

Quantitative Structure–Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure

Hao Zhu,^{†,‡} Todd M. Martin,[§] Lin Ye,^{†,‡} Alexander Sedykh,[‡] Douglas M. Young,[§] and Alexander Tropsha^{*,†,‡}

Carolina Environmental Bioinformatics Research Center, Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Campus Box 7568, 327 Beard Hall, Chapel Hill, North Carolina 27599-7568, and Sustainable Technology Division, National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, Ohio 45268

Received June 8, 2009

Few quantitative structure–activity relationship (QSAR) studies have successfully modeled large, diverse rodent toxicity end points. In this study, a comprehensive data set of 7385 compounds with their most conservative lethal dose (LD₅₀) values has been compiled. A combinatorial QSAR approach has been employed to develop robust and predictive models of acute toxicity in rats caused by oral exposure to chemicals. To enable fair comparison between the predictive power of models generated in this study versus a commercial toxicity predictor, TOPKAT (Toxicity Prediction by Komputer Assisted Technology), a modeling subset of the entire data set was selected that included all 3472 compounds used in TOPKAT's training set. The remaining 3913 compounds, which were not present in the TOPKAT training set, were used as the external validation set. QSAR models of five different types were developed for the modeling set. The prediction accuracy for the external validation set was estimated by determination coefficient R^2 of linear regression between actual and predicted LD₅₀ values. The use of the applicability domain threshold implemented in most models generally improved the external prediction accuracy but expectedly led to the decrease in chemical space coverage; depending on the applicability domain threshold, R^2 ranged from 0.24 to 0.70. Ultimately, several consensus models were developed by averaging the predicted LD₅₀ for every compound using all five models. The consensus models afforded higher prediction accuracy for the external validation data set with the higher coverage as compared to individual constituent models. The validated consensus LD₅₀ models developed in this study can be used as reliable computational predictors of in vivo acute toxicity.

1. Introduction

Chemical toxicity can be associated with many hazardous biological effects such as gene damage, carcinogenicity, or induction of lethal rodent or human diseases. It is important to evaluate the toxicity of all commercial chemicals, especially the high production volume (HPV)¹ compounds as well as drugs or drug candidates, since these compounds could directly affect human health. To address this need, standard experimental protocols have been established by the chemical industry, pharmaceutical companies, and government agencies to test chemicals for their toxic potential. For example, a so-called “Standard Battery for Genotoxicity Test” was established by the International Conference on Harmonization, U.S. Environmental Protection Administration (EPA), U.S. Food and Drug Administration (FDA), and other regulatory agencies. This test includes one bacterial reverse mutation assay (e.g., *Salmonella*

typhimurium mutation test), one mammalian cell gene mutation assay (e.g., mouse lymphoma cell mutation test), and one in vivo micronucleus test. The test battery varies slightly for pharmaceutical compounds, industrial compounds, and pesticides. The current strategies and guidelines for toxicity testing have been described in a recent review (1).

Although the experimental protocols for toxicity testing have been developed for many years and the cost of compound testing has been reduced significantly, computational chemical toxicology continues to be a viable approach to reduce both the amount of effort and the cost of experimental toxicity assessment (2). Significant savings could be achieved if accurate predictions of potential toxicity could be used to prioritize compound selection for experimental testing, especially for testing in vivo.

Many quantitative structure–activity relationship (QSAR) models have been developed for different toxicity end points to address this challenge (3–6). A summary of several models reported in earlier publications on acute rodent toxicity is given in Table 1. There are several shortcomings of earlier toxicity QSAR models that should be pointed out. Most of these studies included a relatively small number of congeneric compounds, and as a result, they had limited applicability for compounds outside of the modeling set. Very few successful QSAR models have been reported for predicting in vivo toxicity end points that are applicable to the diverse compounds of environmental interest (5, 7, 8). For instance, Enslein and co-workers (9, 10) developed multilinear regression models using large, diverse

* To whom correspondence should be addressed. Tel: 919-966-2955. Fax: 919-966-0204. E-mail: alex_tropsha@unc.edu.

[†] Carolina Environmental Bioinformatics Research Center.

[‡] University of North Carolina at Chapel Hill.

[§] U.S. Environmental Protection Agency.

¹ Abbreviations: AD, applicability domain; HPV, high production volume; k NN, k nearest neighbors; LOO-CV, leave-one-out cross-validated; MAE, mean absolute error; NIH, National Institutes of Health; QSAR, quantitative structure–activity relationship; R^2 , coefficient of determination; RF, random forest; TOPKAT, Toxicity Prediction by Komputer Assisted Technology; U.S. EPA, U.S. Environmental Protection Agency; U.S. FDA, U.S. Food and Drug Administration.

Table 1. Details of Previous QSAR Studies of Acute Rodent Toxicity

year	source	class(es) studied	N_{mod}^a	statistical method ^b	validation set used?
1978	ref 9	multiple	425	MLR	yes
1983	ref 10	multiple	1851	MLR	yes
1985	ref 30	alcohols	68	BR	no
1987	ref 31	multiple	147	MLR	no
1991	ref 32	amines and anilines	26/33	MLR	no
1996	ref 33	amides	44	MLR/NN	no
1998	ref 34	alcohols	95	E	yes
1999	ref 35	organo-phosphorus	49	MLR/NN	yes
2006	ref 36	chlorosilanes	10	LR	no
2006	ref 37	organo-phosphorus	38	CoMFA	no
2007	ref 38	multiple	49	LR	yes
2007	ref 39	substituted benzenes	28	MLR	yes

^a Size of the modeling data set. ^b LR, linear regression; MLR, multilinear regression; NN, neural network; BR, bilinear regression; E, expert system; and CoMFA, comparative molecular field analysis.

training sets (425 and 1851 chemicals, respectively), but these models had relatively poor external prediction power, yielding an R^2 value of 0.33 for the large test set.

Indeed, accurate prediction of toxicity for compounds that were not used for model development is a very challenging problem. QSAR models are generally more applicable for the analysis of small data sets of similar compounds with a simple mechanism of action (e.g., congeneric molecules binding to the same receptor or inhibiting the same enzyme) and less accurate for larger data set of compounds with complex mechanisms of action. Toxicity prediction is a hard problem because there are multiple underlying mechanisms of action, and the data sets studied in the context of a general end point (e.g., rat LD₅₀) are large and chemically diverse. Furthermore, QSAR models are developed by interpolating the training set data; therefore, they inherently have limited applicability outside of the training set. At the same time, any external prediction implies inherent and, frequently, excessive extrapolation of the training set models. Poor external predictive power of QSAR models could be due to the lack of or incorrect use of external validation during the modeling process. Each statistical method used in QSAR studies has its particular advantages, weaknesses, and practical constraints, so it is important to select the most suitable QSAR methodology for a specific toxicity end point. Thus, the toxicity prediction challenge should be addressed very carefully using rigorous modeling approaches and extensive model validation procedures.

Our recent studies of aquatic toxicity offered potential solutions to some of the above problems (11). A combinatorial QSAR approach was applied to study an aquatic toxicity data set containing 983 diverse organic compounds tested against *Tetrahymena pyriformis* (11). To explain our choice of methodology and terminology, any QSAR modeling effort requires a set of chemical descriptors and a statistical optimization approach to develop the best correlation between values of descriptors and those of biological activity. For any data set, there are several sets of descriptors that could be calculated using different available software packages. Similarly, there are multiple statistical modeling approaches that could be employed with any of the descriptor sets. In the practice of QSAR modeling, there is no standard combination of the descriptor type and model optimization approach that works best for all data sets. In addition, different QSAR methods usually use different definitions of applicability domain (AD) (or in most cases do not use the AD at all). Combinatorial QSAR modeling implies that for a given experimental data set we calculate several sets of descriptors and employ several statistical model-

ing approaches forming all-against-all pairwise combinations of descriptor sets and modeling techniques to develop multiple types of QSAR models. We require that each model must satisfy certain validation criteria. As we demonstrated in the earlier study (11), the consensus models had the highest external prediction power as compared to any individual model used in the consensus prediction. Because the individual models can have differently defined ADs, the consensus method can also afford greater chemical space coverage as well.

In this paper, a similar combinatorial QSAR workflow was employed to study a much larger and more chemically diverse data set (arguably, the largest and most diverse in vivo toxicity data set ever reported in the public domain) containing 7385 unique organic compounds with experimentally determined oral rat acute toxicity. We have explored various QSAR approaches in terms of their ability to develop robust and externally predictive models. The consensus prediction integrating all validated individual models was found to be the most accurate (using an external prediction set) when compared both to each individual model used in the consensus approach and to a popular commercial software, TOPKAT. The consensus models developed in this study could be used as reliable predictors of rodent acute toxicity for chemical compounds. The models will be made available through the ChemBench web portal maintained in our laboratory (<http://chembench.mml.unc.edu>).

Materials and Methods

Data Sets. The rat LD₅₀ data were collected from different sources (12) to form a data set including more than 8000 compounds. The structures of those compounds were verified using the approach discussed by Young's group (13). The quality of the data has been extensively reviewed over the past several years. After inorganic and organometallic compounds, salts, and compound mixtures were removed, the final acute toxicity data set included 7385 unique organic compounds. The original values of LD₅₀ for each compound were expressed as mol/kg; these were converted to $\log[1/(\text{mol/kg})]$ values according to standard QSAR practices. Chemical structures of all compounds and their experimental LD₅₀ values used in this study are available from the authors upon request.

This data set was compared with the training set used to develop the rat acute toxicity predictor available from the commercial Toxicity Prediction by Komputer Assisted Technology (TOPKAT) software. It was found that 3472 out of 7385 compounds were included in the TOPKAT rat LD₅₀ training database. To enable direct comparison of external predictive power for models generated in our studies vs TOPKAT, these 3472 compounds were used as the modeling set and the remaining 3913 compounds as the external validation set.

QSAR Modeling Approaches. Descriptors. Rat LD₅₀ models for the 3472 modeling set compounds were developed with various types of chemical descriptors, including those from the Dragon software v5.4 (14) and a set of descriptors developed previously by Martin and co-workers at the U.S. EPA (15). The latter set consisted of more than 800 descriptors in the following classes: E-state values and E-state counts, constitutional descriptors, topological descriptors, walk and path counts, connectivity, information content, 2D autocorrelation, Burden eigenvalues, molecular properties (such as the octanol-water partition coefficient), kappa, hydrogen bond acceptor/donor counts, molecular distance edge, and molecular fragment counts. There were overlaps between Dragon and EPA descriptors, but both included unique types of descriptors as well. The Dragon descriptors were used for the k nearest neighbor (k NN) and random forest (RF) methods, and the EPA descriptors were used for the hierarchical clustering, FDA MDL QSAR, and nearest neighbor QSAR methods.

Initial use of Dragon yielded more than a thousand chemical descriptors for the training set, which were processed as follows.

First, we removed all descriptors that had zero values or zero variance for all modeling set compounds. Furthermore, redundant descriptors were identified by analyzing correlation coefficients between all pairs of descriptors; if the correlation coefficient between two descriptor types for all modeling set compounds was higher than 0.95, one of them was removed. As a result, the total number of Dragon descriptors used for model building was reduced to 454. The number of EPA descriptors used for model building (for the hierarchical clustering and FDA MDL QSAR methods) varied depending on the size and composition of the training set molecules that were used for model building.

kNN. The *k*NN QSAR method (16) employs the *k*NN classification principle and a variable (i.e., descriptor) selection procedure. Briefly, a subset of *nvar* (number of selected descriptors) descriptors is selected randomly at the onset of the calculations. The *nvar* is set to different values, and the training set models are developed with leave-one-out cross-validation (LOO-CV), where each compound is eliminated from the training set and its LD₅₀ value is predicted as the average activity of *k* most similar molecules, where the value of *k* is optimized as well (*k* = 1–5). The similarity is characterized by Euclidean distance between compounds in multidimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criteria is used to optimize the selection of descriptors. The objective of this method is to optimize *nvar* and *k* values to obtain the best possible LOO-CV q_{abs}^2 , that is, q^2 with the intercept set to zero, by optimizing the *nvar* and *k*. The additional details of the method can be found elsewhere (16).

In developing *k*NN QSAR models, we followed our general predictive QSAR modeling workflow methodology (17), which places special emphasis on model validation. Briefly, we start by dividing the original data set randomly into a (bigger) modeling set and a (smaller) external validation set; the latter is not used for model development at all, and the former is designated as a modeling set. The modeling set compounds are divided multiple times into training/test sets using the Sphere Exclusion approach (18), which ensures that both training and test sets are chemically diverse. The models are developed using training set data, and their performance is characterized with the standard LOO-CV R^2 (q^2) for the training sets and the conventional coefficient of determination R^2 for the test sets; this coefficient is determined for a regression that is forced through the origin of the experimental vs calculated LD₅₀ plot. The model acceptability threshold values of the LOO-CV accuracy of the training sets and the prediction accuracy for test sets were both set at no less than 0.5. Models that did not meet both training and test set cutoff criteria were discarded. Models that passed these threshold criteria were used to predict LD₅₀ values of the external validation set to ensure their external predictive power as discussed in the Results and Discussion section. The detailed discussion of the workflow used to develop validated QSAR models can be found in a recent review (19).

RF. In machine learning, a RF is a predictor that consists of many decision trees and outputs the prediction that combines outputs from individual trees. The algorithm for inducing a RF was developed by Breiman and Cutler (20). In this study, the implementation of the RF algorithm available in R.2.7.1 (21) was used. In the RF modeling procedure, *n* samples are randomly drawn from the original data. These samples were used to construct *n* training sets and to build *n* trees. For each node of the tree, *m* descriptors were randomly chosen from the total 454 Dragon descriptors. The best data split was calculated using these *m* descriptors for each training set. In this study, only the defined parameters (*n* = 500 and *m* = 13) were used for the model development.

Hierarchical Clustering. The hierarchical clustering method utilizes a variation of the Ward's Minimum Variance Clustering Method (22) to produce a series of clusters from the initial training set. For a training set of *n* chemicals, initially there will be *n* clusters. At each step in the clustering process, two clusters are combined so that the increase in variance over all of the clusters in the system is minimized. The change in variance caused by combining clusters *j* and *k* is as follows:

$$\Delta\sigma^2 = \frac{n_j n_k}{n_j + n_k} \sum_{i=1}^d (C_{j,i} - C_{k,i})^2 \quad (1)$$

where n_j = number of chemicals in cluster *j*, $C_{j,i}$ is the centroid (or average value) for descriptor *i* for cluster *j*, and *d* is the number of descriptors in the EPA pool of descriptors (~800) (15). The process of combining clusters while minimizing variance continues until all of the chemicals are lumped into a single cluster. After the clustering is complete, each cluster is analyzed to determine if an acceptable QSAR model can be developed. A genetic algorithm technique is used to select descriptors to build a multilinear regression model for each cluster (15). Similar to the *k*NN approach, each model must achieve a LOO-CV accuracy of 0.5 to be used in making predictions. The predicted value for a given test chemical is calculated using the equally weighted average of the model predictions from the closest cluster from each step in the hierarchical clustering. This method was previously shown to yield the best results for another acute toxicity end point, IGC₅₀ (50% inhibitory concentration of population growth) of *T. pyriformis* (15).

FDA MDL QSAR Method. A QSAR methodology (denoted here as the FDA MDL QSAR method) based on the studies of Contrera et al. (23) was developed earlier (15). For each test chemical, a cluster is constructed using the 30 most similar chemicals from the training set as defined by the cosine similarity coefficient, $SC_{i,k}$, which is calculated as follows

$$SC_{i,k} = \frac{\sum_{j=1}^{\#\text{descriptors}} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{\#\text{descriptors}} x_{ij}^2 \cdot \sum_{j=1}^{\#\text{descriptors}} x_{kj}^2}} \quad (2)$$

where x_{ij} is the value of the *j*-th normalized descriptor for chemical *i* (normalized with respect to all of the chemicals in the original training set) and x_{kj} is the value of the *j*-th descriptor for chemical *k*. The entire pool of approximately 800 EPA descriptors is used to calculate the similarity coefficient in eq 2. A multiple linear regression model is then built for the new cluster using a genetic algorithm-based method, and the toxicity is predicted.

Nearest Neighbor Method. The nearest neighbor method is a simplification of the variable selection *k*NN approach described above. In the nearest neighbor method, the toxicity is simply predicted as the average of the toxicity of the three most similar chemicals from the training set. The similarity is defined in terms of the cosine similarity coefficient (eq 2). In the nearest neighbor method, the entire available descriptor pool is used to characterize molecular similarity (as opposed to a subset of the descriptor pool as in the descriptor selection *k*NN method). To make a prediction, each of the neighbors in the training set must exceed a minimum cosine similarity coefficient of 0.5.

Identification of Outliers in the Data Set. A common problem for most QSAR studies is the existence of compounds that are highly dissimilar to all other compounds in the data set. These compounds are regarded as outliers in the descriptor space and are likely to present problems in establishing SAR trends, which is critical to QSAR modeling. In this study, we have identified and excluded the structural outliers from the modeling at the beginning of the modeling procedure.

For *k*NN and RF modeling procedures, we have developed a method to detect outliers that are dissimilar to other compounds of the data set in the descriptor space. This procedure included the following steps. (1) calculation of the distance or similarity matrix based on the Dragon descriptors of compounds in the descriptor space, (2) finding the nearest neighbors for all compounds in the data set based on a predefined similarity threshold, and (3) identifying those compounds that have no nearest neighbors as outliers.

To measure similarity, each compound *i* is represented by a point in the *M*-dimensional descriptor space (where *M* is the total number

of descriptors) with the coordinates $X_{i1}, X_{i2}, \dots, X_{iM}$, where X_{is} ($s = 1, \dots, M$) are the values of individual descriptors. The molecular dissimilarity between any two molecules i and j is characterized by the Euclidean distance between their representative points. The Euclidean distance d_{ij} between points i and j in M -dimensional space can be calculated as follows (eq 3):

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (3)$$

Compounds with the smallest distance between them are considered to have the highest similarity. The distances (dissimilarity) of compounds in our modeling set are compiled to produce a chemical similarity threshold D_T , calculated as follows (eq 4):

$$D_T = \bar{y} + Z\sigma \quad (4)$$

Here, \bar{y} is the average Euclidean distance between all compounds and their k NNs (k was set to 1 in this procedure) of each compound within the modeling set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the threshold level and was set to 0.5 in this study. The D_T threshold is used to identify outliers as follows. If the distance of a compound to its nearest neighbor in the modeling set exceeds this threshold, this compound is considered an "outlier" and excluded from the modeling set. After excluding 997 structural outliers, the remaining 2475 modeling set compounds were compiled as a new reduced modeling set to develop k NN and RF toxicity models.

It is important to point out that the identification and exclusion of outliers are based only on consideration of chemical similarity but not activity. Thus, the removal of structural outliers could be regarded as a pretreatment of the modeling set using objective chemometric approaches.

For the hierarchical and FDA MDL QSAR methods, a chemical is removed from a cluster if it is both an influential data point (determined by at least two statistical tests, for example, DFFITS, leverage, Cook's distance, and covariance ratio) and an outlier (determined from studentized deleted residual). The details of these procedures are given elsewhere (24).

Model ADs. Defining model ADs is an active area of modern QSAR research (25, 26). Every QSAR model can formally predict the relevant target property for any compound for which chemical descriptors can be calculated. However, because each model is developed using compounds in the training set only [that cover only a small fraction of the entire chemistry (i.e., descriptor) space], the special AD for each model should always be defined. As a consequence, only a certain fraction of compounds in any external data set is expected to fall within the AD. This fraction is therefore referred to as the data set coverage. There are several discussions about model AD in a recent publication (27). In this study, we present a detailed discussion concerning the effect of the AD on model predictivity using much larger modeling/validation sets than any other reported in the literature including our own previous publications.

AD of k NN and RF. The AD of k NN and RF models is calculated from the distribution of similarities between each compound and its k NN in the training set (similarities are computed as Euclidean distances between compounds represented by their multiple chemical descriptors). On the basis of the previous studies, the standard cutoff value to define the AD for a QSAR model places its boundary at one-half of the standard deviation calculated for the distribution of distances between each compound in the training set and its k NNs in the same set. Thus, if the distance of the test compound from any of its k NNs in the training set exceeds the threshold, the prediction is considered unreliable. The detailed description of the algorithm to define this AD is given elsewhere (18, 28).

AD of the Hierarchical Method. Before any cluster model can be used to make a prediction for a test chemical, it must be determined whether the test chemical falls within the AD for the

Table 2. Statistical Results Obtained with All QSAR Models for the External Validation Set of 3913 Compounds

models	R^2	MAE	coverage (%)
k NN	0.66	0.44	19
RF	0.70	0.41	19
hierarchical clustering	0.41	0.58	66
NN	0.24	0.61	97
FDA MDL QSAR	0.29	0.60	95
TOPKAT	0.35	0.59	74

model. The first constraint, the model ellipsoid constraint, checks if the test chemical is within the multidimensional ellipsoid defined by the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model). The model ellipsoid constraint is satisfied if the leverage of the test compound (h_{00}) is less than the maximum leverage value for all of the compounds used in the model (29). The second constraint, the R_{\max} constraint, checks if the distance from the test chemical to the centroid of the cluster is less than the maximum distance for any chemical in the cluster to the cluster centroid. The final constraint, the fragment constraint, stipulates that the chemicals in the cluster must contain at least one example of each of the fragments that are present in the test chemical (15).

AD of the FDA MDL QSAR Method. For the prediction from the cluster model to be valid, several constraints must be met. The first two constraints are the model ellipsoid and fragment constraints described above. The final constraint is that the predicted toxicity value must be within the range of experimental toxicity values for the chemicals used to build the model (15).

AD of the Nearest Neighbor Method. For a prediction from the nearest neighbor method to be made, there must be three chemicals in the training set that are sufficiently similar to the test chemical (the similarity coefficient between each chemical and the test chemical in eq 1 must exceed 0.5).

Results and Discussion

Individual LD₅₀ Models. The statistical parameters of predictions for the external validation set obtained from all five QSAR models developed in this study as well as using TOPKAT are shown in Table 2. It is difficult to compare all models side by side because the underlying approaches used different definitions of AD; therefore, the statistical results are shown for external data sets of different sizes. Indeed, these initial results suggest that the prediction accuracy and chemical space coverage are tightly interlinked, and in general, as expected, higher accuracy is obtained for smaller external data sets within the AD of each model. Models with the most liberally defined AD (and consequently, the highest coverage), that is, NN and

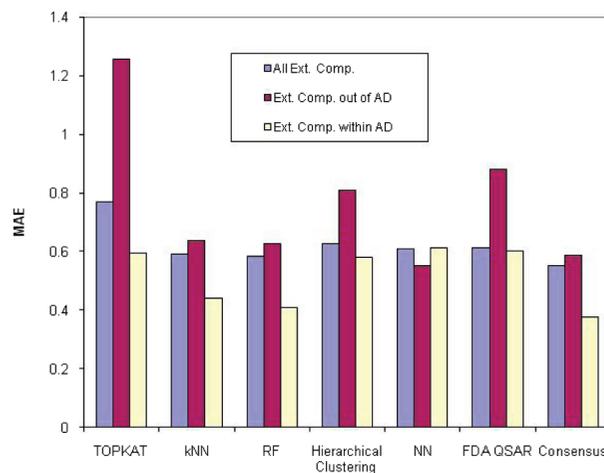


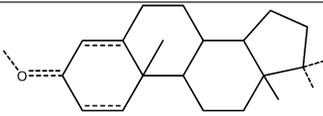
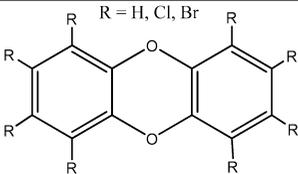
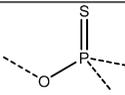
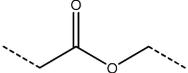
Figure 1. MAEs of five QSAR models for the external validation set.

Table 3. Statistical Results Obtained from All QSAR Models for the External Validation Set of 3913 Compounds

models	TOPKAT AD (2896 compounds predicted, 74% coverage)		hierarchical clustering AD (2583 compounds predicted, 66% coverage)		kNN and RF AD (743 compounds predicted, 19% coverage)	
	R ²	MAE	R ²	MAE	R ²	MAE
kNN	0.41	0.55	0.40	0.57	0.66	0.44
RF	0.33	0.54	0.41	0.56	0.70	0.41
hierarchical clustering	0.33	0.59	0.41	0.58	0.65	0.45
NN	0.35	0.57	0.41	0.58	0.66	0.44
FDA MDL QSAR	0.37	0.57	0.40	0.59	0.64	0.45
TOPKAT	0.35	0.59	0.25	0.70	0.54	0.52
consensus ^a	0.42	0.52	0.48	0.51	0.71	0.39

^a TOPKAT results were not included in the consensus model.

Table 4. Comparison of the Average Experimental LD₅₀ Values for the Modeling Set Compounds, Validation Set Compounds, and the Validation Set Compounds with Large Prediction Errors (MAE > 1.0)

Scaffolds	Modeling set		Validation set		Validation set with large prediction errors	
	N ^a	Aver. LD ₅₀	N	Aver. LD ₅₀	N	Aver. LD ₅₀
	5	2.5	17	3.8	10	4.6
 R = H, Cl, Br	3	8.2	5	5.1	5	5.1
	285	3.6	160	3.5	49	3.9
	83	2.0	124	2.4	15	2.9
All compounds	3,472	2.47	3,913	2.6	520	3.4

^a N is the number of compounds.

FDA MDL QSAR, had the lowest R^2 and the highest mean absolute error (MAE) followed by TOPKAT and hierarchical clustering, which had progressively higher R^2 values (although similar MAE) and smaller coverage. Nevertheless, for these four models, the absolute R^2 values were relatively low, that is, under 0.5. Only two models (k NN and RF) afforded R^2 higher than 0.50 and MAE lower than 0.50 for the external validation set, but the external data set coverage of these two models is the lowest (19%) among all models. It could be argued that for this data set (and perhaps for any large and diverse data set), it is critical to define a rather restrictive AD to achieve the most accurate predictions, as discussed in more detail below.

Effect of the Model AD. All five QSAR approaches implemented method-specific AD except k NN and RF models, which used the same definition of AD. On average, the use of AD improved the performance of individual models, although the improvement came at the expense of the lower chemical space coverage. The direct comparison between individual

models appears difficult due to different definitions of AD and different interplay between coverage and accuracy for relevant models.

Figure 1 shows the distribution of MAE values for the prediction of external validation set for TOPKAT, five individual models, and consensus model developed in this study (see the additional discussion of the consensus model below). The results included the MAE of these models for all external compounds, those located within the AD of each model, and those outside of AD. Notably, all models showed similar predictivity when applied to the entire external set, but the effect of AD was indeed model-specific. Six (TOPKAT, k NN, RF, hierarchical clustering, FDA MDL QSAR, and consensus) out of seven QSAR models that used the AD showed the improvement in the prediction accuracy for external validation set as a result of excluding those compounds outside of the AD. The result of NN practically did not change after applying the AD criteria. This is not surprising given that there were only very few compounds that were outside of the structural AD in this model.

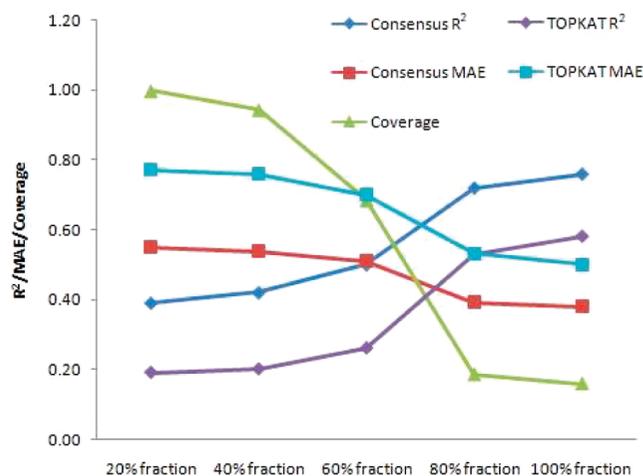


Figure 2. Prediction of external compounds by consensus model and TOPKAT with different consensus prediction fraction levels.

The different predictivity of the external validation set obtained from five QSAR models does not necessarily indicate that statistical approaches or descriptors used to develop these models have greatly different predictive power for this specific toxicity end point. It is noticeable that the resulting predictive accuracy strongly correlates to the model coverage that is decided by the model AD. Once a more restrictive AD was applied, the predictive accuracy improved significantly (Table 2). For this reason, it is interesting to study the performance of each model when the same model AD was implemented. Because only a small number of compounds were out of the ADs of NN and FDA MDL QSAR models, the remaining two model ADs (ADs of hierarchical clustering and *k*NN/RF) and the AD of TOPKAT were used to study the prediction accuracy of each model under the same prediction coverage (Table 3).

When using the same model AD, the prediction coverage of the external prediction set obtained from each individual QSAR models is almost but not exactly the same. This is because there are some compounds (less than 1% of the total external compounds) that cannot be predicted using the hierarchical clustering method even if all of the constraints are relaxed. At similar levels of prediction coverage, the individual predictions using models generated in this study are similar to each other. Interestingly, the results generated using all models are ap-

proximately the same (in terms of R^2 and MAE) when using TOPKAT defined AD, with the *k*NN method arguably showing slightly better performance. However, somewhat surprisingly, with the decrease of the chemical space coverage, most of the individual models developed in this study appear consistently superior to TOPKAT (Table 3). It may be concluded that the prediction accuracy is not sensitive to the statistical approaches employed in this paper but strongly depends on the model AD. Again, as noted above, it could be concluded that the higher accuracy of prediction comes at the expense of reducing the chemical space coverage.

Compounds That Can Not Be Correctly Predicted by Individual Models. There are some compounds that could not be predicted accurately by any of the five individual models. Using MAE > 1.0 as criteria, there are 520 validation set compounds with large prediction errors for any of the individual models. Some specific chemical scaffolds could be identified from these 520 compounds. These scaffolds and the comparison between the average LD₅₀ values of the associated compounds in the modeling set, external validation set, and those validation set compounds that have large prediction errors are listed in Table 4. The average LD₅₀ value of these compounds is 3.4, and it is much higher than that of the compounds in the modeling set (2.47). Therefore, the relatively small fraction of compounds with high values of acute toxicity in the modeling set is a potential reason of the low prediction accuracy for these 520 compounds.

Ten out of 17 steroidlike compounds in the validation set have large prediction errors. As shown in Table 4, the five steroids in the modeling set have lower acute toxicity (average LD₅₀ = 2.5) than these 10 compounds (average LD₅₀ = 4.6). A similar observation is true for the esters. Compounds with the same scaffolds and high acute toxicity need to be added into the modeling set to accurately predict these types of compounds. On the contrary, all five dioxins in the validation set have much lower toxicity (average LD₅₀ = 5.1) than those three in the modeling set (average LD₅₀ = 8.2). Therefore, dioxins with lower acute toxicity need to be added to the modeling set to accurately predict this type of compounds. There is no clear difference between the average LD₅₀ value of 49 thiophosphates with large prediction errors in the validation set and the 285 thiophosphates in the modeling set. However, the activity range

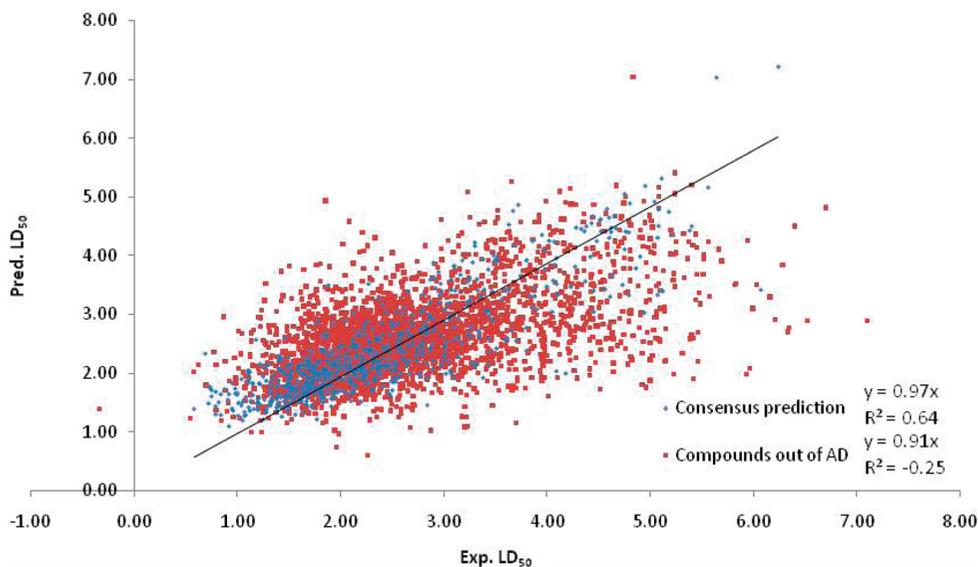


Figure 3. Correlation of experimental and consensus predicted LD₅₀ values when the consensus prediction fraction is 80% (compounds are within AD of four or more individual models).

of these 49 thiophosphates in the validation set is from 1.2 to 6.3, which is much larger than the activity range of 285 thiophosphates in the modeling set, which is from 1.6 to 5.4. For this reason, thiophosphates with both high and low acute toxicity values need to be added to the validation set to improve the model predictivity for this type of compounds. These results indicate the existing shortcomings of the TOPKAT LD₅₀ modeling set. Apparently, the modeling set should be balanced not only in terms of chemical diversity of compounds but also their activity distribution to afford higher external accuracy of models.

Consensus Modeling. The statistical results obtained with individual models indicate that different modeling techniques may have different advantages for predicting the rat oral LD₅₀ of organic compounds. Although the performances of our individual models are comparable or slightly better than that of TOPKAT, it is difficult to judge which model is better than others and which model should be chosen to predict rat acute toxicity potential of new compounds. For this reason, following a strategy that was proven successful in our previous studies (11), a simple consensus model was developed that integrated all of the individual models. In this approach, the LD₅₀ value for each compound is predicted as the arithmetic average of all LD₅₀ values predicted by individual models taking into account the model ADs. Note that additional averaging schemes giving, for example, different weights to different contributing models could be used in principle. However, there has not been sufficient research in the QSAR modeling community into looking for the most optimal scheme for the ensemble QSAR modeling. Thus, we chose the simplest approach in this study. The detailed comparison between consensus predictions and those of other models when using the same AD is listed in Table 3. The data clearly demonstrate that the predictive accuracy of consensus model is higher than that for any individual model. In addition, we used the Wilcoxon test to calculate the *p* values for the differences in MAEs obtained by consensus prediction vs individual methods. Under almost all conditions, the improvement achieved by consensus prediction, as compared with any individual model, is statistically significant (*p* < 0.01), and the only exception is when comparing consensus prediction with RF for the 743 compounds in the AD of RF models (*p* = 0.4).

From the discussion above, it is clear that the AD is an important factor that affects the predictive accuracy of each individual model. In the consensus prediction, model AD was implemented by introducing the concept of "consensus prediction fraction". Because the consensus prediction is the average of predictions using all five models, the fraction of the prediction could be defined as the number of individual model predictions that are available to predict a new compound (due to the AD limitations). Thus, if only one model could predict a compound, the consensus prediction fraction is 20% for this compound. If all five models could make the prediction, the prediction fraction of the consensus model is 100%. Different cutoff values for the prediction fraction could be set to get different prediction accuracy (and different coverage) based on this threshold. Figure 2 shows the change of prediction accuracy of external set, which is indicated by *R*² and MAE, obtained by consensus prediction with different fraction cutoff values. For comparison, the TOPKAT prediction for the same external compounds is also shown in the same Figure 2. Increasing the prediction fraction level increased the prediction accuracy but decreased the prediction coverage. Figure 3 shows the relationship between experimental and consensus-predicted LD₅₀ values when the prediction fraction is 80%. The compounds outside of the AD

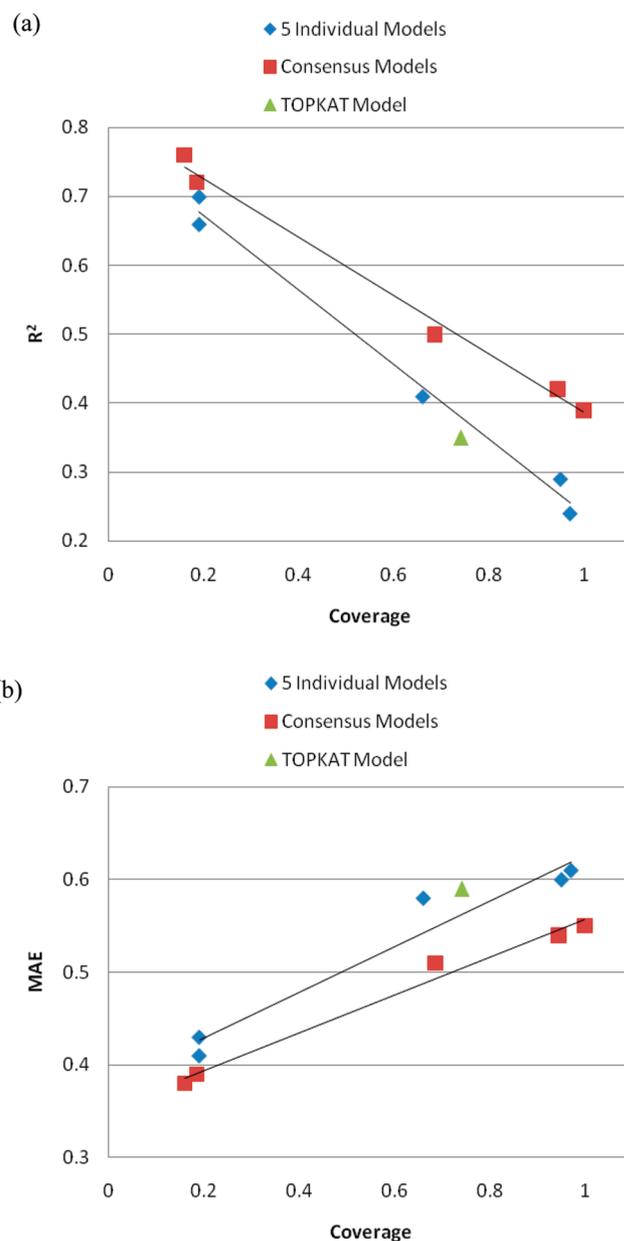


Figure 4. Relationship between prediction coverage and (a) *R*² or (b) MAE for the external compounds.

in this consensus prediction are also shown (Figure 3). Obviously, the removal of outliers improves the correlation. Furthermore, it is also interesting to compare the prediction coverage and accuracy that is indicated by *R*² and MAE. Figure 4 shows the inverse correlation between the coverage and the *R*² (or direct correlation between the coverage and MAE) for all individual models (including TOPKAT) and consensus model (including the results of different prediction fractions). It is clear that the prediction accuracy obtained by this consensus model is higher than that for any individual model under any conditions (Figure 4).

A further understanding of the predictive ability of the models used in this study can be obtained by analyzing compounds for which consensus prediction gave higher accuracy than any of the individual models. It is clear that if all five individual models make similar predictions for a compound, the value from consensus prediction will be similar to any of those generated with individual models. The possible improvement of the prediction accuracy due to the use of consensus prediction could

Table 5. Experimental and Predicted LD₅₀ Values for 10 External Set Compounds That Have the Most Significant Differences in Predicted LD₅₀ Values Using Individual Models

no.	compounds	exp. ^a	RF	kNN	HC ^b	NN	FDA MDLQSAR	cons. ^c	average MAE	cons. MAE
1	4 <i>H</i> -1,3,2-benzodioxaphosphorin-2-amine, <i>N</i> -methyl-6-nitro-, 2-sulfide	3.32	3.81	3.64	4.46	2.99	3.11	3.60	0.50	0.28
2	phosphonothioic acid, ethyl-, <i>O,S</i> -dipropyl ester	5.07	3.43	4.04	4.51	4.49	5.07	4.31	0.76	0.76
3	2-butenitrile	2.13	2.71	2.42	4.07	2.88	3.04	3.02	0.90	0.89
4	phosphorothioic acid, <i>O,O</i> -diethyl <i>S</i> -isopropyl ester	3.07	3.60	3.70	4.71	2.97	2.99	3.60	0.60	0.53
5	isocyanic acid, allyl ester	2.70	2.39	2.12	3.93	2.53	3.10	2.81	0.54	0.11
6	phosphonothioic acid, methyl-, <i>O,S</i> -dipropyl ester	5.21	3.19	3.91	4.71	5.28	4.69	4.36	0.88	0.85
7	dibenzo(b,e)(1,4)dioxin, 1,2,3,4,7,8-hexachloro-	5.64	6.90	6.23	6.34	8.16	7.55	7.04	1.40	1.40
8	mercaptobenil	3.26	2.84	2.55	2.54	3.59	4.71	3.24	0.73	0.02
9	phosphorodichloridic acid, ethyl ester	2.87	2.62	2.78	2.53	3.13	4.93	3.20	0.60	0.33
10	dibenzo- <i>p</i> -dioxin, 1,2,3,7,8-pentachloro-	6.24	7.11	6.27	6.20	8.16	8.36	7.22	1.00	0.98

^a Experimental LD₅₀. ^b HC, hierarchical clustering. ^c Cons., consensus prediction.

be achieved when the individual predictions are different. Table 5 lists 10 compounds, which have the most significant difference between individual predictions.

There are many external validation set compounds (such as 1, 4, 5, and 9 in Table 4) whose individual LD₅₀ predictions include one value with a large deviation from the others, which is usually the one that has the largest prediction error. Therefore, by taking the average for consensus prediction, we could compensate for the large error of such individual result.

On the other hand, compounds 2, 3, 6–8, and 10 show large errors for the majority of their individual predictions. The consensus model is able to make accurate prediction, such as for compound 8, or prediction with moderate error, as for the remaining compounds in the table, because individually predicted LD₅₀ values are both lower and higher than the experimental LD₅₀ value so that the errors to some extent cancel each other. The differences in model predictions arise because they use different descriptors and/or different modeling methods, which could model different aspects of toxicological affects. Thus, the consensus modeling allows for these different effects to be incorporated into a single (and on average, more accurate) prediction.

Conclusions

Several QSAR approaches have been used to develop toxicity models of the largest available set of diverse organic compounds tested for the oral acute toxicity in rats. The resulting models (for the most part incorporating specific ADs) were validated by predicting the toxicity of a large external validation set. It was observed that all models showed somewhat different but comparable performance for the validation set when compared to the commercial toxicity predictor TOPKAT. Formally, the highest accuracies were achieved by kNN and RF approaches ($R^2 = 0.66$ and 0.70 , respectively), but this required a decrease in space coverage (to ca. 19%). However, when the same model AD was implemented, the individual models showed similar performance as applied to the validation set. Here, the use of AD improved the prediction accuracy using individual models but decreased the predictive coverage of the validation set. Notably, with the decrease of the prediction coverage, models developed in this study showed slightly higher prediction accuracy as compared to TOPKAT.

The most significant result of our studies is the demonstrated superior performance of the consensus modeling approach when all models are used concurrently and predictions from individual models are averaged (see Figure 1). The predictive accuracy of the consensus QSAR models was shown to be superior to any individual model when predicting the same set of external compounds. By using different cutoff values for the prediction fraction, trade-offs between the accuracy and the coverage of

consensus prediction results can easily be seen. The predictivity of consensus models was found to be superior to that of TOPKAT when predicting the same external compounds. Finally, these studies indicated that a well-organized modeling set that covers not only a broad chemical space but also broad activity ranges of major chemical scaffolds in this chemical space is necessary to develop successful QSAR toxicity predictors. Additional studies of this data set are ongoing and will be reported in the future. All successful models reported in this paper will be made available for use as LD50 predictors via both the ChemBench web portal (<http://chembench.mml.unc.edu>) and via EPA website (<http://www.epa.gov/nrmrl/std/cppb/qsar/index.html>). Meanwhile, interested researchers can send us any compounds of interest for LD₅₀ prediction.

Acknowledgment. This work was supported, in part, by grants from the National Institutes of Health (NIH) (GM076059 and ES005948) and U.S. EPA (RD832720 and RD833825).

The research described in this article has not been subjected to each funding agency's peer review and policy review and therefore does not necessarily reflect their views, and no official endorsement should be inferred. This manuscript has been reviewed by the U.S. EPA and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The authors declare no competing financial interests.

References

- (1) Putman, D. L., Clarke, J. J., Escobar, P., Gudi, R., Krsmanovic, L. S., Pant, K., Wagner, V. O., III., San, R. H. C., and Jacobson-Kram, D. (2006) Genetic toxicity. In *Toxicological Testing Handbook* (Jacobson-Kram, D., Keller, K. A., Eds.) pp 185–248, Informa Healthcare, New York.
- (2) Hengstler, J. G., Foth, H., Kahl, R., Kramer, P. J., Liliensblum, W., Schulz, T., and Schweinfurth, H. (2006) The REACH concept and its impact on toxicological sciences. *Toxicology* 220 (2–3), 232–239.
- (3) Richard, A. M. (2006) Future of toxicology—Predictive toxicology: An expanded view of “chemical toxicity”. *Chem. Res. Toxicol.* 19 (10), 1257–1262.
- (4) Klopman, G., Zhu, H., Fuller, M. A., and Saiakhov, R. D. (2004) Searching for an enhanced predictive tool for mutagenicity. *SAR QSAR Environ. Res.* 15 (4), 251–263.
- (5) Richard, A. M., and Benigni, R. (2002) AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report. *SAR QSAR Environ. Res.* 13 (1), 1–19.
- (6) Yang, C., Benz, R. D., and Cheeseman, M. A. (2006) Landscape of current toxicity databases and database standards. *Curr. Opin. Drug Discovery Devel.* 9 (1), 124–133.
- (7) Benigni, R., Netzeva, T. I., Benfenati, E., Bossa, C., Franke, R., Helma, C., Hulzebos, E., Marchant, C., Richard, A., Woo, Y. T., and Yang, C. (2007) The expanding role of predictive toxicology: An update on the (Q) SAR models for mutagens and carcinogens. *J. Environ. Sci. Health, Part C: Environ. Carcinog. Ecotoxicol. Rev.* 25 (1), 53–97.

- (8) Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X. Q., Doweiko, A., and Li, Y. (2003) In silico ADME/Tox: Why models fail. *J. Comput.-Aided Mol. Des.* 17 (2–4), 83–92.
- (9) Enslein, K., and Craig, P. N. (1978) A toxicity estimation model. *J. Environ. Pathol. Toxicol.* 2 (1), 115–121.
- (10) Enslein, K., Lander, T. R., Tomb, M. E., and Craig, P. N. (1983) *A Predictive Model for Estimating Rat Oral LD₅₀ Values*, Princeton Scientific Publishers, Princeton.
- (11) Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Oberg, T., Dao, P., Cherkasov, A., and Tetko, I. V. (2008) Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* 48, 766–784.
- (12) National Library of Medicine (2008) ChemIDplus database.
- (13) Young, D., Martin, T., Venkatapathy, R., and Harten, P. (2008) Are the chemical structures in your QSAR correct? *QSAR Comb. Sci.* 27 (11–12), 1337–1345.
- (14) Mauri, A., Consonni, V., Pavan, M., and Todeschini, R. (2006) Dragon software: An easy approach to molecular descriptor calculations. *MATCH* 56 (2), 237–248.
- (15) Martin, T. M., Harten, P., Venkatapathy, R., Das, S., and Young, D. M. (2008) A hierarchical clustering methodology for the estimation of toxicity. *Toxicol. Mech. Methods* 18, 251–266.
- (16) Zheng, W., and Tropsha, A. (2000) Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* 40 (1), 185–194.
- (17) Tropsha, A., and Golbraikh, A. (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* 13 (34), 3494–3504.
- (18) Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee, K. H., and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* 17 (2–4), 241–253.
- (19) Tropsha, A. (2008) Integrated chemo- and bioinformatics approaches to virtual screening. In *Chemoinformatics Approaches to Virtual Screening* (Varnek, A., Tropsha, A., Eds.) pp 295–325, Royal Society of Chemistry, Cambridge, United Kingdom.
- (20) Breiman, L. (2001) Random forests. *Machine Learning* 45 (1), 5–32.
- (21) Dalggaard, P. (2008) *Introductory Statistics with R*, Springer, New York.
- (22) Romesburg, H. C. (1984) *Cluster Analysis for Researchers*, Lifetime Learning Publications, Belmont, CA.
- (23) Contrera, J. F., Matthews, E. J., and Daniel, B. R. (2003) Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol.* 38 (3), 243–259.
- (24) Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004) *Applied Linear Statistical Models*, McGraw-Hill, New York.
- (25) Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D., Schultz, T., Stanton, D. W., van de Sandt, J. J., Tong, W., Veith, G., and Yang, C. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* 33 (2), 155–173.
- (26) Tetko, I. V., Bruneau, P., Mewes, H. W., Rohrer, D. C., and Poda, G. I. (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* 11 (15–16), 700–707.
- (27) Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* 48 (9), 1733–1746.
- (28) Tropsha, A., Gramatica, P., and Gombar, V. K. (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Quant. Struct. Act. Relat. Comb. Sci.* 22, 69–77.
- (29) Montgomery, D. C. (1982) *Introduction to Linear Regression Analysis*, John Wiley and Sons, New York.
- (30) Lipnick, R. L., Pritzker, C. S., and Bentley, D. L. (1985) A QSAR study of the rat LD₅₀ for alcohols. *QSAR Strategies Des. Bioact. Compd., Proc. Eur. Symp. Quant. Struct.-Act. Relat.* 5, 420–423.
- (31) Enslein, K., Tuzzeo, T. M., Borgstedt, H. H., Blake, B. W., and Hart, J. B. (1987) Prediction of rat oral LD₅₀ from *Daphnia magna* LC₅₀ and chemical structure. *QSAR Environ. Toxicol., Proc. Int. Workshop* 2, 91–106.
- (32) Jaeckel, H., and Klein, W. (1991) Prediction of mammalian toxicity by quantitative structure–activity relationships: Aliphatic amines and anilines. *Quant. Struct.-Act. Relat.* 10 (3), 198–204.
- (33) Zakarya, D., Larfaoui, E. M., Boulaamail, A., and Lakhlifi, T. (1996) Analysis of structure-toxicity relationships for a series of amide herbicides using statistical methods and neural network. *SAR QSAR Environ. Res.* 5 (4), 269–279.
- (34) Wang, G., and Bai, N. (1998) Structure-activity relationships for rat and mouse LD₅₀ of miscellaneous alcohols. *Chemosphere* 36 (7), 1475–1483.
- (35) Eldred, D. V., and Jurs, P. C. (1999) Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. *SAR QSAR Environ. Res.* 10, 75–99.
- (36) Jean, P. A., Gallavan, R. H., Kolesar, G. B., Siddiqui, W. H., Oxley, J. A., and Meeks, R. G. (2006) Chlorosilane acute inhalation toxicity and development of an LC₅₀ prediction model. *Inhal. Toxicol.* 18 (8), 515–522.
- (37) Guo, J. X., Wu, J. J., Wright, J. B., and Lushington, G. H. (2006) Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: A molecular modeling study. *Chem. Res. Toxicol.* 19 (2), 209–216.
- (38) Freidig, A. P., Dekkers, S., Verwei, M., Zvinavashe, E., Bessems, J. G., and van de Sandt, J. J. (2007) Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds. *Toxicol. Lett.* 170 (3), 214–222.
- (39) Toropov, A. A., Rasulev, B. F., and Leszczynski, J. (2007) QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: Comparative analysis by MLRA and optimal descriptors. *QSAR Comb. Sci.* 26 (5), 686–693.

TX900189P