Use of Structure Descriptors To Discriminate between Modes of Toxic Action of Phenols

Simon Spycher, Eric Pellegrini, and Johann Gasteiger*

Computer-Chemie-Centrum and Institute of Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received July 1, 2004

Two classification models were developed based on a data set of 220 phenols with four associated Modes of Toxic Action (MOA). Counter-propagation neural networks (CPG NN) and multinomial logistic regression (multinom) were used as classification methods. The combination of topological autocorrelation of empirical π-charge and σ-electronegativity and of surface autocorrelation of hydrogen-bonding potential resulted in a 21-dimensional model that successfully discriminated between the four MOAs. Its overall predictive power was estimated to 92% using 5-fold cross-validation. Subsequently, a simple score for the distance to the training data was used to determine the prediction space of the model and used in an exploratory study on the phenols contained in the open NCI database. The use of a prediction space metric proved indispensable for the screening of such a diverse database. The prediction space covered by the proposed model is still of rather local nature which is either caused by the limited diversity and size of the training set or by the high dimensionality of the descriptors.

1. INTRODUCTION

Toxicity may be one of the most difficult properties to model, especially for high-throughput screening in the drug discovery process,[1] but also for environmental risk assessment of industrial chemicals. The difficulties for developing effective in silico models are caused by two factors. First, the abundance of possible endpoints that will have to be modeled (ranging from in vivo data to specific enzyme tests or microarrays). Second, the scarcity of consistent public data that prevents one from building data sets that span a wide range of the chemical space.

Several approaches are currently applied in order to make quantitative predictions: local models based on sets of clearly defined congeners,[2] global models using flexible machine learning techniques,[3] and models based on common mode of toxic action (MOA). A MOA can be defined as a common set of physiological and behavioral signs characterizing a type of adverse biological response, while a toxic mechanism is defined as the chemical process underlying a given MOA.[4]

If one wants to make predictions for a new compound, a central question for each approach is which quantitative model to apply. For models based on congeners the answer seems rather simple, but for MOA-based models a successful classification into the different MOAs is the prerequisite for successful quantitative prediction. The well-known MOA-based ASTER system (ASsessment Tool for Evaluating Risk) of the US EPA[5] selects the appropriate QSAR for a new compound on the basis of the occurrence of chemical fragments. The limitation of this fragment-based method is to reduce a chemical structure to a specified substructure and ignore the other topological and potential electronic features of the entire compound which may influence its MOA.[6] With the aim to overcome this limitation, several MOA classification studies were published in the last few years.[7−10]

The quantitative prediction and also the classification of reactive MOAs requires descriptors reflecting their reactive nature,[11] which in most cases requires quantum chemical methods. The most frequently used descriptors comprise atomic charges, molecular orbital energies, and superdelocalizabilities,[12] which are usually calculated with semiempirical methods. The calculation of such descriptors is still slow which hampers the screening of large databases. We have developed a set of empirical descriptors that model reactivity[13,14] and therefore should also be applicable to model toxicity involving reactive chemicals.

The MOA data set of Schüürmann et al. was used[10] which is based on the original publication of Aptula et al.[9] but has some small changes and additional descriptors. It contains 220 phenols classified into four MOAs. Several classification models based on various statistical methods had been derived for this data set using quantum mechanical based descriptors and additional whole molecule descriptors.[9,10,15,16]

In the present study, classification models were derived based on empirical atomic physicochemical descriptors encoded by topological autocorrelation, *A* and surface autocorrelation vectors to structure-based vectorial descriptors. These descriptors from now on will be referred to as structure descriptors. Counter-propagation neural networks (CPG NN) and penalized multinomial logistic regression (multinom) were used to build the classification models. This study has the following objectives:

- Development of MOA classifiers based on empirical descriptors.
- Comparison of the predictive power of structure descriptors with descriptors used in earlier studies.
- Application of the models to a potential screening data set.

* Corresponding author phone: (49)-9131-85-26570; fax: (49)-9131-85-26566; e-mail: Gasteiger@chemie.uni-erlangen.de.

10.1021/ci0497915 CCC: $30.25
Published on Web 12/24/2004

2. MATERIALS AND METHODS

2.1. The Data Sets. 2.1.1. Training Set. The MOAs of the 220 phenols[10] were assigned following structural rules developed earlier using growth inhibition assays with *Tetrahymena pyriformis*.[17] First, a short characterization of each MOA and the corresponding structural rules proposed by Schultz et al.[17] will be given.

Uncouplers of oxidative phosphorylation inhibit the coupling between the electron transport and the phosphorylation reactions which take place in mitochondrial membranes. Thus, they inhibit ATP synthesis in energy producing membranes.[18] Uncouplers are hydrophobic weak acids with the ability to transport protons through proton-impermeable membranes (some uncouplers which are not weak acids are reported, but their mechanism is not clear[18]). Schüürmann et al. observed uncoupling to be limited to compounds with pK_a-values between 3.8 and 8.5.[19] The structural rule suggested for uncouplers was as follows: more than one nitro group, or more than one cyano group, or more than three halogen groups, or a single nitro group and more than one halogen group.

Precursors to soft electrophiles (proelectrophiles) are toxicants whose action is based upon their metabolic transformation to an electrophile.[20,21] The transformation takes place by oxidation in a one- or two-electron process. The products have several possible targets.[21] The following rule was suggested for proelectrophiles: two or more hydroxy groups in ortho or para position and at least one unsubstituted aromatic carbon atom or an amino group in ortho or para position to the hydroxy group and at least one unsubstituted aromatic carbon group.

Soft electrophiles alkylate essential protein thiol or amino groups or produce oxidative stress by free radical formation.[9] Suggested mechanisms for the reactivity are either direct reaction with biological nucleophiles or previous reduction of nitro groups to highly reactive nitroso groups.[22,23] The suggested rule was a single nitro group but not more than one halogen atom.

By default, phenols not assigned to any of these three MOAs were assigned to the MOA polar narcotic. The underlying mechanism and the target of polar narcosis is still a matter of debate. It is either caused by the presence of the toxicants in lipid components of the membrane or involves membrane-bound protein target-sites.[24]

To summarize, one should note that there are considerable differences in the current molecular understanding of the four MOAs. While there is a clear understanding of both the mechanism and the target of uncouplers,[25] there is much less understanding of proelectrophiles and soft electrophiles. For these two MOAs the abundance of possible targets causes problems for both classification and quantitative prediction (e.g., correlation of toxicity with the reaction rates with specific nucleophiles such as GSH[26]).

In this study, two compounds of the training set were assigned to a different MOA than in ref 10. 3-Aminophenol and 5-amino-2-methoxyphenol were reassigned as polar narcotics (both previously proelectrophiles, but the amino group is in meta not in para or ortho position). Additionally, 4-acetamidophenol (proelectrophile) and 4-nitrosophenol (soft electrophile) are not covered by the structural rules for their MOAs but were left in their original MOA because of their high excess toxicity.

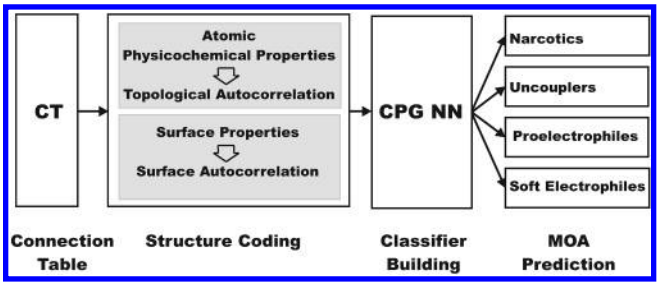

Figure 1. Workflow used for the development of MOA classifiers suitable to screen large structure databases.

Thus, the training data set used in this study contains 220 compounds (155 polar narcotics, 19 uncouplers, 24 proelectrophiles, and 22 soft electrophiles).

2.1.2. Test Sets. Once the final model based on the 220 phenols was established, two additional data sets were used as external tests sets. The first MOA test set was extracted from a publication using the same MOA assignment rules based on *Tetrahymena pyriformis* assays.[27] After removal of the duplicates and compounds with other MOA not covered by the training set a data set containing 25 compounds was obtained. This data set consisted of 17 polar narcotics, 1 uncoupler, and 7 soft electrophiles. A second test set was used for an exploratory study. It consisted of all 3142 monocyclic phenols found among the 248 259 compounds in the open NCI database (release 1, 1999). The subset was cleaned of 261 compounds that were either salts or contained elements other than C,H,O,N,S,P,X (X = halogens), of 9 compounds which could not be processed by the PETRA package,[14] of 13 compounds which match the structural rules of pro-redox cyclers (as defined in ref 17), 263 internal duplicates and 182 compounds contained already in the training set. The final set then comprised 2414 phenols for which the MOA was assigned according to the rules described above. This resulted in a data set of 1770 narcotics, 73 uncouplers, 429 proelectrophiles, and 142 soft electrophiles. The assumption behind this exploratory study is that the rules are valid even for this very diverse data set (with limitations given in the Discussion).

The prediction of several thousands of compounds requires methods which are fast and easy to parametrize and have a high predictive power. A simple scheme of the workflow used to achieve this goal is presented in Figure 1.

The methods are explained in detail in the next three sections.

2.2. Calculation of Atomic Physicochemical Properties. In the present work, atomic physicochemical properties were calculated using PETRA.[14] This program package comprises various empirical methods for the calculation of physicochemical properties in organic molecules. A wide panel of effects is covered ranging from atomic properties (e.g., total, σ- and π-charge distribution, atomic polarizability, lone-pair, σ- and π-electronegativities) to molecular properties (e.g., heat of formation, molecular polarizability).

2.3. Encoding of Atomic Physicochemical Properties. The molecules atomic physicochemical properties were encoded using topological and surface autocorrelation vectors, *A*. They are both vectorial descriptors that integrate a physicochemical property inside an atomic distance-based

function that is sampled on a fixed number of components. Therefore, the number of vector elements is independent of the size of the molecule.

An *A* vector can be expressed in various ways such as topological, 3D, or surface autocorrelation vectors[30−34] depending on the kind of distance that is used for their computation. On a general point of view, they are expressed through the following formula:

$$A(r) = \sum_{i=1}^{N}\sum_{j=i}^{N} p_i p_j \; f(r - r_{ij}) \qquad (1)$$

Depending on the kind of *A* vector, the meaning of the different variables of eq 1 changes a little bit. Indeed, in the case of a topological or 3D distance *A* vector, r_{ij} is the topological or 3D distance between atoms *i* and *j*, and p_i and p_j are an atomic physicochemical property of atoms *i* and *j*. In the case of a *surface A* vector, r_{ij} is the distance between two surface points *i* and *j*, and p_i and p_j are physicochemical properties of points *i* and *j* calculated on the molecular surface. Several kinds of properties can be calculated on the surface of a molecule such as electrostatic potential or hydrophobic and H-bonding potentials. In the present study, the hydrogen-bonding potential (HBP) was used according to a force field formulation of this potential.[35] The nature of the functional form *f* also changes regarding the *A* vector under consideration. In the case of a topological *A* vector, it is a Dirac function that, for each sampled topological distance *r*, considers only the topological distances r_{ij} strictly equal to *r*. For 3D distance *A* or *surface A* vectors, it is a combination of Heavyside functions that, for each sampled real distances *r*, considers a range of distances over *r*. In the present work, topological *A* vectors with various properties *p* and *surface A* vector were used. Hydrogen atoms were excluded for topological autocorrelation vectors. Autocorrelation vectors were always sampled over five topological distances for *A* vector, and nine 3D distance intervals from 1 to 7.75 Å for *surface A* vector, respectively.

Single 3D conformations of the 220 phenols were generated using CORINA.[36] The *A* vectors were calculated with AUTOCORR[37] and the *surface A* vectors with SURFACE.[38] Depending on the atomic properties calculated the standard deviations differed by several orders of magnitude. Therefore each vector element was autoscaled, i.e., mean centered and scaled to unit variance.

2.4. Classification Methods. 2.4.1. Counter-Propagation Neural Networks. Counter-propagation neural networks (CPG NN) are an extension of Kohonen Self- Organizing Maps (SOM) where one or more output layers are added to the Kohonen input layer.[39] These types of neural networks have been applied for both classification studies[40] and for the prediction of continuous values (e.g. IR spectra, H NMR spectra[41−42]). The architecture of CPG NN is illustrated in Figure 2.

The input layers represent the descriptors (X-variables), while the output layers represent the properties to predict (Y-variables). The Y-variables are coded as a matrix of dummy vectors with 1 for actives and 0 for nonactives, i.e., a compound *i* active as uncoupler is coded as $Y_i = (0, 1, 0, 0)$. During training, the weights of both input and output layers are adapted, but only the input layers are used to determine the winning neuron. Thus, the unsupervised character of the SOM is preserved, i.e., the X-variables are organized independently of the Y-variable and therefore cannot be fitted to nonexistent connections between X-variables (descriptors) and Y-variables just by adding additional descriptors. At the end of training, the weights obtained for the output layers are used as the predicted values for the properties under consideration. At this point the architecture shown in Figure 2 becomes clear: each layer of the output block corresponds to one MOA. This allows for studying Kohonen maps of active versus nonactive compounds for each MOA. Predictions for new compounds are made by determining the winning neuron, defined as the neuron with the smallest Euclidian distance between its weight vector and the X-variables of the compound. A rectangular network topology was used, with a size of 11*11 neurons and a training time of 50 epochs. CPG NNs were calculated using SONNIA.[43]

2.4.2. Logistic Regression. Logistic regression is a type of regression analysis where the dependent variable *Y* is a dummy variable (coded 0, 1). A logistic regression model solves the equation

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p \qquad (2)$$

where π is the probability that the event $Y = 1$ occurs, the β are the coefficients of the regression model with p descriptors, *x*. Equation 2 is the so-called logit function. A compound was assigned to the MOA with the highest predicted probability. The Maximum Likelihood Estimation method (MLE) is used in order to derive the regression coefficients, β_i. To reduce overfitting, a penalization term can be added to the Maximum Likelihood Estimation.[44−45] The penalization term has the following formula

$$\log(L^{\lambda}) = \log(L) - \frac{1}{2}\lambda\sum_{i=1}^{p}(s_i\beta_i)^2 \qquad (3)$$

where *L* is the usual likelihood function, λ a predefined penalty, s_i is a scale factor chosen to make $s_i\beta_i$ unitless. In practice, appropriate values for λ can vary in the range of 0.001 to 0.1.[46] In the present study, a value of $\lambda = 0.1$ was chosen as it proved suitable in an earlier study with a data set that was prone to overfitting.[47] Logistic regression can be extended from the basic two classes to a model with *k* classes by using *k*-1 logit functions. The likelihood function used to determine the coefficients of the multinomial regression function β_{ik} is then determined from the summation including all logit functions. The Y-variables are coded as a dummy matrix as described in the subsection above. A detailed description of the method is given in ref 48.

Logistic regression models were calculated using R,[49] with the VR-add-on package containing the extension to multinomial logistic regression models.

2.5. Model Validation. *k*-fold cross-validation was used for model validation.[50] The data set was randomly divided into *k* subsets with equal percentages of each MOA present in each subset.[51] Then a model was fitted taking $k - 1$ of these subsets as a training set and the remaining one as a test set. The procedure is repeated *k* times until each subset

has been used once as a test set. Usually, k takes values between 5 and 10.[46] Due to the random splitting the estimates may vary; therefore, the whole procedure was repeated B times (recommended values vary between 5 and 100). In this study, k was set to five, and B to 19 leading to 95 model fitting runs. The confusion matrix of the average cross-validated predictions was computed at the end of the procedure. From the confusion matrix the overall correct classification rate and sensitivities were derived. Sensitivity is obtained by dividing the number of correctly predicted compounds of a given MOA with the total number of compounds of this MOA.

2.6. Determination of Prediction Space. To determine the space covered by our models, we used a procedure based on the Hotelling's T^2 statistic. First, a PCA was performed on the training set. Then the loadings matrix of the *training* set was used to calculate the scores of the *external* set. Using these scores and a given confidence level α, a compound i of the external set was decided to belong to the prediction space if its Hotelling's score, T_i, satisfied the following criterion

$$T_i^2 < \frac{A(N^2-1)}{N(N-A)}F(p=\alpha) \qquad (4)$$

where $F(p=\alpha)$ is a tabulated value for a F distribution using a confidence level α, A is the number of principal components used to build the Hotelling's test, and N is the number of compounds of the training set and

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{s_a^2} \qquad (5)$$

where s_a^2 is the variance explained by principal component a and t_{ia} is the score of compound i for principal component a.[52] The goal of using this statistical test is to get a value for a compound's distance to the model. One can expect predictions to improve for compounds with small values of T^2 and to deteriorate for compounds with large values. If the T^2 of a given compound falls outside the limit— for a given confidence level—the compound is outside the confidence region, and no predictions for this compound should be made or only with caution.

As PCA is sensitive to outliers, strong outliers in the training data have a strong influence on position and extent of the prediction space. To reduce such effects the following measure was taken. In the first step, the 99% confidence region of the training data was determined. Compounds falling outside this region were then excluded, and the PCA loadings of the remaining training data were recalculated and used to determine the Hotelling's score as described above. This measure leads to a narrower definition of model space and hence to more compounds rejected from prediction but to more robust values for the score.

3. RESULTS

3.1. CPG NN—Models Based on Structure Descriptors.

3.1.1. Models Based on One Physicochemical Property. As an initial test, 6-dimensional topological A vectors were calculated with seven different atomic physicochemical properties. Seven CPG NN models were trained each with one type of property represented with A vectors. Table 1 shows the results of the 5-fold cross-validation for the eight models, with overall correct classification rates ranging from 70.9 to 84.1%.

Table 1. Predictive Power of Seven CPG NN Models Each Using a Different Type of Atomic Physicochemical Descriptor Represented with A Vectors[a]

	correct prediction (%)				
descriptors	overall	N	U	P	S
identity function, A1	70.9	93.6	57.9	0	0
partial charge, q_{tot}	74.6	94.1	31.6	33.3	18.2
π charge,[53] q_π	84.1	98.1	42.1	25.0	86.4
σ charge,[55] q_σ	76.4	94.8	63.2	37.5	0
σ electronegativity,[56] χ_σ	82.3	96.8	78.9	0	72.3
π electronegativity,[54] χ_π	84.1	100.0	68.4	0	77.3
polarizability,[56] α	70.9	97.4	26.3	0	0

[a] Abbreviations: polar narcotics (N), uncouplers of oxidative phosphorylation (U), proelectrophiles (P), soft electrophiles (S).

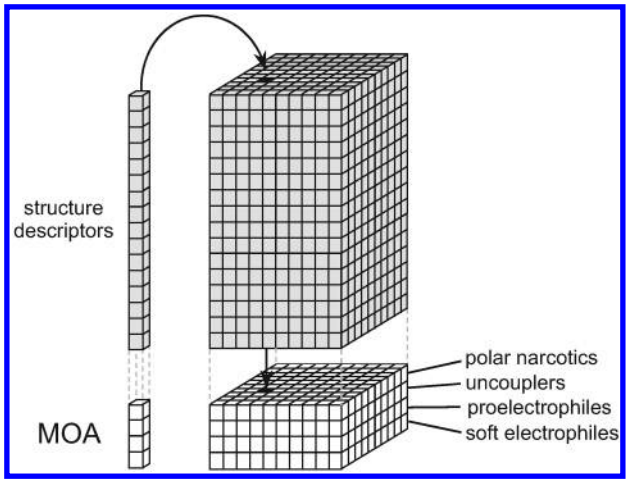

Figure 2. Architecture of the CPG NN for the classification of toxicants with four output layers, one for each MOA.

When the overall correct classification rates are judged the strong prevalence of the 155 polar narcotics needs to be reminded, i.e., a (worthless) classifier predicting all the 220 phenols to be narcotics would result in 70.5% correct predictions. The model based on polarizability, α seems pretty close to such a case. The sensitivities for the specific MOAs U, P, and S vary strongly from one descriptor to another. In the case of proelectrophiles, no descriptor is able to increase the sensitivity.

3.1.2. Models Based on an Additional Physicochemical Property. In a second step, an additional A vector was added to q_π- A, the best performing descriptor in terms of reaching a high sensitivity for one of the three specific MOAs (i.e. uncouplers, proelectrophiles and soft electrophiles). q_π- A could classify 86.4% of the soft electrophiles correctly, but additional descriptors were necessary for uncouplers and proelectrophiles. The descriptor with the second-highest sensitivity was therefore added, namely the χ_σ- A with a sensitivity for uncouplers of 78.9%. Figure 3 shows the four output layers of a network trained with q_π- A and χ_σ- A. It reveals a spread of the polar narcotics over the structural space, while the other MOAs tend to cluster to various extents. The strongest tendency to cluster could be observed for soft electrophiles and the least for proelectrophiles. The largest fraction of the conflicts is caused by proelectrophiles

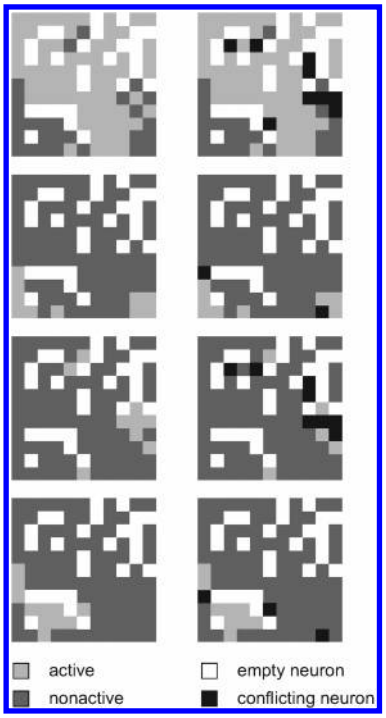

Figure 3. Projections of compounds represented with q_π- and χ_σ- *A* vectors into four maps, corresponding to the four MOA polar narcotics, uncouplers, proelectrophiles, and soft electrophiles (top down). Active molecules in light gray, nonactives in dark gray (colored according to the most frequent output). The maps on the right-hand side additionally show conflicting neurons (black), i.e., neurons containing compounds from different MOAs. White squares indicate empty neurons.

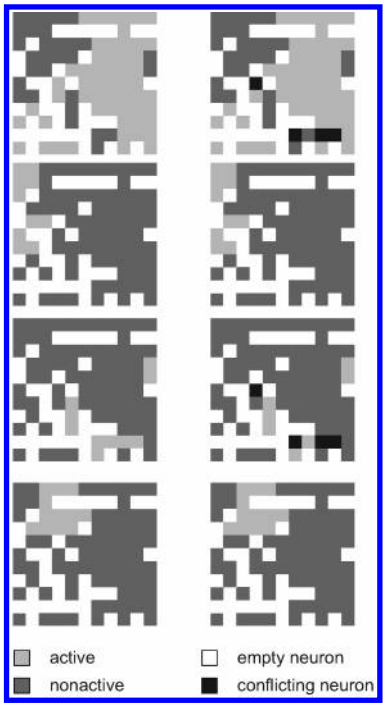

Figure 4. Projections of compounds represented with q_π- *A*, χ_σ- *A*, and *surface A* vectors into four maps, corresponding to the MOAs 1−4, respectively from top to down, polar narcotics, uncouplers, proelectrophiles, and soft electrophiles. Active molecules in light gray, nonactives in dark gray (colored according to the most frequent output). The right-hand side maps are the same than the left-hand side ones but with the conflicting neurons displayed (black). White squares indicate empty neurons.

conflicting with polar narcotics. A typical conflict was between catechol and 2-methoxyphenol. Apparently, for an oxygen or a nitrogen atom, the chosen descriptor set could not retrieve if the atom was a hydrogen bond donor or not.

The estimate of overall predictive power increased only slightly through the addition of χ_σ- *A* to 87.3%. This was due to an increase in sensitivity for uncouplers, which increased from 42.1% (q_π alone) to 68.4% (q_π combined with χ_σ), while sensitivities for proelectrophiles remained low with 29.2%. Thus, the next descriptors should increase sensitivity for this MOA. Analysis of the misclassification of the cross-validated predictions confirmed the pattern observed in the maps: proelectrophiles were predominantly misclassified as polar narcotics (all of the 17 misclassified proelectrophiles) and polar narcotics as proelectrophiles (2 of 3 misclassifications).

3.1.3. Improvement of Proelectrophile Classification by Adding Surface Autocorrelation Descriptors. The problem in discriminating proelectrophiles from polar narcotics had been encountered in previous studies[9,15] as well, and it was the reason that the number of H-bond-donors, N_{Hdon}, was introduced as a descriptor. To keep the concept of hydrogen bond, but to introduce structural information, 9-dimensional *surface A* vectors of hydrogen bonding potential (*surface A*) were calculated. CPG NN models were calculated with a combination of q_π- *A*, χ_σ- *A*, and *surface A*. Figure 4 shows the resulting maps, one for each output layer of the CPG NN, respectively for each MOA.

The most striking difference to Figure 3 is the decrease in conflicts. Uncouplers (row 2) and soft electrophiles (row 4) now form almost perfect clusters with no conflicting neurons. Interestingly, even with a very low number of conflicts the proelectrophiles did not form a distinct cluster. Thus, the CPG NN seemed to model different subgroups of this MOA. A confusion matrix with cross-validated predictions of the resulting models is shown in Table 2.

Table 2. Confusion Matrix of 5-Fold Cross-Validated Predictions of a CPG NN with q_π- *A* and χ_σ- *A* Vectors and *Surface A*

	Act N	Act U	Act P	Act S
Pred N	152	2	5	2
Pred U	1	13	0	1
Pred P	2	0	19	0
Pred S	0	4	0	19
sensitivity	98.1	68.4	79.2	86.4
overall prediction				92.3

The addition of *surface A* dramatically increased the sensitivity for proelectrophiles and also improved the sensitivity for polar narcotics. The overall correct classification rate increased to 92.3%. This combination of two six-dimensional *A* vectors and one nine-dimensional *surface A* vector represents the final 21-dimensional model which was subsequently used in this study.

3.2. Multinomial Logistic Regression-Models Based on Structure Descriptors. The descriptors chosen during the CPG NN model building were also tested with multinom as an alternative classification method. Model parameters used are described in the methods section. When the whole data set was used for fitting the model a correct classification rate of 98.1% was calculated (apparent correct classification), compared to 92.7% estimated with cross-validation. With 21 descriptors for 220 compounds (respectively 176 during 5-fold cross-validation) the number of observations per

Table 3. Confusion Matrix of 5-Fold Cross-Validated Predictions of a Multinomial Logistic Regression Model with q_{π}- *A* and χ_{σ}- *A* Vectors and *Surface A*

	Act N	Act U	Act P	Act S
Pred N	150	0	5	1
Pred U	1	18	0	4
Pred P	4	0	19	0
Pred S	0	1	0	17
sensitivity	96.8	94.7	79.2	77.3
overall prediction				92.7

descriptor is at the lower end, and thus, overfitting can become a problem. However, the difference of 5.4 between apparent and cross-validated correct classification rate is not much higher than the differences obtained using the six descriptors of Schüürmann et al. (cf. subsection 3.3). The confusion matrix with cross-validated predictions of the resulting multinom models is shown in Table 3.

Like in the CPG NN models almost all misclassifications occur either between polar narcotics and proelectrophiles or uncouplers and soft electrophiles. The comparison with Table 2 shows that the difference to the CPG NN model is within this pattern, i.e., more uncouplers are predicted correctly, however, more soft electrophiles are predicted as uncouplers.

3.3. Models Based on Previously Published Descriptors. Models based on structure descriptors were compared to two sets of previously published descriptors. First, models based on descriptors published by Aptula et al. were evaluated.[9] The following five descriptors were used: octanol/water partition coefficient, log *P*; energy of the highest occupied molecular orbital, E_{HOMO}; energy of the lowest unoccupied molecular orbital, E_{LUMO}; negative logarithm of the ionization constant, pK_a; and the number of hydrogen-bond donors, N_{Hdon}. Cross-validation of the CPG NN model resulted in 86.8% correct classifications. The multinom model resulted in 87.7% correct classifications and 91.4% when the whole data set was used for fitting the model (apparent correct classification). Second, the following six descriptors calculated by Schüürmann et al. were evaluated:[10] log *P*, E_{HOMO}, E_{LUMO}, N_{Hdon}, highest electrophilic delocalizability of a carbon atom, D^{E}_{C-max}, and average carbon atom nucleophilic delocalizability, D^{N}_{C-av}. Cross-validation of the CPG NN model resulted in 91.4% correct classifications. The multinom resulted in 91.8% correct classifications with an apparent correct classification of 95.9%. Sensitivities of this multinom model amounted to N = 96.7%, U = 78.9%, P = 87.5%, S = 72.7%.

3.4. Predictions of External Data Sets. The multinom model using the full set of 21 descriptors was tested with the two external data sets described in the methods section. The two external data sets were compiled after the final choice of variables and model parameters was made. As a first step the predictions space spanned by the training set was determined as described in section 2.6. Figure 5 shows the training set scores for the first two principal components (PC) with the ellipse defined by eq 4 calculated for 2 PCs at a 95% confidence level.

The compounds show a good spread over space. However, some compounds with specific MOA are placed outside the prediction space at the chosen confidence level. Uncouplers are specially affected with all 7 dinitrophenols excluded. The figure also shows that the prediction space is defined in a

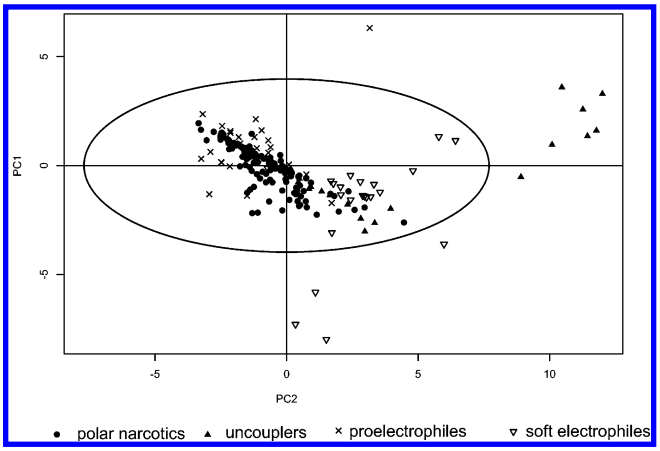

Figure 5. PCA-score plot of the training set with 21 descriptors. The ellipse defines the 95% confidence region.

Table 4. Overall Correct Classification Rate and Sensitivities with Increasingly Large Prediction Spaces[a]

threshold	overall	N	U	P	S
40	83.5 (508)	85.3 (442)	100.0 (1)	70.3 (64)	100.0 (1)
60	80.3 (640)	82.9 (549)	66.7 (3)	63.9 (83)	80.0 (5)
80	76.7 (803)	80.1 (663)	42.9 (7)	59.1 (110)	73.9 (23)
90	74.2 (938)	77.7 (773)	45.5 (11)	55.5 (119)	68.6 (35)
95	72.6 (1024)	76.3 (843)	41.7 (12)	52.7 (131)	68.4 (38)
99	69.0 (1208)	73.5 (989)	42.9 (14)	45.0 (160)	64.4 (45)
99.9	65.5 (1419)	71.1 (1130)	46.7 (15)	39.4 (216)	58.6 (58)
full set	56.2 (2414)	65.0 (1770)	37.0 (73)	30.1 (429)	34.5 (142)

[a] The number of included compounds is in parentheses.

global and rather conservative way. For the analysis of the external data sets a prediction space was defined using the first 5 PCs of the training set to calculate the Hotelling's score (eq 5).

The first external data set consisted of 25 compounds extracted from the study of Cronin et al.[27] Intuitively these compounds can be described as being structurally close to the training data. Two compounds fell outside the 95% confidence region defined by the Hotelling's score, which confirms that the majority of the compounds are close to the training set. The multinom model classified 84% of the 25 compounds correctly and 91% correctly if the two rejected compounds (both misclassified) are not counted.

The second external set consisted of the 2414 phenols extracted from the open NCI-database. This selection can be described as being structurally extremely diverse, and thus, one can expect much more conflicts between predictions and the previously assigned MOA. This allowed us to study how the quality of the predictions changed with increasing distance to the training set compounds. This effect was studied by gradually increasing the confidence region from a very low value of 40% up to 99.9%. The larger the region the more compounds are assigned to be inside the models prediction space. Table 4 shows the overall correct classifications and the sensitivities at different choices of the prediction space including the full external data set.

Two issues are covered by Table 4: first, the behavior of the Hotelling's T^2 statistic used to determine the prediction space and second, the quality of the predictions themselves. The total number of compounds included in the models prediction space ranged from 508 to 1419 of the totally 2414 compounds, i.e., a large fraction of this external test set

cannot be described by the model, e.g., the majority of the compounds containing sulfur and all compounds containing phosphorus (with both elements not present in the training set). The original proportions of the test set, 73.3% polar narcotics, 3.0% uncouplers, 17.8% proelectrophiles and 5.9% soft electrophiles were more or less maintained even when a small prediction space was chosen. The only exception was the uncouplers which on most levels were overproportionally excluded from the model space with a proportion of only 1% or less. At the 99% level 14 uncouplers were selected and 59 excluded. All 55 dinitro- or trinitrophenols fell outside the prediction space which caused the low number of uncouplers included. Concerning the quality of the predictions, it is obvious that both overall correct classification and sensitivities follow the expected trend to decrease with increasing distance to the model. For the narrowest definition of prediction space the sensitivities correspond to the estimates obtained with cross-validation. However, for larger confidence regions the sensitivities of the specific MOAs rapidly decreases, especially for uncouplers. The reason for the low sensitivity for the 14 uncouplers included in the 99% confidence region was the large number of compounds combining structural features from several MOA. There were, e.g., 6 uncouplers with two hydroxy groups and 4 halogen groups which were predominantly predicted as proelectrophiles. If these cases of overlapping MOA-features are left out, the sensitivity for uncouplers increases to 75% however with only 8 compounds remaining in the test set. The reason for the decreasing sensitivity of proelectrophile classification is the size dependence of the surface autocorrelation coefficient. The training set proelectrophiles consisted mainly of phenols with few small substituents (while the training set polar narcotics also contained phenols with long alkyl chains). As phenols with large substituents were increasingly included with larger confidence regions, the misclassification of such compounds as polar narcotics was the main reason for the decrease in sensitivity for proelectrophiles. The sensitivity of soft electrophiles remains fairly constant over the first few confidence regions (up to the 95% confidence region) and is quite close to cross-validation estimate of 77.3%. The most predominant structural pattern of misclassified compounds was the presence of a nitro group plus an electron withdrawing group at the ring which lead to misclassifications as uncouplers.

4. DISCUSSION

This study should achieve three objectives: 1. development of a classifier suitable for database screening, 2. comparison of this classifier with existing approaches, and 3. application of the classifier in exploratory study on a large data set.

The approach of building MOA classifiers presented in this study and illustrated by Figure 1 proved to be straightforward and fast. The processing of a structure through the workflow takes less than a second of calculation time on a desktop machine (PIII, 2.2 GHz). The estimates of the predictive power determined with cross-validation are high. For the full 21 descriptor model with q_π- *A*, χ_σ- *A*, and hydrogen bonding potential (*surface A*) they amounted to 92.3 and 92.7% overall correct classification with CPG NN, respectively, with multinomial logistic regression. These high estimates were additionally supported by the good predictions for the small external set compiled from literature. Thus, the first stated goal to use empirical descriptors which can be rapidly calculated to build a model with high predictive power is met. The remaining misclassifications between uncouplers and soft-electrophiles do not necessarily indicate a model weakness but could point toward an overlap of these two MOAs. This would be in agreement with specific in vitro tests for uncouplers where considerable activity could be observed for compounds such as 4-nitrophenol which is a soft electrophile in the present data set.[58] Overlapping MOAs were also suggested by Katrizky et al.[23]

The estimates for the models based on the six descriptors of Schüürmann et al. amounted to 91.4% for CPG NN and 91.8% for multinom, respectively. The differences to models based on empirical descriptors are very small and do not allow for judging which models have higher predictive power. The same is valid for the sensitivities of the three specific MOAs where differences between all models amount usually to differences of one or two compounds (with the exception of uncouplers in the multinom model of Table 3). The important finding is, however, that for this data set it was possible to replace quantum-chemical descriptors with empirical descriptors coded with autocorrelation.

The similarity of the predictions of structure descriptors and previously published descriptors suggests that these contain similar information. Schüürmann et al. developed in their study a hierarchical scheme which descriptors contribute to separate which MOAs.[10] On the top level of their hierarchy they used E_{LUMO} to separate two MOA groups from each other, i.e., one group with low E_{LUMO} values (uncouplers and soft electrophiles) from one with higher values (polar narcotics and proelectrophiles). A similar pattern could be observed for the model based on q_π- *A* and χ_σ- *A*. In these models, misclassifications occurred either between polar narcotics and proelectrophiles or between uncouplers and soft electrophiles, while the two groups of MOA were well separated. To investigate the relation between these descriptors a PLS model with the Y-variable E_{LUMO} and the 12 X-variables q_π- *A* and χ_σ- *A* was fitted. With 4 components a coefficient of determination, R^2 of 0.84, and a cross-validated R^2 of 0.81 were obtained. Thus, the descriptors do not only lead to similar predictions in classifications but correlate well also as continuous variables. Both types of descriptors are a global measure of electrophilicity decrease in the following order uncouplers > soft electrophiles ≫ proelectrophiles > polar narcotics. Also the subsequent separation within these two blocks partly follows the line of argumentation of Schüürmann et al. They used E_{HOMO}, N_{Hdon}, and D^E_{C-max} for separating polar narcotics from proelectrophiles, while in this study the same end was accomplished using *surface A*.

Surface A leads to major improvements in the classification of proelectrophiles because it describes hydrogen-bond acceptor or donor patterns that occur in the molecule. In that sense, it contains similar information as N_{Hdon} used in previous studies.[9,10] The question for both descriptors is how to relate them to the underlying mechanisms. The case for N_{Hdon} as a descriptor is made on the basis that the distinct feature of polar narcotics is their combination of hydrophobicity and hydrogen bond donor capacity.[10] Thus, the *surface A* vector also seems to contain this information. On the other

hand, Aptula et al. noted that N_{Hdon} might simply reflect the structural rule for the a priori assignment of proelectrophiles, since it essentially counts the number of hydroxy and amino groups.[9] Although both N_{Hdon} and *surface A* vectors highly correlate with the response variable, it is hard to prove that they reflect the underlying mechanisms. The only way to prove it would be to find additional data especially less "typical" compounds for each MOA.

The investigation with the two external sets proved to be particularly interesting in several aspects. The first aspect is the prediction space defined by the methodology based on the Hotelling's T^2 statistic of PCA scores. The conclusions for this aspect are as follows: First, the definition of a prediction space is essential for any model, if it is used to screen a large and diverse database. It prevents that predictions are made for compounds not covered by the training set. Second, the number of compounds included in the prediction space was surprisingly small even for large confidence regions. In other words, the models developed are still rather local in nature. Two factors cause this effect: the first factor is that the training set spans a limited part of the chemical space taken by the 2414 phenols and the second factor is that the high dimensionality of structure descriptors automatically leads to a low sample density in the multidimensional space. The latter phenomenon is commonly referred to as the "curse of dimensionality".[50] Third, the ellipsoid formed by the definition of predictions space is a rather global model which does not capture local clusters toward the outer limits. Such an effect occurred with uncouplers where all dinitrophenols were excluded from the model space because of their q_π-autocorrelation vectors exhibited very high values. This deficiency should by addressed in future studies, e.g. by using bootstrapped distance corrections[57] which can take into account such local effects and also take into account characteristics of a classifier.

The second aspect is the change of the prediction quality with increasing size of the confidence regions. Although sensitivities were high for the narrowest choice of prediction space, it dropped surprisingly fast, considering that as an initial guess one would allow all predictions up to a 95% confidence level. Two reasons can cause this effect. First, the violation of assumptions behind a statistical test, i.e., the Hotelling's T^2 test used here cannot be understood directly as a probability of error, but just as a score and the appropriate limit has to be determined using a test set. Second, the rules used to assign the MOA have their limits and cannot be applied to all compounds of the extremely diverse external test set. For example phenols with carboxy-groups assigned by the rules as uncouplers are probably not enough lipophilic to enter the mitochondrial membrane,[18,25] or on the other hand some well-known uncouplers (e.g., 2,6-di-*tert*-butyl-4-(2,2-dicyanovinyl)phenol)[18] are not assigned as such and also many proelectrophiles[59] are not covered by the rules, respectively would be assigned as polar narcotics. Thus, the data set, although suited for explanatory studies, poses some dangers for the development of predictive models since there is a high probability that descriptors are selected that mainly reflect the rules used to build the data set. With increasing distance to the training set for which the rules were derived such conflicts will increase. Therefore, the study with the 2414 phenols keeps an exploratory nature.

Regarding the comparison between structure descriptors used in this study and descriptors used in previous studies their advantages and disadvantages can be summarized as follows: structure descriptors are much simpler and much more rapid to calculate, and they can be used for modeling even when the underlying mechanisms are not known (which is partially the case in the present data set, e.g. for proelectrophiles). However, they cannot be directly linked to a chemical mechanism as easily as whole molecule descriptors e.g., a log *P* value. The exploratory study gave also indications that their rather high dimensionality leads to models with rather local character. It would be interesting to repeat the experiment using the six descriptors proposed by Schüürmann et al. in order to determine whether these lower dimensional models have confidence regions containing more compounds. If this is confirmed, the price for going from the six dimensional space to the 21 dimensions would be that larger training sets are needed to cover the same space.

5. CONCLUSION

In this study MOA classifiers based on empirical structure descriptors were derived. The calculation of such descriptors is very fast which makes them ideally suited for screening of databases. The estimates of predictive power turned out to be equal to previously published quantum mechanical based descriptors. In some cases the information contained in structure descriptors could be directly related to the whole molecule descriptors, which is an additional justification to use empirical structure descriptors. The determination of a model's prediction space proved indispensable for screening studies with structurally diverse databases. In an exploratory study using a large external test set the models were shown to have limited prediction spaces but to have satisfactory quality within this space.

ACKNOWLEDGMENT

Eric Pellegrini would like to thank the Federation of European Biochemical Societies (FEBS) for granting this work. Simon Spycher would like to thank the Bundesministerium für Bildung und Forschung (BMBF) for funding (Systems Biology, project no. FZJ 0313080 C). We also thank Aynur Aptula for providing the data set and for valuable comments.

REFERENCES AND NOTES

(1) Krajcsi, P.; Darvas, F. High-Throughput In Vitro Toxicology. In *High-Throughput ADMETox Estimation*; Darvas, F., Dormàn, G., Eds.; Eaton Publishing: Westborough, MA, 2002; pp 75−81.
(2) Hansch, C. Quantitative approach to biochemical structure−activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.
(3) Wang, J.; Lai, L.; Tang, Y. Data mining of toxic chemicals: structure patterns and QSAR. *J. Mol. Model* **1999**, *5*, 252−262.
(4) Rand, G. M.; Wells, P. G.; McCarty, L. S. Introduction to Aquatic Toxicology. In *Fundamentals of Aquatic Toxicology: Effects, Environmental Fate and Risk Assessment*; Rand, G. M., Ed.; Taylor & Francis: Washington, DC, 1995; pp 3−67.
(5) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(6) Bradbury, S. P. Predicting modes of toxic action from chemical structure: An overview SAR. *QSAR Environ. Res.* **1994**, *2*, 89−104.
(7) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical, and neural

methods for predicting toxic modes of action. *Environ. Toxicol. Chem.* **1998**, *17*, 1056−1064.
(8) Nendza, M.; Müller, M. Discriminating toxicant classes by mode of action: 2. Physico-chemical descriptors. *Quant. Struct.−Act. Relat.* **2000**, *19*, 581−598.
(9) Aptula, A. O.; Netzeva, T. I.; Valkova, I. V.; Cronin, M. T. D.; Schultz, T. W.; Kühne, R.; Schüürmann, G. Multivariate discrimination between modes of toxic action of phenols. *Quant. Struct.−Act. Relat.* **2002**, *21*, 12−22.
(10) Schüürmann, G.; Aptula, A. O.; Kuehne, R.; Ebert, R. U. Stepwise Discrimination between Four Modes of Toxic Action of Phenols in the Tetrahymena pyriformis Assay. *Chem. Res. Toxicol.* **2003**, *16*, 974−987.
(11) Mekenyan, O. G.; Veith, G. D. The Electronic Factor in QSAR: MO−Parameters, Competing Interactions, Reactivity and Toxicity. *SAR QSAR Environ. Res.* **1994**, *2*, 129−143.
(12) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.
(13) Gasteiger, J.; Marsili, M.; Hutchings, M. G.; Saller, H.; Löw, P.; Röse, P.; Rafeiner K. Models for the Representation of Knowledge about Chemical Reactions *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 467−476.
(14) PETRA − Parameter Estimation for the Treatment of Reactivity Applications. Version 3.1, MolNet GmbH, http://www.mol-net.com, http://www2.chemie.uni-erlangen.de/services/petra/index.html, Erlangen.
(15) Ren, S. Determining the mechanisms of toxic action of phenols to Tetrahymena pyriformis. *Environ. Toxicol.* **2002**, *17*, 119−127.
(16) Ren, S.; Kim, H. Comparative Assessment of Multiresponse Regression Methods for Predicting the Mechanisms of Toxic Action of Phenols. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2106−2110.
(17) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Identification of mechanisms of toxic action of phenols to Tetrahymena pyriformis from molecular descriptors. In *Quantitative Structure−Activity Relationships in Environmental Sciences-VII, Proceedings of QSAR 96, Elsinore, DK, June 24−28, 1996*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 329−342.
(18) Terada, H. Uncouplers of oxidative phosphorylation. *Environ. Health Perspect.* **1990**, *87*, 213-218.
(19) Schüürmann, G.; Somashekar, R. K.; Kristen, U. Structure−activity relationships for chloro- and nitrophenol toxicity in the pollen tube growth test *Environ. Toxicol. Chem.* **1996**, *15*, 1702−1708.
(20) Lipnick, R. L. Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Sci. Total Environ.* **1991**, *109−110*, 131−153.
(21) Netzeva, T. I.; Aptula, A. O.; Chaudary, S. H.; Duffy, J. C.; Schultz, T. W.; Schüürmann, G.; Cronin, M. T. D. Structure−activity relationships for the toxicity of substituted poly-hydroxylated benzenes to *Tetrahymena pyriformis*: Influence of free radical formation. *QSAR Comb. Sci.* **2003**, *22*, 575−582.
(22) Roberts, D. W. An analysis of published data on fish toxicity of nitrobenzene and aniline derivatives. QSAR Environ. Toxicol., Proc. Int. Workshop, 2nd. Kaiser, K. L. E., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1987; pp 295−308.
(23) Katritzky, A. R.; Oliferenko, P.; Oliferenko, A.; Lomaka, A.; Karelson, M. Nitrobenzene toxicity: QSAR correlations and mechanistic interpretations. *J. Phys. Org. Chem.* **2003**, *16*, 811−817.
(24) van Wezel, A. P.; Opperhuizen, A. Narcosis due to environmental pollutants in aquatic organisms: residue-based toxicity, mechanisms, and membrane burdens. *Crit. Rev. Toxicol.* **1995**, *25*, 255−279.
(25) Escher, B. I.; Hunziker, R.; Schwarzenbach, R. P.; Westall, J. C. Kinetic Model To Describe the Intrinsic Uncoupling Activity of Substituted Phenols in Energy Transducing Membranes. *Environ. Sci. Technol.* **1999**, *33*, 560−570.
(26) Harder, A.; Escher, B. I.; Schwarzenbach, R. P. Applicability and Limitation of QSARs for the Toxicity of Electrophilic Chemicals. *Environ. Sci. Technol.* **2003**, *37*, 4955−4961.
(27) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49*, 1201−1221.
(28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, N. Y, 1987; p 450.
(29) Karle, J. Applications of Mathematics to Structural Chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 381−390.
(30) Moreau, G.; Broto, P. The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.
(31) Moreau, G.; Broto, P. Autocorrelation of molecular structures. Application to SAR studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.
(32) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(33) Terfloth, L. Calculation of Structure Descriptors. In *Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: Weinheim, 2003; pp 401−437.
(34) Gasteiger, J. A Hierarchy of Structure Representation. In *Handbook of Chemoinformatics*; Gasteiger J., Ed.; Wiley-VCH: Weinheim, 2003; pp 1034−1061.
(35) Vedani, A.; Huhta, D. W. A new force field for modeling metalloproteins. *J. Am. Chem. Soc.* **1990**, *112*, 4759−4767.
(36) CORINA, Version 2.4, MolNet GmbH, http://www.mol-net.com, Erlangen.
(37) AUTOCORR, Version 2.4, MolNet GmbH, http://www.mol-net.com, Erlangen.
(38) SURFACE. Version 1.1, MolNet GmbH, http://www.mol-net.com, Erlangen.
(39) Zupan, J.; Gasteiger, J. Neural Networks in Chemistry and Drug Design; Wiley-VCH: Weinheim, 1999.
(40) Zupan, J.; Novic, M.; Li, X.; Gasteiger, J. Classification of multicomponent analytical data of olive oils using different neural networks. *Anal. Chim. Acta* **1994**, *292*, 219−234.
(41) Selzer, P.; Gasteiger, J.; Thomas, H.; Salzer, R. Rapid access to infrared reference spectra of arbitrary organic compounds: scope and limitations of an approach to the simulation of infrared spectra by neural networks. *Chem.-A Eur. J.* **2000**, *6*, 920−927.
(42) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. Prediction of 1H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.* **2002**, *74*, 80−90.
(43) SONNIA − Self-Organizing Neural Network for Information Analysis. Version 4.1, MolNet GmbH, http://www.mol-net.com, Erlangen.
(44) le Cessie, S.; van Houwelingen, J. C. Ridge estimators in logistic regression. *Appl. Statist.* **1992**, *41*, 191−201.
(45) Harrell, F. E., Jr. *Regression Modeling Strategies − With applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer-Verlag: New York, 2001.
(46) Ripley, B. D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Cambridge, 1996.
(47) Spycher, S.; Nendza, M.; Gasteiger, J. Comparison of Different Classification Methods Applied to a Mode of Action Data Set. *QSAR Comb. Sci.* **2004**, *23*, 779−791.
(48) Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley and Sons: New York, 2000.
(49) R: A language and environment for statistical computing. Version 1.7.1, R Foundation for Statistical Computing, Vienna, Austria, http://www. R-project.org.
(50) Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning; Springer: New York, 2001.
(51) Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec, Canada, August 20−25, 1995. Morgan Kaufmann Publishers: 1995; pp 1137−1145.
(52) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *An Introduction to Multi- and Megavariate Data Analysis*; Umetrics AB: Umea, 2000.
(53) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3222.
(54) Hutchings, M. G.; Gasteiger, J. Residual electronegativity − an empirical quantification of polar influences and its application to the proton affinity of amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.
(55) Gasteiger, J.; Saller, H. Calculation of the charge distribution in conjugated systems by quantification of the mesomerism concept. *Angew. Chem.* **1985**, *97*, 699−701.
(56) Gasteiger, J.; Hutchings, M. G. Quantification of effective polarizability. Applications to studies of X-ray photoelectron spectroscopy and alkylamine protonation. *J. Chem. Soc., Perkin 2* **1984**, 559−564.
(57) Efron, B.; Tibshirani, R. Improvements on Cross- Validation: The .632+ Bootstrap Method. *JASA* **1997**, *92*, 548−560.
(58) Escher, B. I.; Schwarzenbach, R. P. Mechanistic studies on baseline toxicity and uncoupling of organic compounds as a basis for modeling effective membrane concentrations in aquatic organisms. *Aquat. Sci.* **2002**, *64*, 20−35.
(59) Garg, R.; Kurup, A.; Hansch, C. Comparative QSAR: on the toxicology of the phenolic OH moiety. *Crit. Rev. Toxicol.* **2001**, *31*, 223−245.

CI0497915