The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models

Alexander Tropsha[a]*, **Paola Gramatica**[b]*, **Vijay K. Gombar**[c]*

[a] Laboratory for Molecular Modeling, School of Pharmacy, CB# 7360 Beard Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

[b] QSAR and Environmental Chemistry Research Unit, Department of Structural and Functional Biology, University of Insubria, Via J. H. Dunant 3 – 21100 Varese, Italy

[c] GlaxoSmithKline, Metabolic and Viral Diseases' Center of Excellence for Drug Discovery (MV CEDD), Department of Drug Metabolism and Pharmacokinetics (DMPK), 3030 Cornwallis Road, Research Triangle Park, NC 27709, U.S.A.

Abstract

This paper emphasizes the importance of rigorous validation as a crucial, integral component of Quantitative Structure Property Relationship (QSPR) model development. We consider some examples of published QSPR models, which in spite of their high fitted accuracy for the training sets and apparent mechanistic appeal, fail rigorous validation tests, and, thus, may lack practical utility as reliable screening tools. We present a set of simple guidelines for developing validated and predictive QSPR models. To this end, we discuss several validation strategies including (1) randomization of the modelled property, also called Y-scrambling, (2) multiple leave-many-out cross-validations, and (3) external validation using rational division of a dataset into training and test sets. We also highlight the need to establish the domain of model applicability in the chemical space to flag molecules for which predictions may be unreliable, and discuss some algorithms that can be used for this purpose. We advocate the broad use of these guidelines in the development of predictive QSPR models.

1 Introduction

A Quantitative Structure Property Relationship (QSPR) model describes a mathematical relationship between structural attribute(s) and a property of a set of chemicals. The use of such mathematical relationships to predict the target property of interest for a variety of chemicals prior to or in lieu of expensive and labor-intensive experimental measurements has naturally been very enticing. The potential promise of using QSPR models for screening of chemical databases or virtual libraries before their synthesis, appears equally attractive to chemical manufacturers, pharmaceutical companies and government agencies, particularly in times of shrinking resources. Given the growing sizes of chemical databases resulting from combinatorial synthesis on one side and the regulatory and social pressures for timely assessment of health and environmental risks of chemicals on the other, the need for reliable QSPR models is imperative. For instance, environmental agencies in both Europe and the US require reliable data on environmental effects and fate of all industrial chemicals. Traditionally, biological and environmental testing has provided such data; however, the data are available for only a fraction of industrial chemicals and there exist thousands of industrial chemicals that will continue to go untested. Recently, Walker et al. [1–5] have addressed this problem by providing a set of guidelines for developing and using QSPR models for environmental risk assessment.

In order to be reliable and predictive, QSPR models should: (1) be statistically significant and robust, (2) be validated by making accurate predictions for external data sets that were not used in the model development, and (3) have their application boundaries defined. Therefore, in this paper, we argue that, in spite of their high fitted accuracy and apparent mechanistic appeal, some published QSPR

* to receive correspondence (All authors equally contributed to this paper and can be contacted at the following e-mail addresses: alex tropsha@unc.edu; paola.gramatica@uninsubria.it; vijay.k.gombar@gsk.com.)

Key words: Structure-property relationship modeling, model validation, applicability domain, QSAR

Abbreviations: QSPR Quantitative Structure-Property Relationship; MLR, Multiple Linear Regression; R^2, coefficient of determination; q^2, cross-validated explained variance; LOO, Leave -one-out; LMO, Leave -many-out; EPD, Effective Prediction Domain; SOM, Self-organizing maps; KNN, k-nearest neighbors; RSD, Residual standard deviation

1611-020X/03/0104-0069 $ 17.50+.50/0

models fail rigorous validation tests, and, thus, may lack practical utility as reliable screening tools. We suggest that only validated QSPR models can offer a meaningful mechanistic interpretation, especially in the context of design or discovery of novel chemical agents with desired properties. We describe several possible approaches to QSPR model validation, including Y-randomization, robust internal validation strategies such as multiple leave-many-out cross-validations, and external validation. We also discuss algorithms that can be employed in defining the boundaries of model applicability. Finally, we propose a set of simple guidelines that should be followed by QSPR modelers in developing validated and predictive QSPR models.

2 QSPR Development Steps

2.1 Common Practices

The two most commonly practiced stages in the development of a QSPR model are:

1. **Data preparation**, which includes (i) collection and cleaning of target property data; (ii) calculation of molecular descriptors for chemicals with acceptable target properties; and (iii) merging of the property and descriptor values in a manageable SPR database, and
2. **Model generation**, which implies establishing statistically significant relationships between target property and descriptor values.

Data Preparation: The first step of the initial stage, as trivial as it sounds, is extremely important in making sure to select only accurate, precise, and consistent experimental data. The second step encompasses unique numerical representation of molecular structure in terms of molecular descriptors that capture salient compositional, electronic and steric attributes. It is always preferable to use as few explanatory descriptors as possible to facilitate interpretation of the resulting models. It has been emphasized by many authors that one of the most important aspects of QSPR modeling is the ability to interpret the resulting models in physico-chemical or mechanistic terms. Thus, the descriptors that are often found in many QSPR studies mirror fundamental physico-chemical factors that in some way relate to the endpoint(s) under study. Examples of such descriptors, which provide simple insight into the possible mechanism underlying the response, include measures of hydrophobicity (logP, logD), pKa, molecular refraction, accessible solvent area, molecular weight, as well as quantum-chemical descriptors (HOMO, LUMO, etc). Sometimes, a QSPR model is developed only as a tool to predict target properties of untested chemicals. Thus, some researchers employ large sets of molecular descriptors that do not always allow a simple chemical interpretation (e.g, molecular connectivity indices) followed by variable selection using various stochastic sampling algorithms to select most significant descriptors in the process of model development. Finally, a manageable SPR database is created merely for the ease and maneuverability in subsequent analysis and reporting.

Model Generation: The second stage pertains to creating a statistically significant structure-property relationship; the molecular descriptors serve as independent variables and the modeled property as the dependent variable. Many different methods embedded in a variety of computer software packages are available for this purpose; the choice is generally guided by the answers the resulting QSPR screen is expected to provide. Some widely used model development methods include multiple linear regression (MLR), partial least squares (PLS), artificial neural networks (ANNs), and k-Nearest Neighbor (kNN) methods.

The above procedures represent, to a large extent, a standard practice of any QSPR modeling effort, and the researchers' interests and software availability generally determine the specific details in completing these steps. In fact, many QSPR practitioners find these two steps sufficient to arrive at the acceptable model. Most published QSPR models for biological, chemical and environmental effects, generally, do not include validation as an integral component of model development. It has been emphasized by many authors [6–12] that one of the most important aspects of QSAR modeling is the ability to interpret the models in physico-chemical and/or mechanistic sense. Thus, they limit selection of descriptors to only those, generally whole-molecule descriptors that seem to carry some fundamental physico-chemical information that might be related to the modeled property. For examples, Schultz et al. [10] recently reported several so-called mechanism-based QSPR models. For a series of methacrylates the following QSPR for toxicity data to *Tetrahymena pyriformis* was presented:

$$\log(1/\mathrm{IGC}_{50}) = 0.54 \log \mathrm{K}_{\mathrm{ow}} - 8.90\ \mathrm{E\ \ LUMO} - 0.99$$
$$n = 11,\ r^2 = 0.82,\ s = 0.28,\ r^2_{\mathrm{cv}} = 0.64$$

This two-variable model, developed for only 11 compounds, formally satisfies the generally accepted limit of compounds/descriptors ratio of five to one. However, it obviously lacks any external validation; furthermore, the fact that r^2_{cv} is substantially lower than the fitted r^2 indicates that the model is unstable. Thus, this model may be completely unsuitable for prediction purposes, let alone any meaningful mechanistic deductions.

Similarly, Akers et al. [11] published models for toxicity metrics of halogenated aliphatic compounds, with the training sets ranging from 4 to 39 chemicals. They conjecture extensive deduction of mechanistic details about the underlying processes on the basis of these QSPRs of modest fitted accuracy, R^2, and conclude: "the goodness-of-fit is satisfactory for predictive purposes". These mechanistic claims may not hold true if these models were subjected to rigorous validation tests.

Benigni et al. [12] stated that "the use of a limited set of individual parameters with clear mechanistic significance is still the best approach that ensures the optimal comprehension of the results and gives the possibility of performing non-formal validations much superior to those provided by statistics". However, the QSPR models they base these claims on have not been subjected to any validation; the guiding parameters are mere fitted R and R^2.

Several examples of published QSPR models presented above are representative of a large body of QSPR literature where model robustness is characterized in a very limited way, using only parameters of fitted accuracy of prediction such as R or R^2. It is well known that these statistical parameters can not be considered indicators of the predictive power of the model; they just measure how well the model is able to reproduce the response for the training set. For instance, having realized the importance of validation, Devillers [13] and Kaiser et al. [14] made a recommendation that the ECOSAR package [15], whose QSPRs are not validated and have marginal quality, should be either avoided or used with great caution and an awareness of their limits. As mentioned above, one must be well aware that a QSPR model with high fitted accuracy could still have a poor predictive power for new compounds. Thus, making any mechanistic interpretation of such a QSPR model, in the absence of validation, may be simply unreasonable.

2.2 Not-so-common Practices

Since the real utility of a QSPR model is in its ability to accurately predict the modeled property for new chemicals, a realistic assessment of the model's true predictive power must be ascertained. This constitutes the following two additional steps of model development, which are the main focus of this paper.

3. **Model validation**, which implies quantitative assessment of model robustness and its predictive power, and
4. **Definition of the application domain** of the model in the space of chemical descriptors used in deriving the model.

Not many researchers are concerned with these issues when developing and publishing QSPR models. However, if those models that afford formal interpretation are not subjected to rigorous validation, not only may the models be worthless in making reliable predictions but they may also lead to wrong conclusions about the mechanistic information. We now discuss current approaches to model validation that are used by some QSPR practitioners.

Statistical Diagnostics: It is extremely important to ascertain that a QSPR model is of sufficiently high quality and is worth validating for practical applications. For instance, the amount of dependent variable variance explained by an MLR model is expressed by the coefficient of determination, R^2. It is commonly believed that the closer the value to unity, the better the model. However, it should be noted again that R^2 is just a measure of the quality of the fit between model-predicted and experimental values and it does not reflect at all on the predictive power of the model. It is possible that a QSPR model with high R^2 could be a poor predictor, especially if the high R^2 value is the result of low degrees of freedom, variable multicollinearity, statistically insignificant model descriptors, high-leverage points in the training set, etc. In our experience, a regression model with $\boldsymbol{k}$ descriptors and $\boldsymbol{n}$ training set compounds may be acceptable for validation only if the following criteria are satisfied (cf. [16, 17]):

$\boldsymbol{n} > 4\,\boldsymbol{k}$, and

For any of the $\boldsymbol{k}$ descriptors,
The significance level is $p < 0.05$,
The pair-wise correlation coefficient is < 0.9,
The tolerance is > 0.1, and
For any of the $\boldsymbol{n}$ compounds,
The diagonal element of the hat matrix is $< 3\,\boldsymbol{k/n}$,
The squared Cook's distance is < 1.0,
The standardized residual is < 2.5,
The DFFITS is $< \sqrt{(4\,\boldsymbol{k/n})}$,
The DFBETAS is $< \sqrt{(4/\boldsymbol{n})}$, and
The Covariance ratio is in the range $1 - 3\,\boldsymbol{k/n}$ and $1 + 3\,\boldsymbol{k/n}$.

For other model development methods, similar multiple diagnostic checks must be performed to unmask and appropriately treat influential, leverage, and outlier data points before accepting a model for further consideration.

Internal Validation. Many QSPR practitioners limit themselves to internal validation, commonly achieved by the leave-one-out (LOO), and sometimes by the leave-many-out (LMO), cross-validation test applied to the training set. For regression-like QSPRs, the outcome from such a test is a cross-validated correlation coefficient, q^2, which is calculated according to the formula:

$$q^2 = 1 - \frac{\sum_{i=1}^{training} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{training} (y_i - \bar{y})^2} \qquad (1),$$

where y_i, $\hat{y}$ and $\bar{y}$ are, respectively, the measured, predicted, and averaged (over the entire data set) values of the dependent variable; the summations run over all compounds in the training set. Although a small value of q^2 in the LOO or LMO test typically indicates low predictive ability of a model, the opposite is not necessarily true. Nevertheless, many researchers frequently use the value of q^2 as a criterion of both robustness and predictive ability of the model. Many authors consider q^2 greater than 0.5 as an indicator, and sometimes even as the ultimate proof, of the high predictive power of the model and either do not evaluate the model on an external test set or use only a one- or two-compound test set. Recent publications [6–12, 18–24] provide several examples of such cases. In reality, however, it has been shown that there exists no correlation between LOO cross-validated q^2 and the correlation coefficient R^2 between the predicted and observed activities for an external test set [25, 26]. These studies indicated that while high q^2 is the necessary condition for a model to have a high predictive power, it is not a sufficient condition. A recent study has systematically addressed the issue of q^2 being an inadequate characteristic of a model's predictive power [26]. One of the most demonstrative examples is given by a well-known group of ligands of corticosteroid binding globulin [27]. This dataset is frequently referred to as a benchmark [28] for the development and testing of novel QSPR methods. Novellino et al. [29] developed many 3D QSPR models based on the divisions of this data set into training and test sets, but did not observe any direct relationship between the training-set q^2 and the test-set R^2 values. The same conclusion was drawn when non-linear kNN variable selection QSPR method [30] with 2D descriptors was applied to the same dataset as well [26].

As follows from this discussion, current practices of QSPR modeling do not typically include validation as an integral component of model development. We now consider several approaches to validation that in our opinion should become a standard component of any QSPR model development.

3 Strategies for QSPR Model Validation

3.1 Leave-many-out (LMO) Validation

Internal validation can be performed by the leave-many-out (LMO) procedure. By design, model validation by LMO employs smaller training sets than the LOO procedure and can be repeated many more times due to possibility of larger combinations in leaving many compounds out from the training set. The premise being that if a QSPR model has a high average q^2 in LMO validation, we can reasonably conclude that the obtained model is robust.

In a typical LMO validation, $\boldsymbol{n}$ objects of the data set are divided in $\boldsymbol{G}$ cancellation groups of equal size, $\boldsymbol{m_j}$ ($=\boldsymbol{n/G}$). Based on the value of $\boldsymbol{n}$, $\boldsymbol{G}$ is generally selected between 2 and 10. A large number of models are developed with each of the $\boldsymbol{n\text{-}m_j}$ objects in the training set and $\boldsymbol{m_j}$ objects in the validation set. For each corresponding model, $\boldsymbol{m_j}$ objects are predicted and q^2 computed. Ideal expectation is high average q^2.

3.2 Bootstrapping

Bootstrap re-sampling is another approach to internal validation. A tutorial of different bootstrap methods was published recently [31] The basic premise of bootstrap re-sampling is that the data set should be representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the samples were selected randomly [32]. In a typical bootstrap validation, $\boldsymbol{K}$ groups of the size $\boldsymbol{n}$ are generated by a repeated random selection of $\boldsymbol{n}$ objects from the original data set. Some of these objects can be included in the same random sample several times, while other objects will never be selected. The model obtained on the $\boldsymbol{n}$ randomly selected objects is used to predict the target properties for the excluded sample. As in the LMO validation, a high average q^2 in the bootstrap validation is a demonstration of the model robustness.

3.3 Y-randomization Test

This is a widely used technique to ensure the robustness of a QSPR model [33]. In this test, the dependent-variable vector, Y-vector, is randomly shuffled and a new QSPR model is developed using the original independent-variable matrix. The process is repeated several times. It is expected that the resulting QSPR models should generally have low R^2 and low LOO q^2 values. It is likely that sometimes, though infrequently, high q^2 values may be obtained due to a chance correlation or structural redundancy of the training set [34]. If all QSPR models obtained in the Y-randomization test have relatively high R^2 and LOO q^2, it implies that an acceptable QSPR model cannot be obtained for the given data set by the current modeling method.

The following example illustrates the danger of developing a QSPR model not subjected to the Y-randomization test. Recently, a Comparative Molecular Filed (CoMFA) study of 16 antagonists of the dopamine D_2 receptor was published [35]. q^2 values exceeding 0.9 were obtained for the training set and a test set containing three compounds only produced an R^2 of 0.99. The technique used differed from standard CoMFA methodology in that the conformation of each compound was individually adjusted based upon the magnitude of prediction error. This process was repeated until the model could no longer be improved and then CoMFA columns that conflicted with the experimental results were eliminated from the model. This resulted in the generation of models containing only 3% to 7% of the total field information produced within CoMFA (68 to 147 columns used out of 2112 to 2376). With this large number of descriptors there will be a small subset just by mere chance whose variance correlates with the target property for this small training set. A significant bulk of the paper was then devoted to the discussion of the resulting CoMFA contour plots in terms of their guidance for the future design of D_2 antagonists. However, a simple Y-randomization test (Oloff and Tropsha, unpublished) has demonstrated that with a similar technique many models with acceptable values of LOO q^2 could be obtained for the same data set with a randomly created Y-vector. Clearly, there was no real structure-activity relationship and, consequently, any interpretation of this or any other QSAR model created in this fashion is spurious.

3.4 External Validation

Selecting Training and Test Sets: In typical situations, finding new experimentally tested compounds for this purpose is generally difficult. Recourse is, therefore, taken to splitting the available data set into training set, used for establishing the QSPR model, and a test set, for external validation. The underlying goal at this step is to ensure that both the training and test sets separately span the whole descriptor space occupied by the entire data set and the chemical domains in the two sets are not too dissimilar. An ideal splitting, therefore, leads to a test set such that each of its members is close to at least one point of the training set. Developing rational approaches for the selection of training and test sets is an active area of research. The approaches for creating training and test sets span from the straightforward random selection [36, 37] through activity sampling [38, 39] and various systematic clustering techniques [40–46] to the methods of self-organizing maps (SOM) [47, 48], Kennard Stone [49–51] and formal statistical experimental design (factorial and D-Optimal) [52–58].These methods help achieve desirable statistical characteristics of the training and test sets to varying degrees and have certain advantages and disadvantages. In a comparative study [36] of several methods, it was demonstrated that the best models were built when Kennard-Stone and D-optimal designs were used. SOM was better than splitting by random selection.

Several novel algorithms for the rational selection of training and test sets based on diversity sampling were considered recently [59]. Based on our experience, we suggest that the training and the test sets must satisfy the following criteria: (i) representative points of the test and training sets must be close to each other, and (ii) training set must be diverse. A quantitative description of these criteria, based on data set diversity indices, has been introduced earlier [60]. The division of a dataset into the training and test sets is achieved using one of the three closely related sphere-exclusion algorithms [59]. It is recommended [26] that the external test set must contain at least five compounds, representing the whole range of both descriptors and activity of compounds included into the training set. Using several experimental datasets, it was shown [59] that QSPR models built and validated with this approach had statistically better predictive power than models generated with either random or activity ranking-based selection of the test and training sets.

We also have had positive experiences with the D-optimal design algorithms [61, 62], that we also recommend for the general use. This algorithm of statistical Experimental Design selects samples that maximize the $|\mathbf{X}'\mathbf{X}|$ determinant, where $\mathbf{X}$ is the information (variance-covariance) matrix of independent variables (descriptors), or of independent plus dependent variables (descriptors and response). The points maximizing the $|\mathbf{X}'\mathbf{X}|$ determinant are spanned across the whole area occupied by representative points. These points constitute the training set and the points not selected are used as the test set [63]. This algorithm guarantees well-balanced structural diversity and representativity of the entire data space (descriptors plus response).

Assessing predictive power of QSPR models. In order to estimate the true predictive power of a QSPR model, one needs to compare the predicted and observed activities of a sufficiently large external test set of compounds that were not used in the model development [25, 26, 29, 61, 62, 64, 65]. It is the performance accuracy of the QSPR on this test set that determines the actual predictive power a QSPR model. The predictive power of a QSPR model can be conveniently estimated by an external q^2 defined as follows (similar to Eq. 1 for the training set):

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{test} (y_i - \bar{y}_{tr})^2} \tag{2}$$

where y_i and $\hat{y}_i$ are the measured and predicted (over the test set), respectively, values of the dependent variable and $\bar{y}_{tr}$ is the averaged value of the dependent variable for the training set; the summations run over all compounds in the test set.

The use of the following statistical characteristics of the test set was also recommended: [26] (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination [66] (predicted versus observed activities R_0^2, and observed versus predicted activities R'^2_0); (iii) slopes k and k' of the regression lines through the origin. We consider a QSAR model predictive, if the following conditions are satisfied [26]:

$$q^2 > 0.5; \tag{3}$$

$$R^2 > 0.6; \tag{4}$$

$$\frac{(\mathrm{R}^2 - \mathrm{R}_0^2)}{\mathrm{R}^2} < 0.1 \quad \text{or} \quad \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \tag{5}$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15. \tag{6}$$

It was demonstrated [26] that all of the above criteria are indeed necessary to adequately assess the predictive ability of a QSPR model.

3.5. Defining Model Applicability Domain

It needs to be emphasized that no matter how robust, significant and validated a QSPR may be, it cannot be expected to reliably predict the modeled property for the entire universe of chemicals. Therefore, before a QSPR is put into use for screening chemicals, its domain of application must be defined and predictions for only those chemicals that fall in this domain may be considered reliable. Described below are some approaches that aid in defining the applicability domain.

Extent of Extrapolation. For a regression-like QSPR, a simple measure of a chemical being too far from the applicability domain of the model is its leverage, $\boldsymbol{h_i}$ [67], which is defined as:

$$h_1 = x_i^T (X^T X)^{-1} x_i \quad (i = 1,..,n) \tag{7}$$

where x_i is the descriptor row-vector of the query compound, and X is the $n \times k$-1 matrix of k model descriptor values for n training set compounds. The superscript t refers to the transpose of the matrix/vector. The warning leverage h^* is, generally, fixed at $3\,k/n$, where n is the number of training compounds, and k is the number of model parameters. A leverage greater than the warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and, therefore, may not be reliable [61, 62, 68].

Effective Prediction Domain. Similarly, for regression-like models, especially when the model descriptors are significantly correlated, Mandel [69] proposed the formulation of effective prediction domain, EPD. It has been demonstrated, with examples, that a regression model is justified inside and on the periphery of the EPD. Clearly, if a compound is determined to be too far from the EPD, its prediction from the model should not be considered reliable.

Residual Standard Deviation. Another important approach that can be used to evaluate the applicability domain is the degree-of-fit method developed originally by Lindberg et al. [70] and modified recently [71]. According to the original method, the predicted y values are considered to be reliable if the following condition is met:

$$\mathbf{s^2 < s_a{}^2(E_x)}\boldsymbol{F} \tag{8}$$

where s^2 is the residual standard deviation (RSD) of descriptor values generated for a test compound, $s_a{}^2(E_x)$ is the RSD of the X matrix after dimensions (components) a, and F is the F-statistic at the probability level α and $(p - a)/2$ and $(p - a)(n - a - 1)/2$ degrees of freedom. The RSD of descriptor values generated for a test compound is calculated using the following equation:

$$\mathbf{s^2 = ||\,e\,||/(p - a)}, \tag{9}$$

where p is the number of x-variables, a is the number of components, and $||e||$ is the sum of squared residuals e_i expressed as

$$\mathbf{e}_{\boldsymbol{i}} = \mathbf{x}_{\boldsymbol{i}} - \mathbf{x}_{\boldsymbol{i}}\mathbf{BB'}, \tag{10}$$

where x_i is the i-th x-variable, and B and B′ represent the weight matrix and transposed weight matrix of x variables, respectively. Since the lowest possible value of F is 1.00 at $\alpha = 0.10$ (when both degrees of freedom are equal to infinity), we decided to replace F with the degree-of-fit factor f to simplify the above condition. Thus, the modified degree-of-fit condition [71] is as follows: predicted y values are considered to be reliable if

$$\mathbf{s^2 < s_a{}^2(E_x)}\boldsymbol{f}, \tag{11}$$

Similarity Distance. In the case of non-linear kNN QSPR method, since the models are based on chemical similarity calculations, a large similarity distance could signal query compounds too dissimilar to the training set compounds. We have used [30] a cutoff value, $\boldsymbol{D_c}$ (Eq. 12), that defines a similarity distance threshold for external compounds.

$$\mathbf{D_c = Z\boldsymbol{\sigma} + y} \tag{12}$$

Here $\boldsymbol{y}$ is the average and $\boldsymbol{\sigma}$ is the standard deviation of the Euclidean distances of the k nearest neighbors of each compound in the training set and $\mathbf{Z}$ is an empirical parameter to control the significance level, with the default value of 0.5. If the distance from an external compound to its nearest neighbor in the training set is above $\boldsymbol{D_c}$, we label its prediction unreliable.

4 Conclusions

In this paper, we have presented a short, conceptual overview of the current state of the affairs in the QSPR modeling field. It is supported by the literature that the majority of QSPR practitioners are presently satisfied with explanatory models characterized by traditional statistical parameters of fitness calculated for the training set. We certainly recognize the importance of mechanistic interpretation of QSPR models in understanding the SPR phenomenon. However, it must be understood that the QSPR models that have not been validated and therefore may not be predictive may provide completely wrong mechanistic interpretation. Validation is particularly important when predictions serve as the basis of regulation, say in the field of environmental protection, by regulatory authorities. One of the important research challenges in the QSPR modelling remains finding descriptor types, correlation approaches, and adequate statistical characteristics of the training set only that may ensure high predictive power of the models.

In summary, we strongly advocate rigorous validation of QSPR models prior to their practical application or interpretation. We have presented some algorithms that afford effective internal and external validation of QSPR models as well as approaches to define the domains of models' applicability in the chemical space. The following general guidelines have been presented for the developments of statistically robust and predictive QSPR models:

1. Establish an SPR database using reliable quantitative measurements of the target property and a set of molecular descriptors.
2. Divide the underlying dataset into training and test sets using diversity sampling algorithms.

3. Develop training set models using available QSPR methods or commercial software. Characterize these models with internal validation parameters as discussed in this paper and define the applicability domain for each model.
4. Validate training set models using external test set and calculate the external validation parameters as discussed in this paper. Ideally, repeat the procedure of training and test selection and external validation several times to identify the QSPR model for the smallest training set that affords adequate prediction power for the biggest test set.
5. Finally, explore and exploit validated QSPR models for possible mechanistic interpretation and prediction.

We have used the name of one of the most famous plays of Oscar Wilde in the title of this paper to emphasize the need of making validation an integral component of good practices of QSPR modeling. Another quote of Oscar Wilde says, "The public has an insatiable curiosity to know everything, except what is worth knowing." We believe that this quote applies well to the current practices of QSPR modelling. In our opinion, it is indeed worth knowing if a QSPR model has the validated predictive power before it is applied to predict, let alone explain the SPR phenomenon of, biological, pharmaceutical, environmental, or any other property of chemicals. Our one-line philosophy of QSPR modelling, therefore, is: ***First, validate, and then explore.***

5 References

[1] Walker J. D. (Ed.), *Handbook on Quantitative Structure Activity Relationships (QSARs) for Pollution Prevention, Toxicity Screening, Risk Assessment and World Wide Web Applications,* SETAC Press, Pensacola, FL 2002.

[2] Walker J. D. (Ed.), *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Physical Properties, Bioconcentration Potential and Environmental Fate of Chemicals,* SETAC Press, Pensacola, FL 2002.

[3] Walker J. D. (Ed.), *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Endocrine Disruption Potential of Chemicals,* SETAC Press, Pensacola, FL 2002.

[4] Walker J. D. (Ed.), *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Effects of Chemicals on Environmental-Human Health Interactions,* SETAC Press, Pensacola, FL 2002.

[5] Walker J. D. (Ed.), *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Ecological Effects of Chemicals*, SETAC Press, Pensacola, FL 2002.

[6] Cronin, M. T. D., Dearden, J. C., Duffy, J. C., Edwards, R., Manga, N., Worth, A. P., and Worgan, A. D. P., The importance of hydrophobicity and Electrophilicity Descriptors in mechanistically-based QSARs for Toxicological Endpoints, *SAR QSAR Environ. Res. 13*, 167 – 176 (2002).

[7] Moss, G. P., Dearden, J. C., Patel, H., and Cronin, M. T. D., Quantitative Structure-Permeability Relationships (QSPRs) for percutaneous absorption, *Toxicol. in Vitro 16*, 299 – 317 (2002).

[8] Trohalaki, S., Gifford, E., and Pachter, R., Improved QSARs for predictive toxicology of halogenated hydrocarbons, *Comput. Chem. 24*, 421 – 427 (2000).

[9] Schultz, T. W, Bowerrs, G. S., and Cronin, M. T. D., Structure-toxicity relationships for four classes of aliphatic electrophiles to *Vibrio fischeri*, *Marine Environ. Res. 50*, 61 – 81 (2000).

[10] Schultz, T. W., Aptula, A. O., Netzeva, T. I., and Cronin, M. T. D., Quantitative Structure- Activity Relationships for the Toxicity of AliphaticCompounds to *Tetrahymena pyriformis*. A Mechanism of Action Approach, Presented at the QSAR 2002 meeting, May 25 – 29, Ottawa, Canada (2002).

[11] Akers, K. S., Sinks, G. D., and Schultz, T. W., Structure-toxicity relationships for selected halogenated aliphatic chemicals, *Environ. Toxicol. Pharm*. *7*, 33 – 39 (1999).

[12] Benigni R., Giuliani, A., Franke, R.: and Gruska, A., Quantiative-Structure-Activity Relationships of mutagenic and Carcinogenic Aromatic Amines, *Chem. Rev. 100*, 3697 – 3714 (2000).

[13] Devillers, J., QSAR Modeling of Large Heterogeneous Sets of Molecules, *SAR QSAR Environ. Res. 12*, 515 – 528 (2001).

[14] Kaiser, K. L. E., Dearden, J. C., Klein, W., and Schultz, T. W., A Note of Caution to Users of ECOSAR, *Water Qual. Res. J. Canada 34*, 179 – 182 (1999).

[15] ECOSAR, Version 0.99f, Jan 2000, (http://www.epa.gov/oppt/newchems/21ecosar.htm).

[16] Rousseeuw, P. J., and Leroy, A. M., *Robust Regression and Outlier Detection,* John Wiley & Sons, New York 1987.

[17] Belsley, D. A., Kuh, E., and Welsch, R. E., *Regression Diagnostics – Identifying Influential Data and Sources of collinearity*, John Wiley & Sons, New York 1980.

[18] Gironés, X., Gallegos, A., and Ramon, C.-D., Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum similarity Measures as Descriptors in QSAR, *J. Chem Inf. Comput. Sci. 46*, 1400 – 1407 (2000).

[19] Bordás, B., Kömíves, T., Szántó, Z., and Lopata, A., Comparative Three-Dimensional Quantitative Structure-Activity Relationship Study of Safeners and Herbicides, *J. Agricult. Food Chem. 48*, 926 – 931 (2000).

[20] Fan, Y., Shi, L. M., Kohn, K. W., Pommier, Y., and Weinstein, J. N., Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based Studies, *J. Med. Chem. 44,* 3254 – 3263 (2001).

[21] Randić, M., and Basak, S. C., Construction of high-quality structure-property-activity regressions: the boiling points of sulfides, *J. Chem. Inf. Comput. Sci. 40*, 899 – 905 (2000).

[22] Suzuki, T., Ide, K., Ishida, M., and Shapiro, S., Classification of Environmental Estrogens by Physicochemical Properties Using Principal Component Analysis and Hierarchical Cluster Analysis, *J. Chem. Inf. Comput. Sci. 41,* 718 – 726 (2001).

[23] Basak, S. C., and Mills, D., Prediction of mutagenicity utilizing a hierarchical QSAR approach, *SAR QSAR Environ. Res. 12*, 481 – 496 (2001).

[24] Wang, X., Yin, C., and Wang L., Structure-activity relationships and response-surface analysis of nitroaromatics toxicity to the yeast (Saccharomyces cerevisiae), *Chemosphere 46*, 1045 – 1051 (2002).

[25] Kubinyi, H., Hamprecht, F. A., and Mietzner, T., Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices, *J. Med. Chem. 41*, 2553 – 2564 (1998).

[26] Golbraikh, A., and Tropsha, A., Beware of q2!, *J. Mol. Graph. Model. 20*, 269 – 276 (2002).

[27] Cramer III, R. D., Patterson, D. E., and Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape

on binding of steroids to carrier proteins, *J. Amer. Chem. Soc. 110*, 5959 – 5967 (1988).

[28] Coats, E. A., The CoMFA steroids as a benchmark data set for development of 3D QSAR methods, in: Kubinyi, H., Folkers, G., and Martin, Y. C. (Eds.), *3D QSAR in Drug Design.* V.3., Kluwer/ESCOM, Dordrecht (The Netherlands) 1998, pp. 199 – 213.

[29] Novellino, E., Fattorusso, C., and Greco, G., Use of Comparative Molecular Field Analysis and Cluster Analysis in Series Design, *Pharm. Acta Helv. 70*, 149 – 154 (1995).

[30] Zheng, W., and Tropsha, A., Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle, *J. Chem. Inform. Comput. Sci. 40*, 185 – 194 (2000).

[31] Wehrens, R., Putter, H., and Buydens, L. M. C., The bootstrap: a tutorial, *Chemom. Intell. Lab. Systems 54*, 35 – 52 (2002).

[32] Efron, B., and Tibshirani, R. J., *An Introduction to the Bootstrap*, Chapman & Hall, New York 1993.

[33] Wold, S., and Eriksson, L., Statistical Validation of QSAR Results, in: van de Waterbeemd H., (Ed.), *Chemometrics Methods in Molecular Design*, VCH, Weinheim (Germany) 1995, pp. 309 – 318.

[34] Clark, R. D., Sprous, D. G., and Leonard, J. M., Validating Models Based on Large Dataset, in: Höltje, H.-D., and Sippl, W. (Eds.), *Rational Approaches to Drug Design*, *Proceedings of the 13th European Symposium on Quantitative Structure-Activity Relationship,* Aug 27 – Sept 1, 2000, Prous Science Düsseldorf (Germany) 2001, pp. 475 – 485.

[35] Wilcox, R. E., Huang, W.-H., Brusniak, M.-Y. K., Wilcox, D. M., Pearlman, R. S., Teeter, M. M., DuRand, C. J., Wiens, B. L., and Neve, K. A., CoMFA-Based Prediction of Agonist Affinities at Recombinant Wild Type versus Serine to Alanine Point Mutated D2 Dopamine Receptors, *J. Med. Chem. 43*, 3005 – 3019 (2000).

[36] Wu, W., Walczak, B., Massart, D. L., Heuerding, S., Erni, F., Last, I. R., and Prebble, K. A., Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set, *Chemometr. Intell. Lab. Syst. 33*, 35 – 46 (1996).

[37] Yasri, A., and Hartsough, D., Toward an Optimal Procedure for Variable Selection and QSAR Model Building, *J. Chem. Inf. Comput. Sci. 41,* 1218 – 1227 (2001).

[38] Kauffman, G. V., and Jurs, P. C., QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors, *J. Chem. Inf. Comput. Sci. 41,* 1553 – 1560 (2001).

[39] Mattioni, B. E., and Jurs, P. C., Development of Quantitative Structure-Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors, *J. Chem. Inf. Comput. Sci. 42*, 94 – 102 (2002).

[40] Burden, F. R., and Winkler, D. A., Robust QSAR Models Using Bayesian Regularized Artificial Neural Networks, *J. Med. Chem. 42*, 3183 – 3187 (1999).

[41] Burden, F. R., Ford, M. G., Whitley, D. C, and Winkler, D. A., Use of automatic relevance determination in QSAR studies using Bayesian neural networks, *J. Chem. Inf. Comput. Sci. 40*, 1423 – 1430 (2000).

[42] Adams, M. J., *Chemometrics in Analytical Spectroscopy*, The Royal Society of Chemistry, Cambridge (UK) 1995.

[43] Potter, T., and Matter, H., Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases, *J. Med. Chem. 41*, 478 – 488 (1998).

[44] Lajiness, M., Johnson, M. A., and Maggiora, G. M., Implementing Drug Screening Programs Using Molecular Similarity Methods, in: F. J. Lauchere (Ed.), *QSAR: Quantitative Structure-Activity Relationships in Drug Design*., Alan R. Liss Inc., New York 1989, pp. 173 – 176.

[45] Taylor, R., Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals, *J. Chem. Inf. Comput. Sci. 35*, 59 – 67 (1995).

[46] Snarey, M., Terrett, N. K., Willett, P., and Wilton, D. J., Comparison of Algorithms for Dissimilarity-Based Compound Selection, *J. Mol. Graph. Model. 15*, 372 – 385 (1997).

[47] Gasteiger, J., and Zupan, J., Neural Networks in Chemistry, *Angew. Chem. Int. Ed. Engl. 32*, 503 – 527 (1993).

[48] Gramatica, P., Consonni, V., and Todeschini, R., QSAR Study on the Tropospheric Degradation of Organic Compounds, *Chemosphere 38*, 1371 – 1378 (1999).

[49] Kennard, R. W., and Stone, L. A., Computer Aided Design of Experiments, *Technometrics 11*, 137 – 148 (1969).

[50] Bourguignon, B., Deaguiar, P. F., Thorre, K., and Massart, D. L., Application Of Nonlinear-Regression Functions For The Modeling Of Retention In Reversed-Phase Lc, *J. Chromatogr. Sci. 32,* 144 – 152 (1994).

[51] Bourguignon, B., Deaguiar, P. F., Khots, M. S., and Massart, D. L., Optimization in Irregularly Shaped Regions - PH and Solvent Strength in Reversed-Phase High-Performance Liquid-Chromatography Separations, *Anal. Chem. 66*, 893 – 904 (1994).

[52] Sjöström, M., and Eriksson, L., Applications of Statistical Experimental Design, in: van de Waterbeemd H., (Ed.), *Chemometrics Methods in Molecular Design*, VCH, Weinheim (Germany) 1995, pp. 63 – 90.

[53] Eriksson, L., and Johansson, E., Multivariate Design and Modeling in QSAR. Tutorial. *Chemometr. Intell. Lab. Syst. 34*, 1 – 19 (1996).

[54] Carlson, R., *Design and Optimization in Organic Synthesis*, Elsevier, Dordrecht (The Netherlands) 1992.

[55] Martin, E. J., and Critchlow, R. E., Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery, *J. Comb. Chem. 1,* 32 – 45 (1999).

[56] Miller, A., and Nguyen, N.-K., A Fedorov Exchange Algorithm of D-Optimal Design, *Appl. Stat. 43*, 669 – 678 (1994).

[57] Mitchell, T. J., An Algorithm for the Construction of "D-optimal" Experimental Designs, *Technometrics 16*, 203 – 210 (1974).

[58] Mitchell, T. J., An Algorithm for the Construction of "D-optimal" Experimental Designs, *Technometrics 42*, 48 – 54 (2000).

[59] Golbraikh, A., and Tropsha, A., Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Test and Training Set Selection, *J. Comput.-Aided Mol. Des.*, in press.

[60] Golbraikh, A., *J. Chem. Inf. Comput. Sci., 40*, 414 – 425 (2000).

[61] Gramatica, P., and Papa, E., QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors, *Quant. Struct-Act. Relat*., in press.

[62] Gramatica, P., Papa, E., and Pilutti P., QSAR Predictions of Ozone Tropospheric Degradation, *Quant. Struct.-Act. Relat*., in press.

[63] Marengo, E., and Todeschini, R., A New Algorithm for Optimal, Distance-based Experimental Design, *Chemom. Intell. Lab. Syst.*, 37 – 44 (1992).

[64] Norinder, U., Single and Domain Made Variable Selection in 3D QSAR applications, *J. Chemomet. 10*, 95 – 105 (1996).

[65] Zefirov, N. S., and Palyulin, V. A., QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-

Activity Regressions" the Real High Quality QSAR Models?, *J. Chem. Inf. Comput. Sci. 41,* 1022–1027 (2001).

[66] Sachs, L., *Applied Statistics. A Handbook of Techniques*, Springer-Verlag, Heidelberg (Germany) 1982.

[67] Atkinson, A. C., *Plots, transformations and regression,* Clarendon Press, Oxford (UK) 1985, p. 282.

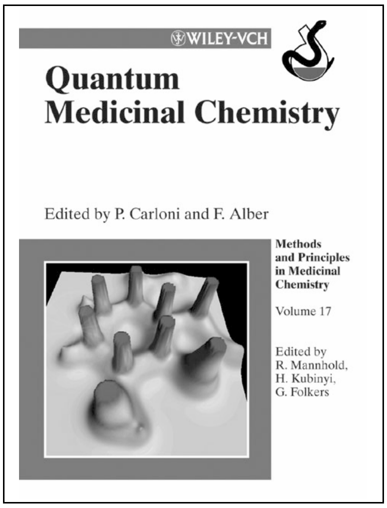
[68] Gramatica, P., Corradi, M., and Consonni, V., Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors, *Chemosphere 41*, 763–777 (2000).

[69] Mandel, J., The Regression Analysis of Collinear Data, *J. Res. Nat. Bur. Stand. 90*, 465–476 (1985).

[70] Lindberg, W., Persson, J.-A., and Wold, S., Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate, *Anal. Chem. 55*, 643–648 (1983).

[71] Cho, S. J., Zheng, W., and Tropsha, A., Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches, *J. Chem. Inf. Comput. Sci. 38*, 259–268 (1998).

Received on: July 17, 2002


WILEY-VCH
Quantum Medicinal Chemistry
Edited by P. Carloni and F. Alber
Methods and Principles in Medicinal Chemistry
Volume 17
Edited by R. Mannhold, H. Kubinyi, G. Folkers