



Published in final edited form as:

Chem Res Toxicol. 2009 December 21; 22(12): 1913–1921. doi:10.1021/tx900189p.

QSAR 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

[§]Sustainable Technology Division, National Risk Management Research Laboratory, Office of Research and Development, United States Environmental Protection Agency

Abstract

Few Quantitative Structure-Activity Relationship (QSAR) studies have successfully modeled large, diverse rodent toxicity endpoints. In this study, a comprehensive dataset of 7,385 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 dataset was selected that included all 3,472 compounds used in the TOPKAT's training set. The remaining 3,913 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 5 models. The consensus models afforded higher prediction accuracy for the external validation dataset 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.

Keywords

acute toxicity; computational toxicology; LD₅₀; oral exposure; QSAR; rat

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

*Corresponding author: Campus Box 7568, 327 Beard Hall, University of North Carolina, Chapel Hill, NC 27599-7568, Telephone: (919) 966-2955, FAX: (919) 966-0204, alex_tropsha@unc.edu.

The authors declare no competing financial interests.

affect human health. To address this need, standard experimental protocols have been established by 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 were 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 endpoints to address this challenge (3–6). The summary of several models reported in earlier publications on acute rodent toxicity are 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 endpoints that are applicable to the diverse compounds of environmental interest (5,7,8). For instance, Enslein and coworkers (9,10) developed multi-linear regression models using large, diverse 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 datasets 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 dataset of compounds with complex mechanisms of action. Toxicity prediction is a hard problem because there are multiple underlying mechanisms of action, and the datasets 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 and 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

¹Abbreviations:

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;
NIEHS, National Institute of Environmental Health Sciences;
NTP, National Toxicology Program;
QSAR, Quantitative Structure-Activity Relationship;
 R^2 , coefficient of determination;
RF, Random Forest;
TOPKAT, Toxicity Prediction by Komputer Assisted Technology;
US EPA, U. S. Environmental Protection Agency;
US FDA, U. S. Food and Drug Administration;

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 endpoint. 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 dataset 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 dataset 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 datasets. In addition different QSAR methods usually use different definitions of applicability domain (or in most cases do not use the applicability domain at all). Combinatorial QSAR modeling implies that for a given experimental dataset we calculate several sets of descriptors and employ several statistical modeling 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. Since the individual models can have differently defined applicability domains, 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 dataset (arguably, the largest and most diverse *in vivo* toxicity dataset ever reported in the public domain) containing 7,385 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>).

2. Methods

2.1. Datasets

The rat LD₅₀ data were collected from different sources (12) to form a dataset including more than 8,000 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 removing inorganic and organometallic compounds, salts, and compound mixtures, the final acute toxicity dataset included 7,385 unique organic compounds. The original values of LD₅₀ for each compound were expressed as mol/kg; these were converted to log(1/(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 dataset 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 3,472 out of 7,385 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 3,472 compounds were used as the modeling set and the remaining 3,913 compounds as the external validation set.

2.2. QSAR Modeling Approaches

2.2.1. Descriptors—Rat LD₅₀ models for the 3,472 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 coworkers at the US 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*NN and random forest 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 of 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 and 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.

2.2.2. *k*NN—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, 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 to 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 leave-one-out cross-validated q^2_{abs} , i.e., q^2 with the intercept set to zero, possible 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), that places special emphasis on model validation. Briefly, we start by dividing the original dataset 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) that 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 leave-one-out cross-validated (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 assure their external predictive power as discussed in the Results section. The detailed discussion of the workflow used to develop validated QSAR models can be found in a recent review (19).

2.2.3. Random Forest—In machine learning, a random forest is a predictor that consists of many decision trees and outputs the prediction that combines outputs from individual trees. The algorithm for inducing a random forest was developed by Breiman and Cutler (20). In this study, the implementation of the random forest algorithm available in R.2.7.1 (21) was used. In the random forest 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.

2.2.4. 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 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 multi-linear 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 endpoint, IGC₅₀ (50% inhibitory concentration of population growth) of *Tetrahymena pyriformis* (15).

2.2.5. 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 thirty most similar chemicals from the training set as defined by the cosine similarity coefficient, $SC_{i,k}$, which is calculated as follows

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

where x_{ij} is the value of the j th normalized descriptor for chemical i (normalized with respect to all 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 equation 2. A multiple linear regression model is then built for the new cluster using a genetic algorithm based method and the toxicity is predicted.

2.2.6. 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 (Equation 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). In order to make a prediction, each of the neighbors in the training set must exceed a minimum cosine similarity coefficient of 0.5.

2.3. Identification of Outliers in the Dataset

A common problem for most QSAR studies is the existence of compounds that are highly dissimilar to all other compounds in the dataset. 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 random forest modeling procedures, we have developed a method to detect outliers that are dissimilar to other compounds of the dataset 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 dataset based on a predefined similarity threshold; 3) identifying those compounds that have no nearest neighbors as outliers.

In order 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 nearest neighbors (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 2,475 modeling set compounds were compiled as a new reduced modeling set to develop k NN and random forest toxicity models.

It is important to point out that the identification and exclusion of outliers is based only on consideration of chemical similarity but not activity. Thus, the removal of structural outliers could be regarded as a pre-treatment 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, e.g. 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).

2.4. Model Applicability Domains

Defining model Applicability Domains (AD) 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, since 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 applicability domain for each model should always be defined. As a consequence, only a certain fraction of compounds in any external dataset is expected to fall within the AD. This fraction is therefore referred to as the dataset 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.

2.4.1. Applicability Domain of k NN and Random Forest—The AD of k NN and Random Forest models is calculated from the distribution of similarities between each compound and its k nearest neighbors in the training set (similarities are computed as Euclidean distances between compounds represented by their multiple chemical descriptors). Based on the previous studies, the standard cutoff value to define the applicability domain 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 nearest neighbors in the same set. Thus, if the distance of the test compound from any of its k nearest neighbors 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).

2.4.2. Applicability Domain 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 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 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 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).

2.4.3. Applicability Domain 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).

2.4.4. Applicability Domain 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 which are sufficiently similar to the test chemical (the similarity coefficient between each chemical and the test chemical in equation 1 must exceed 0.5).

3. Results and Discussions

3.1. 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 and therefore the statistical results are shown for external datasets 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 datasets within the AD of each model. Models with the most liberally defined AD (and consequently, the highest coverage), i.e., NN and FDA MDL QSAR had the lowest R^2 and the highest MAE followed by TOPKAT and Hierarchical Clustering that 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, i.e., 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 dataset coverage of these two models is the lowest (19%) among all models. It could be argued that for this dataset (and perhaps for any large and diverse dataset) it is critical to define a rather restrictive AD in order to achieve most accurate predictions as discussed in more detail below.

3.2. Effect of the Model Applicability Domain

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 additional discussion of the consensus model below) that used the AD for three compound sets: 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.

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 endpoint. It is noticeable that the resulting predictive accuracy strongly correlates to the model coverage that is decided by the model applicability domain. 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 applicability domain was implemented. Since only a small number of compounds were out of the ADs of NN and FDA MDL QSAR models, the remaining two model applicability domains (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 applicability domain, the prediction coverage of the external prediction set obtained from each individual QSAR models are almost but not exactly the same. This is because there are some compounds (less than 1% of the total external compounds) which cannot be predicted using the Hierarchical clustering method even if all 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 approximately the same (in terms of R^2 and MAE) when using TOPKAT defined AD, with *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 increasingly 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 applicability domain. Again, as noted above, it could be concluded that the higher accuracy of prediction comes at the expense of reducing the chemical space coverage.

3.3. 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 steroid-like 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 ten 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 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.

3.4. 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 performance 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 applicability domains. Note that additional averaging schemes giving, e.g., 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, 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 applicability domain 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 applicability domain was implemented by introducing the concept of “consensus prediction fraction”. Since 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 5 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 (Figure 2). 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 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 *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 be achieved when the individual predictions are different. Table 5 lists ten 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, the compounds #2, #3, #6, #7, #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 affects to be incorporated into a single (and on average, more accurate) prediction.

4. 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 applicability domains) 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 *k*NN 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 applicability domain was implemented, the individual models showed similar performance as applied to the validation set. Here, the use of applicability domain 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 dataset are ongoing and will be reported in the future. All successful models reported in this paper will be made available via the ChemBench web portal (<http://chembench.mml.unc.edu>). Meanwhile, interested researchers can send us any compounds of interest for LD₅₀ prediction.

Acknowledgments

This work was supported, in part, by grants from NIH (GM076059 and ES005948) and US 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 US 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.

References

1. Putman, DL.; Clarke, JJ.; Escobar, P.; Gudi, R.; Krsmanovic, LS.; Pant, K.; Wagner, VO., III; San, RHC.; Jacobson-Kram, D. Genetic Toxicity. In: Jacobson-Kram, D.; Keller, KA., editors. Toxicological Testing Handbook. New York: Informa Healthcare; 2006. p. 185-248.
2. Hengstler JG, Foth H, Kahl R, Kramer PJ, Lilienblum W, Schulz T, Schweinfurth H. The REACH concept and its impact on toxicological sciences. *Toxicology* 2006;220(2-3):232-239. [PubMed: 16555380]
3. Richard AM. Future of toxicology--predictive toxicology: An expanded view of "chemical toxicity". *Chem. Res. Toxicol* 2006;19(10):1257-1262. [PubMed: 17040094]
4. Klopman G, Zhu H, Fuller MA, Saiakhov RD. Searching for an enhanced predictive tool for mutagenicity. *SAR QSAR Environ. Res* 2004;15(4):251-263. [PubMed: 15370416]
5. Richard AM, Benigni R. AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report. *SAR QSAR Environ. Res* 2002;13(1):1-19. [PubMed: 12074379]
6. Yang C, Benz RD, Cheeseman MA. Landscape of current toxicity databases and database standards. *Curr. Opin. Drug Discov. Devel* 2006;9(1):124-133.
7. Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, Hulzebos E, Marchant C, Richard A, Woo YT, Yang C. The expanding role of predictive toxicology: An update on the (Q) SAR models for mutagens and carcinogens. *Journal of Environmental Science and Health Part C-Environmental Carcinogenesis & Ecotoxicology Reviews* 2007;25(1):53-97.
8. Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweyko A, Li Y. In silico ADME/Tox: why models fail. *J. Comput. Aided Mol. Des* 2003;17(2-4):83-92. [PubMed: 13677477]
9. Enslein K, Craig PN. A toxicity estimation model. *Journal of Environmental Pathology and Toxicology* 1978;2(1):115-121. [PubMed: 722214]
10. Enslein, K.; Lander, TR.; Tomb, ME.; Craig, PN. A Predictive Model for Estimating Rat Oral LD₅₀ Values. Princeton: Princeton Scientific Publishers; 1983.
11. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J Chem. Inf. Model* 2008;48:766-784. [PubMed: 18311912]
12. National Library of Medicine. ChemIDplus database. 2008.
13. Young D, Martin T, Venkatapathy R, Harten P. Are the chemical structures in your QSAR correct? *QSAR Comb.Sci* 2008;27(11-12):1337-1345.
14. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH* 2006;56(2):237-248.
15. Martin TM, Harten P, Venkatapathy R, Das S, Young DM. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Method* 2008;18:251-266.
16. Zheng W, Tropsha A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci* 2000;40(1):185-194. [PubMed: 10661566]

17. Tropsha A, Golbraikh A. Predictive QSAR Modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des* 2007;13(34):3494–3504. [PubMed: 18220786]
18. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* 2003;17(2–4):241–253. [PubMed: 13677490]
19. Tropsha, A. Integrated chemo- and bioinformatics approaches to virtual screening. In: Varnek, A.; Tropsha, A., editors. *Chemoinformatics Approaches to Virtual Screening*. Cambridge, UK: Royal Society of Chemistry; 2008. p. 295-325.
20. Breiman L. Random forests. *Machine Learning* 2001;45(1):5–32.
21. Dalgaard, P. *Introductory Statistics with R*. Springer; 2008.
22. Romesburg, HC. *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications; 1984.
23. Contrera JF, Matthews EJ, Daniel BR. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol* 2003;38(3):243–259. [PubMed: 14623477]
24. Kutner, MH.; Nachtsheim, CJ.; Neter, J.; Li, W. *Applied Linear Statistical Models*. New York: McGraw-Hill; 2004.
25. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJ, Tong W, Veith G, Yang C. 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* 2005;33(2):155–173. [PubMed: 16180989]
26. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today* 2006;11(15–16):700–707. [PubMed: 16846797]
27. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem. Inf. Model* 2008;48(9):1733–1746. [PubMed: 18729318]
28. Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Quant. Struct. Act. Relat. Comb. Sci* 2003;22:69–77.
29. Montgomery, DC. *Introduction to linear regression analysis*. New York: John Wiley and Sons; 1982.
30. Lipnick RL, Pritzker CS, Bentley DL. A QSAR study of the rat LD₅₀ for alcohols. *QSAR Strategies Des. Bioact. Compd., Proc. Eur. Symp. Quant. Struct.-Act. Relat* 1985;5:420–423.
31. Enslein K, Tuzzeo TM, Borgstedt HH, Blake BW, Hart JB. Prediction of rat oral LD₅₀ from *Daphnia magna* LC₅₀ and chemical structure. *QSAR Environ. Toxicol., Proc. Int. Workshop* 1987;2:91–106.
32. Jaeckel H, Klein W. Prediction of mammalian toxicity by quantitative structure-activity relationships: aliphatic amines and anilines. *Quantitative Structure-Activity Relationships* 1991;10(3):198–204.
33. Zakarya D, Larfaoui EM, Boulaamail A, Lakhlifi T. Analysis of structure-toxicity relationships for a series of amide herbicides using statistical methods and neural network. *SAR QSAR Environ. Res* 1996;5(4):269–279. [PubMed: 9104783]
34. Wang G, Bai N. Structure-activity relationships for rat and mouse LD₅₀ of miscellaneous alcohols. *Chemosphere* 1998;36(7):1475–1483. [PubMed: 9503576]
35. Eldred DV, Jurs PC. Prediction of Acute Mammalian Toxicity of Organophosphorus Pesticide Compounds from Molecular Structure. *SAR & QSAR Environ. Res* 1999;10:75–99. [PubMed: 10491847]
36. Jean PA, Gallavan RH, Kolesar GB, Siddiqui WH, Oxley JA, Meeks RG. Chlorosilane acute inhalation toxicity and development of an LC₅₀ prediction model. *Inhal. Toxicol* 2006;18(8):515–522. [PubMed: 16717023]
37. Guo JX, Wu JJ, Wright JB, Lushington GH. Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: a molecular modeling study. *Chem. Res. Toxicol* 2006;19(2):209–216. [PubMed: 16485896]

38. Freidig AP, Dekkers S, Verwei M, Zvinavashe E, Bessems JG, van de Sandt JJ. Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds. *Toxicol. Lett* 2007;170(3):214–222. [PubMed: 17462838]
39. Toropov AA, Rasulev BF, Leszczynski J. QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by MLRA and optimal descriptors. *QSAR Comb. Sci* 2007;26(5):686–693.

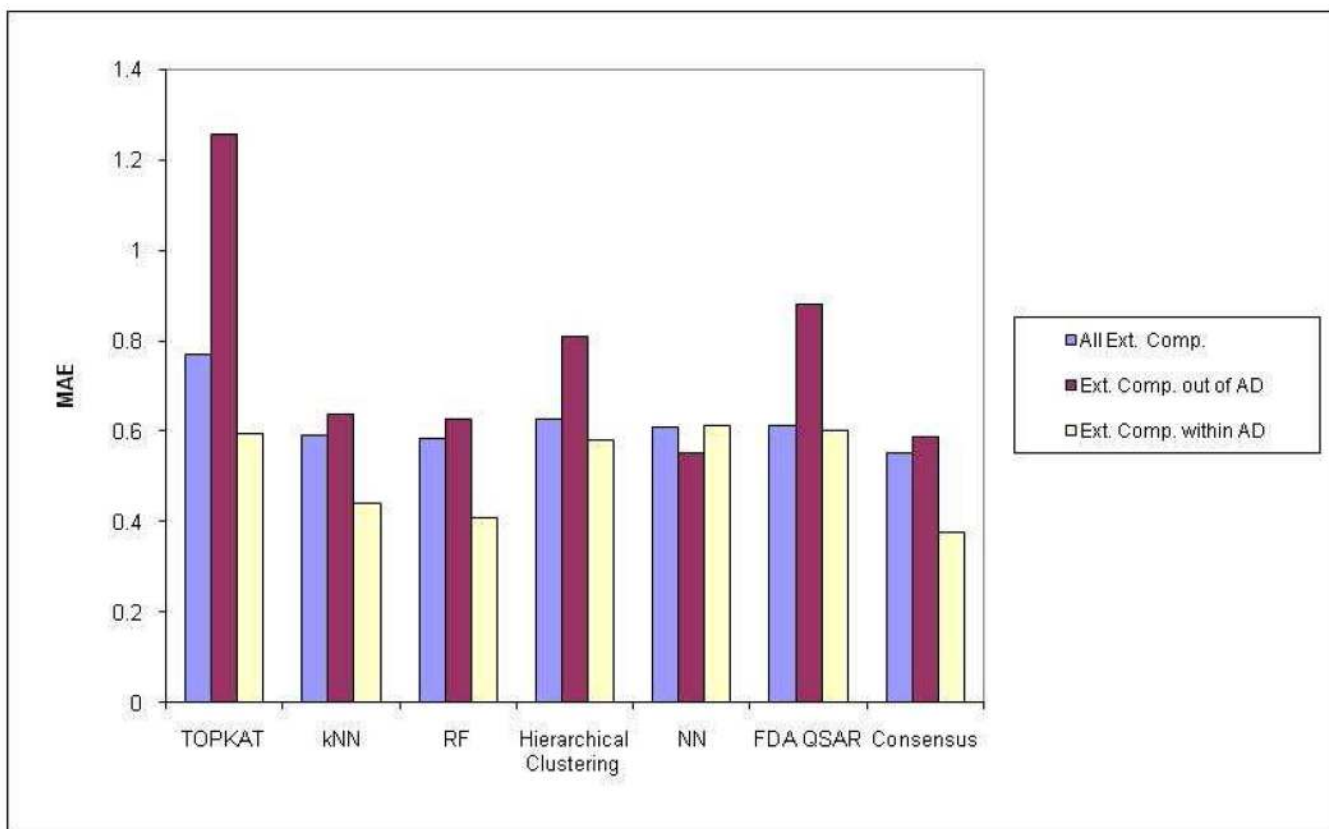


Figure 1. The MAEs of seven QSAR models for the external validation set. The AD of consensus model was defined when the 80% prediction fraction was applied (see text for additional discussion).

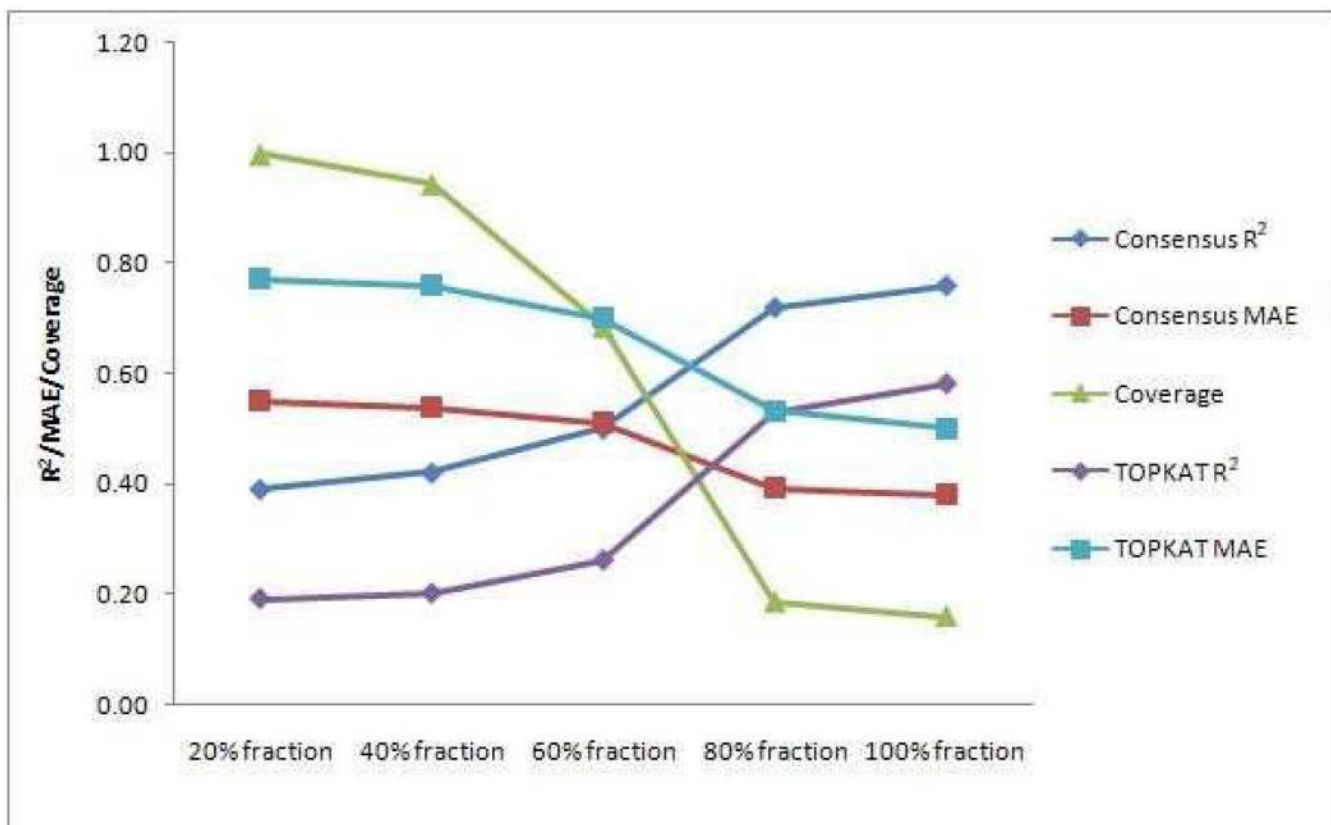


Figure 2. The prediction of external compounds by consensus model and TOPKAT with different consensus prediction fraction levels.

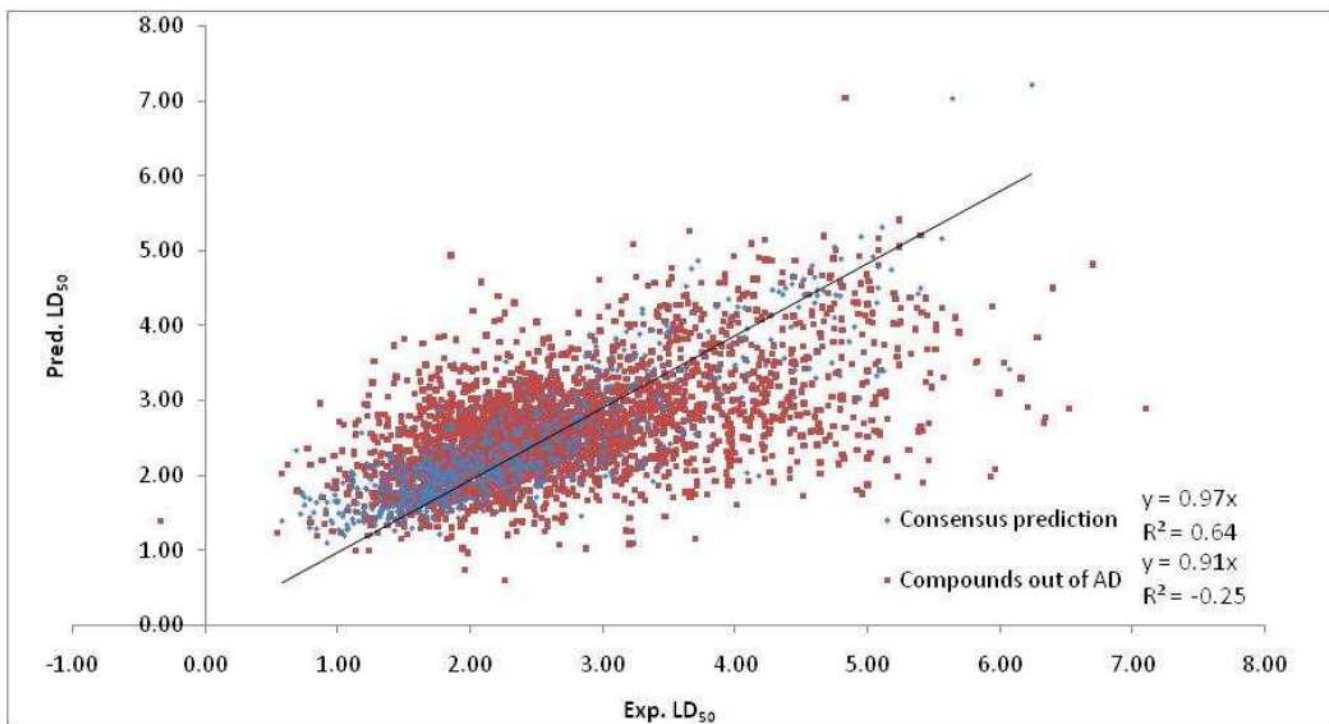


Figure 3. The correlation between experimental and consensus-predicted LD₅₀s when the consensus prediction fraction is 80% (i.e., compounds are within AD of four or more individual models).

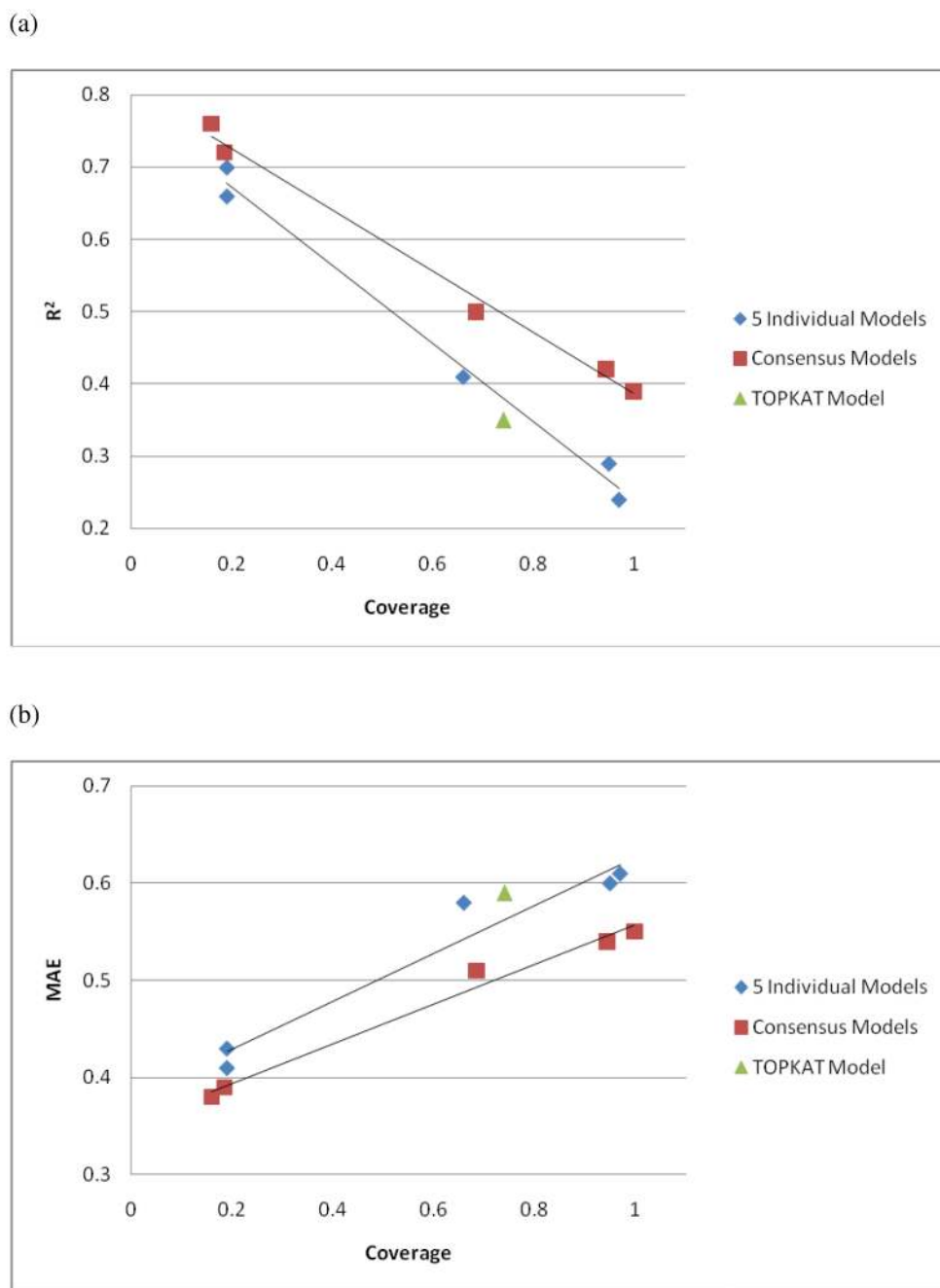


Figure 4. The relationship between prediction coverage and (a) R^2 or (b) MAE for the external compounds.

Table 1

The details of previous QSAR studies of acute rodent toxicity

Year	Source	Class(es) studied	N _{mod} ^a	Statistical method ^b	Validation set used?
1978	Enstein and Craig (9)	Multiple	425	MLR	Yes
1983	Enstein (10)	Multiple	1851	MLR	Yes
1985	Lipnick et al. (30)	Alcohols	68	BR	No
1987	Enstein et al. (31)	Multiple	147	MLR	No
1991	Jaeckel and Klein (32)	Amines and Anilines	26 / 33	MLR	No
1996	Zakarya et al. (33)	Amides	44	MLR/NN	No
1998	Wang and Bai (34)	Alcohols	95	E	Yes
1999	Eldred and Jurs (35)	Organophosphorus	49	MLR/NN	Yes
2006	Jean et al. (36)	Chlorosilanes	10	LR	No
2006	Guo et al. (37)	Organophosphorus	38	CoMFA	No
2007	Freidig et al. (38)	Multiple	49	LR	Yes
2007	Toropov et al. (39)	Substituted benzenes	28	MLR	Yes

^aSize of the modeling dataset.^bLR, linear regression; MLR, multilinear regression; NN, neural network; BR, bilinear regression; E, expert system; CoMFA, comparative molecular field analysis.

Table 2

Statistical results obtained with all QSAR models for the external validation set of 3913 compounds.

Models	R^2	MAE	Coverage (%)
kNN	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

Table 3

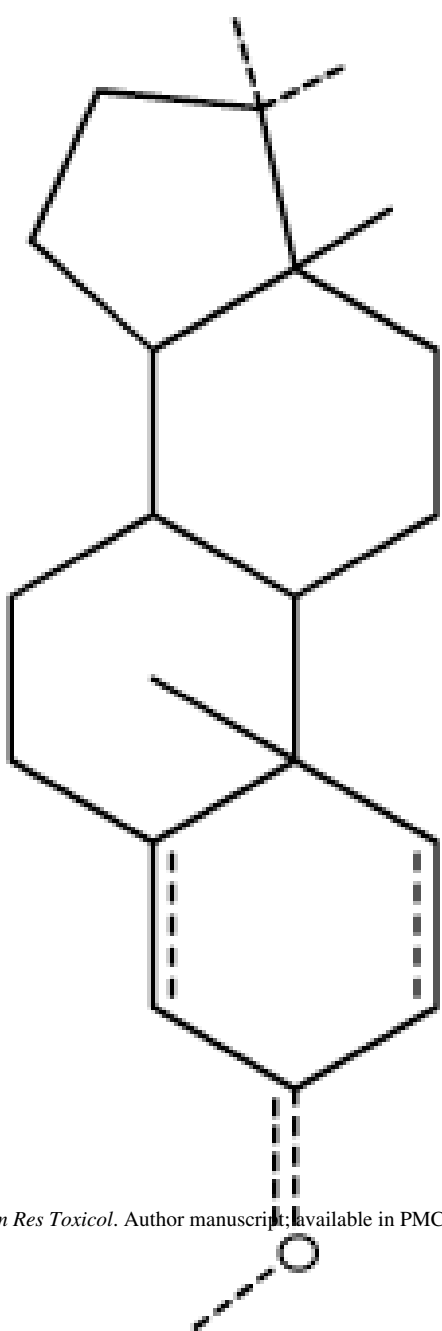
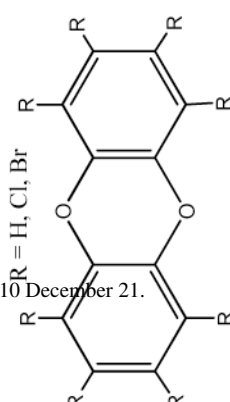
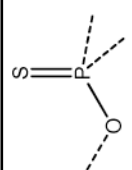
Statistical results obtained from all QSAR models for the external validation set of 3913 compounds.

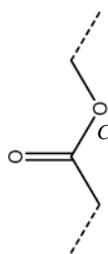
Models	TOPKAT AD (2896 compounds predicted, 74% coverage)		Hierarchical Clustering AD (2583 compounds predicted, 66% coverage)		kNN and RF AD (743 compounds predicted, 19% coverage)	
	R ²	MAE	R ²	MAE	R ²	MAE
kNN	0.41	0.55	0.40	0.57	0.66	0.44
RF	0.33	0.54	0.41	0.56	0.70	0.41
Hierarchical Clustering	0.33	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

^aTOPKAT 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
<p><i>Chem Res Toxicol.</i> Author manuscript; available in PMC 2010 December 21.</p> <p>R = H, Cl, Br</p>	3	8.2	5	5.1	5	5.1
	285	3.6	160	3.5	49	3.9

Scaffolds	Modeling set		Validation set		Validation set with large prediction errors	
	N ^a	Aver. LD ₅₀	N	Aver. LD ₅₀	N	Aver. LD ₅₀
 All compounds	83	2.0	124	2.4	15	2.9
	3,472	2.47	3,913	2.6	520	3.4

^aN is the number of compounds

Table 5

Experimental and predicted LD₅₀ values for ten external set compounds which have the most significant differences in predicted LD₅₀ values using individual models.

#	Compounds	Exp. ^a	RF	kNN	HC ^b	NN	FDA MDL QSAR	Cons. ^c	Aver. MAE	Cons. MAE
1	4H-1,3,2-Benzodioxaphosphorin-2-amine, N-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-, O,S-dipropyl ester	5.07	3.43	4.04	4.51	4.49	5.07	4.31	0.76	0.76
3	2-Butenenitrile	2.13	2.71	2.42	4.07	2.88	3.04	3.02	0.90	0.89
4	Phosphorothioic acid, O,O-diethyl S-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-, O,S-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	Mercarbamil	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-p-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

^aExperimental LD₅₀.

^bHC, Hierarchical Clustering.

^cCons., Consensus prediction