Diagnostic Tools to Determine the Quality of "Transparent" Regression-Based QSARs: The "Modelling Power" Plot

Salvador Sagrado*,† and Mark T. D. Cronin‡

Departamento de Química Analítica, Universitat de València, C/Vicente Andrés Estellés s/n, E-46100 Burjassot, Valencia, Spain, and School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England

Received October 7, 2005

A bivariate plot is presented for comparing two or more QSAR models. It is based on two new statistics associated with a regression model, the "descriptive power" (*Dp*), which is estimated through the relative uncertainty of model coefficients, and the "predictive power" (*Pp*), which is estimated through both the fitted and cross-validated explained variance of the response variable (i.e., biological activity). An algorithm was developed for performing equivalent multiple linear regression and partial-least-squares calculations. The results were validated by comparison with widely accepted commercial software. *Dp* and *Pp* statistics are defined to vary from 0 to 100%, so the modeler has a intuitive impression of the descriptive (i.e., global importance of the selected descriptors) and predictive (i.e., possibility of performing QSAR or just SAR estimations) power. These statistics represent a point in the *Dp* versus *Pp* "modelling power" plot, which facilitates visual multiple models comparison, but also could be used to substitute classical statistics and could even be combined to obtain a unique parameter to define (or compare) the model's quality.

INTRODUCTION

The overall objective of quantitative structure−activity relationships (QSARs) is to find independent variables that give a statistically significant correlation for a series of chemicals to a biological activity of interest.[1] QSARs are of immense practical use in the product development of new chemical entities, as well as for regulatory applications including the prediction of toxicity and fate. However, there are no formal guidelines as to what may constitute a "statistical significant correlation", especially with regard to acceptance criteria that may be required for regulatory use.

There are many statistical approaches to developing a QSAR. Of these, regression analyses are the most commonly applied and are fundamental in their simplicity and transparency. The robustness, in terms of statistical fit and predictivity, of the regression model is an important factor in deciding the utility of a QSAR. Regression-based QSAR implies modeling with a "**y**−**X**" relationship, where the vector **y** represents a given biological activity (response or dependent variable) and **X** represents the matrix of descriptor variables (predictor or independent variables) when related by a particular regression algorithm [i.e., multiple linear regression (MLR), partial least squares (PLS), etc.]. The **y**-response variable can be linked directly to the **X** descriptors through their regression coefficient(s), **b**, conferring "transparency" to the model.[2] Simultaneously, the predictive ability of the model can be assessed by analyzing the model residuals.

It is well recognized that QSARs can be applied for many purposes. With regard to the prediction of the toxicity and fate of chemicals, there is an anticipated increase in use, particularly driven by regulatory needs. New legislation in the European Union, such as the Registration, Evaluation, and Authorisation of Chemicals (REACH) system, as well as the Amendments to the Cosmetics Directive, requires alternatives to traditional animal tests. There are clear advantages in using in silico technologies as part of the process to achieve this. In order for QSARs to be implemented into a regulatory context, a process of acceptance will be required. Acceptance of a QSAR will be at two levels: first, some form of evaluation of the model (which may lead to validation) and, second, being able to place some form of confidence on an individual prediction. Principles for the validation of (Q)SARs have been proposed by the Organisation of Economic Co-operation and Development (OECD) which include an assessment of the goodness-of-fit, robustness, and predictivity of a model.[3] While the principal is defined, guidance is still required as to how a QSAR may be evaluated, particularly with regard to its statistical performance.[4−6] Two main measures of the reliability of a regression-based QSAR are the statistical fit and predictivity associated with the model and the stability of the individual parameters. While many QSAR modelers pay good attention to the quantification of predictive ability, the stability of the variables is seldom considered. This is despite many models being highly multivariate in nature and much being placed in the mechanistic interpretation of variables from their coefficients. In addition to other strategies for model evaluation, resampling techniques make possible the assessment of both the predictive ability (i.e., cross-validation) and the stability of the coefficients (jack-knifing, as a part of the cross-validation process).[7]

Although thousands of QSAR models have been published in the literature, there is still no uniformity and consistency in the presentation of the statistics associated with the model, or even in the definition of those that are critical to permitting a realistic comparison of different models. There have been

* Corresponding author e-mail: sagrado@uv.es. Tel. +34 96 354 4878. Fax. +34 96 354 4953.
† Universitat de València.
‡ Liverpool John Moores University.

10.1021/ci050445c CCC: $33.50
Published on Web 01/31/2006

recent attempts to present and recommend acceptability criteria[8] and to define a (single) global index able to account for the quality and confidence of QSARs.[9] Efforts such as these illustrate the lack of standardization in this regard but will facilitate the comparison of models and progress in this area.

Regression-based QSARs can be well described in terms of their statistical fit, as well as by cross-validation and subsampling techniques, to give estimates of their predictivity. However, recent work has indicated that the "traditional" suite of model statistics may be misleading, especially for complex multivariate problems; therefore, indices such as the root-mean-square error of prediction may be more realistic.[10] This paper introduces two further new intuitive model statistics: (i) the "descriptive power" (*Dp*), which estimates model reliability (stability) by accounting for the estimated uncertainty (i.e., an approximate 95% confidence interval[7]) of the regression coefficients, and (ii) The "predictive power" (*Pp*), which estimates the model's predictive ability by accounting for the explained **y** variance in the cross-validation [i.e., the percentage of the total initial variance in the response variable removed by the models, when they are obtained following the removal of a portion of the data,[7] corrected by the difference between the explained **y** variance for the full (all the data) and the cross-validated models].

These two new statistical measures, when plotted together, form the "modelling power" plot (i.e., the plot of *Dp* vs *Pp*). In addition, when *Dp* and *Pp* are combined, they provide a single index describing the quality of the model, the "modelling power" (*Mp*), that summarizes the descriptive and predictive features of a QSAR. This allows for the description of a single QSAR model or the comparison of two or more QSAR models. As of yet, the use of modeling power has not been demonstrated in QSARs for toxicity.

The aim of this paper, therefore, was to investigate the use of the modeling power plot for regression-based toxicological QSARs. This is to simplify the tasks related to the comparison of models (i.e., variable selection, validation schemes, impact of outliers, number of compounds and data range, MLR vs PLS algorithms, etc.). To investigate the use of the modeling power plot, one of the oldest and most fundamental toxicological databases has been investigated, the narcotic activity to tadpoles [compiled from original data by Abraham et al.[11] and analyzed recently by Agrawal et al].[12] The advantage of this approach is that it forms a historical and well-recognized database, and the mechanism of toxic action (narcosis) is fundamental and its modeling is well understood. It should be noted that there are at least two narcotic mechanisms represented in the database (nonpolar and polar narcosis), so a direct relationship with the octanol−water partition coefficient is not to be expected.

METHODS

Toxicity Data. The concentrations of 123 compounds causing 50% lethality (C_{nar}, in mol dm^{-3}) to the tadpole were taken from ref 12. It should be noted that the data set presented by Agrawal et al.[12] contains duplicate entries (without comment) of six compounds (butan-1-ol, ethanol, methanol, *N*-ethylurethane, octan-1-ol, and propan-1-ol). The duplicate entries for these compounds have different reported experimental toxicity values but the same values for the descriptors. Therefore, there are only 117 individual compounds in this data set. It is beyond the scope of this paper to investigate the implications of duplicate entries. However, for practical purposes and for consistency of comparison of the regression results reported by Agrawal et al.,[12] the duplicate toxicity data have been utilized as replicate toxicity values for these compounds. Therefore, all 123 data were used to develop models in this analysis, with a full appreciation of the problems that this may entail. All of the compounds are assumed to be narcotic in action and cover a broad range of organic structures. The response variable ($\mathbf{y}_{123x1}$) was taken as the inverse of the negative logarithm of the activity in moles per liter (log $1/C_{nar}$).

Physicochemical Properties and Descriptors. This study has utilized two sets of descriptors. The first set is taken from empirical measurements and consists of the five Abraham's solvatochromic parameters (variable numbers 1−5 in Table 1). For simplicity, we have used Abraham's nomenclature for the descriptors: the solute excess molar refractivity in units of cm^3 mol^{-1}/10 (*E*), the solute dipolarity/polarizability (*S*), the overall or summation hydrogen-bond acidity and basicity (*A* and *B* respectively), and the McGowan characteristic volume in units of cm^3 mol^{-1}/100 (*V*). The second set of descriptors consists of five topological indices (variable numbers 6−10 in Table 1). The topological indices used are: the Szeged index (*Sz*), the Wiener index (*W*), the first-order path molecular connectivity index ($^1\chi$), the Balaban index (*J*), and the logarithm of a branching index based on molecular topology [log(*RB*)]. All of the descriptors were taken from Agrawal et al.[12] These variables were combined to form the initial predictor matrix ($\mathbf{X}_{123x10}$).

CALCULATIONS

MLR and PLS analyses were performed using routines developed in MATLAB 5.3 [MATLAB, version 5.3.0.10183 (R11), The Mathworks Inc., Natick, MA]. The results from MLR were compared to those obtained using Statgrahics Plus 5.1 (Statistical Graphics Corp.) to ensure consistency of the algorithms. The results from PLS were compared to those obtained using The Unscrambler, version 7.6 (CAMO ASA), which accounts for most of the criteria (i.e., estimation of **b** coefficient uncertainties and explained variance based on cross-validation) published in Martens and Martens.[7] For simplicity, the given number of latent variables (LVs) in the PLS results has been denoted as $PLS_{(LVs)}$.

The MATLAB algorithm allowed for the calculation of equivalent MLR and $PLS_{(LVs)}$ results. For instance, 11 *b* coefficients were determined from MLR using the original $\mathbf{y}-\mathbf{X}$ data (the classical calculation). When scaled (i.e., autoscaled) $\mathbf{y}-\mathbf{X}$ data was used, 10 standardized *b* values were obtained (without the intercept, b_0; as in the PLS approach), which can be descaled[7] to convert them back into 11 *b* values (which are equivalent to those obtained with the classical calculation). In addition, $PLS_{(LVs)}$ using scaled (i.e. autoscaled) $\mathbf{y}-\mathbf{X}$ data (the conventional calculation) was performed to obtain 10 standardized *b* values, which can be descaled to convert them into 11 *b* values (as in conventional MLR format). Note that the last feature confers the same "transparency" to $PLS_{(LVs)}$ as MLR; in fact, in the present instance, the $PLS_{(10)}$ results were exactly the same as those from MLR.

Table 1. Statistics and Coefficients for Some MLR Models for Tadpole Narcosis[a]

data		model statistics		variable / coefficient	b_0	*E* / b_1	*S* / b_2	*A* / b_3	*B* / b_4	*V* / b_5	*Sz* / b_6	*W* / b_7	$^1\chi$ / b_8	*J* / b_9	log(*RB*) / b_{10}	Mn
No	123	*SE*	0.53	*b*	0.9626	0.9560	−1.3376	−0.5117	−1.7896	0.7709	0.0004	0.0182	0.6475	−0.0556	−0.0679	1
Nv	10	*F*	43.2	*s*(*b*)	0.3551	0.2565	0.2337	0.2397	0.2687	0.4389[b]	0.0010[b]	0.0066	0.1345	0.0926[b]	0.0277	
100*K*	57%	r^2	0.794	100·*t s*(*b*)/*b*	73.1	53.2	34.6	92.8	29.8	112.8	539.7	71.7	41.1	329.9	80.8	
rank	3−5	q^2	0.628	100·*U*(*b*)/*b*	215.7	80.7	97.9	273.0	51.9	705.9	1137.0	121.0	188.3	586.2	163.9	
		RMSEC	0.51													
		RMSECV	0.68													
No	123	*SE*	0.54	*b*	0.7712	1.0463	−1.2475		−1.9686	1.2220		0.0144	0.5087		−0.0528	2
Nv	7	*F*	59.7	*s*(*b*)	0.2539	0.1991	0.2124		0.2578	0.3824		0.0050	0.1143		0.0177	
100*K*	65%	r^2	0.784	100·*t s*(*b*)/*b*	65.2	37.7	33.7		25.9	62.0		68.9	44.5		66.5	
rank	2−4	q^2	0.673	100·*U*(*b*)/*b*	161.3	88.6	76.5		47.8	321.0		65.3	181.0		71.3	
		RMSEC	0.52													
		RMSECV	0.68													
No	123	*SE*	0.58	*b*	1.0079	1.2975	−1.0017		−1.7925	2.4560		0.0120			−0.0379	3
Nv	6	*F*	57.1	*s*(*b*)	0.2676	0.2059	0.2211		0.2746	0.2838		0.0054			0.0187	
100*K*	60%	r^2	0.747	100·*t s*(*b*)/*b*	52.6	31.4	43.7		30.3	22.9		88.4			97.9	
rank	3−3	q^2	0.651	100·*U*(*b*)/*b*	174.4	52.2	72.9		49.3	81.4		132.1			184.7	
		RMSEC	0.56													
		RMSECV	0.66													
No	123	*SE*	0.59	*b*	0.8762	1.1468	−0.9803	0.1542	−1.8714	2.7088						4
Nv	5	*F*	64.7	*s*(*b*)	0.1669	0.1929	0.2236	0.2219	0.2836	0.1676						
100*K*	38%	r^2	0.735	100·*t s*(*b*)/*b*	37.7	33.3	45.2	285.0	30.0	12.3						
rank	3−4	q^2	0.691	100·*U*(*b*)/*b*	63.1	41.5	65.1	234.3	41.9	24.9						
		RMSEC	0.58													
		RMSECV	0.62													
No	123	*SE*	0.59	*b*	0.8884	1.1679	−1.0011		−1.7926	2.7025						5
Nv	4	*F*	81.2	*s*(*b*)	0.1656	0.1900	0.2211		0.2594	0.1670						
100*K*	45%	r^2	0.733	100·*t s*(*b*)/*b*	36.9	32.2	43.7		28.7	12.2						
rank	3−3	q^2	0.693	100·*U*(*b*)/*b*	63.1	40.8	64.2		43.9	25.1						
		RMSEC	0.60													
		RMSECV	0.62													

[a] Symbols: *No*, number of objects, here, compounds (with six duplicated compounds), in the data set; *Nv*, number of variables, here, descriptors, in the data set; *K*, correlation index for **X**,[7] a measure ranged between 0 and 1, of the degree of correlation in a matrix; rank, interval from the minimum to the maximum number of significant principal components of the **X** matrix,[7] a measure of the dimension of **X**; *SE*, standard error of the estimation; *F*, *F* ratio; r^2, squared multiple correlation coefficient calculated from fitted **y** estimates; q^2, squared multiple correlation coefficient calculated from cross-validated **y** estimates; *RMSEC*, root-mean-square error of calibration corresponding to the full model,[4] calculated from fitted **y** estimates; *RMSECV*, root-mean-square error of cross-validation corresponding to the submodels,[4] calculated from cross-validated **y** estimates; *b*, coefficient, b_0 for the intercept and b_1−b_{10} for the 10 variable coefficients; *s*(*b*), standard error of *b*; *t s*(*b*)/*b*, relative uncertainty of *b* estimated from the classic 95% confidence interval; and *U*(*b*)/*b*, relative uncertainty of *b* estimated from jack-knifing.[4] Mn (model number) is used for numbering the models in the text and plots. [b] Nonsignificant variables (probability, $p > 0.05$)

The MATLAB algorithm also allowed for the estimation of the uncertainty of the **b** coefficients, **U**(**b**), using the jack-knife approach,[7] performed during the cross-validation [leave-one-out (LOO) or Venetian blinds] process. The algorithm used here introduces a modification with respect to the cross-validation process used in the Unscrambler software and classical routines written in MATLAB PLS Toolbox. In these routines, after autoscaling the **y**−**X** data, two sample blocks, a calibration sample subset (to build a submodel) and a validation sample subset, are formed in each cross-validation step. Then, the calibration blocks are mean-centered and the validation block is scaled with respect to the corresponding information (the *means* from the calibration block). In the algorithm used in this study, after forming the sample blocks, the calibration block was autoscaled and the validation block was scaled with respect to the corresponding information (*means* and *standard deviations* from the calibration block). Therefore, the final result resembles a future possible use of the full model to make predictions, or in other words, each submodel was used as a definitive full model.

Descriptive Power of a Model. The "descriptive power" is an attempt to quantify the reliability of each descriptor in the model by assessing the stability of the regression coefficients. To achieve this, the estimated overall relative uncertainty of the coefficients was calculated. Specifically, the *Dp* of the model was calculated from the **b** and **U**(**b**) (11 × 1) vectors (including the intercept, b_0), according to

$$Dp = 100\{1 - \text{mean}[|\mathbf{U}(\mathbf{b})/\mathbf{b}|]\} \quad (1)$$

The **b** vector is obtained using the full model (i.e., all the compounds in the dataset). The **U**(**b**) vector is obtained by jack-knifing (in each cross-validation step, the submodel generates a vector of "perturbed respect to **b**" coefficients, **bp**). While traditional jack-knifing compares the individual cross-validation submodels to the average of the submodels, the algorithm used in this study, as suggested by Martens and Martens,[7] compares them to the full model. For each coefficient, b_j (in this example, j goes from 0 to 10), $U(b_j) = 2[\Sigma(b - bp_i)^2(No - 1)/No]^{0.5}$, where i represents the cross-validation step. Note that, in jack-knifing, the partial coefficient perturbations are summed not averaged. The uncertainty of the coefficient, **U**(**b**), can be regarded as an approximate 95% confidence interval.[7]

Ideally, **U**(**b**) could be zero, which corresponds to the maximum descriptive power (*Dp* = 100%). A situation in which the mean[**U**(**b**)/**b**] = 1 was considered unacceptable (*Dp* = 0), and therefore, the minimum *Dp* was set to 0 (there is no interest in discriminating between wholly unacceptable models with negative *Dp* values). Consequently, the range of *Dp* is 0−100%.

Predictive Power of a Model. The "predictive power" quantifies the predictive capability of a model (i.e., the ability to estimate reliable *y* values for compounds not included in the model) by means of the explained **y** variance in the cross-validation process (EVCV). EVCV is the percentage of the total initial variance in the response variable removed by the models (when they are obtained after removing some of the data).[7] If there is little difference between the explained **y** variance in the full data set (EVC) and the cross-validated EVCV, then the model can be assumed to have good predictive capability. The *Pp* of a model is therefore calculated by

$$Pp = 100[\text{EVCV} - (\text{EVC} - \text{EVCV})] \quad (2)$$

The explained **y** variance for models developed on the full and cross-validation data sets (EVC and EVCV, respectively) can be obtained from the corresponding root-mean-squared error values (*RMSEC* and *RMSECV*, respectively).[7] When this is taken into account and eq 2 is rearranged as *Pp* = 100(2EVCV − EVC), it is possible to rewrite eq 2 as

$$Pp = 100[2(1 - RMSECV^2/RMSECV_0^2) - (1 - RMSEC^2/RMSEC_0^2)] \quad (2a)$$

where $RMSEC_0^2$ and $RMSECV_0^2$ correspond to total initial variance (i.e., the situation for $PLS_{(0)}$) and can be estimated by multiplying the **y** variance, $s(\mathbf{y})^2$, by $(No - 1)/No$ and $No/(No - 1)$, respectively.[7] By rearranging, it is again possible to rewrite eq 2 as

$$Pp = 100(1 + \{[No\ RMSEC^2/(No\text{-}1) - 2\ (No\text{-}1)\ RMSECV^2/No]/s(\mathbf{y})^2\}) \quad (2b)$$

Ideally, the maximum value for EVCV could be 1 and the difference between EVC and EVCV could be 0. This corresponds to the maximum possible predictive power (*Pp* = 100%). For QSAR models that are not likely to be statistically acceptable, *Pp* may become negative. In these instances, *Pp* is set to 0 (there is no interest in discriminating between wholly unacceptable models), and thus, the range of *Pp* is 0−100%.

Global Modeling Power of a Model. The global "modelling power" can be examined in two ways:

(i) The "modelling power" plot (*Dp* vs *Pp* plot), in which a model is represented by a single point on the plot, allows for the visualization of the joint descriptive and predictive power of the model. If required, its position can also be compared with prefixed limits. For instance, the algorithm automatically sets the limit to 60% for both *Dp* and *Pp* to discriminate between the QSAR usefulness of the model (>60%) and that of less significance (<60%). Obviously, these limits can be varied by the user; however, ideally, they would be prefixed in model evaluation and validation and may be influenced by the quality of the data being modeled and the intended use and context of the predictions.

(ii) The *Mp*, a numerical index for the global descriptive and predictive power, is calculated by combining *Dp* and *Pp*:

$$Mp = (f_{Dp}Dp + f_{Pp}Pp)/(f_{Dp} + f_{Pp}) \quad (3)$$

where f_{Dp} and f_{Pp} are weighting factors that represent the relative importance that the user wants to confer to the descriptive and predictive aspects, respectively. By default, the algorithm set a value of 1 for both factors, conferring equal (50%) importance to both aspects.

RESULTS AND DISCUSSION

New chemicals legislation within the European Union may require increased use of in silico techniques to predict toxicity. Should such techniques be applied, then models will

require some form of evaluation as to their quality, leading to possible validation in a regulatory context. In addition, the confidence that can be placed in a prediction from such models will need to be assessed individually and within the context of its intended use. A number of statistical criteria have already been recommended for the description of predictive models for toxicity.[8] This paper introduces the new diagnostic tools for the description and evaluation of "transparent" regression-based QSARs, namely, descriptive power and modeling power. *Dp* is related to aspects of the assessment of the uncertainty of individual parameters and to variable selection. These two issues are described in detail by Erikkson et al.[8] The connections between cross-validation and jack-knifing have been associated with PLS (when the assumptions of regression analysis are not fulfilled[8] and classical MLR confidence intervals are not reliable). In the approach presented in this study, estimations of uncertainty and the calculation of *Dp* can be performed in the same manner for both the MLR and PLS models, thus allowing comparison between both algorithms. The descriptive power of a model is a novel and important description of the quality of a model (which is equally, if not more, important than the predictive ability) and one that has been underestimated by QSAR modelers in the past. For instance, although the assessment of the uncertainty of coefficients was considered by Erikkson et al.,[8] it is not considered formally as part of the acceptability criteria for QSAR models.

The predictive ability of a QSAR is normally of the greatest concern. Some reference values have been proposed,[8] including limits for the cross-validated coefficient of determination, $q^2 > 0.5$, and for the difference between fitted and cross-validated values, $r^2 - q^2 < 0.3$. The predictive power statistic accounts for both of these aspects, as it is based on estimations of explained variance.[7] The combination of *Dp* and *Pp* will therefore assist in the definition of the global quality of the model. A further attractive feature of the proposed statistics and the *Dp* versus *Pp* plot is that both indices are in the range of 0−100%, which facilitates their interpretation. Therefore, different studies from a given model, or even different models, can be compared graphically in an intuitive manner.

As biological systems are complex, it is expected that multiple mechanisms are involved in the toxicity of compounds. Thus, linear methods may not always be the most suitable methods to develop QSAR models. The statistics presented in this study are applicable a priori to some nonlinear approaches (i.e., quadratic variables can be used in MLR or PLS algorithms). However, extending the present approach to other nonlinear algorithms deserves more attention. Since the *Pp* statistic is based on predicted response values, it is applicable to any regression algorithm (linear or not). On the other hand, for the calculation of *Dp*, the algorithm must be able to calculate the regression coefficients for all the variables. If the algorithm is not able to do that, then *Dp* cannot be computed. This is a problem of lack of transparency (i.e., lack of information about the importance of variables through their coefficients) of such an algorithm. Therefore, the approach presented is restricted to "transparent" QSAR models for which it is possible to calculate the regression coefficient for the original predictor variables.

Characterizing the Models: Classical Statistical Values. Table 1 shows a limited set of statistical values for some MLR models with different variables. Some of the statistics can be considered "classical" and are frequently used in the QSAR literature [i.e., $s(b)$, *SE*, *F*, r^2, and more recently, q^2]; others have been incorporated from the relatively recent multivariate regression model literature [i.e., $U(b)$ or *RMSECV*].[7] There are many other statistics,[2,7−10] conventional or not, that are not included in Table 1, which illustrates the variety of methods for assessing the quality of a model. However, this does not mean that, in general, all the statistics are satisfactory for comparing two, or more, models. For instance, there is a decrease in "model complexity" (less descriptors are utilized) going from model Mn = 1 to Mn = 5 in Table 1. The accompanying statistics show an unwanted increase in *SE*, but a desired increase in *F*; in addition, with the exception of Mn = 2, there is a decrease in r^2 and an increase in *RMSEC* (a poor trend) but an increase in q^2 and decrease in *RMSECV* (both good trends). While an expert in the field should reach an adequate conclusion, the different results, numeric scales, and trends complicate the assessment of quality for any modeler and model user. In the original paper on modeling tadpole narcosis,[12] the authors concluded that the combined use of some of Abraham's and some of the topological descriptors (models Mn = 2 and 3 in Table 1) are adequate to model tadpole narcosis. However, other authors easily could conclude that using Abraham's descriptors alone (Mn = 4) gives results that are more adequate (based, for instance, on the higher q^2 or lower *RMSECV* values of this model). In addition, a nonquantified aspect of the modeling is the mechanistic interpretation, which is probably easier from the physicochemically based solvatochromic parameters than from those based on topological indices.

In evaluating the models, the regression coefficients and their associated uncertainties have been overlooked (unfortunately, this is a common occurrence in the QSAR literature). In this regard, it is clear from Table 1 that the use of all variables (Mn = 1 in Table 1) is not appropriate, as there are three nonsignificant variables, *V*, *Sz*, and *J*. For these variables, the confidence interval around b, $b \pm t\, s(b)$, includes zero (in other words, $100 \cdot t\, s(b)/b > 100\%$), which is unacceptable. Note that the relative uncertainty of the b coefficients estimated from the classic 95% confidence interval could be used as an acceptability criterion (i.e., $100 \cdot t\, s(b)/b < 100$), but not $s(b)$ since it does not have a relative magnitude. On the other hand, there is evidence in Table 1 that the alternative relative uncertainty of the b coefficients estimated from jack-knifing generates $100 \cdot U(b)/b$ values larger than the corresponding values from $100 \cdot t\, s(b)/b$ for all the variables. This reflects the fact that $U(b)/b$ values are estimated from cross-validation, so they should be more realistic of the reliability of the model and, more specifically, the quality of the coefficients. As before, the largest $100 \cdot U(b)/b$ values correspond to the variables *V*, *Sz*, and *J*, although other (unreliable) variables also have $100 \cdot U(b)/b$ values greater than 100%. This suggests that the model should be simplified, which is consistent with the ranking of the full **X** data statistics (3−5 as in Todeschini et al.[13]). It should be noted that only for the four-variable model (Mn = 5 in Table 1) was a consistent model, in terms, of descriptive power obtained, as all the $100 \cdot U(b)/b$ values were under 100%.

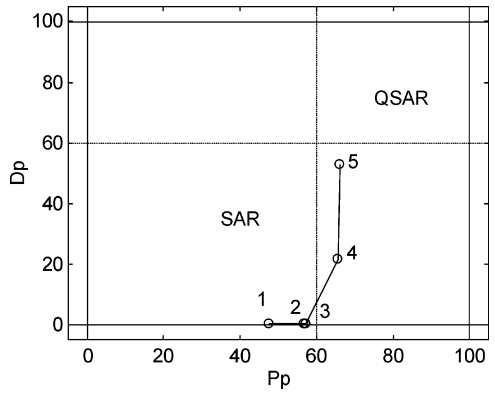

Figure 1. Graphical comparison of models in Table 1 by the "modelling power" plot, based on the "descriptive power" (*Dp*) and the "predictive power" (*Pp*). Legend: Labels indicate the model number (Mn in Table 1).

The "Modelling Power" Plot. A simple way to summarize the most important information held in Table 1 is the use of a plot of *Dp* versus *Pp*—the so-called "modelling power" plot. Figure 1 shows the modeling power plot for the models in Table 1 (labeled by Mn). As can be observed, except for models Mn = 2 and 3 (which have similar *Dp* and *Pp* values), there is a notable change in the location of the points for the models. All models (*Pp* values from ~45% to ~65%) have some predictive ability, if the criterion to make this conclusion is that *Pp* values must be approximately equal to, or larger than, 60%. In contrast, only models Mn = 4 and 5 have some descriptive ability (*Dp* values ~20% and ~50%, respectively) using the same criterion based on the 60% limit. This indicates that only model Mn = 5 is of interest in terms of combining descriptive and predictive power [*Mp* = (66.2% + 52.6%)/2 = 59.4%]. In fact, if the *Dp* and *Pp* limits are fixed at the 60% level, this is the only model that is close to being considered useful for QSAR purposes.

Variable Selection Schemes. The intrinsic simplicity of the "modelling power" plot allows the modeler to perform a large number of analyses during the development of a QSAR. A good example is in the evaluation of models developed using different variable selection techniques. There is currently much debate in the literature about the role of variable selection and possible best methods.[14,15] For instance, Figure 2 shows the application of the *Dp* versus *Pp* plot to compare some variable selection approaches. The results from the MLR-forward-stepwise regression method (performed using Statgraphics) are shown in Figure 2a. The results suggest that only the two-variable model (Abraham's *V* and *B* descriptors) offers a potentially valid QSAR. This model is better, in terms of *Dp* and *Pp*, than those in Figure 1. This process determines that the variable *V* is the most important to model tadpole narcosis (meanwhile, it was not significant in model Mn = 1 in Table 1; see later discussion). Figure 2b shows the plot for the results from the MLR-backward-stepwise regression method (performed using Statgraphics). It can be observed that the final seven-variable model (labeled 3 in Figure 2b) is similar, in terms of modeling power, to that using only two variables (labeled 2 in Figure 2a).

In the previous cases, the conclusions regarding model quality would be affected by the possible impact of using MLR for a data matrix, showing a relatively high intercorrelation and also limited dimensionality (*K* and rank in Table 1). Figure 2c shows the equivalent view of Figure 2a, but using PLS modeling. As would be expected, for models 1−3, the PLS results coincide with those using MLR, since in these cases the maximum number of LVs is used in the PLS model. Clearly, the results of models 4 and 5 using $PLS_{(3)}$ in Figure 2c are better than those in Figure 2a. In fact, model 5 provisionally exhibited the best "modelling power" [*Mp* = (61.6% + 70.0%)/2 = 65.8%], falling into the predetermined area of acceptability for the QSAR. Figure 2d shows the evolution of the $PLS_{(3)}$-variable selection approach (a backward-stepwise scheme that can be performed optionally using The Unscrambler). As in the case of Figure 2b, in each step (after a variable is eliminated from the model), there is an improvement in *Mp*. Note that the final model in Figure 2d coincides with the final model in Figure 2c, as a result of the coincidence in the variables selected by both schemes.

Selecting the Optimal Number of Dimensions for PLS. There is no general consensus in the QSAR community as to whether PLS or MLR is preferable in approaching a particular problem, although some harmonized guidance can be found in Eriksson et al.[8] In general, it is recognized that PLS is more robust than MLR, particularly in the case of a highly multivariate and intercorrelated variable matrices.[2,8] However, some authors have suggested that PLS is less "transparent" than MLR.[2] While this may be so, this point deserves more attention since a format similar to MLR can be used to express the results of $PLS_{(LVs)}$ simply by descaling the standardized coefficients.[7] Moreover, the jack-knifing uncertainty calculations can be expanded to the descaled coefficients, which produces an evident improvement in transparency.

There is a great need to select the appropriate number of LVs in PLS to avoid under- or overfitting the data matrix.[7] Unfortunately, there are no absolute automatic criteria to perform this selection, and a number of different approaches can be used. Figure 3a shows some recommended criteria (based on *RMSECV* or EVCV[7]), which are compared with the *Dp* and *Pp* statistics proposed in this paper. A generally accepted rule is to select the minimum reasonable number of LVs; unfortunately, this is an ambiguous statement and open to subjectivity from the modeler. For instance, it could be interpreted as corresponding to a minimum in *RMSECV* or, equivalently, a maximum in EVCV or, at least, a stabilization of these parameters. Figure 3a indicates that, with some difficulties, the lowest *RMSECV* and highest EVCV is at three LVs; however, some authors may prefer the selection of two LVs. The *Pp* values, based on predictive knowledge, offer information similar to the previous criteria. However, the *Dp* values provide valuable assistance to making a definitive decision, due to the fact that more stable coefficients are obtained for the three-LVs model. This means that the "modelling power" plot should also be used to assist in making such decisions. Figure 3b clearly suggests that the best model is that for $PLS_{(3)}$ and also provides evidence that $PLS_{(5)}$ (equivalent to MLR) produces uncertain coefficients.

Potential Outliers, Number of Compounds, and Validation Schemes. There are a number of methods to identify potential outliers from QSARs for toxicity.[8] The most commonly applied method to identify outliers, however, is to examine the **y** predicted (fitted) versus **y** actual plot, which is a standard plot in QSAR analysis.[8] The information in

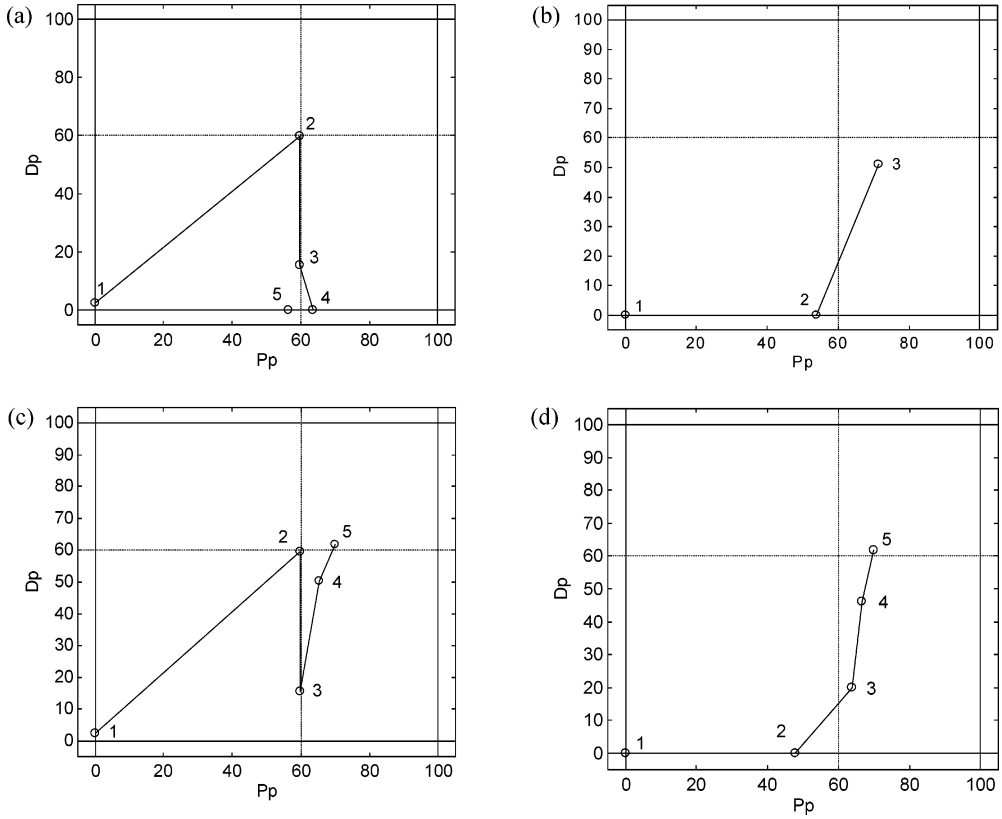

Figure 2. "Modelling power" plot for different variable selection approaches. (a) MLR-forward-stepwise: the variables *V*, *B*, $^1\chi$, *S*, and *E* were included in each model successively (in each step from 1 to 5). (b) MLR-backward-stepwise: starting from 10 variables, excluding the variables *Sz*, *J*, and *V* from the model successively (in each step from 1 to 3). (c) The same variable selection steps as in a but using PLS regression to perform the model ($PLS_{(1)}$ in step 1, $PLS_{(2)}$ in step 2, and $PLS_{(3)}$ in the steps 3, 4, and 5). (d) Variable selection based on a predetermined $PLS_{(3)}$ approach, which started from 10 variables and eliminated successively the variable showing the lowest absolute *b* coefficient with the $b \pm U(b)$ interval including zero, until the remaining $b \pm U(b)$ intervals does not include zero: the variables *A*, *J*, *Sz*, log *RB*, and *W* were eliminated (in each step from 1 to 5). Legend: Labels correspond to the steps during the variable selection process.

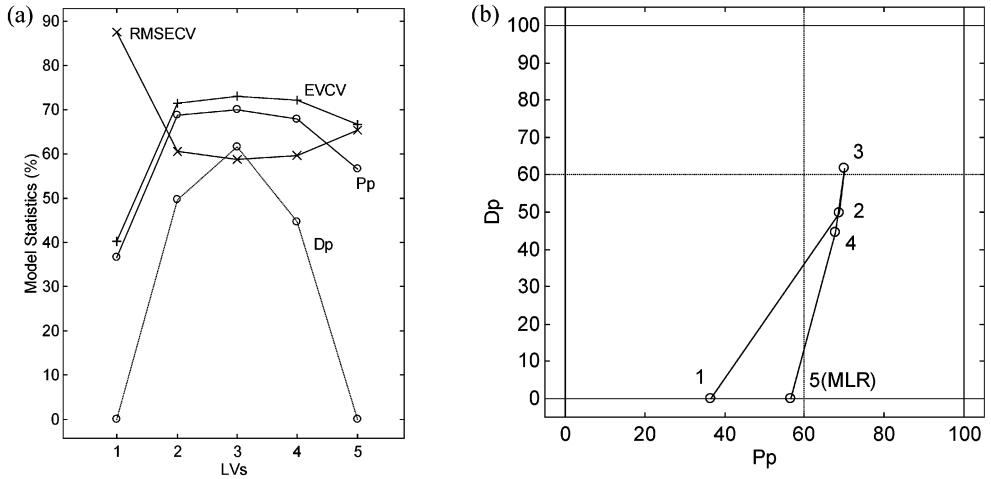

Figure 3. Selection of the optimal number of dimensions in PLS regression for an **X** matrix using the variables *V*, *B*, $^1\chi$, *S*, and *E*. (a) Comparison of some model statistics *RMSECV* (x, solid line), EVCV (+, solid line), *Pp* (o, solid line), and *Dp* (o, dashed line). (b) "Modelling power" plot showing the impact of increasing the number of LVs (used as label) in the PLS model.

this plot can be improved by superimposing the cross-validated predicted **y** values.[7] Alternatively, the corresponding fitted and cross-validated residuals, **e**, can be analyzed. Figure 4a shows these results for the $PLS_{(3)}$ model using all of the available $\mathbf{y}-\mathbf{X}$ data. As can be observed, compounds labeled 27, 98, and 102 exhibit atypical fitted and cross-validated results, which suggest that the "compound's disagreement" is located into the **y** data (i.e., experimental tadpole narcosis is not consistent with the descriptors values, so it is poorly predicted by the model). In contrast, compound 82 exhibits only atypical cross-validated results, so the "compound disagreement" is found in the **X** data (the compound is poorly predicted only by a model in which the compound is excluded). For instance, examination of the autoscaled **X** data reveals that compound 82 has very high relative values for four of the topological descriptors and,

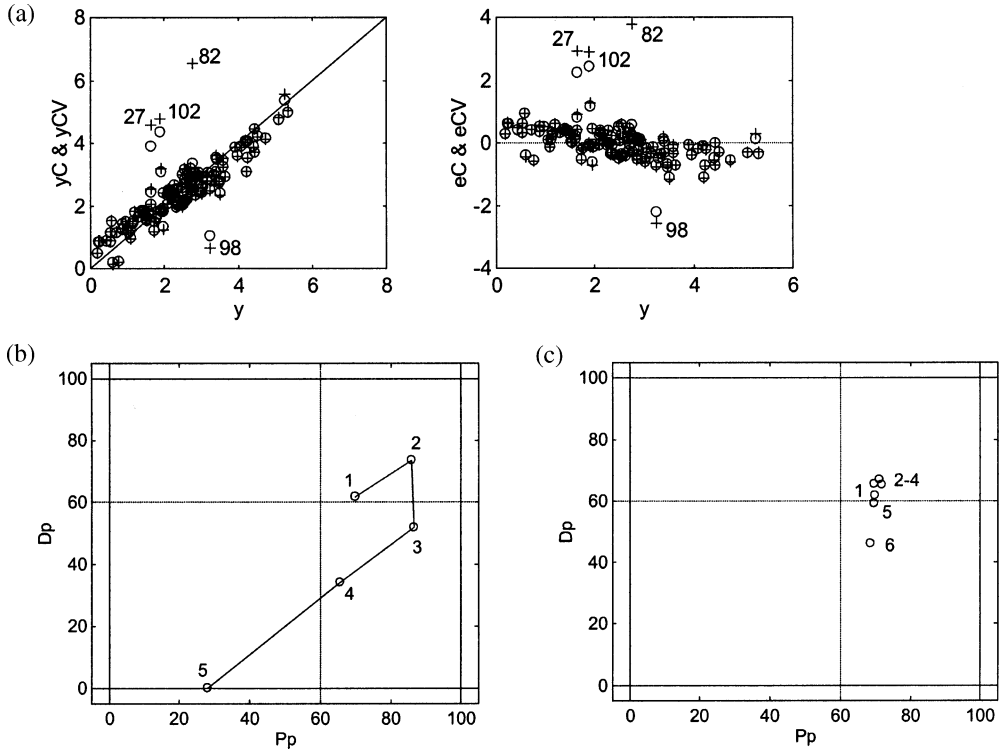

Figure 4. (a) Predicted **y** values, fitted (**yC**; o) and cross-validated (**yCV**; +), vs actual **y** values and residuals, **e** values, fitted (**eC**; o) and cross-validated (**eCV**; +) vs actual **y** values from $PLS_{(3)}$ using all **X** data. Four compounds, labeled 27, 82, 98, and 102 in the original data table,[12] appear as potential outliers. (b) "Modelling power" plot to check the effect of varying the number of compounds on the $PLS_{(3)}$ model with variables *E*, *S*, *B*, *V*, and $^1\chi$ and the initial 123 compounds (label 1): (label 2) eliminating the four potential outliers, thus reducing the set to 119, (label 3) eliminating the four potential outliers and randomly selecting 60 compounds, (label 4) eliminating the four potential outliers and selecting the 60 compounds with the lowest **y** value, and (label 5) eliminating the four potential outliers and selecting the 30 compounds with the lowest **y** value. (c) "Modelling power" plot to check the effect of using different cross-validation schemes on the $PLS_{(3)}$ model with variables *E*, *S*, *B*, *V*, and $^1\chi$ and the initial 123 compounds: (label 1) LOO cross-validation, (label 2) Venetian blind cross-validations selecting 11 subsets (using 112 calibration samples and 11 test samples in each subset), (label 3) Venetian blind cross-validations selecting five subsets (using 98 calibration samples and 25 test samples in each subset), (label 4) Venetian blind cross-validations selecting four subsets (using 92 calibration samples and 31 test samples in each subset), (label 5) Venetian blind cross-validations selecting three subsets (using 82 calibration samples and 41 test samples in each subset), and (label 6) Venetian blind cross-validations selecting two subsets (using 61 calibration samples and 61 test samples in each subset). The **y**−**X** data were reordered randomly before cross-validation.

to a lesser extent, some Abraham's descriptors. It must be remembered that outliers should be removed only if there are good reasons to do so.[2] If mathematical criteria alone are used to eliminate such data, it is considered (and must be declared) as one of the limitations of the QSAR. An in-depth study in this area is out of the scope of this paper. It should also be noted that, in the original study,[12] outliers were not investigated.

In addition to the problems of outliers, another interesting aspect of the QSAR is the number of compounds and the "diversity" (data range) of the **y**−**X** data.[2] Figure 4b illustrates the impact of eliminating outliers, reducing the number of compounds in the calibration set, or the **y** range in the actual example. As expected, eliminating outliers produces an increase in both *Dp* and *Pp* values (from the point labeled 1 to that labeled 2), thus increasing the global *Mp* value. The reduction of the number of compounds following initial attempts at modeling, while maintaining the **y** range (from 119 to 60; points labeled 2 and 3, respectively) results in a deterioration only in the descriptive power. In contrast, a reduction of both the number of compounds and the **y** range (from points labeled 3−5) results in the deterioration of both the descriptive and predictive powers. These effects are expected, but if the "modelling power" plot is used, they can be easily quantified.

There are two main statistical schemes in QSAR to assess predictivity.[2] The first is "external validation", which consists of making predictions for an independent data set not used in the model calibration, and the second is the more popular technique of cross-validation.[2,7] For instance, in this study, a data set (for instance, 60 compounds) carefully chosen to preserve "representativity"[2] could be selected to calibrate the model, and the remaining compounds could be used as the "validation set". However, some drawbacks are observed in this process, such as the loss of "descriptive power" (see Figure 4b) or possible changes in fitted and cross-validated statistics depending on whether one, or more, potential outliers are located in the calibration or the validation set. In contrast, cross-validation means that the final model coefficients are obtained with the entire data set and the effect of potential outliers is compensated for since they appear in both the calibration and cross-validation subsets. On the other hand, there was a strong recommendation to perform the jack-knifing assessment of coefficient stability by means uncertainty estimations.[7]

Figure 4c shows the impact on the "modelling power" plot of different cross-validation schemes, from the popular LOO approach (labeled 1) to the leave-many-out, or Venetian blinds, approach of selecting different subsets of samples. The results may vary slightly when running the process

several times. However, in this instance, it was observed that the LOO results were equivalent to those of the Venetian blinds approach, except for the cases when a low number of subsets had been selected, such as using two subsets (labeled 6 in Figure 4c).

Reporting the QSAR. An example of the classical way of reporting a QSAR is illustrated in the following equation, obtained using an MLR analysis on the tadpole narcosis data, labeled as 2 in Figure 2a. It corresponds to the use of only two of Abraham's descriptors, *V* and *B*, whose results are identical to those obtained using a $PLS_{(2)}$ model:

$$\log(1/C_{nar}) = 0.8133\ (\pm 0.1842) - 2.1440\ (\pm 0.2565)B + 2.7629\ (\pm 0.1875)\ V$$
$$n = 123,\ r^2 = 0.65,\ q^2 = 0.62,\ F = 110.3\ (p < 0.0001),\ SE = 0.67,\ RMSEVC = 0.70 \quad (4)$$

Where *s*(*b*) values are given in parentheses.

Occasionally, some compounds are eliminated to improve statistical fit, or for other mechanistic reasons. For instance, eq 4 can be improved by eliminating four poorly predicted compounds (27, 98, 102, and 121) to give

$$\log(1/C_{nar}) = 0.4872\ (\pm 0.2422) - 2.7233\ (\pm 0.3395)B + 3.4282\ (\pm 0.2584)V$$
$$n = 119,\ r^2 = 0.86,\ q^2 = 0.85,\ F = 350.5\ (p < 0.0001),\ SE = 0.43,\ RMSEVC = 0.44 \quad (5)$$

Where *t s*(*b*) values are given in parentheses to give an estimation of the coefficients' uncertainty.

In many QSARs, some of these statistics do not appear and others are introduced in their place. Guidance is being prepared by the OECD for the validation of QSARs.[3] One of the key issues here will be the correct presentation and description of a model, with readily available statistical criteria (as well as access to the data set etc.). One approach to assist in this direction could be the inclusion of the statistical criteria developed in this investigation, that is,

$$\log(1/C_{nar}) = 0.4872\ (\pm 0.2304) - 2.7233\ (\pm 0.3602)B + 3.4282\ (\pm 0.2528)V$$
$$No = 119,\ N\nu = 2,\ Dp = 77.3\%,\ Pp = 84.9\%,\ Mp = 81.1\% \quad (6)$$

Four compounds were excluded (27, triacetin; 98, hexan-1-ol; 102, decan-1-ol; and 121, acetal) from the LOO cross-validated results [*U*(*b*) values are given in parentheses).

CONCLUSIONS

This study has investigated the use of new diagnostic tools for "transparent" QSARs based around "descriptive power" and "predictive power". These statistical descriptors are simple and easy to calculate. A series of MLR and PLS models were developed for a data set of tadpole narcosis data. These toxicity values represented a number of mechanisms of action and provided a fundamental and multivariatedata matrix to model. The diagnostic tools utilized in this study, including the modeling plot, show particular promise to discriminate between subtly different predictive abilities and the robustness of models. In particular, the "modelling power plot" was shown to be a very useful tool to delineate between models based on different numbers of variables and different variable selection techniques, as well as the optimal number of dimensions in multivariate approaches such as PLS.

The regulatory community requires diagnostic tools to assess the performance and quality of QSARs for toxicity in response to chemicals legislation both within Europe and worldwide. While tools such as the modeling power plot provide useful information, it should be remembered that model evaluation should be context-dependent, that is, dependent on the potential use of the prediction. In addition, the process of using diagnostic tools should not be to identify "winning" QSARs but to evaluate succinctly the strengths and weaknesses of individual models.

ACKNOWLEDGMENT

S.S. acknowledges the Spanish Ministry of Science and Technology (MCYT) and the European Regional Development Fund (ERDF) (Project SAF2002-01330) and the Generalitat Valenciana (research group GR04-02) for financial support.

REFERENCES AND NOTES

(1) Hansch, C. *Comprehensive Medicinal Chemistry*; Pergamon Press: Oxford, U. K., 1990; Vol. 4.

(2) Schultz, T. W.; Cronin, M. T. D. Essential and desirable characteristics of ecotoxicity quantitative structure−activity relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 599−607.

(3) Organisation for Economic Co-Operation and Development (OECD). *Report from the Expert Group on (Quantitative) Structure−Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)-SARs*; OECD Environment Health and Safety Publications Series on Testing and Assessment No. 49; OECD: Paris, 2004; p 206. Available from http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jm-mono(2004)-24 (accessed September 30, 2005).

(4) Jaworska, J. S.; Comber, M.; Auer, C.; van Leeuwen, C. J. Summary of a Workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ. Health Perspect.* **2003**, *111*, 1358−1360.

(5) Worth, A. P.; van Leeuwen, C. J.; Hartung, T. The prospects for using (Q)SARs in a changing political environment − high expectations and a key role for the Commission's Joint Research Centre. *SAR QSAR Environ. Res.* **2004**, *15*, 331−343.

(6) Worth, A. P.; Hartung, T.; van Leeuwen, C. J. The role of the European Centre for the Validation of Alternative Methods (ECVAM) in the validation of (Q)SARs. *SAR QSAR Environ. Res.* **2004**, *15*, 345−358

(7) Martens, H.; Martens, M. *Multivariate analysis of quality. An introduction*; John Wiley & Sons Ltd: Chichester, U. K., 2001.

(8) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluation of classification- and regression-based QSARs. *Environ. Health Prespect.* **2003**, *111*, 1361−1375.

(9) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Evaluation of QSARs for ecotoxicity: a method for assigning quality and confidence. *SAR QSAR Environ. Res.* **2004**, *15*, 385−397.

(10) Aptula, A. O.; Jeliazkova, N. G.; Schultz, T. W.; Cronin, M. T. D. The better predictive model: high q^2 for the training set or low root mean square error of prediction for the test set*? QSAR Comb. Sci.* **2005**, *24*, 385−396.

(11) Abraham, M. H.; Rafols, C. Factors that influence tadpole narcosis − an LFER analysis. *J. Chem. Soc., Perkin Trans.* **1995**, *2*, 1843−1851.
(12) Agrawal, V. K.; Chaturvedi, S.; Abraham, M. H.; Khadikar, P. V. QSAR study on tadpole narcosis. *Bioorg. Med. Chem.* **2003**, *11*, 4523−4533.
(13) Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 13−29.
(14) Ghafourian, T.; Cronin, M. T. D. The impact of variable selection on the modelling of oestrogenicity. *SAR QSAR Environ. Res.* **2005**, *16*, 171−190.
(15) Livingstone, D. J.; Salt, D. W. Variable selection − Spoilt for choice? *Rev. Comput. Chem.* **2005**, *21*, 287−348.

CI050445C