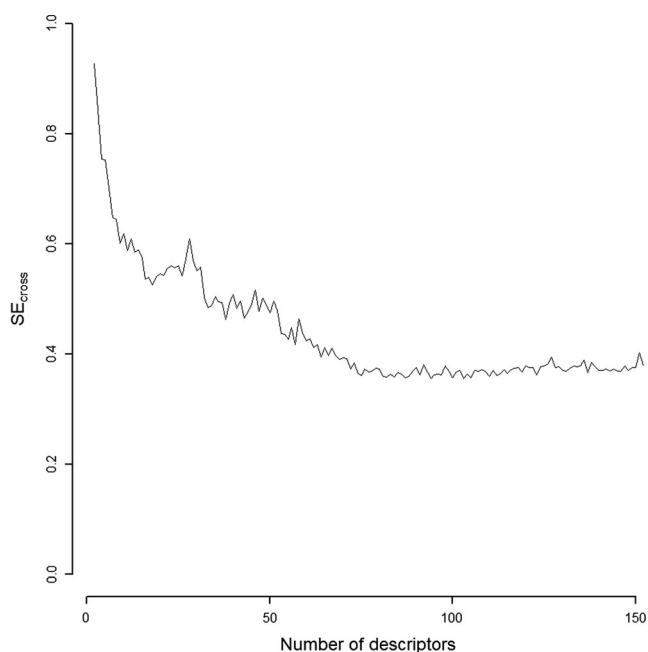
Figs. 3 and 4). Concerning the 33 derivatives associated with outliers, the most toxic derivatives are displayed in Table 4.

Discussion

The definition of a global QSAR model for a dataset of chemicals is a worthwhile endeavor because, for most chemical derivatives, the classification into a specific set is not obvious. The interval of prediction of this model must cover the experimental value and must be a basis for the estimation of risk. For acute toxicities to algal species, even a narcosis model associated with a large set of derivatives is needed. A previous study related to hydrocarbon derivatives, and particularly industrial chemicals such as petroleum products, has

Table 3 Description of some outliers associated with group B (n=310)

<i>phenyl urea</i>	<i>Triazinone</i>	<i>Bipyridylum</i>	<i>Chloroacetamides</i>
 6.16	 7.15	 6.4	 7.37
<i>Diphenylether</i>	<i>Quinolines</i>	<i>Polyaromatics with aniline and nitro functions</i>	
 8.14	 5.65	 8.16	 7.02
Thiols			
 7.01	 5.83		
Reactive species for MOA			
 5.15	 5.66	 5.97	 6.66

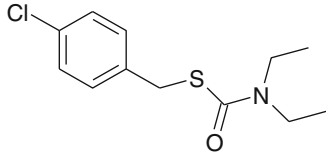
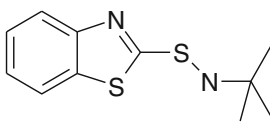
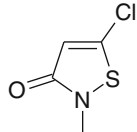
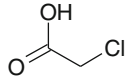
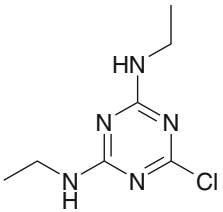
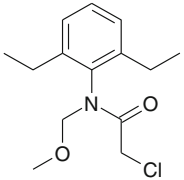
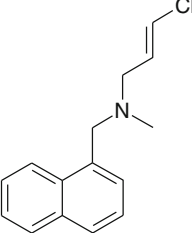
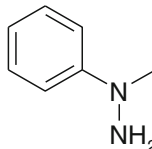
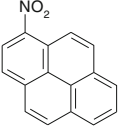
**Fig. 4** Variation of SE_{cross} as a function of the number of components (QSVMR, n=368)

demonstrated that a narcosis target lipid model could be built for algae [10]. A final R^2 value of 0.85 was obtained with a standard deviation of 0.34 for the residuals after exclusion of four outliers.

For our dataset, the initial correlation (equal to 0.4) with $\log K_{ow}$ was clearly observed. With the notion of TR (external equation), 319 derivatives out of 401 were found with a $TR < 10$ and 254 derivatives with a TR between 0.1 and 10. The slope estimate in model (15) was found to be slightly lower than the slopes observed from the previous narcotic equations with an increase in the intercept. As always with $\log K_{ow}$, the QR regression on the overall set (n=401) then on the subset of 336 derivatives (84 %) led to a higher value of the intercept, namely a value around 3 with an interval between 2.35 and 3.9 for different quantiles. These last equations fit the data associated with some classes of derivatives appropriately, such as aniline, which follows a polar narcosis mechanism for the toxic action (aniline ($-\log EC_{50} = -3.35$, $\log P = 0.9$); p-methyl aniline ($-\log EC_{50} = -4.03$, $\log P = 1.39$)).

With QSVMR, for the initial training set (n=277), a low predictive quality of the model was obtained but with three descriptors corresponding to $\log P$, molecular solubility, and

Table 4 Description of the most toxic chemicals among the set of 33 derivatives. The number of derivatives associated with a typical scaffold is indicated in brackets

 6.79 (2)	 6.62 (4)	 6.06 (1)	 6.15 (1)
 6.30 (7)	 7.37 (3)	 6.08 (1)	 7.14 (1)
 8.16 (1)			

Apol. The same type (logP or ALogP, molecular solubility, Apol or polarizability) of optimum descriptors was observed for the set associated with $n=310$ and with $n=336$. A narcotic MOA is a function of the relationship between the activities and the water-octanol partitioning ($\log K_{OW}$). However, octanol is not the optimal middle (as an analog of biological membrane), and the partitioning of the chemical derivatives is very sensitive to the polarity-polarizability factor of the structures [61]. Therefore, it is expected that the polarizabilities of the derivatives are one of the main descriptors associated with $\log K_{OW}$ values. Molecular solubility completes the information by describing the thermodynamic equilibrium between solute and solvent. Thus, for the last two sets ($n=310$ and $n=336$), the optimum relationships were obtained with these three descriptors and with a stability of the predictive quality of the regressions, regardless of the number of descriptors (see Fig. 3).

With QSVMR and in relation with the three descriptors, we always observed two types of evolution in terms of the initial approach. With TR associated with an external equation, derivatives with high toxicities combined with low logP values were suppressed in a first step. Thus, the resulting set (group B) still included hydrophobic reactive derivatives and a series of pesticides. QSVMR on this set led to a relationship with 21 descriptors (optimum value of R^2_{cross}) and a high standard error of estimates (SE_{cross}). The suppression of 18 derivatives from this set ($n=301$) led to a relationship with no difference, excepting the number of descriptors with an optimum for the same three descriptors. With TR defined from

QR, the relationships between logP and the ecotoxicities were analyzed on the basis of the QR slopes and intercepts. With 401 derivatives, the variance of the biological activities in function of logP was observed, but after the selection of the subset of 336 chemicals, no real modification of the variance was observed. With this new set, we obtained a model based on the same three descriptors (QSVMR) with the standard errors of the estimates decreasing by a factor of two, as compared to the previous models. The difference between the two sets (chemicals with $TR > 10$ and others) was understood with a molecular fingerprint (RIF) integrating the molecular shape and the properties on the molecular surfaces. After this classification, a set integrating 91 % of the initial chemicals gave a significant QSAR model with 67 descriptors. To analyze this last situation, different observations could be made: a) 83 % of the derivatives should follow a narcosis mechanism for the MOA with a QSVMR correct for three descriptors ($n=336$); b) the selection (SVM classification with RIF) led to elimination of more than 50 % of the most toxic derivatives ($-\log EC_{50} > 6.5$); c) the remaining derivatives ($n=368$) share a few main properties associated with the molecular surface (RIF descriptors); and d) our descriptors are strongly related to the analysis of the molecular surfaces. Therefore, the QSVMR modeling approach improves the results with an optimum of 67 descriptors. The three descriptors were again integrated to the first descriptors following those initially correlated with biological activities such as molecular surface area, molecular_SASA, Molecular_Weight, Cavity_Volume_DMol3, shadows descriptors (shadow_XY),

Jurs descriptors (Jurs_TASA), followed by other descriptors (PMI_Y, PMI_Z, PMI_Mag). The descriptors associated with Parasurf are also selected in the first ones with a series of properties integrated over the surface: polarizability (POLint), local electro-negativity (ENEGint), local ionization energy (IELint), local hardness (Hardint), local electron affinity (EALint), and molecular electrostatic potential (MEPint). A differentiation between some classes of compounds could be found rapidly with these last descriptors such as bipyridilium (MEPint, the highest value), phthalates and phosphates (IELint, high values), and alkyl halides (IELint, the lowest values with one or two carbons).

Conclusions

These analyses provide evidence for a robust modeling approach based on QR and QSVMR to define a global model. QR, based on a fundamental descriptor in ecotoxicology (logKow), allowed selecting directly a subset of derivatives for which a correct predictive quality was found with QSVMR and three descriptors. Stability of the model, due to its robustness against outliers, was observed with QSVMR, particularly from the R^2_{cross} values. Based on this result, 83 % of our chemicals should have a narcosis mechanism for the MOA. When examining a subset of derivatives, differentiated by the distances from the rotational invariant fingerprint, a correct relationship was obtained for 91 % of our initial dataset by integrating in this case a large number of descriptors related to the molecular surface properties.

Acknowledgments M. Jonathan Villain was supported by a grant from the Région de Bretagne (Region of Brittany) and the French Ministry of Education. The authors thank Laurent Briollais for the improvement of the narrative style of the manuscript and the Agence Nationale de la Recherche (French National Research Agency) (ANR, ANR-07-CP2D-09-02 and Pharm@ecotox) for financial support.

References

- Vogelgesang J (2002) The EC white paper on a strategy for a future chemicals policy. *Altern Lab Anim* 30(Suppl 2):211–212
- Netzeva TI, Pavan M, Worth AP (2008) Review of (quantitative) structure–activity relationships for acute aquatic toxicity. *QSAR Comb Sci* 27(1):77–90
- Chen J, Li X, Yu H, Wang Y, Qiao X (2007) Progress and perspectives of quantitative structure–activity relationships used for ecological risk assessment of toxic organic compounds. *Sci China Ser B Chem* 51(7):593–606
- Aruoja V, Sihtmae M, Dubourguier H-C, Kahru A (2011) Toxicity of 58 substituted anilines and phenols to algae *Pseudokirchneriella subcapitata* and bacteria *Vibrio fischeri*: comparison with published data and QSARs. *Chemosphere* 84(10):1310–1320
- Netzeva T, Manuela P, Worth A (2007) Review of data sources, QSARs and integrated testing strategies for aquatic toxicity. European Communities, Luxembourg
- Hsieh S-H, Hsu C-H, Tsai D-Y, Chen C-Y (2006) Quantitative structure–activity relationships for toxicity of nonpolar narcotic chemicals to *Pseudokirchneriella subcapitata*. *Environ Toxicol Chem* 25(11):2920–2926
- Van Leeuwen CJ, Van Der Zandt PTJ, Aldenberg T, Verhaar HJM, Hermens JLM (1992) Application of QSARs, extrapolation and equilibrium partitioning in aquatic effects assessment. I. Narcotic industrial pollutants. *Environ Toxicol Chem* 11(2):267–282
- Furusjo E, Svenson A, Rahmberg M, Andersson M (2006) The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere* 63(1):99–108
- Fu L, Li JJ, Wang Y, Wang XH, Wen Y, Qin WC, Su LM, Zhao YH (2015) Evaluation of toxicity data to green algae and relationship with hydrophobicity. *Chemosphere* 120:16–22
- McGrath JA, Parkerton TF, Di Toro DM (2004) Application of the narcosis target lipid model to algal toxicity and deriving predicted-no-effect concentrations. *Environ Toxicol Chem* 23(10):2503–2517
- Koenker R, Basset G (1978) Regression quantiles. *Econometrica* 46(1):33–50
- Bradbury SP (1995) Quantitative structure–activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicol Lett* 79(1–3):229–237
- Verhaar H, Leeuwen CV, Hermens J (1992) Classifying environmental pollutants. 1. Structure–activity relationships for prediction of aquatic toxicity. *Chemosphere* 25:471–491
- Ecotox_japan <https://www.env.go.jp/chemi/sesaku/02e.pdf>
- ECB <http://esis.jrc.ec.europa.eu/index.php?PGM=hpv>
- Russom CL, Anderson EB, Greenwood BE, Pilli A (1991) ASTER: an integration of the AQUIRE data base and the QSAR system for use in ecological risk assessments. *Sci Total Environ* 109–110: 667–670
- Faucon JC, Bureau R, Faisant J, Briens F, Rault S (1999) Prediction of the fish acute toxicity from heterogeneous data coming from notification files. *Chemosphere* 38(14):3261–3276
- Wang L, Chen K, Ong Y, Hwang C, Shim J (2005) A simple quantile regression via support vector machine. In: *Advances in natural computation*. Lect Notes Comput Sc 3610:512–520
- Meylan WM, Howard PH (1995) Atom/fragment contribution method for estimating octanol–water partition coefficients. *J Pharm Sci* 84(1):83–92
- Ghose AK (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J Phys Chem* 102:3762–3772
- Tetko IV, Tanchuk VY, Kasheva TN, Villa AE (2001) Estimation of aqueous solubility of chemical compounds using E-state indices. *J Chem Inf Comput Sci* 41(6):1488–1493
- Accelrys (2009) Pipeline Pilot, 7.5 edn. SciTegic Inc, San Diego
- Hahn M (1995) Receptor surface models. 1. Definition and construction. *J Med Chem* 38(12):2080–2090
- Labanowski JK, Andzelm JW (eds) (1991) *Density functional methods in chemistry*. Springer, New York
- Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140(4A):A1133
- Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron–gas correlation energy. *Phys Rev B* 45(23):13244
- Rohrbaugh RH, Jurs PC (1987) Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal Chim Acta* 199:99–109
- Stanton DT, Jurs PC (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal Chem* 62(21):2323–2329
- Ehresmann B, de Groot MJ, Alex A, Clark T (2004) New molecular descriptors based on local properties at the molecular surface and a

- boiling-point model derived from them. *J Chem Inf Comput Sci* 44(2):658–668
30. Parasurf <http://www.ceposinilico.de/products/parasurf.htm>
 31. Clark T, Alex A, Beck B, Chandrasekhar J, Gedeck P, Horn AHC, Hutter M, Martin B, Rauhut G, Sauer W, Schindler T, Steinke T (2008) VAMP. 10.0 edn. Computer-Chemie-Centrum. Universität Erlangen-Nürnberg, Erlangen,
 32. Bottcher CJF, Rip A, Van Belle OC, Bordewijk P (1952) Theory of electric polarization. Elsevier, Amsterdam
 33. Hopfinger AJ (1973) Conformational properties of macromolecules. Molecular biology. Academic, New York
 34. Koenker R (2005) Quantile regression. Cambridge University Press, New York
 35. He X, Shao Q (1996) A general bahadur representation of m-estimators and its application to linear regression with non stochastic designs. *Ann Stat* 24:2608–2630
 36. Durrieu G, Briollais L (2009) Sequential determination of sample size for robust linear regression: application to microarray experimental designs. *J Am Stat Assoc* 104(486):650–660
 37. Dodge Y, Jurečková J (1995) Estimation of quantile density function based on regression quantiles. *Stat Probab Lett* 23:73–78
 38. Koenker R (1994) Confidence intervals for regression quantiles. Springer, New York
 39. Koenker R (1996) Rank tests for linear models. Springer, New York
 40. Gutenbrunner C, Jurečková J, Koenker R, Portnoy S (1993) Tests of linear hypotheses based on regression rank scores. *J Nonparametr Stat* 2:307–333
 41. Parzen M, Wei L, Ying Z (1994) A resampling method based on pivotal estimating functions. *Biometrika* 81:341–350
 42. Biliyas Y, Chen S, Ying Z (2000) Simple resampling methods for censored regression quantiles. *J Econ* 99:373–386
 43. Kocherginsky M, He X, Mu Y (2005) Practical confidence intervals for regression quantiles. *J Comput Graph Stat* 14:41–55
 44. Khmaladze E (1981) Martingale approach in the theory of goodness-of-fit tests. *Theory Probab Appl* 26:240–257
 45. Koenker R, Xiao Z (2002) Inference on the quantile regression process. *Econometrica* 81:1583–1612
 46. Briollais L, Durrieu G (2014) Application of quantile regression to recent genetic and -omic studies. *Hum Genet* 133:951–966
 47. Liu Y, Zou C, Zhang R (2008) Empirical likelihood ratio test for a change-point in linear regression model. *Commun Stat Theory Methods* 37:2551–2563
 48. Kubinyi H (1977) Quantitative structure–activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J Med Chem* 20(5):625–629
 49. Caputo B, Sim K, Furesjo F, Smola A (2002) Appearance-based object recognition using SVMs: which kernel should I use? In: Proc of NIPS workshop on statistical methods for computational experiments in visual processing and computer vision, Whistler
 50. Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw* 17(1): 113–126
 51. Matterna D, Haykin S (1999) Advances in kernel methods. MIT Press, Cambridge, pp 211–241
 52. Thioulouse J, Dray S (2007) Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *J Stat Softw* 22(5):1–14
 53. Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) Kernlab — an S4 package for kernel methods in R. *J Stat Softw* 11(9):1–20
 54. Escher BI, Bramaz N, Eggen RIL, Richter M (2005) In vitro assessment of modes of toxic action of pharmaceuticals in aquatic life. *Environ Sci Technol* 39(9):3090–3100
 55. Vaes WH, Ramos EU, Hamwijk C, van Holsteijn I, Blaauboer BJ, Seinen W, Verhaar HJ, Hermens JL (1997) Solid phase microextraction as a tool to determine membrane/water partition coefficients and bioavailable concentrations in in vitro systems. *Chem Res Toxicol* 10(10):1067–1072
 56. Michielan L, Pireddu L, Floris M, Moro S (2010) Support vector machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals. *Mol Inf* 29(1–2):51–64
 57. Vapnik V (1998) Statistical learning theory. Wiley, New York
 58. DROPPDATA (2009) A guide to pesticides grouped by mode of action. http://www.dropdata.org/RPU/pesticides_MoA.htm
 59. Munday R (1989) Toxicity of thiols and disulphides: involvement of free-radical species. *Free Radic Biol Med* 7(6):659–673
 60. Mavridis L, Hudson BD, Ritchie DW (2007) Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J Chem Inf Model* 47(5):1787–1796
 61. Escher BI, Hermens JL (2002) Modes of action in ecotoxicology: their role in body burdens, species sensitivity, QSARs, and mixture effects. *Environ Sci Technol* 36(20):4201–4217

Supporting information

The data are accessible through the website <http://www.cermn.unicaen.fr> without user registration. The R codes are also available from the authors upon request.