Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 16 (2008) 3395–3407

Bioorganic & Medicinal Chemistry

QSAR modeling of the rodent carcinogenicity of nitrocompounds

Aliuska Morales Helguera,[a,b,c] M. Natália D. S. Cordeiro,[c] Miguel Ángel Cabrera Pérez,[b] Robert D. Combes[d] and Maykel Pérez González[b,e,*]

[a]*Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*

[b]*Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*

[c]*REQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*

[d]*FRAME, Russell & Burch House, 96-98, North Sherwood Street, Nottingham NG1 4EE, UK*

[e]*Department of Organic Chemistry, Vigo University, C.P. 36310, Vigo, Spain*

Received 13 September 2007; revised 8 November 2007; accepted 9 November 2007
Available online 17 November 2007

Abstract—Chemical carcinogenicity is of primary interest, because it drives much of the current regulatory actions regarding new and existing chemicals, and its conventional experimental test takes around three years to design, conduct, and interpret as well as the costs of hundreds of millions of dollars, millions of skilled personnel hours, and several animal lives. Both academia and private companies are actively trying to develop alternative methods, such as QSAR models. This paper reports a QSAR study for predicting carcinogenic potency of nitrocompounds bioassayed in female rats. Several different theoretical molecular descriptors, calculated only on the basis of knowledge of the molecular structure and an efficient variable selection procedure, such as Genetic Algorithm, led to models with satisfactory predictive ability. But the best-final QSAR model is based on the GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) descriptors capturing a reasonable interpretation. In fact, structural features such as molecular shape—linear, branched, cyclic, and polycyclic—and bond length are some of the key factors flagging the carcinogenicity of this set of nitrocompounds. This QSAR model, after removal of one identified nitrocompound outlier, is able to explain around 86% of the variance in the experimental activity and manifest good predictive ability as indicated by the higher q^2s of cross- and external-validations, which demonstrate the practical value of the final QSAR model for screening and priority testing. This model can be applied to nitrochemicals different from the studied nitrocompounds (even those not yet synthesized) as it is based on theoretical molecular descriptors that might be easily and rapidly calculated.

1. Introduction

The process of carcinogenesis is the result of a sequence of stages and complex biological interactions that can be influenced by factors such as age, diet, environment, and hormonal balance. Currently it is estimated that up to 80% of human tumor cases are caused by exogenous (environmental) chemical agents,[1] which has led to considerable efforts and policies aimed at eliminating, or at least drastically reducing, human exposures to potentially carcinogenic agents, particularly chemicals. Chemicals with such activity, together with the related effects of mutagenicity and reproductive toxicity, require extensive assessment as part of the European Commission's new legislation first published as a 'White Paper' on a strategy for a future 'Community Policy for Chemicals'.[2] Under this scheme, the Registration, Evaluation and Authorisation of CHemicals (REACH), all substances currently marketed in Europe in volumes greater than 1 ton per year have to be registered before the end of 2012. It has been estimated that there are some 30,000 such existing substances for which there are important data gaps. The production of such a large quantity of experimental data by using conventional in vivo testing is time-consuming and expensive, and is impracticable if the timetable set by REACH is to be met. In particular, the rodent bioassay for identifying potential human carcinogens is totally inappropriate for the large-scale testing required.[3]

One way to alleviate the above problems is to use alternative, non-animal approaches as tools for generating

Keywords: Nitrocompounds; GETAWAY; Carcinogenicity; QSAR; Genetic Algorithm.

* Corresponding author. Tel.: +53 42 281473; +53 42 281192; fax: +53 42 281130; e-mail: mpgonzalez76@yahoo.es

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.11.029

hazard data useful for, at least, preliminary risk assessment from exposure to chemicals. One such set of tools, being extensively considered within the framework of the EU REACH legislation, comprises *in silico*-based prediction of carcinogenicity, based on structure–activity relationships (SAR) and quantitative structure–activity relationship (QSAR) modeling. QSAR modeling seeks to discover and use mathematical relationships between chemical structure and biological activity, most often by a multi-linear equation that relates molecular properties of the compounds (*descriptors*) to the activity of interest. An extensive number of molecular descriptors exist[4] which can and have been used to model a wide range of toxicities,[5–9] including carcinogenicity.[10–17]

In this connection, Yuta and Jurs[18] in 1981 developed a QSAR to distinguish between carcinogenic and non-carcinogenic aromatic amines in rats. The data set was composed of 157 aromatic amines subdivided into six chemical classes. Biological activity was interpreted as the presence or absence of carcinogenic activity in several organs of rat, according to route of administration and tumor site. In the last decade, Benigni and co-workers have been leading efforts in the field of statistical analysis and modeling of toxicological properties; mainly mutagenesis and carcinogenesis.[19] In their first QSAR analysis of the carcinogenicity of aromatic amines,[19] they investigated the structural factors that influence the gradation of carcinogenic potency in rodents. The carcinogenic potency data set used for this study was derived from the TD_{50} values for rat and mouse reported in the Carcinogenic Potency Data Base (CPDB, see http://potency.berkely.edu/cpdb.htlm). The TD_{50} was defined as follows: for a given target site(s) if there are no tumors in control animals, then the TD_{50} is that chronic dose in mg/kg body wt/day which would induce tumors in half the test animals at the end of a standard lifespan for the species.[20]

In a subsequent report, Franke et al.[10] studied the difference between carcinogenic and non-carcinogenic aromatic amines. For the analysis, 79 compounds taken from the Carcinogenic Potency Data Base were used to generate five QSAR model equations, one for each experimental group (rat and mouse, male and female), and the last one for overall carcinogenicity. The two classes were coded as: 1 = non-carcinogens; 2 = carcinogens. Discriminant analysis was used, and the variables were standardized to compute carcinogenic activity values (w) for these functions. Recently, Benigni et al.[21] presented QSAR models to predict the carcinogenicity of simple and α–β unsaturated aldehydes. They used stepwise linear discriminant analysis to distinguish between active and inactive simple aldehydes.

Also a QSAR carcinogenic model using nitrocompounds was published recently.[13] The authors reported a model validated by leave-one-out cross-validation, with a determination coefficient of 0.666. In addition, this approach enabled the contribution of different fragments to carcinogenic potency to be assessed, thereby making the relationships between structure and carcinogenicity to be transparent. It was found that the carcinogenic activity of the chemicals analyzed was increased by the presence of a primary amine group bonded to the aromatic ring, a manner that was proportional to the ring aromaticity. The nitro group bonded to an aromatic carbon atom is a more important determinant of carcinogenicity than the nitro group bonded to an aliphatic carbon. Finally, the model reported was compared with four other predictive models, but none of these could explain more than 66% of the variance in the carcinogenic potency with the same number of variables, also the authors said that other kind of descriptors should be used to try to improve these predictions.

In this context, a good example of these are the so-called GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) descriptors, proposed by Consonni et al.[22,23] The GETAWAY descriptors are based on matching 3D-molecular geometry (provided by the molecular influence matrix and atom relatedness by molecular topology) with chemical information (provided by using different atomic weightings, such as atomic mass, polarizability, van der Waals volume, and electronegativity). These descriptors have several positive characteristics, including intrinsic 3D information, being based on well-known accepted algorithms and formulae, and possess good predictive power in physicochemical[22] and biological property modeling.[11,22–27]

On the other hand, as potentially toxic chemicals, nitrocompounds are of special concern due their carcinogenic potential. They are widely distributed environmental pollutants found in the workplace, in emissions from diesel and gasoline engines, and on the surface of ambient air particulate matter,[28] where they contribute to local and regional pollution (car exhausts, technological spills). In addition they are suspected to induce tumors in experimental animals, in several organs including the spinal cord, nasal cavity,[29] and the stomach.[30]

The main objective of the present work is to develop a valid QSAR model for predicting the carcinogenicity of environmental nitrocompounds, based on structural information and data for carcinogenic potency, TD_{50}, of a training set of 49 nitrocompounds, as determined in the female rat. We examined the use of regression models along with feature selection algorithms derived from a variety of molecular representations. For this training set, the GETAWAY descriptors provided the best model and exhibited good quality and predictive power, as judged by extensive internal and external validation. Our final model should serve as a useful tool for the preliminary ranking and prioritization of chemicals for carcinogenicity or the synthesis of substitute nitrocompounds with lower carcinogenicity, as required by REACH.

2. Results and discussion

For QSAR modeling, our first goal was to establish a reasonable number of predictor variables and input compounds to ensure a good generalized performance.

This was accomplished by finding regression models for the entire training set (49 nitrocompounds) based on GA selection, in conjunction with the nine sets of molecular descriptors described in Section 4. A combination of all above DRAGON-descriptors and quantum-chemical molecular descriptors was also used for building QSAR models. The results for several combinations of descriptors, that is, for models with sizes ranging from seven to nine variables, are summarized in Supporting information. The statistics reported are R^2, $q^2_{\text{LOO-CV}}$, and q^2_{ext}, the parameters used to evaluate the quality for every single model constructed.

The descriptor combinations (general models: 28–30 in Supporting information) led to better and more efficient predictor models, as manifested by the higher R^2 (model fit) and $q^2_{\text{LOO-CV}}$ (internal validation) values, respectively. However, the external predictivity of GETAWAY descriptors is better than descriptor combinations, as it affords the highest q^2_{ext} values. While the GETAWAY descriptors-based best model was able to predict >66% of the experimental variances of external set, the other models explained <49%. The real utility of QSAR models is the capability of reliable predictions for new chemicals, not used in model development.[31] Hence, we determined that seven to nine theoretical molecular descriptors, in this case the GETAWAY descriptors, would suffice to build an efficient model with good generalized predictivity.

The GETAWAY-based models 9, 18, and 27 seemed to be the models with the greatest predictivity for estimating the TD_{50} of nitrocompounds in female rats (Supporting information). These models will be referred to as Model 1, Model 2, and Model 3, respectively. The meaning of each GETAWAY descriptor used is shown in Table 1. The three models are given below, together with detailed statistics of the MLR analysis as the basis of a more exhaustive analysis undertaken to select the most accurate final QSAR model.

Model 1

$$\begin{aligned}-\log \text{TD}_{50} = {} & -0.251 - 1.880\cdot \text{HATS}_3(\text{u}) + 7.838 \\ & \cdot \text{H}_8(\text{m}) - 50.024\cdot \text{H}_8(\text{v}) + 3.791 \\ & \cdot \text{H}_0(\text{p}) - 6.696\cdot \text{R}_7^+(\text{m}) - 28.576 \\ & \cdot \text{R}_2(\text{v}) + -0.713\cdot \text{R}_5(\text{e})\end{aligned} \tag{1}$$

$$N = 49,\quad R^2 = 81.35,\quad S = 0.429,\quad F = 25.54,\quad p < 10^{-5},\quad \rho = 6.125$$

$$\text{LOF} = 0.302,\quad \text{AIC} = 0.256,\quad \text{FIT} = 1.815$$

$$q^2_{\text{LOO-CV}} = 73.67,\quad S_{\text{LOO-CV}} = 0.467,\quad q^2_{\text{Boot}} = 68.59,\quad R^2_{\text{Scram}} = 0.105$$

$$N_{\text{ext}} = 6,\quad q^2_{\text{ext}} = 51.30,\quad S_{\text{ext}} = 0.560$$

Model 2

$$\begin{aligned}-\log \text{TD}_{50} = {} & -0.156 - 1.855\cdot \text{HATS}_3(\text{u}) - 26.965 \\ & \cdot \text{H}_8(\text{p}) - 31.468\cdot \text{R}_2^+(\text{v}) + 3.921 \\ & \cdot \text{H}_0(\text{p})9.161\cdot \text{HATS}_8(\text{v}) - 4.359 \\ & \cdot \text{HATS}_5(\text{p}) - 6.653\cdot \text{R}_7^+(\text{m}) - 2.368 \\ & \cdot \text{R}_8(\text{u})\end{aligned} \tag{2}$$

$$N = 49,\quad R^2 = 84.66,\quad S = 0.394,\quad F = 27.59,\quad p < 10^{-5},\quad \rho = 5.444$$

$$\text{LOF} = 0.280,\quad \text{AIC} = 0.225,\quad \text{FIT} = 1.940$$

$$q^2_{\text{LOO-CV}} = 77.22,\quad S_{\text{CV}} = 0.434,\quad q^2_{\text{Boot}} = 72.04,\quad R^2_{\text{Scram}} = 0.117$$

$$N_{\text{ext}} = 6,\quad q^2_{\text{ext}} = 66.40,\quad S_{\text{ext}} = 0.532$$

Model 3

$$\begin{aligned}-\log \text{TD}_{50} = {} & 2.040 + 4.335\cdot \text{HATS}_4(\text{m}) - 1.851 \\ & \cdot \text{HATS}_6(\text{m}) - 26.216\cdot \text{H}_8(\text{v}) \\ & + 1.179\cdot \text{HATS}_4(\text{e}) + 2.918\cdot \text{H}_0(\text{p}) \\ & - 6.371\cdot \text{RARS} - 4.926\cdot \text{R}_7^+(\text{m}) \\ & + 4.551\cdot \text{R}_8^+(\text{m}) - 30.295\cdot \text{R}_2^+(\text{v})\end{aligned} \tag{3}$$

$$N = 49,\quad R^2 = 86.56,\quad S = 0.360,\quad F = 30.49,\quad p < 10^{-5},\quad \rho = 4.900$$

$$\text{LOF} = 0.257,\quad \text{AIC} = 0.196,\quad \text{FIT} = 2.094$$

$$q^2_{\text{LOO-CV}} = 77.21,\quad S_{\text{CV}} = 0.422,\quad q^2_{\text{Boot}} = 70.73,\quad R^2_{\text{Scram}} = 0.137$$

$$N_{\text{ext}} = 6,\quad q^2_{\text{ext}} = 65.00,\quad S_{\text{ext}} = 0.523$$

The large F (>25.54) indices and small p (<10^{-5}) values are indicative of the statistical significance of these three models. In addition, the values of the determination coefficients for the regression (R^2 can take values from zero—*no correlation*—to one—*perfect correlation*—), and for LOO-CV, as well as for bootstrapping and external validation (measures of the model predictability), all show that the models displayed an adequate goodness-of-fit and prediction. The former property of Model 3 was greater than that for Model 1, as judged by the R^2 values. The predictive ability (given by $q^2_{\text{LOO-CV}}$, q^2_{Boot}, and q^2_{ext}) was greater for Model 2 than Model 1. However, there was no significant enhancement in the internal and external predictive power of Model 3, compared with Model 2. It is concluded that Model 3 (with nine descriptors) is overfitted to the data, as its predictions are no better than those of the simpler Model 2 (with 8 descriptors).[32] Therefore,

Table 1. Symbols for the GETAWAY descriptors and their definitions

Symbols	Descriptor definition
$HATS_3(u)$	Leverage-weighted autocorrelation of lag 3/unweighted
$HATS_4(e)$	Leverage-weighted autocorrelation of lag 4/weighted by atomic Sanderson electronegativities
$HATS_4(m)$	Leverage-weighted autocorrelation of lag 4/weighted by atomic masses
$HATS_6(m)$	Leverage-weighted autocorrelation of lag 6/weighted by atomic masses
$HATS_5(p)$	Leverage-weighted autocorrelation of lag 5/weighted by atomic polarizabilities
$HATS_8(v)$	Leverage-weighted autocorrelation of lag 8/weighted by atomic van der Waals volumes
$R_8(u)$	*R* autocorrelation of lag 8/unweighted
$H_8(m)$	*H* autocorrelation of lag 8/weighted by atomic masses
$H_0(p)$	*H* autocorrelation of lag 0/weighted by atomic polarizabilities
$H_8(p)$	*H* autocorrelation of lag 8/weighted by atomic polarizabilities
$H_8(v)$	*H* autocorrelation of lag 8/weighted by atomic van der Waals volumes
RARS	*R* matrix average row sum
$R_5(e)$	*R* autocorrelation of lag 5/weighted by atomic Sanderson electronegativities
$R_7^+(m)$	*R* maximal autocorrelation of lag 7/weighted by atomic masses
$R_8^+(m)$	*R* maximal autocorrelation of lag 8/weighted by atomic masses
$R_2^+(v)$	*R* maximal autocorrelation of lag 2/weighted by atomic van der Waals volumes

use of Model 3 is not statistically justified as the regression Model 2 is sufficient to explain the carcinogenic activity of the chemicals in the data set. This could be due to the fact that either the extra descriptor in Model 3: (a) is irrelevant for predicting $-\log TD_{50}$; and/or (b) its predictive power duplicates that of other predictors used in the model. It is, therefore, concluded that the eight-dimensional model, Model 2, is the most accurate one of those used for predicting carcinogenicity in female rats of the range of nitrocompounds involved.

However, further analysis of this regression model should only be conducted after checking the reliability of the pre-adopted assumptions. MLR analysis establishes a linear, additive relation between the molecular descriptors and the underlying biological activity (i.e., carcinogenicity), and, in fact, this is the simplest mathematical form that might be envisaged for the model without any *a priori* information. Nevertheless, by looking at the distribution of the standardized residuals for all cases (Fig. 1a), no specific pattern is seen, thereby reinforcing the idea that the model does exhibit a linear dependence.

The hypothesis of normally distributed residuals can be verified from the normal probability plot of residuals (Fig. 1b), as all the plotted points approximate to a straight line fit. The results from the Kolmogorov–Smirnov (*D*) and Shapiro–Wilk (*W*) tests also confirm the normality of residuals. Both statistics, $D = 0.117$, $p = 0.092$ and $W = 0.959$, $p = 0.090$, are not significant ($p > 0.05$), then the hypothesis that the residual distribution is normal should be accepted.

With regard to the hypothesis of homocedasticity of residuals, Figure 1c shows the plot of standardized residuals versus the predicted $-\log TD_{50}$ values. As can be seen, the points are scattered throughout and do not seem to cluster in any significant way, thus confirming the homogeneity of variance. This plot also provides a check for the no autocorrelation of the residuals.

An aspect deserving special attention is the degree of collinearity of the variables of the model, which can readily be diagnosed by analyzing the cross-correlation matrix. As seen in Table 2, the pairs of descriptors ($R_8(u)$; $H_8(p)$), ($H_0(p)$; $HATS_8(v)$), and ($R_8(u)$; $HATS_8(v)$) are correlated with each other. Rather than deleting any of these descriptors, it is of interest to examine the performance of orthogonal complements in modeling the carcinogenic activity.

Following the Randić technique, we determined orthogonal complements for all variables in Model 2, which in turn were further standardized, to enable derivation of the following best six descriptor equation (Model 4):

Model 4

$$\begin{aligned}-\log TD_{50} = {} & -1.650 - 0.449\cdot{}^1\Omega HATS_3(u) \\ & - 0.367\cdot{}^2\Omega H_8(p) + 0.395\cdot{}^4\Omega H_0(p) \\ & - 0.252\cdot{}^6\Omega HATS_5(p) \\ & - 0.295\cdot{}^7\Omega R_7^+(m) - 0.240\cdot{}^8\Omega R_8(u) \quad (4)\end{aligned}$$

$$N = 49,\quad R^2 = 82.91,\quad S = 0.406,\quad F = 34.01,\quad p < 10^{-5},\quad \rho = 7.000$$

$$LOF = 0.247,\quad AIC = 0.219,\quad FIT = 2.387$$

$$q^2_{LOO\text{-}CV} = 77.47,\quad S_{CV} = 0.432,\quad q^2_{Boot} = 74.51,\quad R^2_{Scram} = 0.085$$

$$N_{ext} = 6,\quad q^2_{ext} = 66.60,\quad S_{ext} = 0.543,$$

where the symbol ${}^i\Omega X$ means the orthogonal complement of variable *X*, while the subscript refers to the order selected for orthogonalizing the variables.

Descriptors ${}^5\Omega HATS_3(u)$ and ${}^3\Omega R_2^+(v)$ have been excluded as they were found to be statistically non-significant. Their omission, however, had little effect on the overall fitness of the model as the statistics are as robust

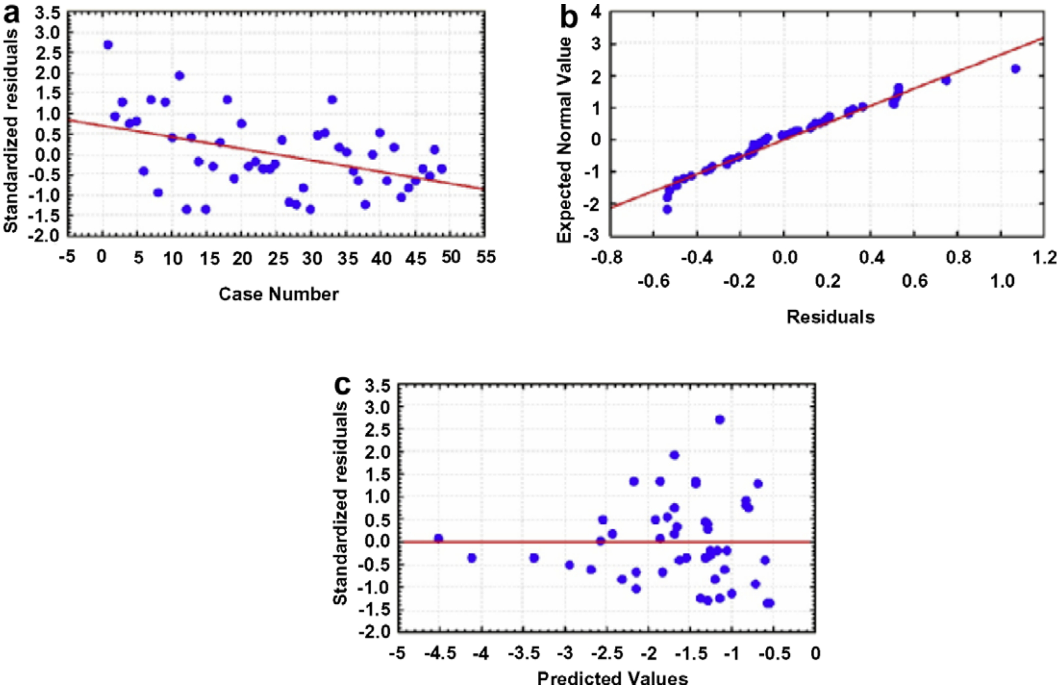

Figure 1. Checking the validity of the MLR assumptions. (a) Linearity of the model: plot of standardized residuals versus cases. (b) Normality: normal probability plot of the residuals. (c) Homocedasticity: plot of standardized residuals versus predicted values.

Table 2. Intercorrelation among the eight descriptors selected as statistically significant by the MLR–GA technique[a]

	HATS$_3$(u)	H$_8$(p)	R$_2^+$(v)	H$_0$(p)	HATS$_8$(v)	HATS$_5$(p)	R$_7^+$(m)	R$_8$(u)
HATS$_3$(u)	1.00	−0.22	0.13	−0.26	−0.22	−0.43	−0.30	−0.22
H$_8$(p)		1.00	0.00	0.40	0.59	0.03	0.16	**0.71**
R$_2^+$(v)			1.00	0.49	0.17	0.14	0.04	0.01
H$_0$(p)				1.00	**0.71**	0.47	0.45	0.57
HATS$_8$(v)					1.00	0.13	0.40	**0.88**
HATS$_5$(p)						1.00	0.32	0.07
R$_7^+$(m)							1.00	0.26
R$_8$(u)								1.00

[a] Significant correlations are marked in bold.

as before, and moreover the predictive ability of the model slightly improved (from $q^2_{\text{LOO-CV}} = 77.22$ to$q^2_{\text{LOO-CV}} = 77.47$; from $q^2_{\text{Boot}} = 72.04$ to $q^2_{\text{Boot}} = 74.51$; from $q^2_{\text{ext}} = 66.40$ to $q^2_{\text{ext}} = 66.60$). Yet there are significant differences between Model 2 and Model 4 as regards the interpretation of the results. By comparing Eq. 2 with Eq. 4, one can see that there are no changes in either the sign of the regression coefficients or of the constant. Nevertheless, the relative contributions of the variables in the orthogonal-descriptor model are different to those in the non-orthogonalized model. Therefore, for purposes of QSAR interpretability, we shall use the orthogonal-descriptor model defined in Eq. 4.

It is also important to search for possible outliers in the training set that might be unduly distorting the regression model. The Williams plot, illustrated in Figure 2A, is sometimes used for detecting outliers, as well as influential chemicals.[33,34]

It showed us that chemical 1 (*N*-methyl-*N*′-nitro-*N*-nitrosoguanidine) is clearly an outlier (standardized residual = 2.910), whereas chemicals 47 (trinitroglycerin), 48 (nitroethane), and 49 (trifluralin) are the influential chemicals in the training set with leverage values higher than 'warning leverage' fixed at $3p/n(h^* = 0.429)$ which were predicted correctly (Fig. 2B), but would be expected to have a disproportionate influence on the regression line. The anomalous behavior of outlier chemical could be due to the following: (1) incorrect experimental input data, (2) the descriptors selected do not capture some relevant structural features present in this molecule and absent in the others, (3) its biological mechanism is different from the remaining chemicals. At this stage, we think that *N*-methyl-*N*′-nitro-*N*-nitrosoguanidine (MNNG) has a different mechanism of carcinogenicity induction, given the presence of both a nitroso group and a nitro group. MNNG does not need to be activated to be genotoxic, unlike the rest of nitrocompounds studied here.[35] It was noticed that the same outlier was detected in a previous QSAR modeling work by Helguera et al.[13] As the regression analysis is seriously biased by this specific compound, it is reasonable to consider removing it from the training set, to

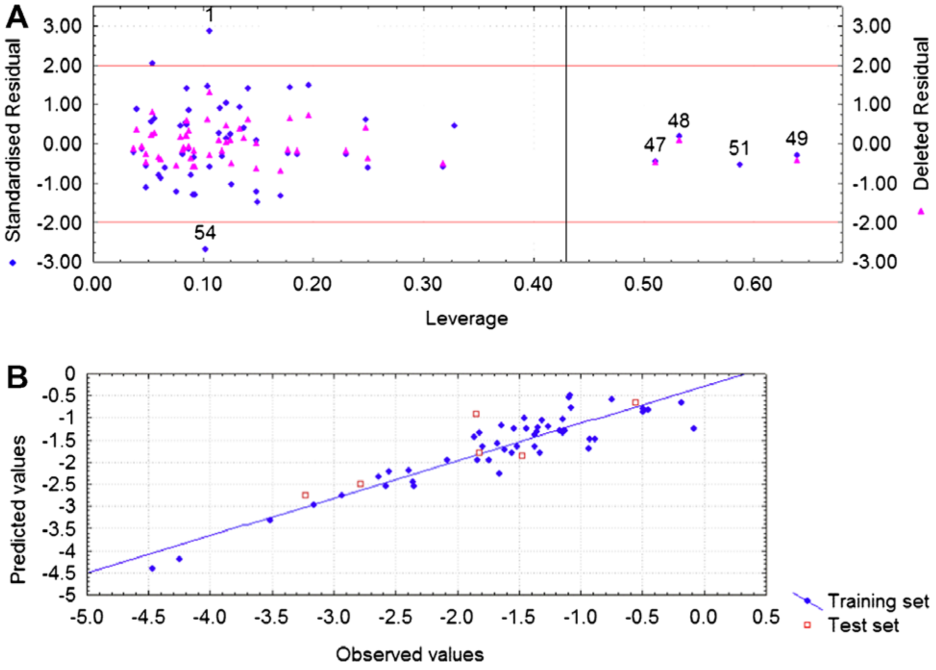

Figure 2. (A) Williams plot, that is, plot of standardized residual versus leverage values, with a warning leverage of 0.429 and taking into account chemicals of training (numbered from 1 to 49) and test (numbered from 50 to 56) sets. (B) Plot of observed versus predicted activity, expressed as $-\log \mathrm{TD}_{50}$, for chemicals of training set.

obtain the model shown below (Model 5). Predicted, observed values, simple residuals, and deleted residuals are given in Table 3.

Model 5

$$\begin{aligned} -\log \mathrm{TD}_{50} = {} & -1.677 - 0.447 \cdot {}^{1}\Omega\mathrm{HATS}_{3}(\mathrm{u}) \\ & - 0.345 \cdot {}^{2}\Omega\mathrm{H}_{8}(\mathrm{p}) + 0.397 \cdot {}^{4}\Omega\mathrm{H}_{0}(\mathrm{p}) \\ & - 0.291 \cdot {}^{6}\Omega\mathrm{HATS}_{5}(\mathrm{p}) \\ & - 0.264 \cdot {}^{7}\Omega\mathrm{R}_{7}^{+}(\mathrm{m}) - 0.227 \cdot {}^{8}\Omega\mathrm{R}_{8}(\mathrm{u}) \quad (5) \end{aligned}$$

$$N = 48, \quad R^2 = 85.91, \quad S = 0.361, \quad F = 41.7,$$
$$p < 10^{-5}, \quad \rho = 6.857$$

$$\mathrm{LOF} = 0.198, \quad \mathrm{AIC} = 0.179, \quad \mathrm{FIT} = 2.955$$

$$q^2_{\mathrm{LOO\text{-}CV}} = 80.97, \quad S_{\mathrm{CV}} = 0.388,$$
$$q^2_{\mathrm{Boot}} = 77.96, \quad R^2_{\mathrm{Scram}} = 0.081$$

$$N_{\mathrm{ext}} = 6, \quad q^2_{\mathrm{ext}} = 71.10, \quad S_{\mathrm{ext}} = 0.488$$

Undoubtedly, the new model exhibits the best overall predictive performance, providing an excellent correlation between the observed $-\log \mathrm{TD}_{50}$ data and the orthogonalized and standardized GETAWAY descriptors. It accounts for around 86% of the variance in that both the data and all included descriptors are significant ($p < 10^{-5}$). Also, its superiority with respect to Model 4 is manifested by the lower LOF and AIC and higher FIT. Moreover, it shows better predictive ability than Model 4, as indicated by the higher cross-validated q^2 values (Model 5: $q^2_{\mathrm{LOO\text{-}CV}} = 80.97$ and $q^2_{\mathrm{Boot}} = 77.96$; Model 4: $q^2_{\mathrm{LOO\text{-}CV}} = 77.47$ and $q^2_{\mathrm{Boot}} = 74.51$) and higher q^2 value of external test set (Model 5: $q^2_{\mathrm{ext}} = 71.10$; Model 4: $q^2_{\mathrm{ext}} = 66.40$). It important to note that the Model 5 provides predictions that are expected to be reliable for 5 out of 6 chemicals of test set that fall in the applicability domain (Fig. 3A). Thus, chemical **51** (2,3,5,6-tetrachloro-4-nitroanisole) has high leverage and therefore the carcinogenic potency value should be considered as extrapolated by the model and potentially unreliable. If only the predictions that fit the Model 5 applicability domain were considered in the calculation of explained variance and external standard deviation, these values will be $q^2_{\mathrm{ext}} = 71.02$ and $S_{\mathrm{ext}} = 0.534$. In the Williams plot, it is also possible to identify chemical **54** (*ortho*-nitroanisole) of validation set as an outlier for response *Y*, in spite of this, it is inside *X* descriptor applicability domain of the model and its predicted toxicity value should also be considered in the calculation of the explained variance; q^2_{ext} and external standard deviation error of the prediction; S_{ext}. Predicted, observed values and simple residuals of this test set are given in Table 4.

Model 5 shows the importance of considering different structural aspects of the chemicals in the training set for predicting the carcinogenic potential of nitrocompounds in the female rat. In fact, polarizability and atomic mass appear, in this case, to be the most relevant physicochemical parameters for describing $-\log \mathrm{TD}_{50}$, which were used as atomic weights within the GETAWAY descriptors. ${}^{1}\Omega\mathrm{HATS}_{3}(\mathrm{u})$, the most important descriptor, is formally similar to the Moreau–Broto autocorrelation descriptor[4] of a topological structure with lag 3, but it also takes into account 3D-molecular geometry by using the leverage values as atom weights.[23]

Table 3. Observed, predicted, and residual values of the 48 nitrocompounds used for obtaining the final QSAR model (Model 5; Eq. 5)

Compound	Name	Carcinogenic potency[a]			RES[b]	RES_{del}[c]
		TD_{50}	P_{obs}	P_{pred}		
1	*N*-Methyl-*N′*-nitro-*N*-nitrosoguanidine	1.210	−0.083		Outlier	
2	Dimethylnitramide	2.842	−0.454	−0.948	0.495	0.563
3	2-(2,2-Dimethylhydrazino)-4-(5-nitro-2-furyl)thiazole	1.538	−0.187	−0.601	0.414	0.482
4	2-Amino-5-(5-nitro-2-furyl)-1,3,4-thiadiazole	3.120	−0.494	−0.826	0.332	0.363
5	2-Hydrazino-4-(*p*-nitrophenyl) thiazole	3.145	−0.498	−0.846	0.348	0.403
6	1,2-Dihydro-2-(5-nitro-2-thienyl) quinazolin-4(3*H*)-one	5.558	−0.745	−0.607	−0.137	−0.157
7	4,6-Diamino-2-(5-nitro-2-furyl)-*S*-triazine	7.697	−0.886	−1.570	0.684	0.764
8	4,6-Dimethyl-2-(5-nitro-2-furyl) pyrimidine	11.907	−1.076	−0.898	−0.178	−0.190
9	2,4-Dinitro-6-*tert*-butylphenylmethanesulfonate	8.357	−0.922	−1.463	0.541	0.591
10	2-Amino-5-(5-nitro-2-furyl)-1,3,4-oxadiazole	13.971	−1.145	−1.424	0.279	0.305
11	l-5-Morpholinomethyl-3-[(5-nitrofurfurylidene)amino]-2-oxazolidinone·HCl	8.665	−0.938	−1.645	0.707	0.747
12	2-Hydrazino-4-(5-nitro-2-furyl)thiazole	12.510	−1.097	−0.642	−0.455	−0.502
13	1-Nitropyrene	13.468	−1.129	−1.339	0.210	0.231
14	Formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide	13.925	−1.144	−1.059	−0.084	−0.093
15	4-(2-Hydroxyethylamino)-2-(5-nitro-2-thienyl)quinazoline	12.139	−1.084	−0.712	−0.372	−0.449
16	2,4-Dinitrotoluene	23.664	−1.374	−1.439	0.065	0.081
17	4-Morpholino-2-(5-nitro-2-thienyl) quinazoline	14.692	−1.167	−1.354	0.187	0.217
18	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]formamide	21.195	−1.326	−1.584	0.258	0.322
19	*N*-([3-(5-Nitro-2-furyl)-1,2,4-oxadiazole-5-yl]-methyl)acetamide	20.620	−1.314	−1.109	−0.205	−0.216
20	1-[(5-Nitrofurfurylidene)amino]-2-imidazolidinone	23.464	−1.370	−1.775	0.405	0.422
21	4-Methyl-1-[(5-nitrofurfurylidene)amino]-2-imidazolidinone	22.418	−1.351	−1.318	−0.033	−0.034
22	2,2,2-Trifluoro-*N*-[4-(5-nitro-2- furyl)-2-thiazolyl]acetamide	18.164	−1.259	−1.178	−0.082	−0.085
23	2-Amino-4-(5-nitro-2-furyl)thiazole	27.699	−1.442	−1.084	−0.358	−0.440
24	Nitrobenzene	47.681	−1.678	−1.592	−0.086	−0.094
25	Acetone[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazone	22.721	−1.356	−1.169	−0.188	−0.205
26	5-Nitro-2-furaldehyde semicarbazone	32.907	−1.517	−1.879	0.362	0.394
27	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]acetamide	28.630	−1.457	−0.975	−0.482	−0.506
28	5-(5-Nitro-2-furyl)-1,3,4-oxadiazole-2-ol	43.682	−1.640	−1.267	−0.374	−0.404
29	*N*-[5-(5-Nitro-2-furyl)-1,3,4-thiadiazol-2-yl]acetamide	34.773	−1.541	−1.258	−0.284	−0.302
30	*p*-Nitroaniline	65.521	−1.816	−1.403	−0.413	−0.455
31	8-Nitroquinoline	54.836	−1.739	−1.839	0.100	0.105
32	Azathioprine	36.428	−1.561	−1.782	0.221	0.234
33	AF-2	45.932	−1.662	−2.304	0.642	0.784
34	*N*,*N′*-[6-(5-Nitro-2-furyl)-*S*-triazine-2,4-diyl]bisacetamide	41.798	−1.621	−1.813	0.192	0.217
35	3-(5-Nitro-2-furyl)-imidazo(1,2-alpha) pyridine	68.938	−1.838	−1.842	0.003	0.004
36	1-(2-Hydroxyethyl)-3-[(5-nitrofurfurylidene)amino]-2-imidazolidinone	62.261	−1.794	−1.643	−0.152	−0.167
37	1,2-Dimethyl-5-nitroimidazole	120.458	−2.081	−1.947	−0.134	−0.185
38	*trans*-2-[(Dimethylamino)methylimino]-5-[2-(5-nitro-2-furyl)vinyl]-1,3,4-oxadiazole	73.052	−1.864	−1.255	−0.608	−0.716
39	Methylnitramide	377.360	−2.577	−2.683	0.106	0.125
40	2-Amino-5-nitrothiazole	223.922	−2.350	−2.142	−0.208	−0.277
41	5-Nitro-2-furamidoxime	245.455	−2.390	−2.221	−0.169	−0.190
42	1-[(5-Nitrofurfurylidene)amino]hydantoin	228.001	−2.358	−2.433	0.075	0.086
43	Metronidazole	356.989	−2.553	−2.135	−0.418	−0.484
44	5-Nitro-2-furanmethanediol diacetate	431.796	−2.635	−2.383	−0.252	−0.277
45	Chloramphenicol	860.336	−2.935	−2.333	−0.602	−0.882
46	2-Nitropropane	3288.692	−3.517	−3.290	−0.227	−0.295
47	Trinitroglycerin	1448.787	−3.161	−3.080	−0.081	−0.167
48	Nitroethane	29174.093	−4.465	−4.516	0.051	0.109
49	Trifluralin	17567.456	−4.245	−4.152	−0.092	−0.259

[a] Carcinogenic activity estimated as TD_{50} (chronic dose in μmol/kg of body weight per day inducing tumors in 50% of the test animals at the end of a lifetime) and then log-transformed, that is, $P = -\log TD_{50}$.

[b] RES = $P_{obs} - P_{pred}$.

[c] Deleted residuals.

$^{4}\Omega H_0(p)$, $^{2}\Omega H_8(p)$, the second and third most important descriptors, are modifications too of the Moreau–Broto autocorrelation descriptor defined for path of length zero and eight, respectively, but provide information on the degree of interaction between atom pairs.[23] These are followed by the leverage-weighted $^{6}\Omega HATS_5(p)$ descriptor of lag 5 and finally the *R* maximal autocorrelation of lag 7, weighted by atomic masses ($^{7}\Omega R_7^{+}(m)$) and *R* autocorrelation of lag 8/unweighted ($^{8}\Omega R_8(u)$).

The data obtained above from our final regression model have been analyzed to provide some information relevant to determining the mode of action of the chemicals used, in terms of structural features that could be associated to carcinogenicity of the data set. As carcinogenic potency is expressed as $-\log TD_{50}$, negative descriptor contributions increase the TD_{50} value and decrease the carcinogenic activity and vice versa. On this basis, and according to Model 5, $^{1}\Omega HATS_3(u)$,

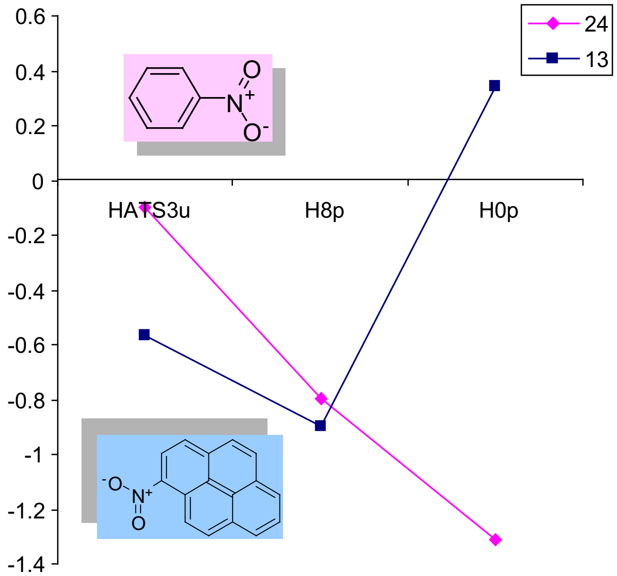

Figure 3. Profiles of $HATS_3(u)$, $H_8(p)$, and $H_0(p)$ for nitrobenzene (chemical 24) and 1-nitropyrene (chemical 13).

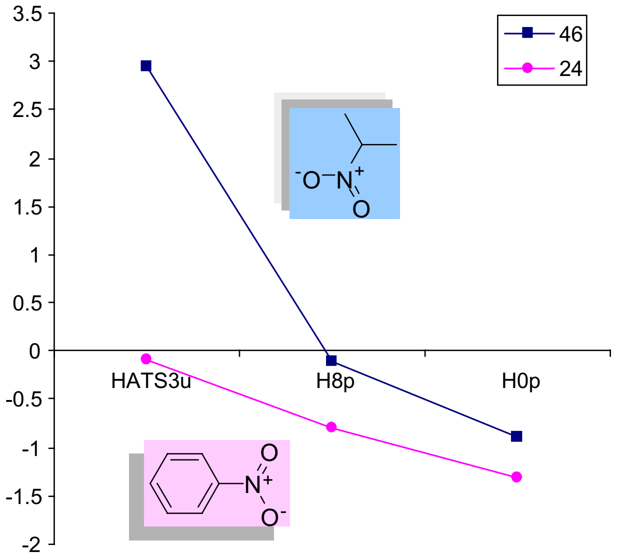

Figure 4. Profiles of $HATS_3(u)$, $H_8(p)$, and $H_0(p)$ for nitrobenzene (chemical 24) and 12-nitropropane (chemical 46).

${}^{2}\Omega H_8(p)$, ${}^{6}\Omega HATS_5(p)$, ${}^{7}\Omega R_7^+(m)$, and ${}^{8}\Omega R_8(u)$ contribute negatively to carcinogenicity, while ${}^{4}\Omega H_0(p)$ makes a positive contribution. To estimate the significance of each descriptor in predicting carcinogenicity, comparisons have been made between particular nitrocompound species (Figs. 3–6) by using the three most important descriptors of Model 5.

As seen in Figure 3, ${}^{1}\Omega HATS_3(u)$ is sensitive to molecular size, and its value decreases as the number of rings increases, for instance from nitrobenzene (chemical 24, $TD_{50}(obs) = 47.681$) to 1-nitropyrene (chemical 13, $TD_{50}(obs) = 13.468$). A similar tendency is observed for ${}^{2}\Omega H_8(p)$ and ${}^{8}\Omega R_8(u)$ following the opposite trend descriptor ${}^{4}\Omega H_0(p)$. As Figure 4 suggests, the most important descriptor—${}^{1}\Omega HATS_3(u)$—seems to be sensitive to molecular branching and cyclization, as its value decreases substantially from that for 2-nitropropane (chemical 46, $TD_{50}(obs) = 3288.692$) to that for nitrobenzene (chemical 24, $TD_{50}(obs) = 47.681$). Figure 5 shows the variation of the most important descriptors for two structures, 2-nitropropane (chemical 46, $TD_{50}(obs) = 3288.692$) and dimethylnitramide (chemical 2, $TD_{50}(obs) = 2.842$). It can be seen that ${}^{1}\Omega HATS_3(u)$ increases from dimethylnitramide to 2-nitropropane, since it varies with bond length. Finally, Figure 6 suggests that the degree of molecular branching is important since it varies in extent between the structures of methylnitramide (chemical 39, $TD_{50}(obs) = 377.360$) and dimethylnitramide (chemical 2, $TD_{50}(obs) = 2.842$), and all descriptor values are then changed by this structural feature, although the change is the most pronounced for ${}^{1}\Omega HATS_3(u)$ and ${}^{7}\Omega R_7^+(m)$. In summary, we conclude that structural features such as molecular shape (linear, branched, cyclic, and polycyclic) and bond length could be important structural features contributing to the carcinogenicity of this set of nitrocompounds. However, further work is required to take account of the possible involvement of biotransformation of these chemicals to carcinogenic forms in vivo, by deriving models for some of the suspected key metabolites.

3. Conclusions

The relationship between the chemical structure of nitrocompounds and their carcinogenicity in female rats has been investigated with the principal objective of

Table 4. Names, CAS numbers, observed, predicted, and residual values of nitrocompounds used as test set in this QSAR study

Compound	Name	Carcinogenic potency[a]				
		CAS	TD_{50}	P_{obs}	P_{pred}	RES[b]
50	HC blue no. 1	2784-94-3	1680.55	−3.23	−2.74	0.49
51	2,3,5,6-Tetrachloro-4-nitroanisole	2438-88-2	65.65	−1.82	−1.85	−0.04
52	Nithiazide	139-94-6	605.88	−2.78	−2.46	0.32
53	5-Nitroacenaphthene	602-87-9	30.02	−1.48	−2.00	−0.52
54	*ortho*-Nitroanisole	91-23-6	69.87	−1.84	−1.00	0.84
55	Tetranitromethane	509-14-8	3.60	−0.56	−0.88	−0.33

[a] Carcinogenic activity estimated as TD_{50} (chronic dose in μmol/kg of body weight per day inducing tumors in 50% of the test animals at the end of a lifetime) and then log-transformed, that is, $P = -\log TD_{50}$.

[b] $RES = P_{obs} - P_{pred}$.

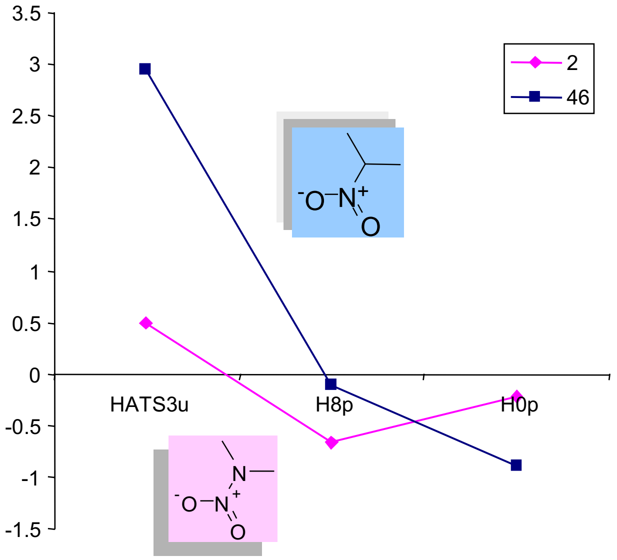

Figure 5. Profiles of $HATS_3(u)$, $H_8(p)$, and $H_0(p)$ for 2-nitropropane (chemical 46) and dimethylnitramide (chemical 2).

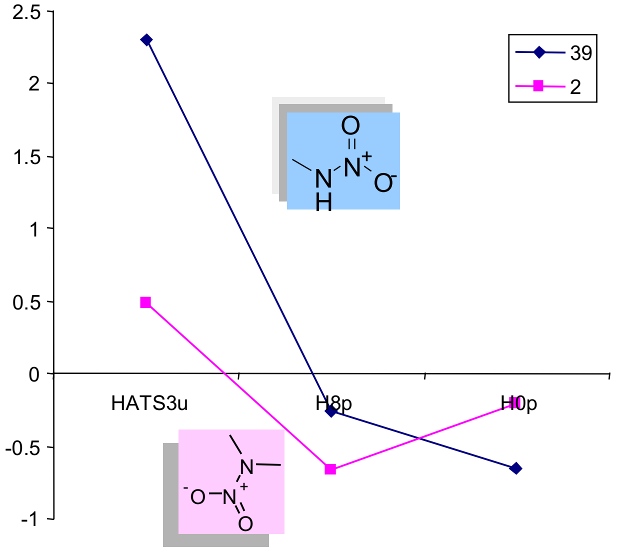

Figure 6. Profiles of $HATS_3(u)$, $H_8(p)$, and $H_0(p)$ for methylnitramide (chemical 39) and dimethylnitramide (chemical 2).

developing QSAR models for setting testing priorities, and for screening of putative new chemical molecules before their synthesis. The use of several different theoretical molecular descriptors, calculated only on the basis of knowledge of the molecular structure, and an efficient variable selection procedure, such as Genetic Algorithm, led to models with satisfactory predictive ability for carcinogenicity.

The most accurate QSAR model was based on GETAWAY molecular descriptor capturing a reasonable interpretation. In fact, structural features such as molecular shape—linear, branched, cyclic, and polycyclic—and bond length are some of the key structural factors flagging the carcinogenicity of this set of nitrocompounds. The practical value of the final QSAR model for screening and priority testing has been demonstrated. This model has better statistical parameters than others reported in the literature which make use of very similar data sets.[13] Moreover, it is able to detect and interpret key sub-structures in the biological activity like different softwares such as TOPS-MODE approach. Likewise, this is one of the first QSAR models published about the carcinogenicity of nitrocompounds, as far as we know, that can be used to detect potential carcinogenicity in similar compounds by using a rational design of sub-structures.

This model can be applied to novel nitrochemicals as it is based on theoretical molecular descriptors that might be easily and rapidly calculated. Finally, it must be underlined that the predicted data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was obtained.[33]

4. Experimental

4.1. Data set

A set of 55 aromatic and aliphatic nitrocompounds taken was used as the training set of chemicals. These had been experimentally assayed for carcinogenic potency (TD_{50}) in female rats.[20] For a given target site(s), and in the absence of tumors in control animals, TD_{50} is taken to be the chronic dose (in mg/kg of body weight per day) that induces tumors in half of the test animals at the end of a standard lifespan for the species.[20] Thus, a low value of TD_{50} indicates a potent carcinogen, whereas a high value reflects a weak carcinogen. The lowest TD_{50} values reported for each chemical, expressed in μmol/kg of body weight per day and log-transformed ($-\log TD_{50}$), was used in the following QSAR modeling. In order to obtain validated QSAR models, the data set was divided into the training and test sets (49 training and 6 test sets). Ideally, this division is performed such that points representing both the training and test sets are distributed within the whole descriptor space occupied by the entire data set, and each point of the test set is close to at least one point of the training set. K-means cluster analysis (k-MCA) was used in splitting the set of compounds to guarantee this distribution.[12,36–38] Tables 3 and 5 give a complete list of the chemicals along with the Simplified Molecular Input Line Entry Specification (SMILES) code, the Chemical Abstract Service (CAS) Registry Number, and experimental data for each chemical.

4.2. Molecular descriptors

Our models are based on nine different sets of descriptors with a long history of usage in structure–activity and structure–property correlation,[11,22,23,26,27,39–45] which are available in the DRAGON software package (version 2.1).[46] These sets of molecular descriptors were computed after optimizing the geometry of each molecule by using the quantum-chemical semi-empirical method, Austin Model 1 (AM1), implemented in MOPAC 6.0.[47]

Table 5. Names, CAS numbers, and SMILES of nitrocompounds used in this QSAR study

Compound	Name	CAS number	SMILES
1	*N*-Methyl-*N*′-nitro-*N*-nitrosoguanidine	70-25-7	N=C(N(N=O)C)N[N+](=O)[O−]
2	Dimethylnitramide	4164-28-7	O=[N+](N(C)C)[O−]
3	2-(2,2-Dimethylhydrazino)-4-(5-nitro-2-furyl)thiazole	26049-69-4	C1(N=C(SC=1)NN(C)C)C2=CC=C(O2)[N+]([O−])=O
4	2-Amino-5-(5-nitro-2-furyl)-1,3,4-thiadiazole	712-68-5	N([H])C1SC(C2OC([N+]([O−])=O)=C(C=2[H])[H])=NN=1
5	2-Hydrazino-4-(*p*-nitrophenyl) thiazole	26049-70-7	C1(N=C(SC=1)NN)C2=CC=C(C=C2)[N+]([O−])=O
6	1,2-Dihydro-2-(5-nitro-2-thienyl)quinazolin-4(3*H*)-one	33389-33-2	C1=CC=C2C(=C1)NC(NC2=O)C3=CC=C(S3)[N+]([O−])=O
7	4,6-Diamino-2-(5-nitro-2-furyl)-*S*-triazine	720-69-4	NC1=NC(=NC(=N1)C2=CC=C(O2)[N+]([O−])=O)N
8	4,6-Dimethyl-2-(5-nitro-2-furyl) pyrimidine	59-35-8	C1(=CC(=NC(=N1)C2=CC=C(O2)[N+]([O−])=O)C)C
9	2,4-Dinitro-6-*tert*-butylphenylmethanesulfonate	29110-68-7	CS(=O)(=O)OC1=C(C=C(C=C1C(C)(C)C)[N+]([O−])=O)[N+](=O)[O−]
10	2-Amino-5-(5-nitro-2-furyl)-1,3,4-oxadiazole	3775-55-1	O1C(=NN=C1C2OC(=CC=2)[N+](=O)[O−])N
11	l-5-Morpholinomethyl-3-[(5-nitrofurfurylidene)amino]-2-oxazolidinone. HCl	3031-51-4	O=[N+]([O−])C(O3)=CC=C3/C=N/N1C(OC(CN2CCOCC2)C1)=O
12	2-Hydrazino-4-(5-nitro-2-furyl)thiazole	26049-68-3	NNC1=NC(=CS1)C2=CC=C(O2)[N+]([O−])=O
13	1-Nitropyrene	5522-43-0	O=[N+](C1=CC=C2C3=C4C(=CC=C13)C=CC=C4C=C2)[O−]
14	Formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide	3570-75-0	[O−][N+](=O)C1=CC=C(O1)C2=CSC(=N2)NNC=O
15	4-(2-Hydroxyethylamino)-2-(5-nitro-2-thienyl)quinazoline	33389-36-5	N1C(=NC(=C2C=CC=CC=12)NCCO)C3=CC=C(S3)[N+](=O)[O−]
16	2,4-Dinitrotoluene	121-14-2	CC1=C(C=C(C=C1)[N+](=O)[O−])[N+](=O)[O−]
17	4-Morpholino-2-(5-nitro-2-thienyl)quinazoline	58139-48-3	C1=CC=C2C(=C1)N=C(N=C2N3CCOCC3)C4=CC=C(S4)[N+]([O−])=O
18	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]formamide	24554-26-5	[O−][N+](=O)C1=CC=C(O1)C2=CSC(=N2)NC=O
19	*N*-([3-(5-Nitro-2-furyl)-1,2,4-oxadiazole-5-yl]-methyl)acetamide	36133-88-7	O=C(C)NCC1=NC(=NO1)C2=CC=C(O2)[N+]([O−])=O
20	1-[(5-Nitrofurfurylidene)amino]-2-imidazolidinone	555-84-0	[O−][N+](=O)C1=CC=C(O1)C=NN2CCNC2=O
21	4-Methyl-1-[(5-nitrofurfurylidene)amino]-2-imidazolidinone	21638-36-8	C1(NC(CN1/N=C/C2=CC=C(O2)[N+](=O)[O−])C)=O
22	2,2,2-Trifluoro-*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]acetamide	42011-48-3	[N+]([O−])(=O)C1OC(=CC=1)C2N=C(SC=2)NC(=O)C(F)(F)F
23	2-Amino-4-(5-nitro-2-furyl)thiazole	38514-71-5	C1(N=C(SC=1)N)C2=CC=C(O2)[N+]([O−])=O
24	Nitrobenzene	98-95-3	O=[N+](C1=CC=CC=C1)[O−]
25	Acetone[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazone	18523-69-8	C(/C)(C)=N\NC1=NC=C(S1)C2=CC=C(O2)[N+](=O)[O−]
26	5-Nitro-2-furaldehyde semicarbazone	59-87-0	O=[N+](C1=CC=C(O1)/C=N/NC(=O)N)[O−]
27	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]acetamide	531-82-8	[N+]([O−])(=O)C1OC(=CC=1)C2N=C(SC=2)NC(=O)C
28	5-(5-Nitro-2-furyl)-1,3,4-oxadiazole-2-ol	2122-86-3	C1(C2OC(=NN=2)O)OC([N+](=O)[O−])=CC=1
29	*N*-[5-(5-Nitro-2-furyl)-1,3,4-thiadiazol-2-yl]acetamide	2578-75-8	CC(=O)NC1=NN=C(S1)C2=CC=C(O2)[N+]([O−])=O
30	*p*-Nitroaniline	100-01-6	O=[N+](C1=CC=C(C=C1)N)[O−]
31	8-Nitroquinoline	607-35-2	O=[N+](C1=CC=CC2=CC=CN=C12)[O−]
32	Azathioprine	446-86-6	CN1C(=C(N=C1)[N+](=O)[O−])SC2=NC=NC3=C2N=CN3
33	AF-2	3688-53-7	NC(=O)/C(=C/C1=CC=C(O1)[N+](=O)[O−])C2=CC=CO2
34	*N*,*N*′-[6-(5-Nitro-2-furyl)-*S*-triazine-2,4-diyl]bisacetamide	51325-35-0	C1(=NC(=NC(=N1)NC(C)=O)C2=CC=C(O2)[N+](=O)[O−])NC(C)=O
35	3-(5-Nitro-2-furyl)-imidazo(1,2-alpha) pyridine	75198-31-1	C1/N2C(/C=C\C=1)=NC=C2C3=CC=C(O3)[N+]([O−])=O
36	1-(2-Hydroxyethyl)-3-[(5-nitrofurfurylidene)amino]-2-imidazolidinone	3/3/5036	N(=C/C1OC(=CC=1)[N+](=O)[O−])/N2C(=O)N(CC2)CCO
37	1,2-Dimethyl-5-nitroimidazole	551-92-8	C1(N(C(C)=NC=1)C)[N+](=O)[O−]
38	*trans*-2-[(Dimethylamino)methylimino]-5-[2-(5-nitro-2-furyl)vinyl]-1,3,4-oxadiazole	55738-54-0	[N+]([O−])(=O)C1OC(=CC=1)/C=C/C2O/C(=N/CN(C)C)NN=2
39	Methylnitramide	598-57-2	CN[N+](=O)[O−]
40	2-Amino-5-nitrothiazole	121-66-4	O=[N+](C1=CN=C(S1)N)[O−]
41	5-Nitro-2-furamidoxime	772-43-0	O1C(=CC=C1[N+](=O)[O−])/C(=N/O)N
42	1-[(5-Nitrofurfurylidene)amino]hydantoin	67-20-9	O=C1N(CC(=O)N1)/N=C/C2=CC=C(O2)[N+](=O)[O−]
43	Metronidazole	443-48-1	N1(C(=CN=C1C)[N+](=O)[O−])CCO
44	5-Nitro-2-furanmethanediol diacetate	92-55-7	CC(=O)OC(C1=CC=C(O1)[N+](=O)[O−])OC(=O)C
45	Chloramphenicol	56-75-7	O[C@@H](C1=CC=C(C=C1)[N+](=O)[O−])[C@@H](NC(=O)C(Cl)Cl)CO
46	2-Nitropropane	79-46-9	CC([N+](=O)[O−])C
47	Trinitroglycerin	55-63-0	O=[N+](OC(CO[N+](=O)[O−])CO[N+](=O)[O−])[O−]
48	Nitroethane	79-24-3	O=[N+](CC)[O−]
49	Trifluralin	1582-09-8	O=[N+](C1=C(C(=CC(=C1)C(F)(F)F)[N+](=O)[O−])N(CCC)CCC)[O−]

In addition, in order to provide energy information, four quantum-chemical descriptors, energy of highest occupied molecular orbital (EHOMO) and lowest unoccupied molecular orbital (ELUMO), local dipole index and polarization calculated by the semi-empirical molecular orbital program MOPAC[47] (AM1 for geometry optimization), are added as electronic descriptors. The nature of the descriptors is given below.

4.2.1. Constitutional descriptors (Cons). A set of 34 constitutional descriptors, including molecular weight, van der Waals volume, atomic electronegativities and polarizabilities, number of atoms, non-H atoms, covalent bonds, multiple bonds, bond orders, aromatic ratio, number of double and triple bonds, aromatic bonds, as well as different types of (*n*-membered) rings and benzene-like rings.[4]

4.2.2. BCUT descriptors (BCUT). These descriptors are derived from the positive and negative eigenvalues of the adjacency matrix of the target molecule, weighting the diagonal elements with atom weights.[4,48]

4.2.3. Galvez topological charges indices descriptors (Gal). This set consists of 21 descriptors representing the first 10 eigenvalues (absolute values) obtained from a corrected adjacency matrix.[4]

4.2.4. 2D-autocorrelation descriptors (2D-A). This set consists of 96 descriptors calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag). The 2D-autocorrelations by Moreau–Broto (ATS), Moran (MATS), and Geary (GATS) Algorithms are calculated from lag 1 to lag 8 for four different weighting schemes.[4,49]

4.2.5. Randić molecular profile descriptors (Ran). A set of 41 descriptors derived from the distance distribution moments of the geometry matrix, defined as the average row sum of its entries raised at the *k*th power and normalized by *k*![4]

4.2.6. Geometrical descriptors (Geo). A set of 29 descriptors consisting of different kinds of conformation-dependent descriptors based on molecular geometry.[4] These include descriptors such as the 3D-Wiener index, the 3D-Balaban index, the 3D-Harary index average geometric distance degree, the D/D index, the average distance/distance degree gravitational index G1, the gravitational index G2 (bond-restricted), and the radius of gyration (mass weighted).

4.2.7. RDF descriptors (RDF). This set consists of 135 descriptors obtained by radial basis functions centered on different interatomic distances (from 0.5 to 15.5 Å). By including characteristic atomic properties of the involved atoms, the RDF codes might be used in different tasks to fit the requirements of the information to be modeled.[4]

4.2.8. WHIM descriptors (WHIM). A set of 99 descriptors obtained as statistical indices of the atoms projected onto the three principal components obtained from weighted covariance matrices of the atomic coordinates.[4]

4.2.9. GETAWAY descriptors. Two sets of molecular descriptors, that in theory are closely related, have been devised. These are the *H*-GETAWAY and the *R*-GETAWAY descriptors. The first set consists of 107 *H*-GETAWAY descriptors calculated from the leverage matrix obtained by the centered atomic coordinates, that is, the *molecular influence* matrix (*H*), while the remaining 90 *R*-GETAWAY descriptors are calculated from the *influence/distance* (*R*) matrix.

The molecular influence matrix is defined in terms of the molecular matrix, M, $H = M\cdot(M^{\mathrm{T}}\cdot M)^{-1}\cdot M^{\mathrm{T}}$, which has as many rows as the number of atoms in the molecule (hydrogen atoms included) and three columns corresponding to the Cartesian coordinates x, y, z of each atom in the optimized molecular structure. Notice that atomic coordinates are assumed to be calculated with respect to the geometrical center of the molecule in order to obtain translational invariance.[23] The diagonal elements h_{ii} of this *molecular influence* matrix, called leverages, encode atomic information and represent the 'influence' of each atom in determining the whole shape of the molecule. In fact, mantle atoms always have higher h_{ii} values than atoms near the center of the molecule. The off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interact with the *i*th atom. A negative sign for the off-diagonal element means that the two atoms occupy opposite molecular regions with respect to the center, and hence there is a low mutual degree of accessibility. On the other hand, matrix R, a symmetrical matrix whose elements resemble the single terms in the sums of the gravitational indices, is defined as $[R]_{ij} \equiv \left[\frac{\sqrt{h_{ii}\cdot h_{jj}}}{r_{ij}}\right], i \neq j$, where h_{ii} and h_{jj} are the leverages of the two considered atoms and r_{ij} their geometric distance. Obviously, the diagonal elements of matrix R are zero, and the largest values of its off-diagonal elements derive from the most external atoms (i.e., with high leverages) and simultaneously next to each other in the molecular space (i.e., small interatomic distance).

Finally notice that, in many of these H and R descriptors, the molecule atoms are weighted in such a way as to account for atomic mass, polarizability, van der Waals volume, and electronegativity, with the aim of incorporating relevant chemical information.

4.3. Modeling technique

The objective was to obtain a mathematical function (Eq. 6) that best describes carcinogenic potency, P ($= -\log \mathrm{TD}_{50}$), as a linear combination of the predictor X-variables (descriptors), with the coefficients a_k. Such coefficients were optimized by means of multi-linear regression (MLR) analysis, implemented in version 1.0-2004,[50] using the chemicals under study.

$$P = a_1X_1 + a_2X_2 + \cdots + a_kX_k + a_0 \qquad (6)$$

An exhaustive search for the best regressions within a wide set of descriptors requires extensive computational resources and is time-consuming, given the extremely high number of possible descriptor combinations.

4.4. Feature selection

The Genetic Algorithm (GA) approach was used as the variable selection method.[51,52] Starting from a population of 100 random models with a number of variables equal to or less than a user-defined maximum value, the algorithm explores new combinations of variables, selecting them by a mechanism of population evolution involving processes analogous to biological reproduction/mutation. The models based on the selected subsets of variables were tested and evaluated by the cross-validated explained variance (q^2), and only the best quality models were retained in the population undergoing the evolution procedure. The variables for the obtained models were found to be highly significant, within a 95% confidence level.

In any multiple linear-based QSAR it is desirable that the variables included in the model are not interrelated to each other. Highly correlated variables clearly contain redundant information that might be more usefully encoded by a single variable. Further, and most importantly from the point of view of a QSAR model, correlated independent variables lead to multi-colinearity, which can cause problems in interpreting the individual estimated coefficients. One very useful and informative approach of avoiding multi-colinearity is the *orthogonal descriptors* technique suggested by Randić some years ago.[53,54] In the Randić's approach, after choosing a starting descriptor, subsequent descriptors are added only as their orthogonal complements to the descriptors already present. This approach has the advantages that: (a) the regression coefficients are stable (i.e., they do not change as new descriptors are added); and (b) the new information supplied by each additional descriptor is clearly distinguishable in the final equation statistics. In order to address the problem of multi-colinearity, we have applied Randić's approach by inserting the variables in descending order based on their relative contributions to q^2_{LOO}, and then pursuing their orthogonalization. The resulting orthogonal-descriptor model was standardized afterwards.

4.5. Model evaluation

Several diagnostic statistical tools were used for evaluating our model equations, in terms of the criteria *goodness-of-fit* and *goodness-of-prediction*. Measures of *goodness-of-fit* have been estimated by standard statistics such as determination coefficient, R^2; the standard deviation, S; the Fisher's statistic, F; as well as the ratio between the number of compounds and the number of adjustable parameters in the model, known as ρ statistics. We have also checked the validity of the pre-adopted parametric assumptions (i.e., the linearity of the model, normality, homocedasticity, and no correlation of the residuals and non-multi-colinearity between the descriptors), which is another important aspect in the application of multiple linear statistical-based approaches.[55] *Goodness-of-prediction* of the final model has been assessed by means of internal cross-validation (CV), basically by leave-one-out (LOO-CV), bootstrap, scrambling techniques, and external validation (verified by q^2_{ext}).[4] In LOO-CV, the statistics adopted for estimating the predictive ability of models are q^2 and standard deviation of LOO-CV (S_{CV}). In *scrambling*, the objective is to verify models with chance correlations, thus the quality of models is checked by R^2, after which the sequence of response vectors has been randomly modified 300 times, while *bootstrap* is a procedure of building training and evaluation sets, that is repeated 5000 times, all squared difference between the true response and predicted response of each molecule from the evaluation set are collected in *PRESS* and the average predictive power is expressed as q^2_{Boot}. The external predictability of QSAR models was checked, by comparing the predicted and the actual data, evaluating the prediction errors, and computing the external standard deviation (S_{ext}) and q^2_{ext}. In Table 3 is given a list of six nitrochemicals ($N_{\mathrm{ext}} = 6$) that constitute the external set together with CAS Registry Number and experimental data for each chemical. It is important to remark that these nitrochemicals never were used in QSAR model development.

Apart from the classical regression parameters listed above, we analyzed other important statistics, namely the Kubinyi function (FIT), the Akaike's information criterion (AIC), and the *Friedman's* lack-of-fit function (LOF).[4] These gave us enough criteria for comparing models with different parameters, numbers of variables, and chemicals.

In summary, good overall quality of the models is indicated by a large F (significance of the models), FIT, and ρ values; small AIC and LOF (overfitting) values; R^2 (goodness-of-fit) and q^2 (predictability) values close to one. In the case of R^2_{Scram}, this should have a value close to zero, as it checks random correlations.

The presence of outliers and chemicals very structurally influential in determining model parameters [i.e., compounds with high leverage value (h) greater than $3p/n$ (h^*), where n is the number of training chemicals, p is the number of model parameters (intercept and descriptors), and h^* 'warning leverage'] was verified by Williams plot.[33,34] Also the reliability of the predicted data with regard to chemical domain was verified by the leverage approach: the prediction for chemicals of the test set must be considered reliable only for those chemicals that fall within the applicability domain on which the model was obtained.

Acknowledgments

The authors thanks to Prof. Reinaldo Molina Ruiz by the useful comments and suggestions. The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) (**SFRH/BD/22692/2005**) and the Cuban Higher Education Ministry (R&D project number 6.181-2006) for financial support.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2007.11.029.

References and notes

1. Tomatis, L.; Huff, J.; Hertz-Picciotto, I.; Sandler, D. P.; Bucher, J.; Boffetta, P.; Axelson, O.; Blair, A.; Taylor, J.; Stayner, L.; Barrett, J. C. *Carcinogenesis* **1997**, *18*, 97.
2. Commission of the European Communities. White Paper on the Strategy for a future Chemicals Policy, 2001.
3. Knight, A.; Bailey, J.; Balcombe, J. *ATLA* **2006**, *34*, 19.
4. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
5. González, M. P.; Diaz, H. G.; Cabrera, M. A.; Ruiz, R. M. *Bioorg. Med. Chem.* **2004**, *12*, 735.
6. Gonzalez-Diaz, H.; Tenorio, E.; Castanedo, N.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 1523.
7. Helguera, A. M.; González, M. P.; Briones, J. R. *Polymer* **2004**, *45*, 2045.
8. González-Diaz, H.; Marrero, Y.; Hernandez, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castanedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
9. Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614.
10. Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. *Carcinogenesis* **2001**, *22*, 1561.
11. Helguera, A. M.; Duchowicz, P. R.; Cabrera, M. A.; Castro, E. A.; Cordeiro, N.; González, M. P. *Chemometr. Intell. Lab. Syst.* **2006**, *81*, 180.
12. Helguera, A. M.; Cabrera Perez, M. A.; González, M. P. *J. Mol. Model.* **2006**, *12*, 769.
13. Helguera, A. M.; Perez, M. A.; Combes, R. D.; Gonzalez, M. P. *Toxicology* **2006**, *220*, 51.
14. Helguera, A. M.; Cabrera Perez, M. A.; González, M. P.; Ruiz, R. M.; Gonzalez-Diaz, H. *Bioorg. Med. Chem.* **2005**, *13*, 2477.
15. Helguera, A. M.; Cabrera, M. A.; Combes, R. D.; González, M. P. *Curr. Comput. Aided Drug Des.* **2005**, *1*, 237.
16. Contrera, J. F.; Matthews, E. J.; Daniel Benz, R. *Regul. Toxicol. Pharmacol.* **2003**, *38*, 243.
17. Helguera, A. M.; González, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189–202.
18. Yuta, K.; Jurs, P. C. *J. Med. Chem.* **1981**, *24*, 241.
19. Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. *Chem. Rev.* **2000**, *100*, 3697.
20. Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L. *Environ. Health Perspect.* **1999**, *107*, 527.
21. Benigni, R.; Passerini, L.; Rodomonte, A. *Environ. Mol. Mutagen.* **2003**, *42*, 136.
22. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693.
23. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
24. González, M. P.; Teran, C.; Teijeira, M.; Besada, P. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2641.
25. González, M. P.; Teran, C.; Teijeira, M.; Gonzalez-Moa, M. J. *Eur. J. Med. Chem.* **2005**, *40*, 1080.
26. Saiz-Urra, L.; Gonzalez, M. P.; Fall, Y.; Gomez, G. *Eur. J. Med. Chem.* **2006**.
27. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Besada, P. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2641.
28. IARC. In Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans: Lyon, France, 1989; Vol. 46, p 458.
29. Scherf, H. R.; Frei, E.; Wiessler, M. *Carcinogenesis* **1989**, *10*, 1977.
30. Zaidi, N. H.; O'Connor, P. J.; Butler, W. H. *Carcinogenesis* **1993**, *14*, 1561.
31. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
32. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
33. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. *Altern. Lab. Anim.* **2005**, *33*, 155.
34. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Podda, G.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15*, 2544.
35. Rostkowska, K.; Zwierz, K.; Różański2, A.; Moniuszko-Jakoniuk, J.; Roszczenko, A. *Pol. J. Environ. Stud.* **1998**, *7*, 321.
36. Gonzalez, M. P.; Dias, L. C.; Helguera, A. M.; Morales, Y. R.; de Oliveira, L. G.; Torres, G. L.; Gonzalez-Dı́az, H. *Bioorg. Med. Chem.* **2004**, *12*, 4467.
37. González, M. P.; Helguera, A. M.; Cabrera, M. A. *Bioorg. Med. Chem.* **2005**, *13*, 1775.
38. González, M. P.; Gonzalez Diaz, H.; Molina Ruiz, R.; Cabrera, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
39. Fatemi, M. H.; Goudarzi, N. *Electrophoresis* **2005**, *26*, 2968.
40. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Gonzalez-Moa, M. J. *Eur. J. Med. Chem.* **2005**, *40*, 1080.
41. Panek, J. J.; Jezierska, A.; Vracko, M. *J. Chem. Inf. Model.* **2005**, *45*, 264.
42. Grodnitzky, J. A.; Coats, J. R. *J. Agric. Food Chem.* **2002**, *50*, 4576.
43. Gramatica, P.; Pilutti, P.; Papa, E. *Atmos. Environ.* **2003**, *37*, 3115.
44. Deconinck, E.; Xu, Q. S.; Put, R.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. *J. Pharm. Biomed. Anal.* **2005**, *39*, 1021.
45. Papa, E.; Battaini, F.; Gramatica, P. *Chemosphere* **2005**, *58*, 559.
46. Todeschini, R.; Consonni, V.; Pavan, M. 2002.
47. Frank, J. MOPAC (Computer software), version 6.0; Seiler Research Laboratory, US Air Force Academy, Colorado, Springs Co., 1993.
48. González, M. P.; Teran, C.; Teijeira, M.; Besada, P.; Gonzalez-Moa, M. J. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3491.
49. Broto, P.; Moreau, G.; Vandicke, C. *Eur. J. Med. Chem.* **1984**, *19*, 79.
50. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. TALETE srl: Milano, 2004.
51. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 285.
52. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 393.
53. Randić, M. *New J. Chem.* **1991**, *15*, 517.
54. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
55. Pestana, M.; Gageiro, J. Análise de dados para Ciências Sociais. A Complementaridade do SPSS.; Lisboa: Edições Sílabo, 2000.