Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training−Test Sets and Consensus Modeling

Paola Gramatica,* Pamela Pilutti, and Ester Papa

Department of Structural and Functional Biology, QSAR and Environmental Chemistry Research Unit, University of Insubria, via Dunant 3, 21100 Varese, Italy

Received February 26, 2004

The rate constant for hydroxyl radical tropospheric degradation of 460 heterogeneous organic compounds is predicted by QSAR modeling. The applied Multiple Linear Regression is based on a variety of theoretical molecular descriptors, selected by the Genetic Algorithms-Variable Subset Selection (GA-VSS) procedure. The models were validated for predictivity by both internal and external validation. For the external validation two splitting approaches, D-optimal Experimental Design and Kohonen Artificial Neural Networks (K-ANN), were applied to the original data set to compare the two methodologies. We emphasize that external validation is the only way to establish a reliable QSAR model for predictive purposes. Predicted data by consensus modeling from different models are also proposed.

INTRODUCTION

Volatile Organic Compounds (VOCs) are a class of organic chemicals largely present in the troposphere because of their vapor pressure. Chemical and biological transformations, and degradations, play a major role in the transport and mobility of such chemicals in the environment. An indirect measure of the persistence of organic compounds in the atmosphere, and therefore a necessary parameter in environmental exposure assessment,[1] is the rate at which these compounds react; however, the experimental determination of such reaction rates is difficult, time-consuming, and expensive.

The dominant chemical process of chemicals in the gas-phase is their reaction with OH radicals, NO_3 radicals, and ozone. Indeed the hydroxyl radical is the key reactive species in the troposphere, where it reacts with practically every organic compound,[2] and this has led to many experiments to determine the rate constants of the OH radical reaction with chemicals. Although extensive experimental work has been carried out in recent years, the rate constants of only approximately 750 organic compounds (from among the thousands of potential organic pollutants in the environment) have been measured for the gas-phase reaction with OH radicals.[3] Since the number of compounds released into the atmosphere from antropogenic and biogenic sources, or formed in the atmosphere, greatly exceeds the amount of experimentally available data, it would be very desirable to have methods to estimate oxidation rate constants. Considering the growing number of organic compounds of anthropogenic origin, this need is particularly urgent as such methods could lead to the rapid recognition of safe/high risk organic chemicals and, additionally, could be usefully used for the planning and development of new safer organic chemicals.

* Corresponding author phone: +39-0332-421573; fax: +39-0332-421554; e-mail: paola.gramatica@uninsubria.it, http://www.qsar.it.

In recent years, several QSAR\QSPR (Quantitative Structure−Activity\Property Relationships) models predicting oxidation rate constants with $OH^{\bullet}$, $NO_3^{\bullet}$, and ozone for many heterogeneous compounds have been published, and the different approaches concerning molecular description and adopted methodology have been compared.[1,3−14]

The most used method[15] for estimating tropospheric degradation by hydroxyl radicals is Atkinson's fragment contribution method.[4] This method generally produces the lowest residual sum of squares (RSS) error for the majority of the examined compounds if, for instance, it is compared with Klamt's QSPR model based on quantum chemical descriptors.[7] However, as Bakken and Jurs[8] pointed out, this result could be biased as Atkinson used many of these compounds to calculate fragment contribution parameters.

Other estimation methods for OH radical reaction rates have been developed using different properties (such as bond dissociation energy, ionization potential, etc.) or theoretical molecular descriptors.[3,8,10] All these methods have been fairly successful in obtaining prediction models, but little work has been done to examine model predictivity and chemical domain of application over a wide range of compounds, especially for new chemicals. Bakken and Jurs,[8] among the few, published an elegant verification of model predictivity by external validation of QSAR models, applied to heterogeneous chemicals: they developed their models using theoretical molecular descriptors selected by evolutionary optimization techniques.

In effect, any QSAR model must be validated for its predictivity before it can be used to predict the response of additional chemicals. Validating QSAR with external data (i.e. data not used in the model development), although the most demanding, is the best method of validation.[16,17] However the availability of an independent external validation set of several compounds is rare in QSAR. Thus, the input data set must be adequately split by experimental design or other splitting procedures[17,18] into representative training and test sets.

10.1021/ci049923u CCC: $27.50
Published on Web 08/03/2004

This paper proposes new MLR-QSAR models of OH radical reaction rate for a wide and heterogeneous data set of 460 volatile organic compounds (VOCs). The peculiarity of these models, based on different theoretical molecular descriptors selected by Genetic Algorithm as a variable subset selection procedure, is their applicability to heterogeneous chemicals, and their validation for predictive purposes by both internal and external validation. External validation was performed by splitting the original data set by two different methods: the statistical Experimental Design procedure (D-optimal distance) and by the Kohonen Artificial Neural Networks (K-ANN); this was done to verify the impact structural heterogeneity (in chemicals split into training and test sets) has on model performance.The predicted data were verified for reliability in the chemical domain of applicability. Finally, predictions by consensus modeling, that considers more than just one QSAR model, are proposed and commented on.

EXPERIMENTAL DATA AND CHEMOMETRIC METHODS

Experimental Data. The experimental data of the OH radical degradation rate constants of 460 heterogeneous organic compounds were obtained from the literature[19] (see Table S1 in the Supporting Information). The selected data were for reactions at 25 °C and 1 atm; all the rate constants, reported in $cm^3\ s^{-1}\ molecule^{-1}$, have been transformed to logarithmic units and multiplied by -1 to obtain positive values.

The data set includes alkanes, alkenes, alcohols, halogenated chemicals, amines, aromatics, and other functional groups.

Theoretical Molecular Descriptors. The molecular descriptors for the given compounds were mainly calculated using *DRAGON* software[20] on the (x,y,z)-atomic coordinates of the minimal energy conformations determined by the MM+ method in *HYPERCHEM* Package.[21] A total of 1308 molecular descriptors of differing types were calculated to describe compound structural diversity. The descriptor typology is as follows: (a) 0D-48 constitutional (atom and group counts); (b) 1D-121 functional groups, (c) 1D-120 atom centered fragments, (d) 2D-140 topological, (e) 2D-64 BCUTs, (f) 2D-21 Galvez Indices from the adjacency matrix, (g) 2D-96 various auto-correlations from the molecular graph, (h) 2D-33 connectivity index, (i) 2D-47 information index, (l) 2D-44 eigenvalue-based indices m) 3D-99 WHIMs,[22] and (n) 3D-197, recently proposed, GETAWAY descriptors.[23] The list of these molecular descriptors, and their meaning, is provided with literature references by the *DRAGON* package; the calculation procedure is explained in detail, with related literature references, in the *Handbook of Molecular Descriptors.*[24] Furthermore, to provide energy information, the following electronic descriptors are added: three quantum-chemical descriptors (Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO) energies, HOMO−LUMO gap), calculated by the semiempirical molecular orbital program *MOPAC* (PM3 Hamiltonian for geometry optimization) in the software *HYPERCHEM.*[21]

Constant values and descriptors found to be correlated pairwise were excluded in a prereduction step (one of any two descriptors with a K correlation greater than 0.95 was removed to reduce redundant and nonuseful information), thus 470 molecular descriptors underwent subsequent variable selection.

Principal Component Analysis. Principal Component Analysis (PCA) for data exploration was performed on autoscaled data by the *SCAN*[25] and *STATISTICA*[26] packages.

Model Building, Variable Selection, and Validation of Internal Predictivity. Multiple linear regression (MLR) analysis and variable selection were performed by the software *Moby Digs*[27] using the *Ordinary Least Squares* regression (*OLS*) method for the modeling and the *GA-VSS* (Genetic Algorithm − Variable Subset Selection) method[28] for the variable selection. Genetic Algorithm was applied to the input set of 470 molecular descriptors for each chemical and the related response in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals. The parameter that is optimized is the cross-validated correlation coefficient R^2_{cv} or Q^2_{LOO} (*leave-one-out*), calculated by the formula

$$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_{i/i})^2}{\sum (y_i - \bar{y})^2}$$

where y_i, $\hat{y}_i$, 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.

Particular attention was devoted to the collinearity of the selected molecular descriptors: in fact, to avoid multicollinearity without, or with, "apparent" prediction power (due to chance correlation), regression was calculated only for variable subsets with an acceptable multivariate correlation with response, by applying the *QUIK* rule (Q Under Influence of *K*).[29] The acceptable models are only those with a global correlation of [X+y] block (K_{XY}) greater than the global correlation of the X block (K_{XX}) variable, X being the molecular descriptors, and *y* the response variable. The collinearity in the original set of molecular descriptor results in many similar models that more or less yield the same predictive power (in *MOBY-DIGS* software 100 models of different dimensionality). Therefore, when there were models of similar performance those with higher ΔK ($K_{XY} - K_{XX}$) were selected and further verified.

The application of Genetic Algorithms, repeated several times, yielded the same population of 100 satisfactory regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors (with 1−2 variables, models developed by the all-subset-models procedure in order to explore all the low dimension combinations). The number of descriptors was subsequently increased one by one and new models formed. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree.

To avoid overestimation of model internal predictive power and to verify the predictivity stability of the models, stronger cross-validation by the *leave-many-out* procedure (repeated 5000 times, with 50% of objects left out from the training set at each step) was performed, and the corresponding Q^2_{LMO} was calculated.

The proposed models were also checked for reliability and robustness by permutation testing: new models were recalculated for randomly reordered response (Y scrambling). It is expected that the models applied on the data set with randomized response should have significantly lower Q^2 values than the proposed ones.

Chemical Domain. QSAR models must always be verified for their applicability with regard to chemical domain, to produce reliable predicted data for chemicals that are not too structurally dissimilar. The Williams plot, obtained by the *SCAN* package,[25] verified the presence of *outliers* (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, $\pm 3\sigma$) and chemicals very *influential* in determining model parameters. Also the data predicted by the models were verified for reliability by their *leverage*, so that only predicted data for chemicals belonging to the chemical domain of the training set would be proposed. In fact, *leverage* can be used as a quantitative measure of the model applicability domain suitable for evaluating the degree of extrapolation: it represents a sort of compound "distance" from the model experimental space. Prediction for a compound having a high *leverage* value ($h > h^*$, the critical value being $h^* = 3p'/n$, where p' is the number of the model variables plus one, and n the number of the objects used to calculate the model) must be considered as unreliable. Conversely, when the compound has a *leverage* value lower than the critical one, the probability of accordance between the predicted and the actual values is as high as that for the training set chemicals.[30] The chemical domain of the studied chemicals in the models was verified by the *leverage* approach to verify prediction reliability.[16,31]

Standard Deviation Error in Prediction (SDEP), Standard Deviation Error in Calculation (SDEC), Standard Error of Estimate (s), the intercorrelation of the selected descriptors (K_{XX}), and the correlation of the X block with response (K_{XY}) are also reported, together with the coefficient of determination (R^2).

External Validation: Splitting Training/Test. To obtain compounds for external validation, the available set of chemicals was split into a training set and an external validation set. The training set selection was performed by the software *DOLPHIN* [32] using a D-optimal experimental design[33−36] (Marengo-Todeschini algorithm) and by the software *KOALA*[37] using the Kohonen Artificial Neural Network (K-ANN).[38,39]

The Marengo−Todeschini algorithm[33] is an algorithm for optimal distance based on experimental design that does not require any preliminary hypothesis about a regression model. The best set of compounds is defined through a fast exchange algorithm where, in each cycle, substitution provides the maximum increase in the minimum distance between currently selected compounds. Such an algorithm provides a final distribution of the most dissimilar compounds selected from the set of allowed candidates. Regression models were developed on the selected training set, and once the models were established, predictions were made for the remaining molecules under study (test set).

The splitting of the data set, realized by Kohonen Artificial Neural Network (K-ANN)[38] using the package *KOALA*,[37] takes advantage of the clustering capabilities of K-ANN, allowing the selection of a meaningful training set and a representative validation set. The 41 most significant principal components, calculated from each group of *DRAGON* molecular descriptors, were used to describe the relevant structural information of the chemicals. This structural information and the response were used as variables to build a Kohonen map (12×12 neurons, 500 epochs). At the end of 500 epochs of the net training, similar chemicals fall within the same neuron, i.e., they carry the same information. To select the training set of chemicals, it is assumed that the compound closest to each neuron centroid is the most representative of all the chemicals within the same neuron. Thus, the selection of the training set chemicals was performed by the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, were used for the test set.

Predictivity. The predictive power of the regression model developed on the selected training set is estimated on the predictions of test set chemicals, by the external Q^2 that is defined[40]

$$Q^2_{ext} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_{tr})^2}$$

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

RESULTS AND DISCUSSION

Molecular Modeling. The aim of this work is the development and proposal of a single and validated QSAR model for the prediction of a hydroxyl radical rate constant for a wide and heterogeneous set of organic compounds, using only theoretical molecular descriptors. The great advantage of theoretical descriptors is that defined software can calculate them homogeneously for all chemicals, also for not yet synthesized chemicals, the only need being a hypothesized chemical structure. The application of a single and general QSAR model for all compounds could make it unnecessary to have a priori knowledge of the degradation process of each chemical. To find a relationship between k_{OH} and the structural features of the chemicals, a wide set of theoretical molecular descriptors that takes into account different structural features (monodimensional, bidimensional, and three-dimensional) was used. Information regarding the energetics of the studied reaction was also taken into account by adding quantum-chemical descriptors, chosen from among the most commonly used descriptors and those always relevant to radical reaction modeling (HOMO, LUMO, and the HOMO−LUMO gap). As modeling input variables we used many different types of molecular descriptors, this was to give us the possibility of capturing all the relevant structural features really related to response. As we cannot have a priori knowledge of which descriptors, and which particular combinations with others, are related to the studied response and are able to be used in models for prediction aims, we applied *Genetic Algorithms* [27,28] as the variable selection procedure (GA-VSS) to select only the best combinations of those most relevant to obtaining models with the highest predictive power for k_{OH}.

Internal and External Validation. The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but is mainly their possibility of predictive application. For this reason the model calculations were performed by GA-VSS maximizing the explained variance in prediction (Q^2_{LOO}) (see Experimental Section).

In addition, to avoid the danger of overfitting and the possibility of overestimating model predictivity by using only Q^2_{LOO}[41,42] the internal predictivity stability of the models was verified using the *leave-many-out* (Q^2_{LMO}) procedure, as is strongly recommended for QSAR modeling.[16,31,40,43] The robustness of the proposed models and their predictivity was guaranteed by the stability of the Q^2_{LMO}, even when 50% of the training compounds were randomly left out.

Finally, Y-randomization was applied to exclude the possibility of chance correlation, i.e., fortuitous correlation without any predictive ability. It gave the following results: the random models, performed using a scrambled order of the experimental rate constant, were found to have significantly lower R^2 and Q^2 than the original models (R^2 range: 1.1% to 7.2%; Q^2 range: −3.6% to 2.8%) corroborating the statistical reliability of the actual models.

We examined all the most important internal validation criteria, namely cross-validation (LOO and LMO) and Y-scrambling, that appear to be necessary conditions, though still not sufficient, for the model to have high predictive power. In fact we,[16] like other authors,[42,43] are strongly convinced, from personal experience, that models with high apparent predictivity, highlighted only by internal validation methods, could be unpredictive when verified on new chemicals not used in developing the model. Thus, for a stronger evaluation of model applicability for prediction on new chemicals, external validation of the models was always performed. The effective predictive capability of a model was evaluated by the "external" validation procedure, i.e., by comparing the predictions made for molecules excluded from the model generation step with their actual experimental activity.

Given a single data set, which is a typical situation in QSAR modeling, external validation can only be achieved by splitting the original data set into a training set, used to establish the QSAR model, and a test set, to evaluate model performance. The approaches for creating training and test sets range from straightforward random selection, through various clustering techniques, to the methods of Kohonen maps[38,44−46] and formal statistical experimental design (factorial and D-Optimal).[33−36] The underlying goal at this step is to ensure that both the training and the test sets separately span the whole descriptor space occupied by the entire data set, and the chemical domain in the two data sets is not too dissimilar.

Two techniques were used in this work to select the test set and compare the results: Experimental Design[32,33] and Kohonen Artificial Neural Network (K-ANN).[38,39] The Experimental Design procedure (here D-optimal design) provides a strategy for selecting the most dissimilar molecular structures in a data set, taking into account the complete structural information and also the response value, thus, the chemicals in the training set represent the breadth, or variety, of all existing chemicals within that domain (i.e. diversity) and the test set is within the training set domain. K-ANN, taking advantage of the clustering capabilities, ensures that both sets are homogeneously distributed within the entire area of the descriptor space; in this case the chemicals in both sets, selected to maximize the coverage of the descriptor space (i.e. representativity), represent the depth of distribution of all existing chemicals. The selected training chemicals are those with the minimal distance from the centroid of each cell in the top map. In this case, the representative points of the test set are close (in the same cell of the top map) to representative points of the training set in the multidimensional descriptor space.[17,18,47]

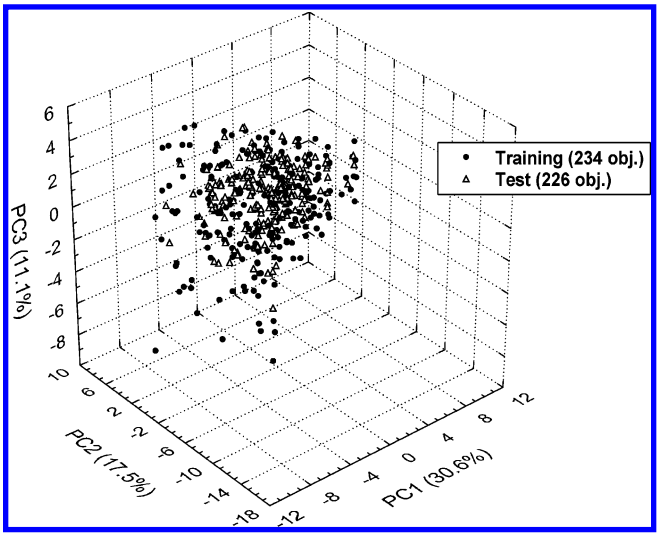

Figure 1. Principal Component Analysis on the 41 "more significant" principal components, calculated from each group of *DRAGON* molecular descriptors, (PC1−PC2−PC3: EV% = 59.3%) for the data set chemicals split by experimental design in the training (●) and test sets (△).

A recent paper compared these two methodologies in toxicity modeling,[17] but the splitting was realized only on the two descriptors selected a priori, in a mechanistic approach, in the proposed QSAR model (hydrophobicity/electrophilicity plane).

The two splitting methodologies were applied here, in our statistical approach, to all the relevant structural information in the original molecular descriptor sets: the 41 significant principal components of molecular descriptors were calculated from each group of *DRAGON* descriptors and the response.

Both methodologies were applied using strong splitting: about 50% of the chemicals in each set (234 chemicals in the training set and 226 in the test).

Figures 1 and 2 show the Principal Components Analysis of the data set for the two splittings, where the first three principal components explain 59.3% of the structural variance in the two chemical sets. The difference in the two splittings is immediately evident from the two figures. Figure 1 shows the validation set to be completely contained within the training set, confirming the Experimental Design characteristic of selecting the most dissimilar structures for the training set. Instead, in Figure 2 the splitting by K-ANN is better balanced, the chemical composition of the training and test sets being more similar.

Also the response range highlights this different behavior. For D-Optimal-based splitting the range of response in the training set is 9.44−15.70 log unit ($\bar{y}$ = 11.40 log unit), while

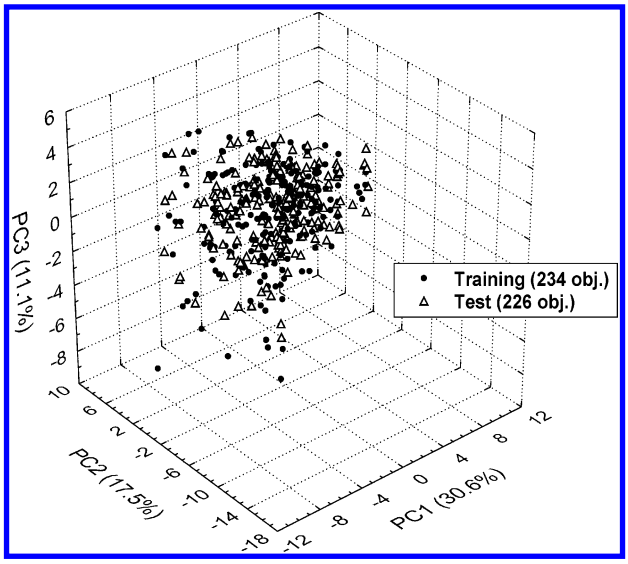

Figure 2. Principal Component Analysis on the 41 "more significant" principal components, calculated from each group of *DRAGON* molecular descriptors, (PC1−PC2−PC3: EV% = 59.3%) for the data set chemicals split by K-ANN in the training (●) and test sets (△).

the range of the validation set response is 9.65−14.05 log unit ($\bar{y}$ = 10.96 log unit). For the K-ANN approach the range of response in the training set is 9.44−15.70 log unit ($\bar{y}$ = 11.19 log unit), while the range of the validation set response is 9.50−14.77 log unit ($\bar{y}$ = 11.18 log unit).

As our QSAR approach is statistical and based on Genetic Algorithm selection of variables modeling studied response, the collinearity in the original set of variables results in a population of similar models that more or less yield the same predictive power based on different sets of molecular descriptors. All the models developed on these two training sets and used to predict values for the test chemicals have high predictive performance, verified with both internal and external validation.

The best predictive model we selected in this population of 100 models (no. variables = 1÷4), when using the D-Optimal algorithm for the splitting, was

$$\text{-logk(OH)} = 5.15 - 0.66\ \text{HOMO} + 0.33\ \text{nX} - 0.37\ \text{CIC0} + 0.13\ \text{nCaH}$$

$$\text{n(training)} = 234,\ \text{n(test)} = 226,\ R^2 = 82.8\%,$$
$$Q^2 = 81.9\%,\ Q^2_{\text{LMO(50\%)}} = 81.6\%,\ Q^2_{\text{EXT}} = 82.6\%,$$
$$s = 0.478,\ \text{SDEP} = 0.484,\ \text{SDEC} = 0.473,$$
$$K_{\text{XX}} = 33.8\%,\ K_{\text{XY}} = 44.6\%$$

Figure 3 shows the regression line of the above proposed model, the 3σ interval is reported as dotted lines, and Table S1 in the Supporting Information gives the data predicted by this model.

The selected model was preferred because, from among those with similar performance, it had fewer influential chemicals and no outlier. Compared with the other chemicals in the data set, 1.1.1.2.2-pentafluoroethane, hexafluorobenzene, and propylpentafluorobenzene are the influential chemicals with higher leverage value in the training set, probably because of their high fluorine content. No outliers or influential chemicals were present in the test set.

The best predictive model, selected in the population of 100 models (no. variables = 1 ÷ 4) and using K-ANN-based splitting, was the one with the higher external predictivity and with the same descriptors as the previous splitting:

$$\text{-logk(OH)} = 5.00 - 0.68\ \text{HOMO} + 0.35\ \text{nX} - 0.39\ \text{CIC0} + 0.13\ \text{nCaH}$$

$$\text{n(training)} = 234,\ \text{n(test)} = 226,\ R^2 = 82.8\%,$$
$$Q^2 = 81.6\%,\ Q^2_{\text{LMO(50\%)}} = 81.0\%,\ Q^2_{\text{EXT}} = 81.3\%,$$
$$s = 0.427,\ \text{SDEP} = 0.436,\ \text{SDEC} = 0.422,$$
$$K_{\text{XX}} = 32.7\%,\ K_{\text{XY}} = 43.5\%$$

This model was selected as it had fewer influential chemicals and only one outlier (chloromethane) in the training set. The influential chemicals with high leverage value in the training set are hexafluorobenzene and propylpentafluorobenzene, trifluoromethane and 1.1.1.2-tetrafluoroethane, while 1.1.1.2.2-pentafluoroethane and 1.1.1-trifluoroethane are test set chemicals outside model domain applicability and thus with predicted data that could be unreliable. It is evident that the highly fluorinated chemicals have a strong structural peculiarity that the model is not able to capture. Chloromethane is the sole outlier of the test set, and there are two possible explanations: either the experimental input data was wrong or the descriptors selected in the model failed to capture some relevant structural feature present in this small molecule and absent in others. At this stage it is impossible to verify either statement.

Increasing the number of descriptors by 1−5 variables led to a 2% increase in model performance, that does not justify the increased model complexity.

The two reported models are very similar, being based on the same molecular descriptors; their differences lie only in the regression coefficients that are related to the training set compositions derived from the two splittings. The models are stable and predictive both internally (high value of cross-validation parameters Q^2 and $Q^2_{\text{LMO(50\%)}}$) and externally Q^2_{EXT}.

Again it is possible to highlight that splitting by D-optimal design gives a more optimistic idea of external predictivity (Q^2_{EXT} almost always higher than Q^2_{LOO}), as can be expected for a test set lying completely within the descriptor space of the training chemicals. In fact, splitting by D-optimal design gives models with completely interpolated predictions, whereas for some chemicals the prediction using K-ANN splitting could be interpreted as extrapolations.

The molecular descriptors (selected by Genetic Algorithm) in the models are all highly informative of different aspects of the studied reaction. In the above equations the descriptors are reported in decreasing order of significance, according to standardized coefficient values (reported in brackets after the descriptor abbreviation). The best descriptor is the energy of the highest occupied molecular orbital (HOMO, −0.68). This descriptor is a measure of the nucleophilicity of a molecule, highly reactive chemicals having high HOMO energy, while the number of halogen atoms (nX, 0.36) and the number of unsubstituted aromatic C (sp2) (nCaH, 0.28) probably encode different points of attachment in hydroxyl radical reactions; the CIC0 (−0.32) is a complementary information content index,[48,49] related to the differences in

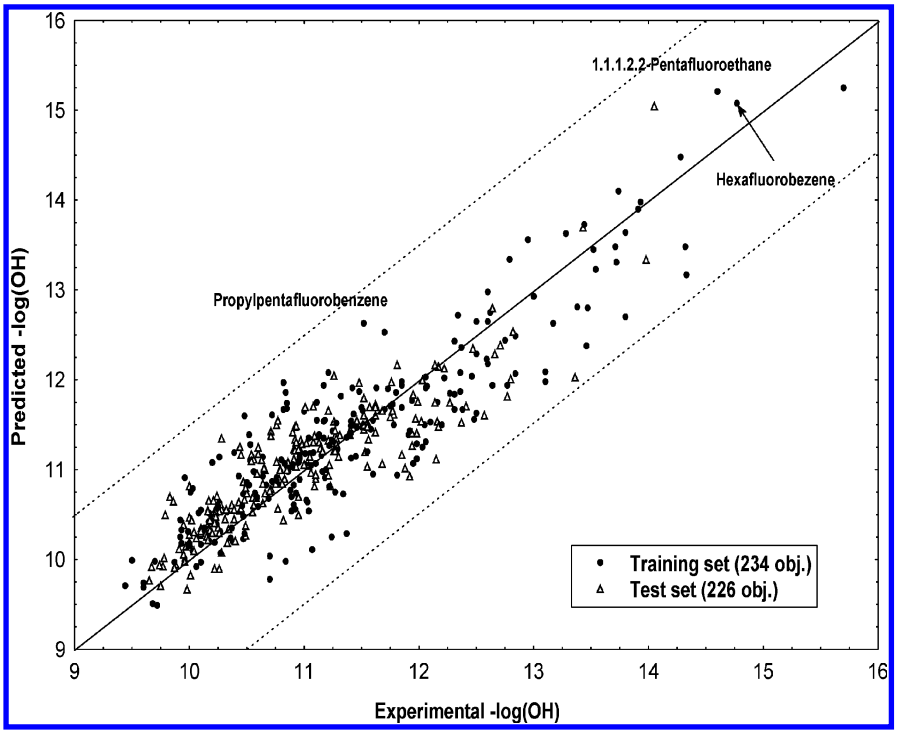

Figure 3. Regression line for the externally validated model by experimental design splitting. The log k(OH) values for the training and test set chemicals are labeled differently, and the outliers and influential chemicals are highlighted (see text). The dotted lines indicate the 3σ interval.

the atomic distribution and the molecular dimension in the studied chemicals.

When considering the population of 40 models (no. variables = 4), the range of Q^2 is 80.3% to 81.9% and the range of $Q^2_{LMO}(50\%)$ is 79.8% to 81.4%, while the range of Q^2_{EXT} is 76.8% to 83.3% (Q^2_{EXT} mean = 80.7%) for splitting with the D-optimal algorithm; the range of Q^2 is 80.1% to 82.1% and the range of $Q^2_{LMO}(50\%)$ is 79.4% to 81.5%, while the range of Q^2_{EXT} is 74.4% to 81.3% (Q^2_{EXT} mean = 78.7%) for splitting with K-ANN. The Q^2_{EXT} values confirm the high predictive ability of all the QSAR models in the population selected by GA-VSS; these models are also reliable for use on new chemicals not employed in the developed model, with the condition of belonging to the same chemical domain. Comparison of the Q^2_{EXT} values for the two splitting methodologies indicates that the response is well predicted in all cases, and the difference between these methodologies for this data set is not big, even though the range of external predictivity is always more optimistic in D-optimal splitting.

Consensus Modeling. As mentioned above, the application of the Genetic Algorithm-Variable Subset Selection (GA-VSS) procedure to capture the most relevant variables in modeling response provides a large set of possible models with nearly equivalent predictive performance. The models employ a variety of descriptors, so that they reflect different aspects of molecular structure. An individual QSAR model may overemphasize some aspects, underestimate others, and ignore many important features completely. Thus it seems reasonable that a consensus QSAR model, which can be derived by calculating an average result for representative individual models, might provide better predictive ability than the majority of individual models. This consensus modeling might take into account several peculiar aspects of some particular structures contemporaneously. The availability of several possible models, equally reliable for response prediction, highlights the need for methods able to preserve both model quality and diversity for model comparison.[50−53]

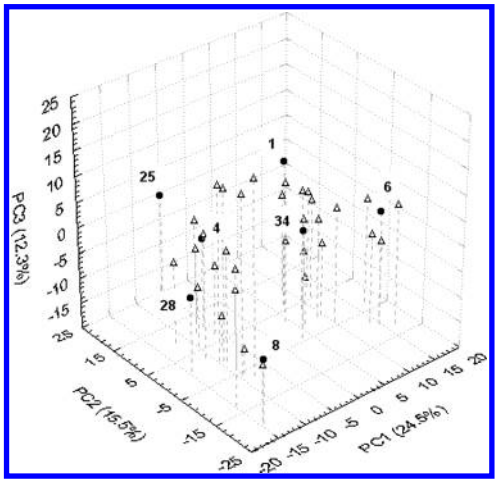

Figure 4. Principal Component Analysis on model residuals (PC1−PC2−PC3: EV% = 52.3%). The models used for consensus modeling are highlighted differently (●).

For the comparison of different QSAR models it is useful to examine their variability in predicting responses.[52] An intuitive comparison emerges from a loading plot derived by applying PCA to model residuals of 40 models. In this graphical representation (Figure 4) the loadings associated with the first three principal components, accounting for 52.3% of residual variance, are reported. Significantly different models (e.g. with dissimilar residual profiles) are distant in this graph, while similar models (e.g. giving similar predicted values) are clustered and redundant. On the basis of different structural descriptions an average/consensus model can be derived from dispersed models corresponding to different prediction schemes.

Table 1.

ID mod.	no. var.	obj. tr.	obj. test	descriptors	Q^2	R^2	SDEP	SDEC	$Q^2_{(LMO)50\%}$	Q^2_{EXT}
1	4	234	226	Homo; BELm2; nCaH; nX	82.1	83.2	0.430	0.416	81.5	80.2
4	4	234	226	Homo; nX; CIC0; nCaH	81.6	82.8	0.436	0.422	81.0	81.3
6	4	234	226	Homo; BELm2; nCaH; ATS2m	81.5	82.5	0.437	0.426	81.0	77.9
8	4	234	226	Homo; Me; Dm; nCaH	81.3	82.3	0.439	0.427	80.8	78.1
25	4	234	226	Homo; nX; nCaH; nC	80.5	81.7	0.449	0.435	79.8	80.9
28	4	234	226	Homo; AMW; nCaH; BEHp1	80.4	81.5	0.451	0.438	79.8	78.9
34	4	234	226	Homo; BELm2; AMW; nBnz	80.2	81.3	0.453	0.439	79.6	79.9
		234	226	consensus model		84.4		0.401		82.1

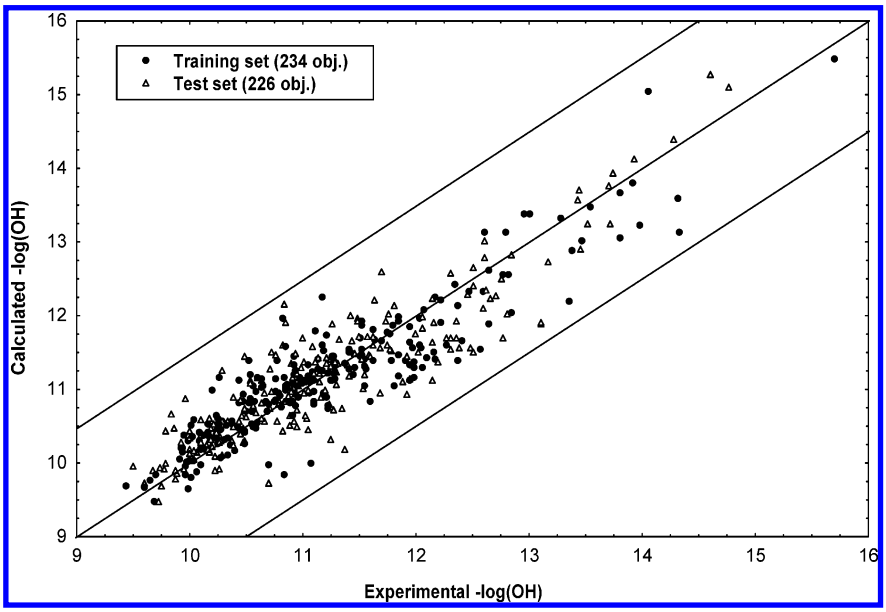

Figure 5. Regression line of the consensus model (K-ANN splitting). The log k(OH) values for the training and test set chemicals are labeled differently. The dotted lines indicate the 3σ interval.

A consensus prediction for the OH degradation rate constant is calculated by averaging the predicted values from seven individual models (numbered in Figure 4 according to the predictivity order in model population, listed in Table 1); this is an arbitrary choice, and other choices are possible. A new methodology for the selection of the most dissimilar models, based on the Hamming distance, was recently proposed.[53]

The population of models selected for this approach is that obtained on the K-ANN split training set; this is because the external predictivity of such derived models is more reasonable than that of models derived from the D-optimal split, where the models are too optimistic in their external predictivity. The models selected as being the more representative for the consensus modeling are those which are the most dissimilar in the PCA graph, the model previously selected as the best (no. 4) can be considered a sort of "intermediate" model in the PC1/PC2 space of Figure 4. It can also be verified immediately that the difference in the residuals derives, as expected, from the difference in the molecular descriptors: indeed, the selected models are the most dissimilar in molecular descriptor composition, although this population of models is characterized by a relatively homogeneous set of descriptors.

In addition to the descriptors of the best models, reported above (HOMO, nX, CIC0, and nCaH) and present in most of the models, the descriptors selected in the consensus modeling are those most abundant in the GA-population and thus the most relevant in modeling the response, in their alternative combinations: the number of benzene-like rings (nBnz); the average molecular weight (AMW); the number of carbon atoms (nC); the mean of atomic Sanderson electronegativity scaled on carbon atom (Me); the 3D-WHIM total accessibility index (Dm),[22] the 2D-Broto-Moreau autocorrelation of a topological structure (ATS2m),[54] and the 2D-BCUTs descriptors[55] (BELm2, weighted by atomic masses and BEHp1, weighted by atomic polarizabilities).

Comparison of the statistical parameters calculated for the consensus model with those obtained from 40 individual models provides evidence of the superiority of consensus modeling. Fitting ability and predictive ability are better than any individual model ($R^2 = 84.4\%$; $Q^2_{EXT} = 82.1\%$ on the test set selected by the K-ANN design.)

In Figure 5 the regression line plotting the predicted values by the consensus model and the experimental data is reported: no outliers are present neither in the training nor in the test set.

To verify the chemical domain of the new consensus model and the distribution of the studied chemicals in this new multidimensional space, the chemicals are plotted in a Principal Components graph, obtained by PCA of the selected molecular descriptors (Figure 6).

From this PCA plot it is evident that the chemicals with smaller leverage value (less influential) are more grouped in the central part of the graph, while the more influential chemicals (with higher leverage value) are more isolated and mainly at the border of the molecular descriptors domain.

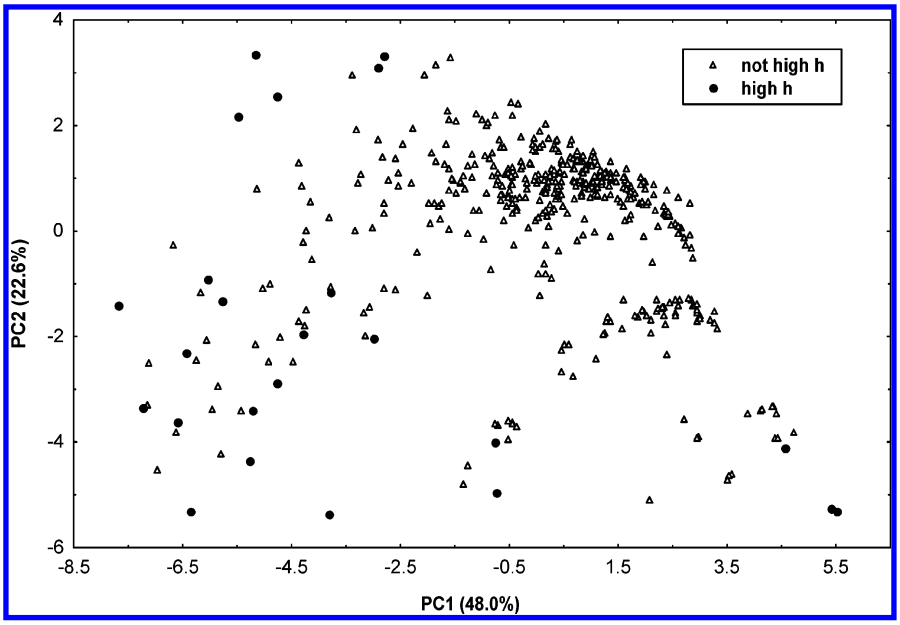

Figure 6. Principal Component Analysis on the selected molecular descriptors for consensus modeling (PC1−PC2: EV% = 70.6%). Influential chemicals with higher leverage (h) values are highlighted differently (●).

It is also interesting to note that all the models selected for the consensus modeling predicted some chemicals with practically the same rate constant, while for other chemicals the rate constants were predicted differently by each selected model (see the Δ value among the predictions of the selected models in Table S1 in the Supporting Information, range of Δ 0.05−1.45). The chemicals in the former case can be defined as "prediction safe", while those in the latter can be defined as "prediction sensitive", being selectively related to the structural information included in different models. The range of predicted values among the models can be used to highlight "prediction sensitive" chemicals and therefore those of greatest concern (higher Δ)). If the rate constants are detected experimentally, "prediction sensitive" compounds could be useful in selecting the best among several possible models, and in interpreting the model's molecular descriptors in terms of the particular structural features characterizing "prediction sensitive" compounds.

CONCLUSIONS

In summary, multivariate models have been proposed to predict the degradation rate constant by the hydroxyl radical of a wide set of heterogeneous chemicals. The models are developed by Genetic Algorithm selection of theoretical molecular descriptors from among a wide set of calculated descriptors. The advantage of the theoretical descriptors is that they can be homogeneously calculated for all the chemicals by a defined software, even before their synthesis. The proposed models have good stability, robustness, and predictivity when verified by internal validation (cross-validation by LOO and LMO) and also external validation. Two methodologies are used to split the original chemical data set into training and test sets, the statistical D-optimal Experimental Design and the Kohonen Artificial Neural Network (K-ANN), and the consequences on the derived model predictivity are compared. D-optimal design, where the most dissimilar chemicals are always selected for the training set, leads to models with more optimistic predictive performance than models developed on the training set selected by K-ANN. The chemical applicability domain of the models and the reliability of the predictions are always verified by the leverage approach. A selection of the most dissimilar models in the population of GA-models allows the proposal of consensus predictions and the highlighting of "prediction safe" or "prediction sensitive" chemicals depending on their independence, or not, of the model choice.

ACKNOWLEDGMENT

We wish to thank the University of Insubria for a fellowship granted to Pamela Pilutti.

Supporting Information Available: Information on data set, experimental and predicted data, training and test set chemicals, and descriptor values. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Medven, Z.; Güsten, H.; Sábljic, A. Comparative QSAR Study on Hydroxyl Radical Reactivity with Unsaturated Hydrocarbons: PLS Versus MLR. *J. Chemom.* **1996**, *10*, 135−147.

(2) Atkinson, R. Atmospheric Chemistry of VOCs and NO_x. *Atmos. Environ.* **2000**, *34*, 2063-2101.

(3) Meylan, W. M.; Howard P. H. A Review of Quantitative Structure−Activity Relationship Methods for the Prediction of Atmospheric Oxidation of Organic Chemicals.*Environ. Toxicol. Chem.* **2003**, *22* (8), 1724−1732.

(4) Atkinson, R. A Structure−Activity Relationship for the Estimation of Rate Constants for the Gas-Phase Reactions of OH Radicals with Organic Compounds. *Int. J. Chem. Kinet.* **1987**, *19*, 799−828.

(5) Sabljic, A.; Güsten, H. Predicting the Night-Time NO_3 Radical Reactivity in the Troposphere. *Atmos. Environ.* **1990**, *24A*, 73−78.

(6) Müller, M.; Klein, W. Estimating Atmospheric Degradation Processes by SARs. *Sci. Total Environ.* **1991**, *109*, 261−273.

(7) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Oxygenated Compounds Based on Molecular Orbital Calculations. *Chemosphere* **1996**, *32* (4), 717−726.

(8) Bakken, G.; Jurs, P. Prediction of Hydroxyl Radical Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1064−1075.

(9) Gramatica, P.; Consonni, V.; Todeschini, R. QSAR Study on the Tropospheric Degradation of Organic Compounds. *Chemosphere* **1999**, *38*, 1371−1378.

(10) Güsten, H. Predicting the Abiotic Degradability of Organic Pollutants in the Troposphere. *Chemosphere* **1999**, *38* (6), 1361−1370.

(11) Pompe, M.; Veber, M. Prediction of Rate Constants for the Reaction of O_3 with Different Organic Compounds. *Atmos. Environ*. **2001**, *35*, 3781−3788.
(12) Gramatica, P.; Pilutti, P.; Papa, E. Ranking of Volatile Organic Compounds for Tropospheric Degradability by Oxidants: a QSPR Approach. *SAR QSAR Environ. Res.* **2002**, *13* (7−8), 743−753.
(13) Gramatica, P.; Pilutti, P.; Papa, E. QSAR prediction of ozone tropospheric degradation. *QSAR Comb. Sci.* **2003**, *22*, 364−373.
(14) Gramatica, P.; Pilutti, P.; Papa, E. Predicting the NO_3 Tropospheric Degradability of Organic Pollutants by Theoretical Molecular Descriptors. *Atmos. Environ.* **2003**, *37* (22), 3115−3124.
(15) *AOPWIN*, Ver.1.90. Environmental Protection Agency, U.S.A., 2000.
(16) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−76.
(17) Schultz, T. W.; Netzeva, T. I.; Cronin M. T. D. Selection of Data Sets for QSARs: Analyses of *Tetrahymena* Toxicity from Aromatic Compounds. *SAR QSAR Environ. Res.* **2003**, *14* (1), 59−81.
(18) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational Selection of Training Sets for the Development of Validated QSAR Models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241−253.
(19) Atkinson, R. Kinetics and Mechanisms of the Gas-Phase Reactions of the Hydroxyl Radical with Organic Compounds. *J. Phys. Ref. Data* **1989**, Monograph 1.
(20) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON − Software for the calculation of molecular descriptors*. Version 5.0 for Windows, Talete srl, Milan, Italy, 2004.
(21) *HyperChem*. Release 7.03 for Windows, Molecular Modeling System. Hypercube, Inc., Gainesville, Florida, U.S.A., 2002.
(22) Todeschini, R.; Gramatica, P. The WHIM Theory: New 3D Molecular Descriptors for QSAR in Environmental Modelling. *SAR QSAR Environ. Res.* **1997**, *7*, 89−115.
(23) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlation and Similarity/Diversity Analysis by GETAWAY Descriptors. Part 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693−705.
(24) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(25) *SCAN − Software for Chemometric Analysis*. Rel. 1.1 for Windows, Minitab, U.S.A., 1995.
(26) *STATISTICA*. Rel. 5.1 for Windows, StatSoft, Inc., U.S.A., 1987.
(27) Todeschini, R.; Consonni, V.; Pavan, M. *MOBY DIGS − Software for multilinear regression analysis and variable subset selection by genetic algorithm*. Ver. 1.2 for Windows, Talete srl, Milan, Italy, 2002.
(28) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.
(29) Todeschini, R.; Maiocchi, A.; Consonni, V. The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom. Int. Lab. Syst.* **1999**, *46*, 13−29.
(30) Atkinson, A. C. *Plots, Transformations and Regression;* Clarendon Press: Oxford, 1985.
(31) Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.; McDowell, R. M.; Gramatica, P. Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Regression Based and Classification QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361−1375.
(32) Todeschini, R.; Mauri, A. *DOLPHIN−Software for optimal distance-based experimental design.* Ver. 2.0 for Windows, Talete srl, Milan, Italy, 2002.
(33) Marengo, E.; Todeschini, R. A New Algorithm for Optimal Distance − Based Experimental Design. *Chemom. Int. Lab. Syst.* **1992**, *16*, 37−44.
(34) Sjostrom, M.; Eriksson, L. *Applications of Statistical Experimental Design, in Chemometric Methods in Molecular Design;* Van de Warerbeemd H., VCH: 1995; Vol. 2, pp 63−90.
(35) Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. Tutorial. *Chemom. Int. Lab. Syst.* **1996**, *34*, 1−19.
(36) Carlson, R. *Design and Optimization in Organic Synthesis*; Elsevier: 1992.
(37) Todeschini, R.; Consonni, V. *KOALA-Software for Kohonen Artificial Neural Networks.* Rel. 1.0 for Windows, Talete srl, Milan, Italy, 2001.
(38) Zupan, J.; Novic, M.; Ruisánchez, I. Kohonen and Counter propagation Artificial Neural Networks in Analytical Chemistry. *Chemom. Int. Lab. Syst.* **1997**, *38*, 1−23.
(39) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.
(40) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.
(41) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.
(42) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(43) Wold, S.; Eriksson, L. *Chemometric Methods in Molecular Design*; VCH: Germany, 1995.
(44) Pötter, T.; Matter, H. Random or Rational Design? Evolution of Diverse Compound Subsets from Chemical Structure Data Set. *J. Med. Chem.* **1998**, *41*, 478−488.
(45) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.
(46) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Regularized Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.
(47) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training Set Selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.
(48) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; RSP-Wiley: Chichester, U.K., 1983.
(49) Magnuson, V. R.; Harriss, D. K.; Basak, S. C. *Studies in Physical and Theoretical Chemistry*; King, R. B., Ed.; Elsevier: Amsterdam, The Netherlands, 1983; pp 178−191.
(50) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256−1262.
(51) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.
(52) Sutherland, J. J.; Weaver, D. F. Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1028−1036.
(53) Todeschini, R.; Consonni, V.; Pavan, M. A Distance Measure Between Models, a Tool for Similarity/Diversity Analysis of Models Population. *Chemom. Int. Lab. Syst.* **2004**, *70*, 55−61.
(54) Broto, P.; Devillers, J. *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: London, U.K., 1990.
(55) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.

CI049923U