Journal of Hazardous Materials 304 (2016) 233–241

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

# A best comprehension about the toxicity of phenylsulfonyl carboxylates in *Vibrio fischeri* using quantitative structure activity/property relationship methods

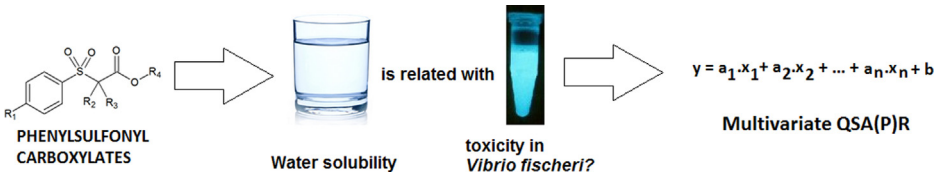
Eduardo Borges de Melo[a,*], João Paulo Athaíde Martins[b], Eduardo Hösel Miranda[a], Márcia Miguel Castro Ferreira[c]

[a] *Western Paraná State University, UNIOESTE, Cascavel, PR, Brazil*
[b] *Instituto de Educação Superior de Brasília, IESB, Brasília, DF, Brazil*
[c] *Institute of Chemistry, University of Campinas, UNICAMP, Campinas, SP, Brazil*

## HIGHLIGHTS

- Aromatic sulfones: chemicals used in agrochemical and pharmaceutical industries.
- A QSAR study about their environmental toxicity was carried out.
- It also assessed the relationship between this endpoint and the water solubility.
- The models, including by PLS2, indicate some relation between the endpoints.
- However, the results indicate that the toxicity is due to enzyme inhibition.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 28 July 2015
Received in revised form 17 October 2015
Accepted 22 October 2015
Available online 28 October 2015

*Keywords:*
Aquatic toxicity
Water solubility
Environmental risk
Partial least squares
Quantitative structure-activity relationships

## ABSTRACT

Aromatic sulfones comprise a class of chemicals used in agrochemical and pharmaceutical industries and as floatation and extractant agents in petrochemical and metallurgy industries. In this study, new QSA(P)R studies were carried out to predict the toxicity against *Vibrio fischeri* of a set of 52 aromatic sulfones. The same approach was used to evaluate the relationship between these endpoint and the water solubility, another important environmental endpoint. The study resulted in models of good statistical quality and mechanistic interpretation with a possible correlation between the two endpoints, but the toxic effect is also likely to depend on other physicochemical properties. The use of the PLS2, a method not commonly used in QSA(P)R studies, also produced models of greater reliability, and the relationship between the two endpoints was reinforced to some degree. These results are useful for better understanding the process by which these compounds exert their environmental toxicity, thus aiding in the development of industrially useful compounds with less potential environmental damage.



## 1. Introduction

Developmental toxicity involves several issues and adopts a variety of experimental methods. The complexity, time, and cost of the experiments and the late recognition of the importance have resulted in a low number of available studies [1]. Nevertheless, information on toxicity of chemicals is required in order to estimate

* Corresponding author at: Cell Dept. of Pharmacy, 2069 Universitária St, 85819110, Cascavel, PR, Brazil.
*E-mail address:* eduardo.b.de.melo@gmail.com (E.B. de Melo).

http://dx.doi.org/10.1016/j.jhazmat.2015.10.047
0304-3894/

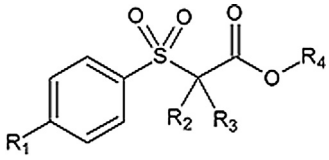



**Fig. 1.** Basic structure of the dataset.

the risk of chemical substances on living organisms. A detrimental effect of a chemical can be expressed after a short-term and/or a long-term exposure [2].

Thus, the rising interest in finding alternative tests for chemicals has led to the investigation of several options. The bacterial assays are based in part on the fact that bacteria are an integral part of the aquatic ecosystem, the easiness and speed of the assays, and the assumption that most chemicals exert their effects by interfering with common cellular processes. Among the available tests, the Microtox test measures the reduction in light output of natural luminescence of marine bacteria *Vibrio fischeri* upon exposure to the toxicant under study [3].

Some physicochemical properties, especially those related to partition processes, are also related to toxicity, because this property may be dependent of the pharmacokinetics of the chemicals [4]. However, experimental determination of these properties can also be a laborious, time consuming, and expensive procedure. Such situation and the existence of a vast amount of new molecules are a challenge for chemical databases, which have to remain constantly updated [5].

Nowadays, application of quantitative structure-activity/property relationship (QSA[P]R) methods plays a crucial role in ecotoxicological modeling. The goal is to develop acceptable correlation between molecular structure and a endpoint of interest of a set of chemicals [6]. An important initiative, the legislation of Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) considers the QSA(P)R as a preliminary tool related to the safety of chemicals [7].

A class of compounds of great industrial interest comprises aromatic sulfones. These compounds are used as reactants in agrochemical and pharmaceutical industries and also as flotation agents and extractants in petrochemical industry and metallurgy. As a consequence, they can impact the environment [8,9]. Hence, many authors performed several QSA(P)R studies on two datasets related to aromatic sulfones [8–14]. Thus, the objective of this work is to obtain QSA(P)R models related to the toxicity to *V. fischeri* ($pEC_{50}$) of phenylsulfonyl carboxylates (Fig. 1 and Fig. S1), and explore the relationship between this endpoint and the logarithm of water solubility ($Log S_w$), an important environmental endpoint related with the partition processes.

## 2. Material and methods

### 2.1. Data set

The data set of 52 phenylsulfonyl carboxylates was originally aggregated from Liu et al. [8], which provided toxicity values against *V. fischeri* for 56 samples (1–56). In turn, Liu et al. [12] provided values for Log*Sw* (samples 1–28 and 37–60). A subset of this data (1–28) has been widely used for the development of QSA(P)R models of environmental interest (including Log*Sw* and toxicity) [11].

### 2.2. Molecular descriptors

The SMILES strings were obtained for each compound. The use of this approach aimed to acquire the descriptors in a quick and easy way, as well as more easily reproduced. These strings were used to generate descriptors of several classes (see Supplementary material). The total number of descriptors was reduced to 224 descriptors for matrix of $pEC_{50}$, and 194 for matrix of $Log S_w$ by: (i) eliminating the constant and near-constant variables; (ii) eliminating the variables with standard deviation less than 0.001; (iii) removing variables with pair correlation larger than or equal to 0.90; and (iv) eliminating those that presented the absolute value of correlation coefficient, $|r|$, with each endpoint lower than 0.2. The SMILES notations were also used to generate values for calculated $Log S_w$.

### 2.3. Variable selection

Selection of the most important descriptors was carried out with the ordered predictors selection (OPS) algorithm [15–20]. This approach rearranges data matrix according to informative vectors, placing the most important descriptors in the first columns. Then, successive partial least squares (PLS) regressions are built by increasing the number of descriptors, with the goal of finding the set that create the best latent variables (LVs) for correlation with the endpoints under study [11,15,21]. Three informative vectors were used: correlation vector, regression vector, and their product pointwise. The obtained models were classified according to their coefficient of determination of leave-one-out cross-validation ($Q^2_{LOO}$). The process is repeated in an iterative manner, until the best combination between $Q^2_{LOO}$ and number of descriptors/LVs was obtained according to the user criteria.

### 2.4. Model building

Currently, the number of descriptors available for QSAR studies is very large. Thus, the risk of obtain models with highly correlated descriptors, a serious problem in multivariate calibration, is enhanced. In this scenario, the use of PLS is suitable. This projection method is a regression approach for modeling a relationship between dependent ($Y$) and independent ($X$) variables, that reduces the dimensionality of the data, while retaining most of the variation generating latent variables (LV). PLS models both $X$ and $Y$ simultaneously to find LVs from the original descriptors in $X$ that will predict the latent variables in $Y$. Therefore, even if several descriptors are selected for a model, the number of LVs is in general much smaller. The new LVs are orthogonal, and thus the problem of collinearity is also avoided [22].

PLS1 use one response at a time, while PLS2 can handle several responses, simultaneously. In PLS modeling, the data is divided into two groups of variables, $X$ and $Y$. The objective in PLS modeling is to model $X$ in such a way that the information in $Y$ can be predicted as well as possible. PLS maximizes the covariance between matrices $X$ and $Y$. PLS2 is used when more than one dependent variable is under study, and its application is justified only if the responses in $Y$ matrix are correlated. Thus, it is difficult to find QSAR/QSPR studies using PLS2 regression due to the generally low correlation among response under investigation.

As the numerical range of each selected descriptor may be very different, the data in QSAR should normally be pre-processed by the auto-scaling scheme, which was applied herein. This procedure was also carried out during the variable selection step [21,23].

The relationship between $pEC_{50}$ and calculated $Log S_w$, with the first as a dependent and the second as a predictor variable was also checked using univariate regression. Experimental $Log S_w$ were not considered as dependent variable to allow the use of obtained model for prediction purposes. It was also verified if the model of toxicity was statistically improved if the calculated $Log S_w$ was added to the model. Finally, models were obtained for $pEC_{50}$ and $Log S_w$ by combining the selected descriptors in the initial stage of the study and by applying PLS2 to a set of 48 samples (1–28

and 37–56) that had experimental $pEC_{50}$ and $LogS_w$. The number of descriptors of the PLS2 models was reduced based on the importance of auto-scaled values of regression coefficients. The *r* between $pEC_{50}$ and $LogS_w$ of 0.791 was considered sufficiently high to justify its application.

The presence of outliers was verified by analyzing the graph of studentized residuals ($\sigma$) versus leverages ($h$) [24]. Those compounds with $\sigma > 2.5$ and values of $h$ higher than the cutoff limit, $h^*$ ($h^* = 3p/n$; *p:* number of latent variables; *n:* number of samples), are seen as detrimental to the quality of the model and therefore can be considered outliers. In cases where $h < h^*$ but $\sigma \approx 2.5$, the compounds were also analyzed for their possible influence as outliers in the model.

In the PLS2 modeling, the outliers were checked by the distance to the model in the *X* and in the *Y* matrices containing the dependent variables (DModX and DModY) [22]. For the training set, this may correspond to samples which are detrimental to the quality of the model. For the test set, samples can be classified as anomalous when they are predicted outside the domain defined by the training set (extrapolation). That also assists in determining the adequacy of the predicted results in the model's applicability domain (AD) [1]. This strategy was also used to determine the AD of all models in the next section, the validation step.

### 2.5. Validation

The most used approach for validation of QSAR models involves internal and external validation. The approaches for calculating the recommended statistics parameters for these steps are available in previously published studies [23,25,26]. All obtained models have undergone the same quality assessments.

During internal validation, the explained variance of the model is checked using tests that measure the quality of the data fit, the significance of the model, and the predicted variance using the LOO cross-validation: coefficient of determination ($R^2$), the associated root mean square error of calibration (*RMSEC*), *F*-ratio test (95% confidence interval, $\alpha = 0.05$), the $Q^2_{LOO}$, the root mean square error of cross-validation (*RMSECV*), and the scaled average and difference values of the RmSquare metrics in the LOO cross-validation (average $r^2_m$(LOO)-scaled and $\Delta r^2_m$(LOO)-scaled) [25,26]. In this step, the robustness (low sensitivity of the model to small and deliberate variations) and the chance correlation (explained and predicted variances are not due to spurious correlation) should also be checked [1].The leave-*N*-out cross-validation (LNO, $N = 1$ to 13, repeated six times for each *N*) was used to check the robustness [23]. Chance correlation was assessed by the ***y***-randomization test using the approach suggested by Eriksson et al. (2003): the $|r|$ between the original and each randomized vector ***y*** was obtained, and two regression lines were constructed, both using these correlations in the x-axis, one with $R^2$ in the *y*-axis, and other with the $Q^2$LOO in the *y*-axis.

For external validation, the data set was split into training and test sets for each endpoint using hierarchical cluster analysis (HCA) [27], where the entire endpoints range was represented. Employing the test set, the predictability was tested using the coefficient of determination of external validation ($R^2_{pred}$), the associated root mean square error of prediction (*RMSEP*), the Golbraikh–Tropsha's slopes ($k$ and $k'$), the absolute difference of the determination coefficient of the linear relation between the observed and predicted values without an intercept ($R^2_0$), and the predicted and observed values without an intercept ($R'^2_0$) ($|R^2_0 - R'^2_0|$), and by the average $r^2_m$(pred)-scaled and $\Delta r^2_m$(pred)-scaled [25,26,28].

### 2.6. Software

The 2D structures and SMILES strings were built in the ChemSketch (www.acdlabs.com). The obtention of molecular descriptors and the three fist steps of variable reduction were carried out using the Dragon 6 (http://www.talete.mi.it). The calculated $LogS_w$ (ALOGpS) was obtained by the ALOGPS2.1 (http://www.vcclab.org/lab/alogps). The QSAR Modeling [21] (http://lqta.iqm.unicamp.br) was used for the last step of the variable reduction, variable selection, internal validation, outlier detection, LNO cross-validation, and *y*-randomization test. The RmSquare metrics were obtained in http://aptsoftware.co.in/rmsquare BuildQSAR [29] was used for univariate regression between $pEC_{50}$ and ALOGpS. The PLS2 study was conducted in XLSTAT 2014 demonstration version (http://www.xlstat.com). The LNO cross-validation and **y**-randomization test for PLS2 models were carried out in Matlab (www.mathworks.com) with an in house algorithm. Pirouette 4 (http://www.infometrix.com) was used for the test set selection.

## 3. Results and discussion

### 3.1. PLS models

In the initial step, were built models both for $pEC_{50}$ and $LogS_w$. The models obtained by the use of OPS were refined using Pirouette 4. The final models for both endpoints were based on four molecular descriptors (Table 1) and two LVs. For model 1, two LVs cumulate 88.039% of variance (LV1: 54.493%; LV2: 33.546%). For model 2, the cumulative variance was 86.470% (LV1: 65.093%; LV2: 21.377%).

$$pEC_{50} = -4.301 + 3.587^*(SpMAX\_A) - 0.792^*(ATS4m) - 0.018^*(CATS\_2D\_07\_LL) - 0.663^*(SpMin8\_Bh(s)) \quad (1)$$

$$LogS_w = -4.193 - 0.018^*(P\_VSA\_v\_3) + 4.305^*(GATS2e) - 2.503^*(SpMin8\_Bh(s)) - 0.596^*(SM02\_AEA(dm)) \quad (2)$$

The internal validation (Table 2 and Table S1) shows that both models provided adequate explained and predicted information and were statistically significant. The maintenance of the descriptors' signs in relation to individual *r* with endpoints (Table S2) indicates that both models are self-consistent [23]. The difference between $R^2$ and $Q^2_{LOO}$ is very low, an indication that both models were not over-fitted [23,30]. Both models were approved in the *y*-randomization test and in the LNO cross-validation. In the first test (Fig. 2A and B), the intercepts for both models were within the recommended limits (lower than 0.3 for the explained and 0.05 for the predicted variances) [31]. The second test (Fig. 2C and D) indicated that both models can be considered robust, with average $Q^2_{LNO}$ (0.856 and 0.870) close to their respective $Q^2_{LOO}$ (0.860 and 0.870). The standard deviation for each *N* value was small, with maximum in $Q^2_{L11O}$ for both models (0.025 for model 1 and 0.013 for model 2) [23].

External validation is necessary for a model to be considered realistic and applicable to a prediction [32]. The test sets were formed by representative compounds within the variation range of endpoints and with a good structural variability. The results (Table 2) show that the models have high external predictive power. The $R^2_{pred}$, a test that is usually used to measure quality of the external predictive power, was higher than the threshold (0.5) for both models. The $k$, $k'$, $|R^2_0 - R'^2_0|$, average $r^2_m$(pred)-scaled, and $\Delta r^2_m$(pred)-scaled values were also within the recommended limits [25,26].

The QSA(P)R model is a reductionist model, and it is inevitably associated with the limitations related to chemical structures and the corresponding molecular descriptors and with the mechanisms related to each of the studied endpoints to generate reliable predictions. Thus, the AD is very useful to define boundaries, whereby

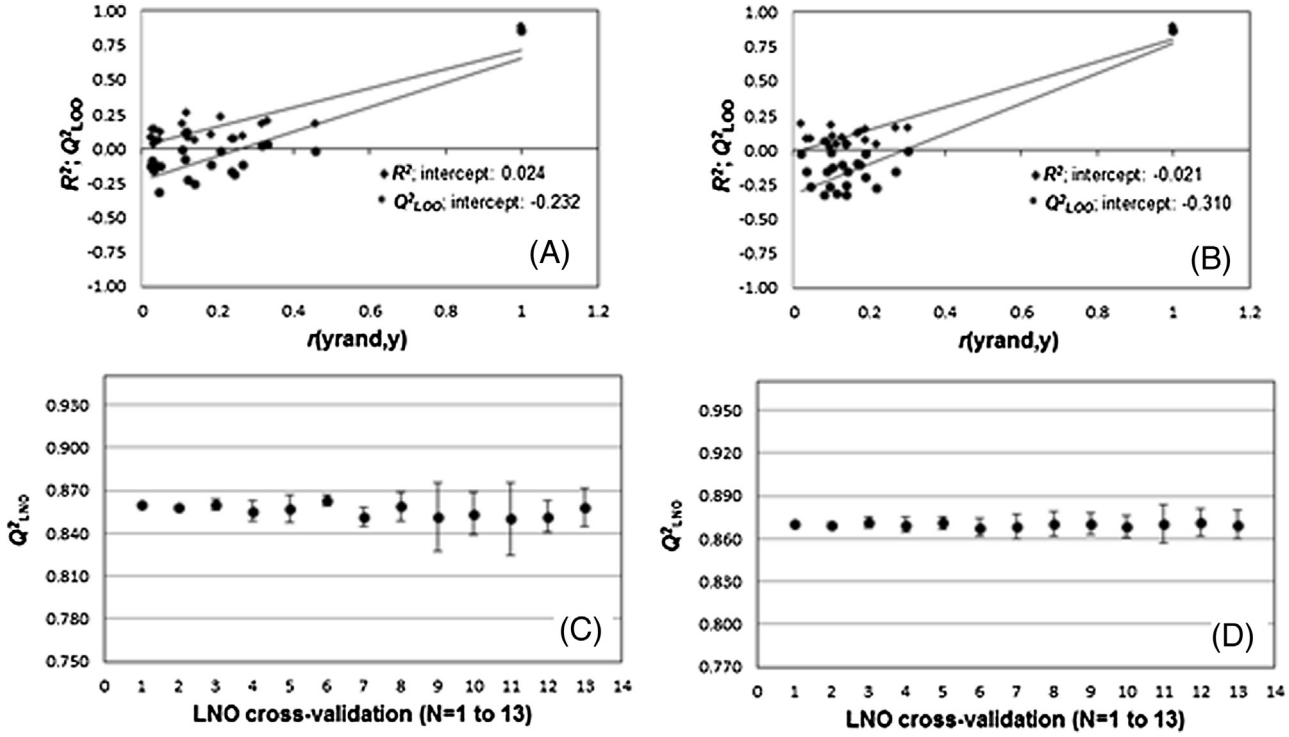



**Fig. 2.** Results of the ***y***-randomization test (A and B) and leave-*N*-out (LNO) cross-validation (B and C).

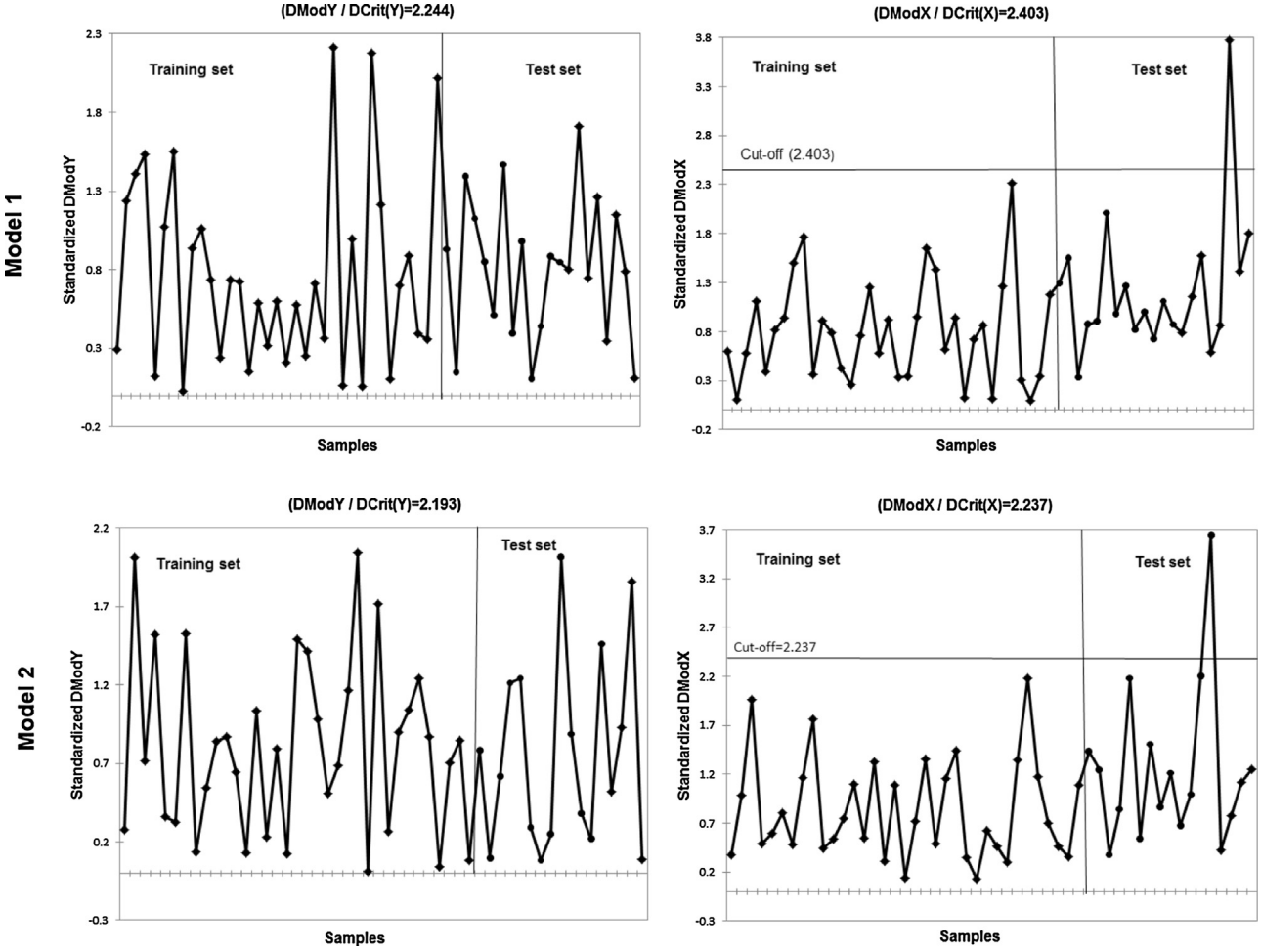



**Fig. 3.** Applicability domains (DModY and DModX method) of the training and the test sets of models 1 and 2.

**Table 1**
Descriptors selected in the obtained models.

| Symbol | Descriptor | Class |
|---|---|---|
| ATS4m | Broto–Moreau autocorrelation lag 4 weighted by mass | 2D autocorrelations |
| GATS2e | Geary autocorrelation lag 2 weighted by Sanderson electronegativity | |
| CATS_2D_07_LL | CATS2D Lipophilic–Lipophilic lag 07 | CATS2D fingerprint |
| P_VSA_v_3 | P_VSA-like on van der Waals volume, bin 3 | P_VSA-like descriptors |
| SM02_AEA(dm) | Spectral moment order 2 from augmented edge adjacency matrix weighted by dipole moment | Edge adjacency indices |
| SpMAX_A | Lovasz–Pelikan index | 2D matrix-based descriptors |
| SpMin8_Bh(s) | Lowest eigenvalue 8 of Burden matrix weighted by I-state | Burden eigenvalues |

**Table 2**
Results of the internal and external quality for all models presented in this study.

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Expected result |
|---|---|---|---|---|---|---|---|
| $n$[a] | 41 | 37 | 41 | 40 | 35 | 35 | – |
| $p$[b] | 2 | 2 | – | 2 | 2 | 2 | – |
| Internal quality | | | | | | | |
| $R^2$ | 0.884 | 0.890 | 0.704 | 0.934 | 0.804 | 0.854 | >0.6 |
| *RMSEC* | 0.162 | 0.394 | 0.262 | 0.121 | | | Lowest possible |
| $F$ (table value at $\alpha = 0.05$) | 144.988 (3.245[c]) | 138.954 (3.276[d]) | 92.958 (4.091[e]) | 261.843 (3.252[f]) | 65.633 (3.295[g]) | 93.589 (3.295[g]) | Higher than table value |
| $Q^2_{LOO}$ | 0.860 | 0.870 | 0.680 | 0.919 | 0.756 | 0.822 | >0.5 |
| *RMSECV* | 0.172 | 0.412 | 0.269 | 0.128 | 0.214 | 0.423 | Lowest possible |
| Average $r^2$m(LOO)-scaled | 0.786 | 0.805 | 0.558 | 0.886 | 0.729 | 0.794 | > 0.5 |
| $\Delta r^2$m(LOO)-scaled | 0.089 | 0.087 | **0.219** | 0.049 | 0.151 | 0.119 | <0.2 |
| $R^2 - Q^2_{LOO}$ | 0.024 | 0.020 | 0.024 | 0.015 | 0.048 | 0.032 | Lower than 0.2–0.3 [23]; lower than 0.1 [30] |
| External quality | | | | | | | |
| $R^2_{pred}$ | 0.873 | 0.865 | 0.641 | 0.845 | 0.825 | 0.736 | >0.5 |
| *RMSEP* | 0.150 | 0.398 | 0.250 | 0.166 | 0.192 | 0.376 | Lowest possible |
| $k$ | 0.987 | 0.987 | 1.027 | 1.01 | 0.959 | 1.002 | 0.85 < k < 1.15 |
| $k'$ | 1.002 | 1.004 | 0.943 | 0.976 | 1.023 | 0.991 | 0.85 < k < 1.15 |
| $\lvert R^2_0 - R'^2_0 \rvert$ | 0.014 | 0.007 | **0.412** | 0.031 | 0.169 | 0.028 | >0.3 |
| Average $r^2_m$(pred)-scaled | 0.816 | 0.812 | **0.433** | 0.777 | 0.554 | 0.677 | >0.5 |
| $\Delta r^2_m$(pred)-scaled | 0.093 | 0.044 | **0.292** | 0.124 | 0.198 | 0.035 | <0.2 |
| Overall statistics | | | | | | | |
| Average $r^2_m$(overall)-scaled | 0.810 | | | 0.861 | | | >0.5 |
| $\Delta r^2_m$(overall)-scaled | 0.092 | | | 0.068 | | | <0.2 |

[a] Number of samples in the training set.
[b] Number of latent variables or descriptors.
[c] For $p = 2$ and $n$-$p$-1 = 38.
[d] For $p = 2$ and $n$-$p$-1 = 34.
[e] $p = 1$ and $n$-$p$-1 = 39.
[f] $p = 2$ and n-p-1 = 37.
[g] For $p = 2$ and $n$-$p$-1 = 3.

the obtained predicted endpoints of the test set can be trusted with confidence [24]. Both DModX and DModY method did not determine the AD for compound 50 in model 1 and model 2 (Fig. 3), which is an unwanted outcome. However, both models adequately predicted 93.3% of the test set ($n = 14$), which can be considered acceptable.

Both models are of reasonable statistical quality. However, it is always desirable to obtain a model where the molecular descriptors are related with the endpoint. Interpreting a QSA(P)R model in terms of the contribution of molecular descriptors is not straightforward and therefore, it is in general a difficult task [33]. Geometric (not used) and topological descriptors are generally more difficult to interpret than, for example, Log P and constitutional descriptors.

One of the selected descriptors of the simplest interpretation is CATS2D_07_LL. The CATS2D descriptors define pharmacophore points, and the selected descriptor is related to two lipophilic acceptor points (L) at the topological distance of 7 (i.e., a distance of seven bonds) [34]. The higher the number of lipophilic groups in the molecules, the lower the toxic tendency of the compound.

SpMAX_A is a molecular branching index that is calculated for special types of trees such as path or star. Halder et al. [35] have suggested that a negative coefficient for this descriptor indicates that lesser branching in the overall structure may favor the endpoint under study. This result is in agreement with Roy and Ghosh [14] who proposed that the toxicity of a compound decreases with increased branching.

Weighting factors are usually used for topological descriptors to encode in a numerical form the information about atom and bond properties that are not represented in a simple molecular graph [26]. ATS4m and P_VSA_v_3 were weighted by steric properties, and both have negative coefficients. Since the weighting factor is directly proportional to the result of the descriptor, this indicates that larger molecules tend to be less toxic and less soluble in water. Roy and Ghosh [14] also proposed that the toxicity of these compounds tends to decrease with the increase in their size.

Electronic properties were used in three descriptors: SpMin8_Bh(s), GATS2e, and SM02_AEA(dm). The value of descriptor GATS2e tends to increase with increasing the number of electronegative atoms in a molecule [36]. In SM02_AEA(dm), where dm is the dipole moments of the chemical bonds [36], the negative coefficient indicates that the presence of unsaturated bonds, which have higher dm, can be detrimental to the toxic activity, also similar to Roy and Ghosh proposal [14].

It is noteworthy that electronic weighting factors predominate in model 2, indicating the importance of the formation of electrostatic interactions between the molecules and water for an adequate solubility. This is in agreement with Liu et al. [12], whose 3D-QSPR study suggested that great electrostatic interaction of negatively charged atoms with water molecules increases the solvation of solute. However, this type of interaction also appears between a ligand and its site of biological action, which may explain the presence of the weighting by electrotopological intrinsic state (s) in SpMin8_Bh(s). This property may be viewed as the ratio of the $\pi$ and the lone pair electron count [37]. The negative coefficient indicates that substitutions with phenyl group are detrimental to the toxic activity.

Interestingly, the increase of hydrophobicity and size are detrimental to the toxic effect, suggesting that these compounds do not cause toxicity by narcosis or nonspecific toxicity, a physicochemical process directly proportional to these two properties [38]. Therefore, as suggested by Liu et. al. [8,9], is possible that these compounds cause specific toxicity in the Microtox test by inhibition of luciferase, the enzyme responsible for the bioluminescence of *V. fisheri*.

The results of the validation and evaluation of the models were compared with previously published studies [8–10,12–14]. Information about each model is available in Table S3. However, because these studies were not standardized, it was possible to perform only a simple comparison between the metrics used in each study with the equivalent ones used herein. It is important to point out that the obtained results are not an indication that a particular model is better or worse than the other.

In addition, the results obtained from model 1 indicated that, despite having the smallest explained variance compared to previously published models, it has predicted the highest variance among all other models; it was slightly higher (by 0.012 units) than the a previously MLR model [14]. Similarly, higher internal prediction was observed in model 2 compared to the models published by Liu et al. [12]. Goodarzi and Freitas [10] did not provide the results for LOO cross-validation and therefore their comparison was not possible.

Unlike other previously published models, model 1 and 2 were both tested for internal quality related to robustness and presence of chance correlation and for external quality, and both models were extensively validated, including by the test of the AD of the predictions. Only a study [10] carried out tests of external quality, which were, considering the coefficient $R^2$pred, inferior to model 1 (0.078 and 0.005 units). The results, particularly those related to the external validation and applicability domain, suggest that models 1 and 2 are more reliable compared to the previously published models.

### 3.2. Univariate model: relation between $pEC_{50}$ and $LogS_w$

Given that the Microtox test for *V. fischeri* is carried out in aqueous medium and that water solubility is inversely proportional to the partition coefficient [39,40], the following question was raised: is there a correlation between the Log$S_w$ and $pEC_{50}$? To answer this question, initially we built a model using $pEC_{50}$ as a dependent and Log$S_w$ as an independent variable. However, the values for both variables were available for only 48 samples, and the use of experimental Log$S_w$ as a descriptor limits the size and structural representation of the dataset and the possible use for prediction purposes. Thus, we used a calculated Log$S_w$ (ALOGpS):

$$pEC_{50} = 0.401(\text{ALOGpS}) + 3.059 \quad (3)$$

The results (Table 2) show that ALOGpS alone explained 70.5% and predicted 68% of the variance, showed adequate results for ***y***-randomization and LNO cross-validation tests (Fig. S2). However, the model was approved only in one of the $r^2_m$(LOO) tests. Moreover, the external validation did not approve the model in all tests (Table 2). These results show a correlation between the aqueous solubility and $pEC_{50}$, but not sufficient for prediction purposes. These results are also plausible from a biological standpoint, since toxicity endpoints are nothing more than "biological activities": the toxicity does not depend only on the quantity of molecules that will penetrate the bacteria, but also how these molecules will interact with cell structures.

### 3.3. Combination of models 1 and 3

The good internal results of model 3 allowed combining ALOGpS with model 1 and generating a model that could explain in more detail the toxicity of the studied compounds. The model 4, also self-consistent [23] (Table S4), was constructed with two LV (LV1 59.895%; LV2 29.262%) and showed an outlier (31). Thus, model 4 appears to be able to explain and predict larger amounts of information than model 1. Moreover, with the inclusion of information from ALOGpS, auto-scaled coefficient of CATS_2D_LL was changed (−0.239– −0.144). A test showed that this descriptor could be eliminated without affecting the model. Thus, model 4 also has four descriptors. As models 1 and 4 were constructed based on the same amount of LVs, both have similar tabulated *F*; hence, it was possible to compare the significance of equations, where model 4 is almost two-fold more significant than model 1. Model 4 also shows adequate results for internal prediction ($Q^2_{LOO}$ = 0.919; predicted values and residuals are given in Table S5), *y*-randomization, and LNO cross-validation tests (Fig. 4). The largest standard deviation obtained during LNO cross-validation test at only 0.010 for $Q^2_{L10O}$ and $Q^2_{L13O}$ can be considered negligible, and the difference between $Q^2_{LOO}$ and the average $Q^2_{LOO}$ was only 0.01.

$$pEC_{50} = 0.183(\text{ALOGpS}) + 2.876(\text{SpMAX\_A}) - 0.505(\text{ATS4m}) - 0.632(\text{SpMin8\_Bh(s)}) - 3.003 \quad (4)$$

The quality of the external prediction was equivalent to model 1, with $R^2_{pred}$ only 0.028 units lower. Since the main objective of the QSA(P)R model is prediction, the $r^2_m$(overall) metrics was applied. One application of this metrics is to aid in selecting the most suitable prediction model when several models are obtained [41]. Despite the small difference between the obtained results (Table 2), it can be proposed that model 4 provides the highest overall prediction and can be considered more suitable for prediction.

The signal of the ALOGpS in model 4 indicate that increasing the aqueous solubility of the derivatives leads to more toxic compounds. The maintenance of information regarding the interpretation of models 1 and 2 ensures higher level of reliability for model 4. The AD results (Fig. 5) showed that three compounds, two in DModY (**7** and 36) and one in DModX (50), are not in the AD, which is equivalent to 20% of the test set. Although this can still be considered a good result, such increased number of samples outside the AD can be related to the poorer quality of the variance for external predictions due to the use of the descriptor ALOGpS. Furthermore, the behavior for sample 50 as an extrapolation is not unexpected, since the same behavior was observed in models 1 and 2.

### 3.4. Multivariate PLS2 study

The final analysis of the correlation between Log$S_w$ and $pEC_{50}$ was performed using PLS2, which led to models 5 and 6. Both models were successful in all tests (Figs. 6 and 7, and Supplementary material, Table S6). Both models also feature self-consistency

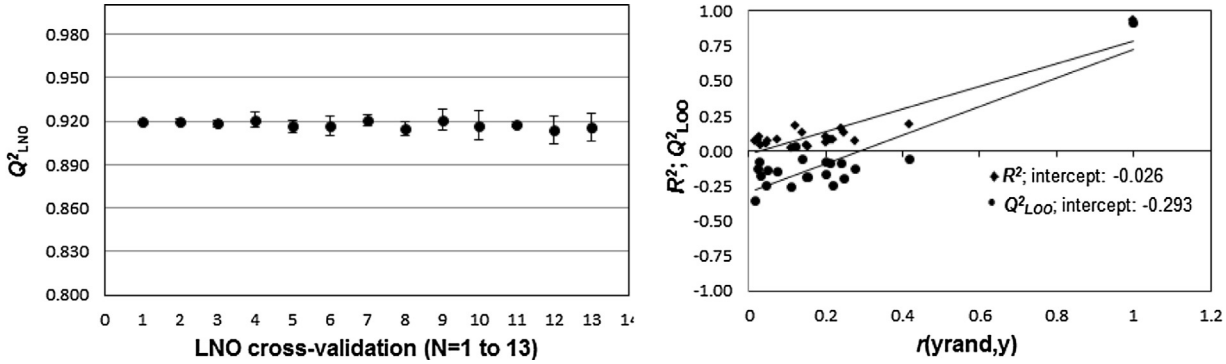



**Fig. 4.** Results of the *y*-randomization and leave-*N*-out (LNO) cross-validation tests for model 4.

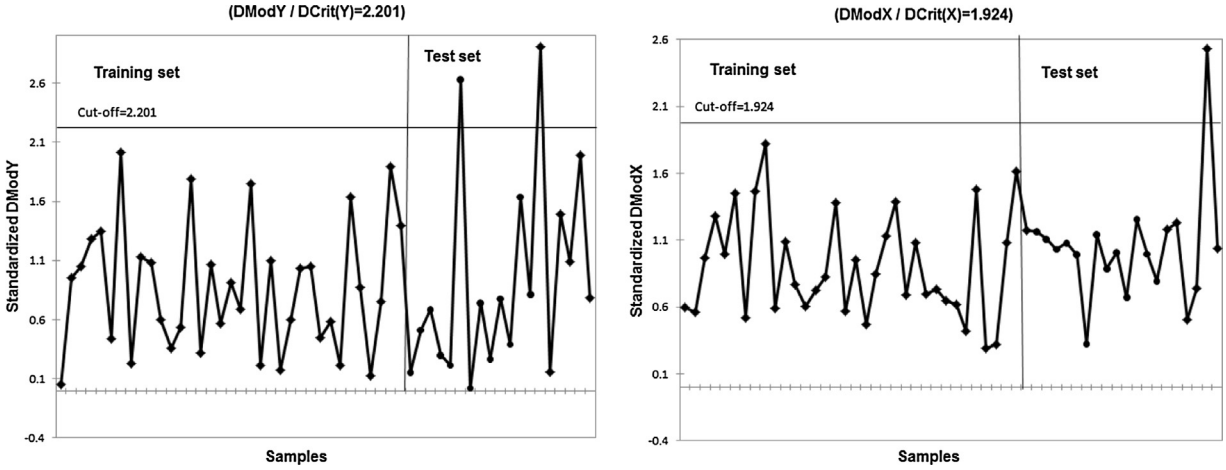



**Fig. 5.** Applicability domains (DModY and DModX method) of the training and the test set of model 4.

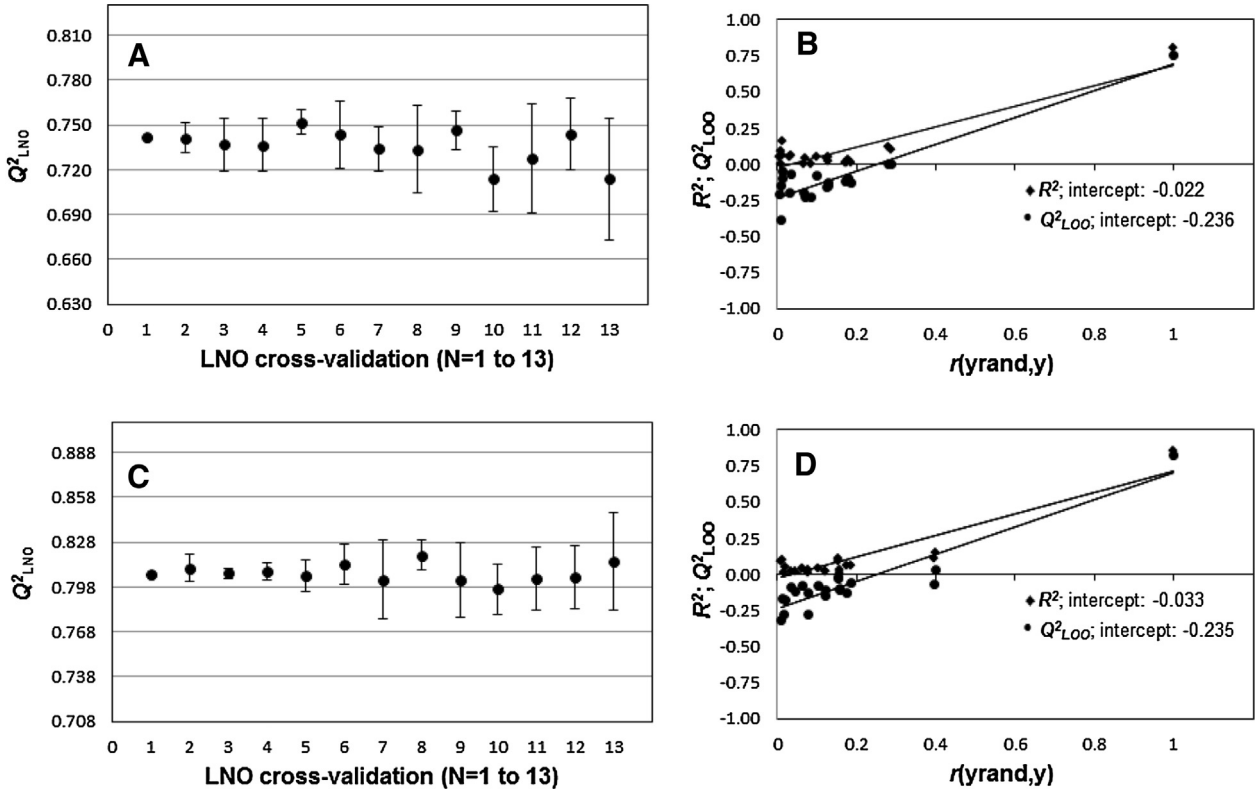



**Fig. 6.** Results of the *y*-randomization test (A and B) and leave-*N*-out (LNO) cross-validation (B and C)—models 5 and 6.

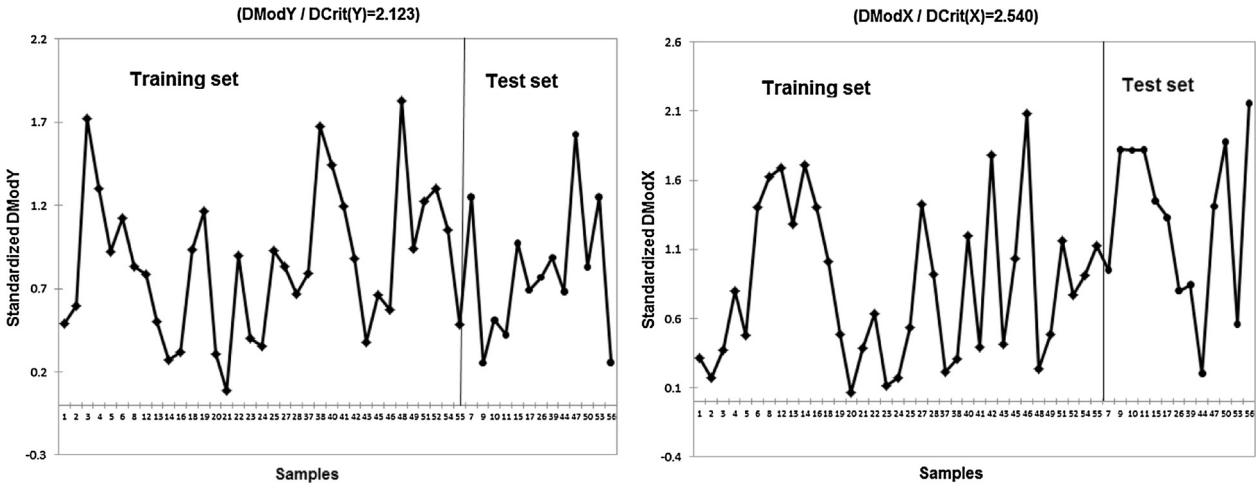



**Fig. 7.** Applicability domains (DModY and DModX method) of the training and the test sets obtained in PLS2 modeling. A single plot is created for both endpoints because the PLS2 regression analyses were carried out simultaneously.

[23] (Table S7). Of the original set, three descriptors were eliminated based on their auto-scaled regression coefficients, and the new models has descriptors with the same sign for the coefficients and independent term, indicating that both endpoints are actually interrelated.

$$pEC_{50} = 4.603 - 0.005(P\_VSA\_v\_3) - 0.352(SM02\_AEA(dm)) - 0.599(ATS4m) - 0.009(CATS2D\_07\_LL) \quad (5)$$

$$Log S_w = 2.944 - 0.010(P\_VSA\_v\_3) - 0.455(SM02\_AEA(dm)) - 1.347(ATS4m) - 0.059(CATS2D\_07\_LL) \quad (6)$$

However, this relationship is not direct. The absolute values of the auto-scaled coefficients (Table S6) show that the order of importance is not the same. In model 5, the order is SM02_AEA(dm) > ATS4m > P_VSA_v_3 > CATS2D_07_LL. In model 6, is CATS2D_07_LL > ATS4m > P_VSA_v_3 > SM02_AEA(dm). This change may have two explanations: (i) the correlation between the endpoints is 0.795, indicating that, although related, there is other relevant information for each phenomenon; and (ii) while there is a predominance of a molecular descriptor weighted by dipole moment in model 5, which can be interpreted as the importance of some polarization phenomena in a ligand site, the descriptor that predominates in model 6 may be related to solubilization, since its sign is negative, indicating that higher descriptor value (i.e., more hydrophobic) means its lower solubility in water. Thus, given these potential explanations, the PLS2 models are capable of showing correlation between the endpoints, with a cautionary note that these are different phenomena.

Another important result in the PLS2 study is related to the AD. It is expected that a single model is capable of explaining both endpoints, as suggested by obtaining a single DModX and DModY for both endpoints. Finally, the results of the external prediction of the PLS2 regression fit in the AD. Therefore, despite some statistical results being lower than the results obtained in other models (Table 2), this method was able to generate suitable regressions for external prediction, i.e., for samples not used in the modeling process.

## 4. Conclusions

In this work, we have shown that the OPS method is a useful tool in the construction of QSA(P)R models for toxicity against *V. fischeri*, an important environmental organism, using only molecular descriptors calculated by SMILES strings. Models 1 show good internal and external quality, and also provide satisfactory results for the AD. In addition, information found in previous studies strengthens the hypothesis that smaller and less branched compounds tend to be less toxic. The correlation between $pEC_{50}$ and water solubility (as a molecular descriptor, ALOGpS) reveals that the toxicity is not only explained by this property, but the combination of this property and others previously selected produces a significant model, but with more samples outside the AD. Finally, PLS2, the method that generated the most reliable external predictions, indicates that $pEC_{50}$ and $LogS_W$ are actually related, since the equations with same descriptors resulted in coefficients with same signal. However, probably other physicochemical characteristics also contribute to the environmental toxicity of phenylsulfonyl carboxylates.

## Acknowledgments

E.B.M. and E.H.M: CAPES and FAPPR (grant 2010/7354); J.P.A.M.: IESB; M.M.C.F.: CNPq.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhazmat.2015.10.047.

## References

[1] OECD, 2007. Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. Retrieved November 20, 2012, from <http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2>.
[2] T.I. Netzeva, M. Pavan, A.P. Worth, Review of (quantitative) structure–activity relationships for acute aquatic toxicity, QSAR Comb. Sci. 27 (2008) 77–90.
[3] M.T.D. Cronin, T.W. Schultz, Validation of *Vibrio fisheri* acute toxicity data: mechanism of action-based QSARs for non-polar narcotics and polar narcotic phenols, Sci. Total Environ. 204 (1997) 75–88.
[4] J. Chen, L. Wang, Acute toxicity of alkyl (1-phenylsulfonyl) cycloalkane-carboxylates to photobacterium phosphoreum and quantitative structure-activity relationship study based on the AM1Hamiltonian, Toxicol. Environ. Chem. 57 (1996) 17–26.
[5] E.B. de Melo, M.M.C. Ferreira, Nonequivalent effects of diverse Log*P* algorithms in three QSAR studies, QSAR Comb. Sci. 28 (2009) 1156–1165.
[6] K. Roy, I. Mitra, On the use of the metric rm2 as an effective tool for validation of QSAR models in computational drug design and predictive toxicology, Mini Rev. Med. Chem. 12 (2012) 491–504.

[7] B. Bhhatarai, P. Gramatica, Modelling physico-chemical properties of (benzo) triazoles, and screening for environmental partitioning, Water Res. 45 (2011) 1463–1471.
[8] X. Liu, Z. Yang, L. Wang, Three-dimensional quantitative structure–activity relationship study for phenylsulfonyl carboxylates using CoMFA and CoMSIA, Chemosphere 53 (2003) 945–952.
[9] X. Liu, Z. Yang, L. Wang, CoMFA of the acute toxicity of phenylsulfonyl carboxylates to *Vibrio fischeri*, SAR QSAR Environ. Res. 14 (2003) 183–190.
[10] M. Goodarzi, M.P. Freitas, PLS and N-PLS-based MIA-QSTR modelling of the acute toxicities of phenylsulphonyl carboxylates to *Vibrio fischeri*, Mol. Simul. 36 (2010) 953–959.
[11] E.B. de Melo, Modeling physical and toxicity endpoints of alkyl (1-phenylsulfonyl) cycloalkane-carboxylates using the Ordered Predictors Selection (OPS) for variable selection and descriptors derived with SMILES, Chemom. Intell. Lab. Syst. 118 (2012) 79–87.
[12] X. Liu, Z. Yang, L. Wang, Three-dimensional, quantitative-structure-property-relationship study of aqueous solubility for phenylsulfonyl carboxylates using comparative-molecular-field analysis and comparative-molecular-similarity-índices analysis, Water Environ. Res. 77 (2005) 519–524.
[13] K. Roy, G. Ghosh, QSTR with extended topochemical atom indices: 4. Modeling of the acute toxicity of phenylsulfonyl carboxylates to *Vibrio fischeri* using principal component factor analysis and principal component regression analysis, QSAR Comb. Sci. 23 (2004) 526–535.
[14] K. Roy, G. Ghosh, QSTR with extended topochemical atom indices. Part 5: modeling of the acute toxicity of phenylsulfonyl carboxylates to *Vibrio fischer*i using genetic function approximation, Bioorg. Med. Chem. 13 (2005) 1185–1194.
[15] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48.
[16] E.B. de Melo, J.P.A. Martins, T.C.M. Jorge, M.C. Friozi, M.M.C. Ferreira, Mutivariate QSAR study on the antimutagenic activity of flavonoids against 3-NFA on Salmonella typhimurium TA98, Eur. J. Med. Chem. 45 (2010) 4562–4569.
[17] E.G. Barbosa, K.F.M. Pasqualoto, M.M.C. Ferreira, The receptor-dependent LQTA-QSAR: application to a set of trypanothione reductase inhibitors, J. Comput. Aided Mol. Des. 26 (2012) 1055–1065.
[18] E.B. de Melo, M.M.C. Ferreira, Four-dimensional structure activity relationship model to predict HIV-1 integrase strand transfer using LQTA-QSAR methodology, J. Chem. Inf. Model. 52 (2012) 1722–1732.
[19] N.B.H. Lozano, V.G. Maltarollo, K.C. Weber, K.M. Honorio, R.V.C. Guido, A.D. Andricopulo, A.B.F. da Silva, Molecular features for antitrypanosomal activity of thiosemicarbazones revealed by OPS-PLS QSAR studies, Med. Chem. 8 (2012) 1045–1056.
[20] J.P.A. Martins, E.G. Barbosa, K.F.M. Pasqualoto, M.M.C. Ferreira, LQTA-QSAR: a new 4D-QSAR methodology, J. Chem. Inf. Model. 49 (2009) 1428–1436.
[21] J.P.A. Martins, M.M.C. Ferreira, QSAR modeling: a new open source computational package to generate and validate QSAR models, Quim. Nova 36 (2013) 554–560.
[22] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[23] R. Kiralj, M.M.C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (2009) 770–787.
[24] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for Koc prediction, J. Mol. Graphics Model. 25 (2007) 755–766.
[25] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm2 metrics for validation of QSPR models, Chemom. Intell. Lab. Syst. 107 (2011) 194–205.
[26] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, 2nd ed., Wiley-VCH, Weinheim, 2009.
[27] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, 1998.
[28] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz, M.T.D. Cronin, The better predictive model: high q2 for the training set or low root mean square error of prediction for the test set? QSAR Comb. Sci. 24 (2005) 385–396.
[29] D.B. de Oliveira, A.C. Gaudio, BuildQSAR: a new computer program for QSAR analysis, QSAR Comb. Sci. 19 (2001) 599–601.
[30] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. new intercomparable thresholds for different validation criteria and the need for scatter plot inspection, J. Chem. Inf. Model. 52 (2012) 2044–2058.
[31] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.
[32] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemom. 24 (2010) 194–201.
[33] F. Luan, A. Melo, F. Borges, M.N.L.D.S. Cordeiro, Affinity prediction on A3 adenosine receptor antagonists: the chemometric approach, Bioorg. Med. Chem. 19 (2011) 6853–6859.
[34] U. Fechner, L. Franke, S. Renner, P. Schneider, G. Schneider, Comparison of correlation vector methods for ligand-based similarity searching, J. Comput. Aided Mol. Des. 17 (2003) 687–698.
[35] A.K. Halder, N. Adhikari, T. Jha, Structural findings of 2-phenylindole-3-carbaldehyde derivatives for antimitotic activity by FA-sMLR QSAR analysis, Chem. Biol. Drug Des. 75 (2010) 204–213.
[36] Dragon User Guide, version 6.0, 2011. Talete srl., Italy. (<http://www.talete.mi.it/help/dragon_help/index.html>).
[37] L.B. Kier, L.H. Hall, The electrotopological state: structure modeling for QSAR and database analysis, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, United Kingdom, 1999, pp. 491–562.
[38] R.A. Frank, H. Sanderson, R. Kavanagh, B.K. Burnison, J.V. Headley, K.R. Solomon, Use of a (quantitative) structure?activity relationship [(Q)SAR] model to predict the toxicity of naphthenic acids, J. Toxicol. Environ. Health A 73 (2010) 319–329.
[39] C. Hansch, J.E. Quinlan, G.L. Lawrence, Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids, J. Org. Chem. 33 (1968) 347–350.
[40] J. Wang, T. Hou, Recent advances on aqueous solubility prediction, Comb. Chem. High Throughput Screen. 14 (2011) 328–338.
[41] P.P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, Molecules 14 (2009) 1660–1701.