EPA/600/R-16/058
https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

User’s Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool)

A Program to Estimate Toxicity from Molecular Structure

User’s Guide for T.E.S.T.
(Toxicity Estimation Software Tool)

by

T. Martin
U.S. EPA/National Risk Management Research
Laboratory/Sustainable Technology Division, Cincinnati, OH
45268

Sustainable Technology Division
National Risk Management Research Laboratory
Cincinnati, Ohio, 45268

Notice/Disclaimer

The U.S. Environmental Protection Agency, through its Office of Research and Development, funded and conducted the research described herein under an approved Quality Assurance Project Plan (Quality Assurance Identification Number S-14987-QP-1-0). It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Foreword

The U.S. Environmental Protection Agency (US EPA) is charged by Congress with protecting the Nation's land, air, and water resources. Under a mandate of national environmental laws, the Agency strives to formulate and implement actions leading to a compatible balance between human activities and the ability of natural systems to support and nurture life. To meet this mandate, US EPA's research program is providing data and technical support for solving environmental problems today and building a science knowledge base necessary to manage our ecological resources wisely, understand how pollutants affect our health, and prevent or reduce environmental risks in the future.

The National Risk Management Research Laboratory (NRMRL) within the Office of Research and Development (ORD) is the Agency's center for investigation of technological and management approaches for preventing and reducing risks from pollution that threaten human health and the environment. The focus of the Laboratory's research program is on methods and their cost-effectiveness for prevention and control of pollution to air, land, water, and subsurface resources; protection of water quality in public water systems; remediation of contaminated sites, sediments and ground water; prevention and control of indoor air pollution; and restoration of ecosystems. NRMRL collaborates with both public and private sector partners to foster technologies that reduce the cost of compliance and to anticipate emerging problems. NRMRL's research provides solutions to environmental problems by: developing and promoting technologies that protect and improve the environment; advancing scientific and engineering information to support regulatory and policy decisions; and providing the technical support and information transfer to ensure implementation of environmental regulations and strategies at the national, state, and community levels.

Abstract

This guide provides an introduction into QSAR (Quantitative Structure Activity Relationship) models, a detailed description of the QSAR methodologies in TEST, a description of the experimental datasets, a detailed analysis of the validation results for the external test sets, and step-by-step instructions for using the software.

Table of Contents

1. Introduction

Quantitative Structure Activity Relationships (QSARs) are mathematical models that are used to predict measures of toxicity from physical characteristics of the structure of chemicals (known as molecular descriptors). Acute toxicities (such as the concentration, which causes half of fish to die) are one example of toxicity measures, which may be predicted from QSARs. Simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

$$Toxicity = ax_1 + bx_2 + c$$

where x_1 and x_2 are the independent descriptor variables and *a*, *b*, and *c* are fitted parameters. The molecular weight and the octanol-water partition coefficient are examples of molecular descriptors.

QSAR toxicity predictions may be used to screen untested compounds in order to establish priorities for expensive and time-consuming traditional bioassays designed to establish toxicity levels. When conditions do not permit traditional bioassays, QSARs are an alternative to bioassays for estimating toxicity. In addition, QSAR models are useful for estimating toxicities needed for green process design algorithms such as the Waste Reduction Algorithm [1].

The Toxicity Estimation Software Tool (T.E.S.T.) has been developed to allow users to easily estimate toxicity using a variety of QSAR methodologies. T.E.S.T allows a user to estimate toxicity without requiring any external programs. Users can input a chemical to be evaluated by drawing it in an included chemical sketcher window, entering a structure text file, or importing it from an included database of structures. Once a chemical has been entered, its toxicity can be estimated using one of several advanced QSAR methodologies. The program does not require molecular descriptors from external software packages (the required descriptors are calculated within T.E.S.T.).

1.1. Toxicity Endpoints

T.E.S.T allows you to estimate the value for several toxicity end points:

1. 96 hour fathead minnow LC_{50} (concentration of the test chemical in water in mg/L that causes 50% of fathead minnow to die after 96 hours)
2. 48 hour *Daphnia magna* LC_{50} (concentration of the test chemical in water in mg/L that causes 50% of *Daphnia magna* to die after 48 hours)
3. 48 hour *Tetrahymena pyriformis* IGC_{50} (concentration of the test chemical in water in mg/L that causes 50% growth inhibition to *Tetrahymena pyriformis* after 48 hours)
4. Oral rat LD_{50} (amount of chemical in mg/kg body weight that causes 50% of rats to die after oral ingestion)
5. Bioaccumulation factor (ratio of the chemical concentration in fish as a result of absorption via the respiratory surface to that in water at steady state)
6. Developmental toxicity (whether or not a chemical causes developmental toxicity

effects to humans or animals)

7. Ames mutagenicity (a compound is positive for mutagenicity if it induces revertant colony growth in any strain of *Salmonella typhimurium*)

T.E.S.T. allows you estimate several physical properties:

1. Normal boiling point (the temperature in °C at which a chemical boils at atmospheric pressure)
2. Density (the density in g/cm^3)
3. Flash point (the lowest temperature in °C at which it can vaporize to form an ignitable mixture in air)
4. Thermal conductivity (the property of a material in units of mW/mK reflecting its ability to conduct heat)
5. Viscosity (a measure of the resistance of a fluid to flow in cP defined as the proportionality constant between shear rate and shear stress)
6. Surface tension (a property of the surface in dyn/cm of a liquid that allows it to resist an external force)
7. Water solubility (the amount of a chemical in mg/L that will dissolve in liquid water to form a homogeneous solution)
8. Vapor pressure (the pressure of a vapor in mmHg in thermodynamic equilibrium with its condensed phases in a closed system)
9. Melting point (the temperature in °C at which a chemical in the solid state changes to a liquid state)

1.2. QSAR Methodologies

T.E.S.T allows you to estimate toxicity values using several different advanced QSAR methodologies [2]:

- **Hierarchical method:** The toxicity for a given query compound is estimated using the weighted average of the predictions from several different models. The different models are obtained by using Ward’s method to divide the training set into a series of structurally similar clusters. A genetic algorithm based technique is used to generate models for each cluster. The models are generated prior to runtime.
- **FDA method**: The prediction for each test chemical is made using a new model that is fit to the chemicals that are most similar to the test compound. Each model is generated at runtime.
- **Single model method**: Predictions are made using a multilinear regression model that is fit to the training set (using molecular descriptors as independent variables) using a genetic algorithm based approach. The regression model is generated prior to runtime.
- **Group contribution method**: Predictions are made using a multilinear regression model that is fit to the training set (using molecular fragment counts as independent variables). The regression model is generated prior to runtime.
- **Nearest neighbor method**: The predicted toxicity is estimated by taking an average of the 3 chemicals in the training set that are most similar to the test

chemical.

- **Consensus method**: The predicted toxicity is estimated by taking an average of the predicted toxicities from the above QSAR methods (provided the predictions are within the respective applicability domains).
- **Random forest method:** The predicted toxicity is estimated using a decision tree which bins a chemical into a certain toxicity score (i.e. positive or negative developmental toxicity) using a set of molecular descriptors as decision variables. *The random forest method is currently only available for the developmental toxicity endpoint*. The random forest models for the developmental toxicity endpoint were developed by researchers at Mario Negri Institute for Pharmacological Research as part of the CAESAR project [3].
- **Mode of action method:** The predicted toxicity is estimated using a two-step process. In the first step the mode of action is determined from the linear discriminant analysis model with the highest score. In the second step the toxicity is estimated using the multilinear regression model corresponding to the predicted mode of action. *The mode of action method is currently only available for the 96 hour fathead minnow LC_{50} endpoint.*

T.E.S.T provides multiple prediction methodologies so one can have greater confidence in the predicted toxicities (assuming the predicted toxicities are similar from different methods). In addition, some researchers may have more confidence in particular QSAR approaches based on personal experience. The QSAR methodologies above are described in more detail in the Theory section. The advantages and disadvantages of the different QSAR methods are given in Table 1.2.

Table 1.2. Advantages and disadvantages of the QSAR methods in T.E.S.T.

Method	**Advantages**	**Disadvantages**
Hierarchical	• Can produce more reliable predictions since predictions are made from multiple models	• Cannot provide external estimates of toxicity for compounds in the training set
Single model	• Single transparent model can be easily viewed/exported • The model does not need to rely on clustering the chemicals correctly	• Since the model is fit to the entire dataset it may incorrectly predict the trends in toxicity for certain chemical classes • Cannot provide external estimates of toxicity for compounds in the training set
Group contribution	• Single transparent model can be easily viewed/exported • Estimates of toxicity can be made without using a computer program	• The model doesn't correct for the interactions of adjacent fragments • Since the model is fit to the entire dataset it may incorrectly predict the trends in toxicity for certain chemical classes • Cannot provide external estimates of toxicity for compounds in the training set
FDA	• Can generate a new model based the closest analogs to the test compound • Always provides an external prediction of toxicity	• Predictions sometimes take longer since it has to generate a new model each time
Nearest neighbor	• Provides a quick estimate of toxicity • Allows one to determine structural analogs for a given test compound • Always provide an external prediction of toxicity	• It does not use a QSAR model to correlate the differences between the test compound and the nearest neighbors • Was shown to achieve the worst prediction results during external validation
MOA	• Provides a more biologically relevant estimate of acute aquatic toxicity which provides greater confidence in the prediction for toxicologists	• Size of the training set is reduced • Prediction error may be compounded by the fact that the mode of action must be predicted correctly
Consensus	• Was shown to achieve the best prediction results during external validation	• Cannot provide external estimates of toxicity for compounds in the training set

2. THEORY

2.1. Molecular Descriptors

Molecular descriptors are physical characteristics of the structure of chemicals such as the molecular weight or the number of benzene rings. The overall pool of descriptors in the software contains 797 2-dimensional descriptors. The descriptors include the following classes of descriptors: E-state values and E-state counts, constitutional descriptors, topological descriptors, walk and path counts, connectivity, information content, 2d autocorrelation, Burden eigenvalue, molecular property (such as the octanol-water partition coefficient), Kappa, hydrogen bond acceptor/donor counts, molecular distance edge, and molecular fragment counts. *The complete list of descriptors and their sources from the literature are described in the Molecular Descriptors Guide*.

The descriptors were calculated using computer code written in Java. The basis of the molecular calculations was the Chemistry Development Kit [4]. The Chemistry Development Kit (CDK) is a Java library for structural chemo- and bioinformatics [5]. The descriptor values were validated using MDL QSAR [6], Dragon [7], and Molconn-z [8]. The descriptor values were generally in good agreement (aside from small differences in the descriptor definitions for descriptors such as the number of hydrogen bond acceptors).

2.2. QSAR Methodologies

2.2.1. Hierarchical Clustering

The hierarchical clustering method utilizes a variation of Ward's Method [9] to produce a series of clusters from the training set. Clusters are subsets of chemicals from the overall set, which possess similar properties. An example of a hierarchical clustering for a hypothetical training set with five chemicals is given in Figure 2.2.1.

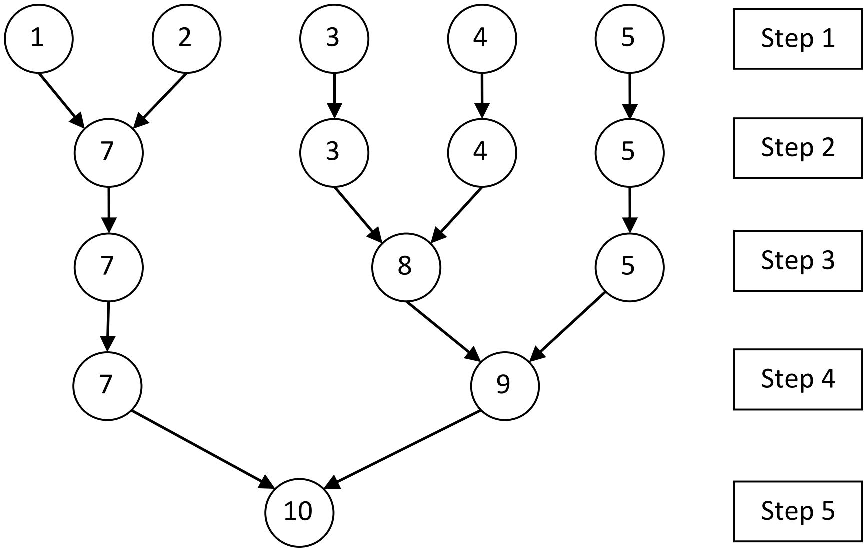


Figure 2.1.1. Hierarchical clustering with five chemicals

For a training set of *n* chemicals, initially there will be *n* clusters (each cluster contains one chemical). The overall variance in the system at a given step *l* is defined to be the sum of the variances of the individual clusters:

$$V(l) \equiv \sum_{k=1}^{m} v(k,l) \qquad (1)$$

where $v(k,l)$ is the variance (in terms of the molecular descriptors) for cluster *k* at step *l*:

$$v(k,l) \equiv \sum_{i=1}^{n_k} \sum_{j=1}^{d} \left(x_{ij} - C_j\right)^2 \qquad (2)$$

where n_k is the number of chemicals in the *k*th cluster, *d* is the number of descriptors in the overall descriptor pool, x_{ij} is the normalized descriptor *j* for chemical *i*, and C_j is the centroid or average value for descriptor *j* for cluster *k*:

$$C_j = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij} \qquad (3)$$

Each step of the method adds two of the clusters together into one cluster so that the increase in variance over all clusters in the system is minimized:

$$\min \Delta V(l+1) \equiv V(l+1) - V(l) = v(k', l+1) - v(k_1, l) - v(k_2, l) \qquad (4)$$

where clusters k_1 and k_2 join together at step l to make cluster k' at step $l+1$. The process of combining clusters 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 can be developed. Each cluster undergoes evaluation using a genetic algorithm technique to determine an optimal descriptor set for characterizing the toxicity values of the chemicals within that cluster. The maximum number of descriptors allowed for a given cluster will be $n_k / 5$ because the recommended ratio of compounds to variables should be at least 5 [10, 11] for reasonably small probability for chance correlations. The genetic algorithm used in this study was taken from the Weka statistical package, version 3.5.1 [12, 13].

The genetic algorithm is used to maximize the adjusted fivefold leave many out cross validation coefficient ($q^2_{adj,LMO}$):

$$q^2_{adj,LMO} = 1 - \left[\frac{\sum_{i=1}^{n_k} \left(\hat{y}_i - y_{\exp,i}\right)^2 / \left(n_k - p - 1\right)}{\sum_{i=1}^{n_k} \left(y_{\exp,i} - \bar{y}_{\exp}\right)^2 / \left(n_k - 1\right)} \right] \qquad (5)$$

where $\hat{y}_i$ and $y_{\exp,i}$ are the predicted and experimental toxicity values for chemical *i*, $\bar{y}_{\exp}$ is the average experimental toxicity for the chemicals in the cluster, and *p* is the number of parameters in the model. The predicted toxicity values are calculated by dividing the dataset into five folds (a fold is a subset of the training set). The toxicities of the chemicals in each fold ($\hat{y}_i$) are predicted using a multiple linear regression model fit to the chemicals in the other folds. The five fold q^2 was used instead of the traditional q^2 LOO (leave-one-out) inside the genetic algorithm because it yields a significant degree of computational savings for large cluster sizes. The $n_k - p - 1$ term penalizes models that include extra parameters that do not significantly increase the predictive power of the model (by decreasing the value of $q^2_{adj,LMO}$).

During the optimization process the models are checked for outliers. A chemical is determined to be an outlier if at least two statistical tests (e.g., DFFITS, leverage, Cook’s distance, and covariance ratio) indicate that the chemical represents an influential data point and if the chemical represents an outlier in terms of the studentized deleted residual [14]. If a chemical is determined to be an outlier, the chemical is deleted from the cluster and the genetic algorithm descriptor selection is repeated. The process of model building via the genetic algorithm and outlier removal is repeated until no outliers are detected in the optimized model. *For binary endpoints such as Ames mutagenicity, outliers were not removed because this had the potential to produce clusters with all positive or all negative chemicals. In addition the outlier statistical tests described above may not apply to binary endpoints.*

Once the iteration for the optimum model has been completed, the q^2 LOO value for the model is calculated. If the q^2 LOO is greater than or equal to 0.5, the model is considered to be valid (see pg 67 of Erikkson et al. [15]). If the q^2 LOO is less than 0.5, the model from the cluster is not used to make predictions for test compounds. For binary endpoints, the validity of a model is determined from the concordance LOO instead of q^2 LOO. Concordance is the fraction of all compounds that are predicted correctly (i.e., experimentally active compounds that are predicted to be active and experimentally inactive compounds that are predicted to be inactive). If the concordance LOO is greater than or equal to 0.8, the model is considered to be valid. In addition both the leave-one-out sensitivity and specificity must be at least 0.5 to avoid using models which are heavily biased to predict either active or inactive scores. Sensitivity is the fraction of experimentally active compounds that are predicted to be active. Specificity is the fraction of experimentally inactive compounds that are predicted to be inactive.

The predicted toxicity ($\hat{y}$) for a test chemical is given by the weighted average for all the valid predictions [16]:

$$\hat{y} = \frac{\sum_{j=1}^{nvc} w_j \hat{y}_j}{\sum_{j=1}^{\#valid\,clusters} w_j} \tag{6}$$

where $\hat{y}_j$ and w_j are prediction and weight for the *j*th model and *nvc* is the number of valid cluster model predictions. If the mean toxicity is given by the maximum likelihood estimator of the mean of the probability distributions, the weight values are given by [16]

$$w_j = \frac{1}{se_j^{\ 2}} \tag{7}$$

where se_j is the standard error for the *j*th prediction given by

$$se_j = \sqrt{\sigma_j^{\ 2}(1 + h_{00})} \tag{8}$$

where $\sigma_j^{\ 2}$ is given by

$$\sigma_j^{\ 2} = \frac{\sum_{i=1}^{n_j} (\hat{y}_i - y_{\exp,i})^2}{n_j - p_j - 1} \tag{9}$$

where n_j is the number of chemicals in cluster model *j* and p_j is the number of model parameters for model *j*. h_{00}, the leverage for the test chemical, is given by

$$h_{00} = X_o^T (X^T X)^{-1} X_0 \tag{10}$$

where X_0 is the vector of model descriptor values for the test compound. *For binary endpoints such as Ames mutagenicity, the predictions were made using equal weighting of the individual predictions (i.e. $w_j = 1$ in equation 6) because weighting by the standard error (see equation 7) did not improve the external prediction accuracy.*

The square of the standard deviation for the prediction from multiple models ($\sigma_\mu^{\ 2}$) can be approximated as

$$\sigma_{\mu}{}^{2} = \frac{\overline{\sigma^2}}{nvc} = \left(1\big/_{nvc}\right)\frac{\sum_{j=1}^{nvc} w_j se_j{}^2}{\sum_{j=1}^{nvc} w_j} = \left(1\big/_{nvc}\right)\frac{\sum_{j=1}^{nvc}\left(1\big/_{se_j{}^2}\right)se_j{}^2}{\sum_{j=1}^{nvc}\left(1\big/_{se_j{}^2}\right)} = \frac{1}{\sum_{j=1}^{nvc}\left(1\big/_{se_j{}^2}\right)} \tag{11}$$

The uncertainty ($\hat{u}$) in the overall prediction for the test chemical is given by

$$\hat{u} = t_{1-\alpha/2,nvc}\sigma_{\mu} = t_{1-\alpha/2,\,nvc-1}\sqrt{1/\sum_{j=1}^{nvc}\frac{1}{se_j{}^2}} \tag{12}$$

where t is the t-statistic, α = 0.1 (90% confidence interval), and se_j is the standard error for the jth prediction. The prediction interval is obtained by adding and subtracting the uncertainty from the predicted toxicity:

$$\hat{y} - \hat{u} \le Toxicity \le y + \hat{u} \tag{13}$$

The prediction interval indicates that one is 90% confident that the actual toxicity is between $\hat{y} - \hat{u}$ and $\hat{y} + \hat{u}$.

The prediction uncertainty for a given cluster model is given by [17]

$$u_j = t_{1\text{-}\alpha/2,n_j-p-1}\sqrt{\sigma^2(1+h_{00})} \tag{14}$$

The uncertainty is a function of the quality of the regression model (from the σ^2 parameter) and the distance (in the descriptor space of the model) between the test chemical and the chemicals in the cluster used to build the model (from the h_{00} parameter).

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 domain of applicability for the model. The applicability domain is defined using several different constraints. 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 [17]. The second constraint, the Rmax 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 distance is defined in terms of the entire pool of descriptors (instead of just the descriptors appearing in the model):

$$distance_i = \sum_{j=1}^{d}\left(x_{ij} - C_j\right)^2 \tag{15}$$

where $distance_i$ is the distance of chemical i to the centroid of the cluster.
The last constraint, the fragment constraint, is that the compounds in the cluster have to have at least one example of each of the fragments contained in the test chemical. For example if one was trying to make a prediction for ethanol, the cluster must contain at least one compound with a methyl fragment (-CH_3 [aliphatic attach]), one compound with a methylene fragment (-CH_2 [aliphatic attach]), and one compound with a hydroxyl fragment (-OH [aliphatic attach]). This constraint was added to

avoid situations where a chemical might have a similar backbone structure to the chemicals in a given cluster but has a different functional group attached. For example if a given cluster contained only short-chained aliphatic amines one would not want to use it to predict the toxicity of ethanol. If a chemical contains a fragment that is not present in the training set, the toxicity cannot be predicted. The fragment constraint can be removed by checking the **Relax fragment constraint** checkbox. *For binary endpoints such as Ames mutagenicity, the fragment constraint was not employed since it did not improve the external prediction accuracy and decreased the prediction coverage.*

In the current version of the software, the predictions are made using the closest cluster from each step in the hierarchical clustering (in terms of the distance of the chemical to the centroid of the cluster defined above). The rationale behind this approach is that one would like to follow the hierarchical clustering process, selecting the best model from each step. In order for the prediction from the model to be used it must be statistically valid and meet the constraints defined above. If the closest cluster for a given step does not have a statistically valid model (or violates any of the constraints), no prediction is used from that step. If the closest cluster for a given step in the clustering process is the same as the closest cluster from a previous step, it is not used again in the prediction of toxicity.

2.2.2. FDA Method

The Food and Drug Administration (FDA) method is based on the work of Contrera and coworkers [18]. In this method, predictions for each test chemical are made using a unique cluster (constructed at runtime) which contains structurally similar chemicals selected from the overall training set. This is in contrast to the Hierarchical method, where the predictions are made using one or more clusters that were constructed *a priori* using Ward's method.

Contrera and coworkers constructed the training cluster by selecting 15-20 chemicals, which had at least a cosine similarity coefficient of 75% with the test chemical. The cosine similarity coefficient, $SC_{i,k}$, is given by

$$SC_{i,k} = \frac{\sum_{j=1}^{\#descriptors} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{\#descriptors} x_{ij}^{2} \cdot \sum_{j=1}^{\#descriptors} x_{kj}^{2}}} \tag{16}$$

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*. A multiple linear regression model is then built for the new cluster using a genetic algorithm and the toxicity is predicted. The advantage of this method is that the training cluster is tailored to fit the test chemical. In addition, the test chemical is never present in the cluster model, which allows one to make external predictions for training set chemicals. The disadvantage of this method is that a new model has to be generated at runtime (which takes somewhat longer than computing the toxicity from preexisting models).

In this version of the software, clusters are constructed using the 30 most similar chemicals from the training set in terms of the cosine similarity coefficient. However, a minimum similarity coefficient of 75% is not required for membership in the training cluster. Previously, it was determined that this constraint did not increase the predictive performance of the methodology [2]. For a prediction to be valid, the cluster must not violate the model ellipsoid and fragment constraints described above. In addition, the predicted toxicity value must be within the range of experimental toxicity values for the chemicals used to build the model. This additional constraint was added to avoid potentially erroneous predictions. *However this constraint was not utilized for binary toxicity endpoints such as Ames mutagenicity since predicted values less than 0 or greater than 1 do not invalidate the prediction result.*

Again, for a cluster to have a valid predictive model, the LOO q^2 must be at least 0.5. If the model for the cluster is invalid or the prediction violates one of the constraints, the cluster size is increased incrementally (up to a maximum of 75 chemicals) until a valid prediction can be made. If a prediction cannot be made using a cluster with 75 chemicals, no prediction is made.

2.2.3. Single model

In the single model approach, a single multiple linear regression model is fit to the entire training set. The model is generated using techniques and constraints similar to those for the hierarchical method (except that the training cluster contains the entire training set). The advantage of this approach is that a simple transparent model can be developed which does not rely on clustering the chemicals correctly. The disadvantage of this approach is that sometimes an overall model cannot correctly correlate the toxicity for every chemical class [19]. For example the single model might be able to correctly describe the trend of linearly increasing toxicity for a series of normal alcohols (i.e. 1-propanol, 1-butanol,1-pentanol, …), but it may incorrectly describe the trend for a series of normal acids (i.e. propanoic acid, butanoic acid, pentanoic acid, …) that does not increase linearly.

2.2.4. Group contribution

The group contribution approach is based on the group contribution approach of Martin and Young [20]. Fragment counts (such as the number of methyl and hydroxyl groups in a compound) are used to fit a multiple linear regression model to the entire data set. A genetic algorithm approach is not used to reduce the number of parameters in the model because the approach tries to characterize the contribution from all the fragments appearing in the training set. The only constraint on the fragments appearing in the final model is that there must be at least three molecules in the training set that contain each fragment. If a fragment appears less than three times in the training set, it is deleted from the list of fragments and all the chemicals containing this fragment are removed from the training set. After the multiple linear regression is performed, the model is checked for outliers. If outliers are detected, they are removed and the regression is performed again. The process is repeated until no more outliers are found. Similar to the hierarchical methodology, predictions are made using the model ellipse and fragment constraints.

The advantage of this approach is a single transparent model can be developed whose descriptors can be determined from visual inspection of the molecular structure of the test compound. The disadvantage of this approach is that it assumes that the contribution of each fragment does not depend on the presence of nearby fragments in the molecule.

2.2.5. Nearest neighbor

In the nearest neighbor approach, the predicted toxicity is simply the average of the toxicities of the three most similar chemicals (structural analogs) in the training set. In order to make a prediction, each of the structural analogs must exceed a certain minimum cosine similarity coefficient (SCmin). SCmin was set at 0.5 so that the prediction coverage was similar to the other QSAR methods [2]. The nearest neighbor method provides a quick external estimate of toxicity (the test chemical is never present in the selected set of analogs). The disadvantage of the nearest neighbor method is that the structural differences between the test chemical and its structural analogs are not accounted for.

2.2.6. Mode of action

In the mode of action (MOA) method, the toxicity is predicted using a two-step process [21, 22]. In the first step, the MOA is predicted using a series of linear discriminant analysis (LDA) models. The predicted MOA is given by the LDA model, which yields the highest score. In order for a predicted MOA to be valid, the maximum score must be at least 0.5. In addition, the model ellipsoid and Rmax constraints must be satisfied. In the second step, the toxicity is predicted using the multilinear regression model, which corresponds to the predicted MOA. Again, the model ellipsoid and Rmax constraints must be satisfied for the toxicity model for a prediction to be within the domain of applicability. *The fragment constraint is not employed for the MOA method*. The advantage of the MOA method is that it provides a more biologically relevant estimate of acute aquatic toxicity, which can greater confidence in the prediction for toxicologists. The disadvantages of this method are that the size of the training set is reduced (which reduces the chemical space covered by the model) and that the prediction error may be compounded by the fact that the mode of action must be predicted correctly.

2.2.7. Consensus

In the consensus method, the predicted toxicity is simply the average of the predicted toxicities from the other QSAR methodologies (taking into account the applicability domain of each method)[23]. If only a single QSAR methodology can make a prediction, the predicted value is deemed unreliable and not used. This method typically provides the highest prediction accuracy since errant predictions are dampened by the predictions from the other methods. In addition, this method provides the highest prediction coverage because several methods with slightly different applicability domains are used to make a prediction.

2.3. Validation Methods

2.3.1. Statistical external validation

The predictive ability of each of the QSAR methodologies was evaluated using statistical external validation [24]. In version 2.0 of the TEST software, the data set was divided into training and test sets using the Kennard-Stone rational design algorithm [25-28]. Starting in version 3.0, random selection was used to develop the training and test sets because it was felt that using Kennard-Stone method yields an overly optimistic estimate of predictive ability (because the test compounds are always within the model calibration domain). *For the developmental toxicity endpoint, however, the training and test sets were taken from the datasets used in CAESAR* [3]. This was done so that the CAESAR random forest model could be incorporated into the TEST software.

A QSAR model has acceptable predictive power if the following conditions are satisfied [29]:

$$q^2 > 0.5; \tag{17}$$

$$R^2 > 0.6; \tag{18}$$

$$\frac{\left(R^2 - R_o^2\right)}{R^2} < 0.1 \text{ and } 0.85 \leq \text{k} \leq 1.15 \tag{19}$$

where q^2 is the leave one out correlation coefficient for the training set, R^2 is correlation coefficient between the observed and predicted toxicities for the test set, R_o^2 is correlation coefficient between the observed and predicted toxicities for the test set with the Y-intercept set to zero (where the regression line is given by Y=kX).

The prediction accuracy will be evaluated in terms of equations 18 and 19. In addition the accuracy will be evaluated in terms of the RMSE (root mean square error), and the MAE (mean absolute error) for the test set. It has been demonstrated that q^2 (the leave one out correlation coefficient for the training set) is not correlated with R^2 for the test set [30]. The prediction coverage (fraction of chemicals predicted) must be considered because the prediction accuracy (in terms of R^2 and RMSE) can sometimes be improved at the sacrifice of the prediction coverage.

For binary (active/inactive) toxicity endpoints such as developmental toxicity, the prediction accuracy is evaluated in terms of the fraction of compounds that are predicted accurately. The prediction accuracy is evaluated in terms of three different statistics: concordance, sensitivity, and specificity. Concordance is the fraction of all compounds that are predicted correctly (i.e. experimentally active compounds that are predicted to be active and experimentally inactive compounds that are predicted to be inactive). Sensitivity is the fraction of experimentally active compounds that are predicted to be active. Specificity is the fraction of experimentally inactive compounds that are predicted to be inactive.

3. EXPERIMENTAL DATA SETS

3.1. 96 hour fathead minnow LC_{50} data set

The fathead minnow LC_{50} endpoint represents the concentration in water, which kills half of fathead minnow (*Pimephales promelas*) in 4 days (96 hours). The data set for this endpoint was obtained by downloading the ECOTOX aquatic toxicity database[31].

The database was then filtered using the following criteria:

- The ECOTOX “Media Type” field = “FW” (fresh water)
- The ECOTOX “Test Location” field = “Lab” (laboratory)
- The ECOTOX “Conc 1 Op (ug/L)” field cannot be <, >, or ~ (i.e. use only discrete LC_{50} values)
- The ECOTOX “Effect” field = “Mor” (mortality)
- The ECOTOX “Effect Measurement” field = “MORT” (mortality)
- The ECOTOX “Exposure Duration” field = “4” (4 days or 96 hours)
- Compounds can only contain the following element symbols: C, H, O, N, F, Cl, Br, I, S, P, Si, As
- Compounds must represent a single pure component (i.e. salts, undefined isomeric mixtures, polymers, or mixtures were removed)

The LC_{50} values were taken from the “Conc 1 (ug/L)” field in ECOTOX. For chemicals with multiple LC_{50} values, the median value was used.

In version 2.0 of T.E.S.T., 10 compounds in this dataset possessed 2d isomers (the structures were equivalent in terms of their molecular connectivity). In version 3.0, only one isomer was kept, using the average toxicity value. In version 4.0, all isomers were kept since the presence of the isomers had negligible impact on the external prediction statistics. The final fathead minnow LC_{50} data set contained 823 chemicals. For use in QSAR modeling, the experimental values in μg/L were converted to –Log_{10} (LC_{50} mol/L).
For the hierarchical, single model, group contribution, FDA, Nearest neighbor, and Consensus methods, the data set were divided randomly into a training set (80% of the overall set) and a test set (20% of the overall set). *For the mode of action method, chemicals with a known MOA (372 chemicals) were placed in the training set while the remaining chemicals (440 chemicals) were placed in the test set[22]. Thus, the results for the mode of action method will have to be considered separately*.

3.2. 48 hour *Daphnia magna* LC_{50} data set

The *Daphnia magna* LC_{50} endpoint represents the concentration in water, which kills half of *D. magna* (a water flea) in 48 hours. The data set for this endpoint was obtained from the ECOTOX aquatic toxicity database[31]. The database was filtered using the same criteria as those for the 96 hour fathead minnow LC_{50}. The final *D. magna* LC_{50} data set contained **353** chemicals. The modeled endpoint was $-\text{Log}_{10}$ (LC_{50} mol/L).

3.3. 40 hour *Tetrahymena pyriformis* IGC_{50} data set

The *Tetrahymena pyriformis* IGC_{50} endpoint represents the 50% growth inhibitory concentration of the *T. pyriformis* organism (a protozoan ciliate) after 40 hours. The IGC_{50} training set was obtained from Schultz and coworkers [23, 32-69]. The final *T. pyriformis IGC*$_{50}$ data set contained 1792 chemicals. The modeled endpoint was $-\text{Log}_{10}$ (IGC_{50} mol/L).

3.4. Oral rat LD_{50} data set

The oral rat LD_{50} endpoint represents the amount of the chemical (mass of the chemical per body weight of the rat) which when orally ingested kills half of rats. The dataset for this endpoint was obtained by downloading records from the ChemIDplus database [70]. 13548 records were obtained by using the following search criteria:

- "Test" = LD50
- "Species" = rat
- "Route" = oral

The list of chemicals was filtered using the following criteria:

- Only chemicals with discrete LD_{50} values were used (i.e. chemicals with LD_{50} values with ">" or "<" were removed)
- Compounds can only contain the following element symbols: C, H, O, N, F, Cl, Br, I, S, P, Si, or As
- Compounds must represent a single pure component (i.e. salts, undefined isomeric mixtures, polymers, or mixtures were removed)

In version 2.0 of T.E.S.T., the final dataset consisted of 7392 chemicals. 87 compounds in this dataset possessed 106 2d isomers. In version 3.0, only one isomer was kept, using the average toxicity value. In version 4.0 and greater, all isomers were kept because the presence of the isomers had negligible impact on the external prediction statistics. The final oral rat LD_{50} data set contained 7413 chemicals. The modeled endpoint was the $-\text{Log}_{10}$ (LD_{50} mol/kg).

3.5. Bioconcentration factor data set

The bioconcentration factor (BCF) is defined as the ratio of the chemical concentration in biota as a result of absorption via the respiratory surface to that in water at steady state [71]. Data were compiled from several different databases [72-75]. The final dataset consists of 676 chemicals (after removing salts, mixtures, and ambiguous compounds). The modeled endpoint was the Log_{10}(BCF).

3.6. Developmental toxicity data set

The developmental toxicity is defined as whether or not a chemical causes developmental toxicity effects in humans and animals. Developmental toxicity includes any effect interfering with normal development, both before and after birth. A dataset of 293 chemicals was created by Arena and Coworkers [76, 77] by combining data from the Teratogen Information System (TERIS) [78] and FDA guidelines [79]. The developmental toxicity values were taken from the revised binary toxicity values developed for the CAESAR project [3]. One chemical, Azatguiorube, was removed because structural information could not be found for this chemical. The final dataset consists of 285 chemicals (after removing salts, mixtures, and ambiguous compounds).

3.7. Ames mutagenicity data set

In the Ames test, frame-shift mutations or base-pair substitutions can be detected by exposure of histidine-dependent strains of Salmonella typhimurium to a test compound. When these strains are exposed to a mutagen, reverse mutations that restore the functional capability of the bacteria to synthesize histidine enable bacterial colony growth on a medium deficient in histidine (revertants). A compound is classified Ames positive if it significantly induces revertant colony growth in at least one of out of five strains. A dataset of 6512 chemicals was compiled by Hansen and coworkers from several different sources [80, 81]. The final dataset consists of 5743 chemicals (after removing salts, mixtures, ambiguous compounds, and compounds without CAS numbers).

3.8. Normal boiling point

The normal boiling point is defined as the temperature at which a chemical boils at atmospheric pressure. The data set for this endpoint was obtained from the boiling point data contained in EPI Suite [82]. Forty-one chemicals were removed from the data set because they were previously shown to be badly predicted and had experimental values which were significantly different (>50K) from other sources such as NIST[83] and LookChem [84]. The final data set contained 5759 chemicals. The modeled property was the boiling point in °C.

3.9. Density

The density is defined as mass per unit volume. The data set for this endpoint was obtained from the density data contained in LookChem [84]. The data set was restricted to chemicals with boiling points greater than 25°C (or the boiling point was unavailable). The data set was further restricted to chemical with densities > 0.5 and < 5 g/cm^3. The final dataset consisted of 8909 chemicals. Data from LookChem are not peer reviewed but the set is very large and thus provides a large degree of structural diversity. The modeled property was density in g/cm^3.

3.10. Flash point

The flash point is defined as the lowest temperature at which a chemical can vaporize to form an ignitable mixture in air. A dataset of 8362 chemicals was compiled from lookchem.com [84]. Chemicals with flash points greater than 1000°C were omitted from the data set. The modeled property was the flash point in °C.

3.11. Thermal conductivity

Thermal conductivity is defined as a materials ability to conduct heat. The thermal conductivity at 25°C for 442 chemicals was obtained from Jamieson and Vargaftik [85, 86]. Thermal conductivity values were obtained from Jamieson and Vargaftik as follows:

- If a value is available at 25°C this value is used
- If an experimental value is not available, a value is extrapolated to 25°C (as long as the closest data point is within 10°C of 25°C)
- If the temperature coefficient is not available (or only a single data point is available), the thermal conductivity of the nearest data point is used (as long as the closest data point is within 10°C of 25°C)
- Only data with a quality grade of A or B (preferably grade A) in Jamieson were used. The thermal conductivities for the chemicals in common between Jamieson and Vargaftik agreed rather well (R2 = 0.95 for 381 compounds). The modeled property was the thermal conductivity in mW/mK.

3.12. Viscosity

Viscosity is a measure of the resistance of a fluid to flow in cP defined as the proportionality constant between shear rate and shear stress). The viscosity at 25°C for 557 chemicals was obtained from Viswanath and Riddick [87, 88]. The viscosity values were obtained from Viswanath and Riddick were obtained as follows:

1. If a value is available at 25°C this value is used
2. If an experimental value is not available, a value is extrapolated to 25°C (as long as the closest data point is within 10°C of 25°C) using the following empirical correlation:

$$\log_{10} viscosity = A + B/T$$

Extrapolation was used in order to expand size of the overall dataset. The modeled property was $\log_{10}$(viscosity cP).

3.13. Surface tension

Surface tension is a property of the surface of a liquid that allows it to resist an external force. The surface tension at 25°C for 1416 chemicals was obtained from the data compilation of Jaspar [89]. The experimental values (at 25°C) are estimated using an empirical correlation, which is fit to experimental data from Jaspar:

$$\text{surface tension} = A - BT$$

The estimated experimental surface tension value is only used if the closest experimental data point is within 10°C of 25°C. The modeled property was the surface tension in dyn/cm.

3.14. Water solubility

Water solubility is defined as the amount of chemical that will dissolve in liquid water to form a homogeneous solution. A dataset of 5020 chemicals was compiled from the database in EPI Suite [82]. Chemicals with water solubilities exceeding 1,000,000 mg/L were omitted from the overall dataset. In addition, data were limited to data points that are within 10°C of 25°C. The water solubility is an important property because sometimes the predicted LC50 values for aquatic species can exceed the water solubility. The modeled property was −Log10(water solubility mol/L).

3.15. Vapor pressure

Vapor pressure is defined as the pressure of a vapor in mmHg in thermodynamic equilibrium with its condensed phases in a closed system. The vapor pressure at 25°C for 2511 chemicals was obtained from the database in EPI Suite [82]. The modeled property was Log10(vapor pressure mmHg).

3.16.Melting point

Melting point is the temperature, in °C, at which a chemical in the solid state changes to a liquid state. The melting point for 9385 chemicals was obtained from the database in EPI Suite [82]. The modeled property was Log10(vapor pressure mmHg).

4. VALIDATION RESULTS

4.1. 96 hour fathead minnow LC_{50}

4.1.1. Statistical External Validation

The consensus approach achieved the best results in terms of all the prediction statistics (see Table 4.1.1). The hierarchical method achieved the best results of any of the individual QSAR methods. Statistics highlighted in pink represent predictions where a condition in equation 18 or 19 was not met. Models, which do not meet these conditions, are not invalid, per se, but should be used with caution. The predicted values for the test set for the fathead minnow LC50 endpoint for the consensus method are given in Figure 4.1.1.

Table 4.1.1. Prediction results for the fathead minnow LC_{50} test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.710	0.075	0.966	0.801	0.574	0.951
Single Model	0.704	0.134	0.960	0.803	0.605	0.945
FDA	0.626	0.113	0.985	0.915	0.656	0.945
Group contribution	0.686	0.123	0.949	0.810	0.578	0.872
Nearest neighbor	0.667	0.080	1.001	0.876	0.649	0.939
Consensus	0.728	0.121	0.969	0.768	0.545	0.951

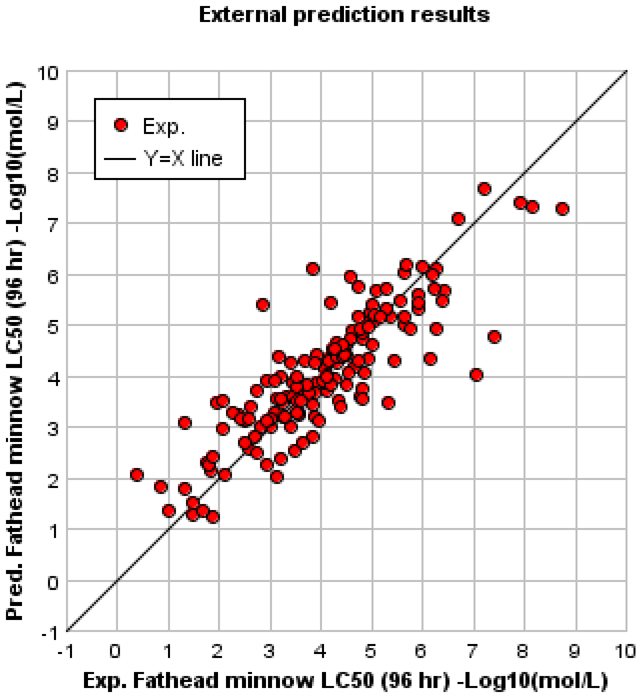


Figure 4.1.1. Experimental vs predicted values for the fathead minnow LC_{50} test set

4.1.2. *Statistical External Validation for mode of action method*

The mode of action method yields slightly worse results than the hierarchical and single model methods (see Table 4.1.2). The results for the hierarchical and single model methods are worse than those from section 4.1.1 because the training set used to fit the models was smaller.

Table 4.1.2. Prediction results for the fathead minnow LC_{50} test set using the MOA method

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical	0.612	0.242	0.990	0.847	0.611	0.907
Single model	0.575	0.141	0.993	0.920	0.640	0.902
Mode of action	0.543	0.049	0.949	0.978	0.678	0.834

4.2. 48 hour *Daphnia magna* LC_{50}

4.2.1. *Statistical External Validation*

The consensus method achieved the best results in terms of both prediction accuracy and coverage (see Table 4.2.1). The prediction results for the consensus method are given in Figure 4.2.1.

Table 4.2.1. Prediction results for the *D. magna* LC_{50} test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical	0.695	0.151	0.981	0.979	0.757	0.886
Single Model	0.697	0.152	1.002	0.993	0.772	0.871
FDA	0.565	0.257	0.987	1.190	0.909	0.900
Group contribution	0.671	0.049	0.999	0.803	0.620	0.657
Nearest neighbor	0.733	0.014	1.015	0.975	0.745	0.871
Consensus	0.739	0.118	1.001	0.911	0.727	0.900

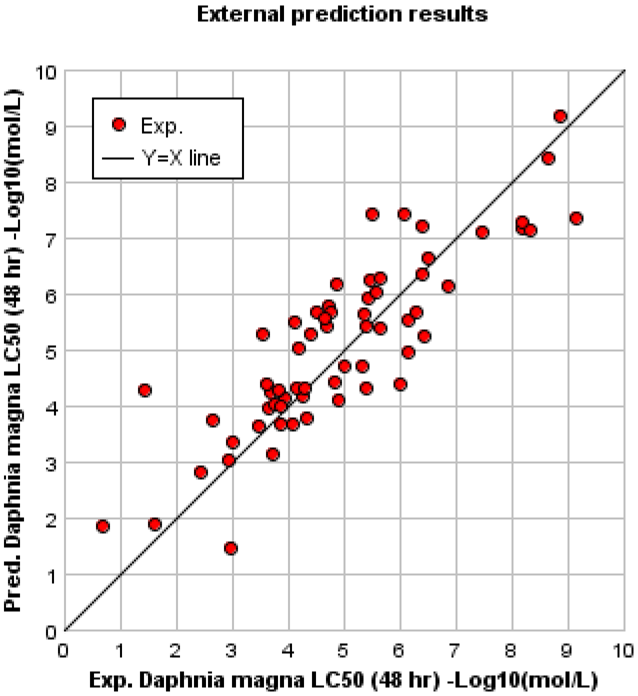


Figure 4.2.1. Experimental vs predicted values for the fathead minnow LC_{50} test set

4.3. *Tetrahymena pyriformis* 50% growth inhibitory concentration (IGC_{50})

4.3.1. Statistical External Validation

Again, the consensus method achieved the best results (see Table 4.3.1). The R^2 value for the consensus method in version 4.1 of TEST was slightly lower than the value for version 4.0. This is because the data set has been expanded to include a wider variety of chemical classes. The prediction results for the consensus method are given in Figure 4.3.1.

Table 4.3.1. Prediction results for the *T. pyriformis* IGC_{50} test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.719	0.023	0.978	0.539	0.358	0.933
FDA	0.747	0.056	0.988	0.489	0.337	0.978
Group contribution	0.682	0.065	0.994	0.575	0.411	0.955
Nearest neighbor	0.600	0.170	0.976	0.638	0.451	0.986
Consensus	0.764	0.065	0.983	0.475	0.332	0.983

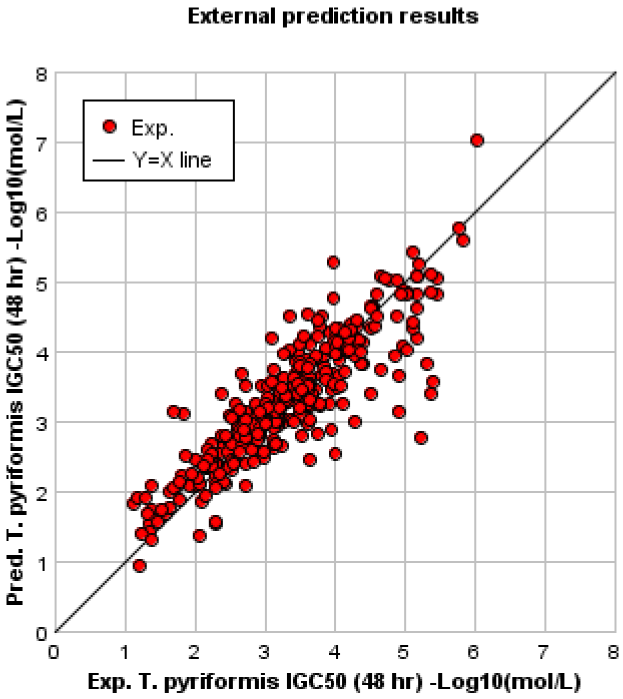


Figure 4.3.1. Experimental vs predicted values for the *T. pyriformis* IGC_{50} test set

4.4. Oral rat LD_{50} dataset

4.4.1. Statistical External Validation

It was not possible to develop a single model or a group contribution model that fit the entire training set (see Table 4.4.1). The consensus method achieved the best results in terms of both prediction accuracy and prediction coverage. The prediction statistics for this endpoint were not as good as those for the other endpoints. This is not surprising because this endpoint has a higher degree of experimental uncertainty and has been shown to be more difficult to model than other endpoints [90]. The prediction results for the consensus method are given by in Figure 4.4.1.

Table 4.4.1. Prediction results for the oral rat LD_{50} test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.578	0.184	0.969	0.650	0.460	0.876
FDA	0.557	0.238	0.953	0.657	0.474	0.984
Nearest neighbor	0.557	0.243	0.961	0.656	0.477	0.993
Consensus	0.626	0.235	0.959	0.594	0.431	0.984

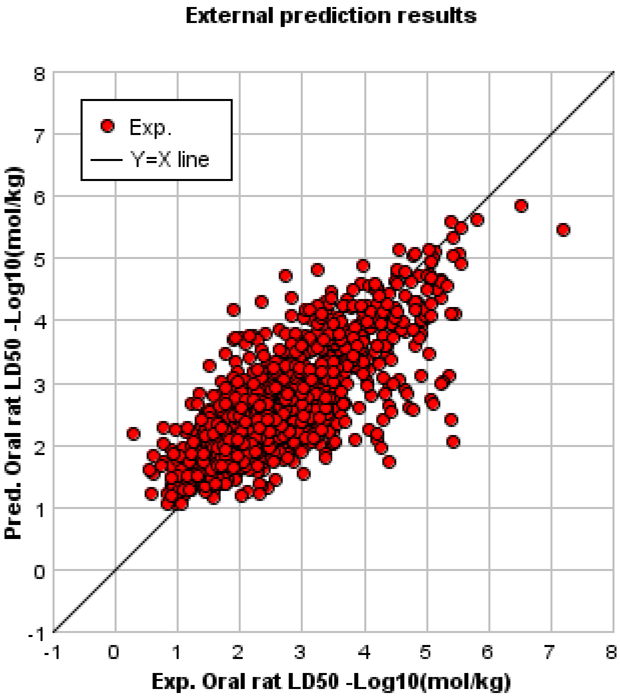


Figure 4.4.1. Experimental vs predicted values for the oral rat LD_{50} test set

4.5. [1]Bioaccumulation factor (BCF)

4.5.1. Statistical External Validation

Again, the consensus method yielded the best statistics if one considers both prediction accuracy and coverage (see Table 4.5.1.). The prediction results for the consensus method are given in Figure 4.5.1.

Table 4.5.1. Prediction results for the BCF test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.734	0.019	0.888	0.712	0.541	0.926
Single Model	0.742	0.083	0.901	0.684	0.543	0.926
FDA	0.705	0.036	0.905	0.746	0.571	0.911
Group Contribution	0.675	0.187	0.888	0.760	0.622	0.874
Nearest neighbor	0.609	0.100	0.931	0.884	0.604	0.948
Consensus	0.760	0.066	0.900	0.661	0.513	0.926

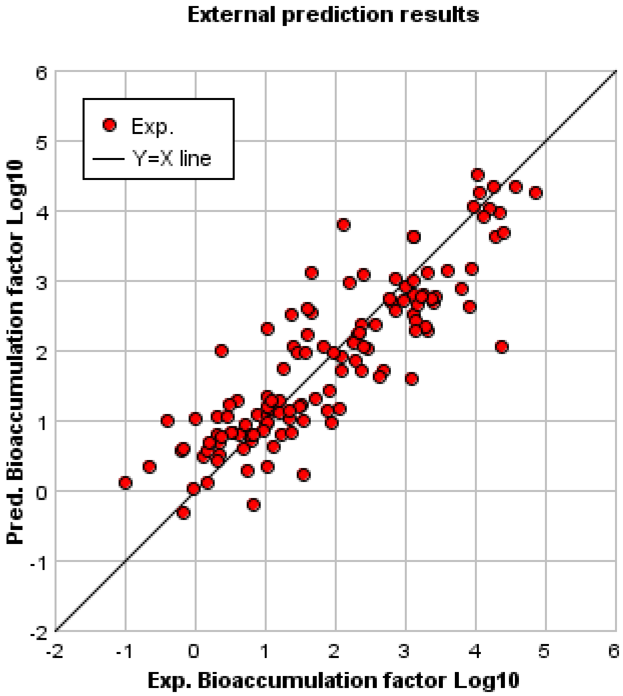


Figure 4.5.1. Experimental vs predicted values for the BCF test set

The BCFBAF (bioconcentration factor bioaccumulation factor) module (v. 3.00) of US EPA's EPI Suite software package [82] yielded an R^2 value of 0.766 and MAE of 0.50 (for the same chemicals that were able to be predicted by the consensus method). Thus, the predictions for the consensus method are comparable to those from EPI Suite. However, this may not be a fair comparison because some of the chemicals in the prediction set may have appeared in the training set for the BCF model in EPI Suite.

4.6. Developmental toxicity

4.6.1. Statistical External Validation

The consensus method achieved the best results for the EPA developed QSAR methods (in terms of prediction accuracy and coverage) (see Table 4.6.1). The CAESAR random forest method achieved similar results to the EPA Consensus model (the concordance was higher but the coverage was lower). All of the methods achieved appreciably higher prediction sensitivities than specificities. This is acceptable for regulatory applications because it is desired to minimize the number of false negatives.

Table 4.6.1. Prediction results for the reproductive toxicity test set

Method	Concordance	Sensitivity	Specificity	Coverage
Hierarchical	0.724	0.829	0.471	1.000
Single Model	0.732	0.850	0.438	0.966
FDA	0.724	0.780	0.588	1.000
Nearest neighbor	0.795	0.844	0.667	0.759
Consensus	0.793	0.902	0.529	1.000
Random Forest	0.852	0.949	0.600	0.931

4.7. Ames mutagenicity

4.7.1. Statistical External Validation

Again, the consensus method achieved the best prediction accuracy (concordance) and prediction coverage (see Table 4.7.1). The single model and group contribution methods could not be applied to this endpoint. All of the methods achieved a nice balance of prediction sensitivity and specificity.

Table 4.7.1. Prediction results for the Ames mutagenicity test set

Method	Concordance	Sensitivity	Specificity	Coverage
Hierarchical	0.763	0.776	0.746	0.956
FDA	0.775	0.766	0.787	0.961
Nearest neighbor	0.770	0.783	0.752	0.990
Consensus	0.790	0.789	0.791	0.995

4.8. Normal boiling point

4.8.1. Statistical External Validation

The consensus method achieved the best statistics in terms of both prediction accuracy and coverage (see Table 4.8.1). In general, the prediction statistics for the physical properties were excellent. The prediction results for the consensus method are given in Figure 4.8.1.

Table 4.8.1. Prediction results for the normal boiling point test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.949	0.001	0.991	18.700	10.613	0.935
FDA	0.936	0.002	0.991	21.431	12.214	0.988
Group contribution	0.897	0.002	0.997	27.554	17.000	0.977
Nearest neighbor	0.877	0.005	0.968	29.967	19.754	0.988
Consensus	0.947	0.002	0.987	19.403	11.460	0.986

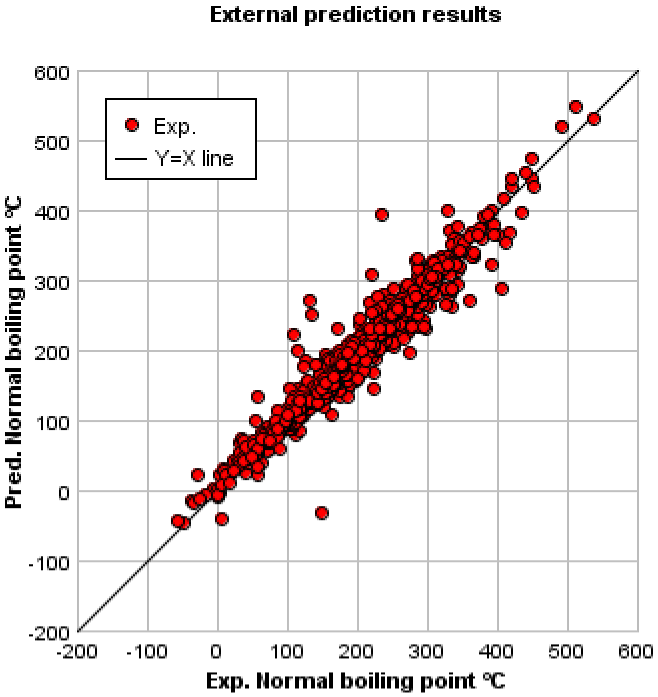


Figure 4.8.1. Experimental vs predicted values for the normal oiling point test set

4.9. Density

4.9.1. Statistical External Validation

For this property, the hierarchical and FDA methods gave a slightly higher R^2 value than the consensus method (see Table 4.9.1.). However, the consensus method yielded a near 100% prediction coverage. The prediction results for the consensus method are given in Figure 4.9.1.

Table 4.9.1. Prediction results for the density test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.972	0.001	0.997	0.052	0.026	0.942
FDA	0.968	0.001	0.993	0.057	0.031	0.992
Group contribution	0.872	0.005	0.997	0.116	0.071	0.992
Nearest neighbor	0.859	0.021	0.978	0.121	0.073	0.997
Consensus	0.956	0.005	0.991	0.068	0.038	0.996

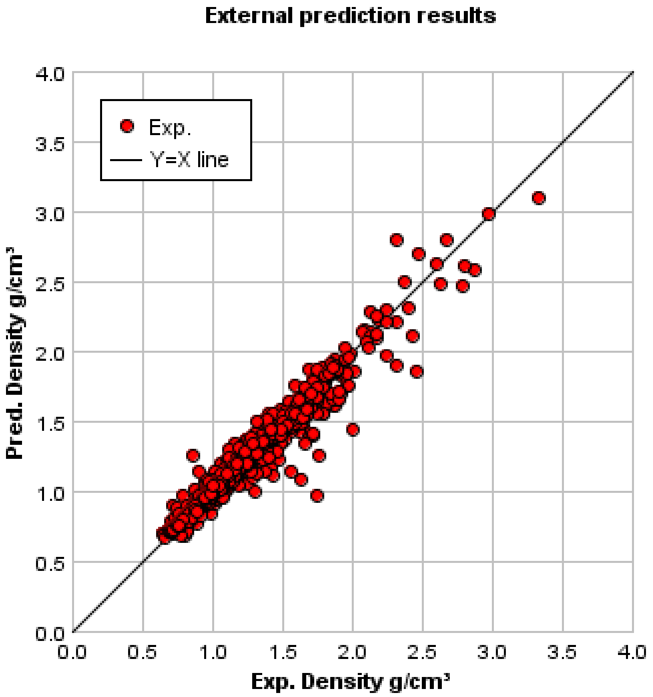


Figure 4.9.1. Experimental vs predicted values for the density test set

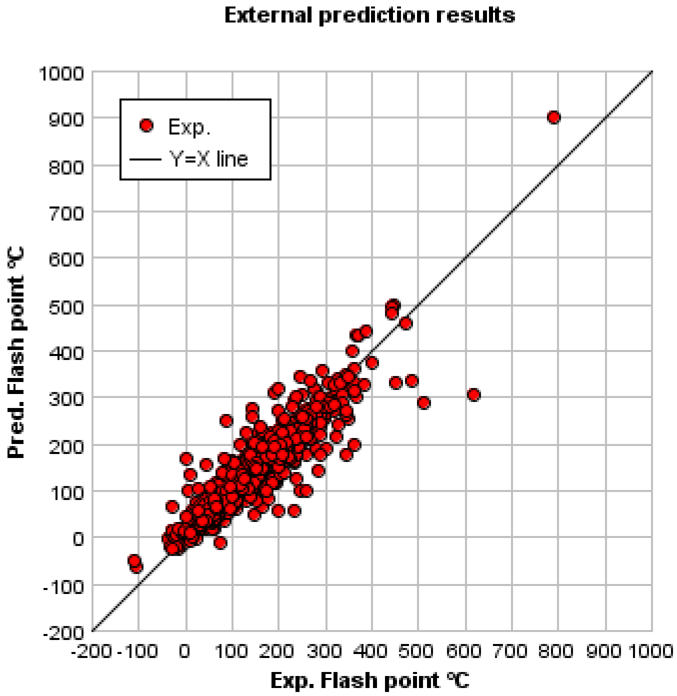
4.10.Flash point

4.10.1. Statistical External Validation

For this property, the consensus method gives the best results in terms of prediction accuracy and coverage (see Table 4.10.1). The prediction results for the consensus method are given in Figure 4.10.1.

Table 4.10.1. Prediction results for the flash point test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical	0.871	0.008	0.962	28.898	16.749	0.924
FDA	0.853	0.010	0.960	31.481	19.227	0.989
Group contribution	0.834	0.009	0.968	33.630	20.426	0.987
Nearest neighbor	0.801	0.018	0.925	36.833	23.832	0.993
Consensus	0.879	0.011	0.953	28.503	16.908	0.992



4.10.1. Experimental vs predicted values for the flash point test set

4.11. Thermal conductivity

4.11.1. Statistical External Validation

For this property, the hierarchical method gives similar results to the consensus method (see Table 4.11.1). The prediction results for the consensus method are given in Table 4.11.1.

Table 4.11.1. Prediction results for the thermal conductivity test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.906	0.025	0.996	11.024	6.731	0.956
Single Model	0.890	0.031	0.992	11.864	8.524	0.956
FDA	0.845	0.000	1.018	16.406	9.008	0.967
Group contribution	0.803	0.088	0.979	15.898	9.825	0.911
Nearest neighbor	0.884	0.021	1.004	12.832	8.449	0.978
Consensus	0.892	0.010	1.005	12.413	7.046	0.967

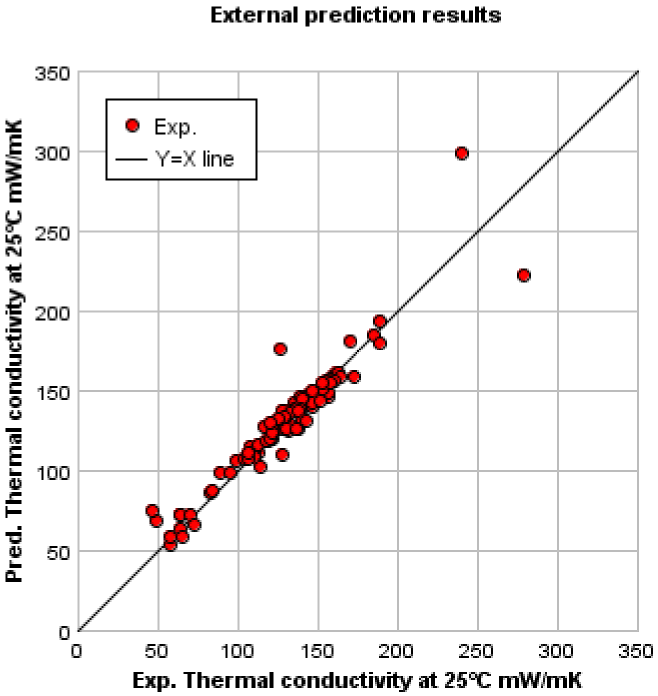


Figure 4.11.1. Experimental vs predicted values for the thermal conductivity test set

4.12. Viscosity

4.12.1. Statistical External Validation

For this property, the consensus method gives the best results if you consider both prediction accuracy and coverage (see Table 4.12.1). The low *k* values for this endpoint can be attributed to the two possible outliers in the test set that fall below the Y=X line. The prediction results for the consensus method are given Figure 4.12.1.

Table 4.12.1. Prediction results for the viscosity test set

Method	*R*²	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical	0.868	0.001	0.809	0.214	0.131	0.929
Single Model	0.644	0.010	0.625	0.346	0.217	0.929
FDA	0.868	0.003	0.875	0.207	0.142	0.929
Group contribution	0.888	0.001	0.831	0.200	0.113	0.814
Nearest neighbor	0.757	0.009	0.726	0.289	0.194	0.920
Consensus	0.876	0.004	0.778	0.215	0.125	0.929

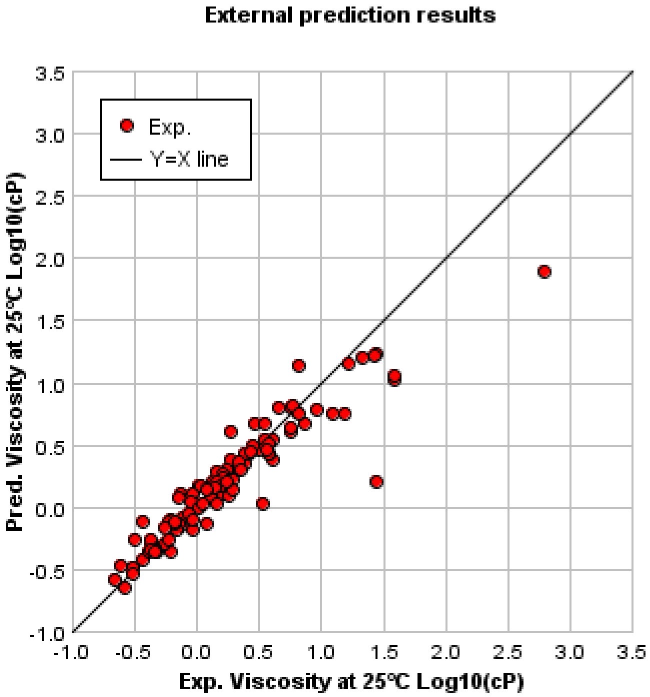


Figure 4.12.1. Experimental vs predicted values for the viscosity test set

4.13. Surface tension

4.13.1. Statistical External Validation

For this property, the consensus method gives the best results in terms of prediction accuracy and coverage (see Table 4.13.1(. The prediction results for the consensus method are given Figure 4.13.1.

Table 4.13.1. Prediction results for the surface tension test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.929	0.016	0.989	1.792	1.037	0.919
FDA	0.890	0.015	0.992	2.219	1.297	0.979
Group contribution	0.794	0.044	0.986	2.933	2.114	0.926
Nearest neighbor	0.759	0.068	0.973	3.317	1.923	0.936
Consensus	0.903	0.027	0.987	2.112	1.317	0.968

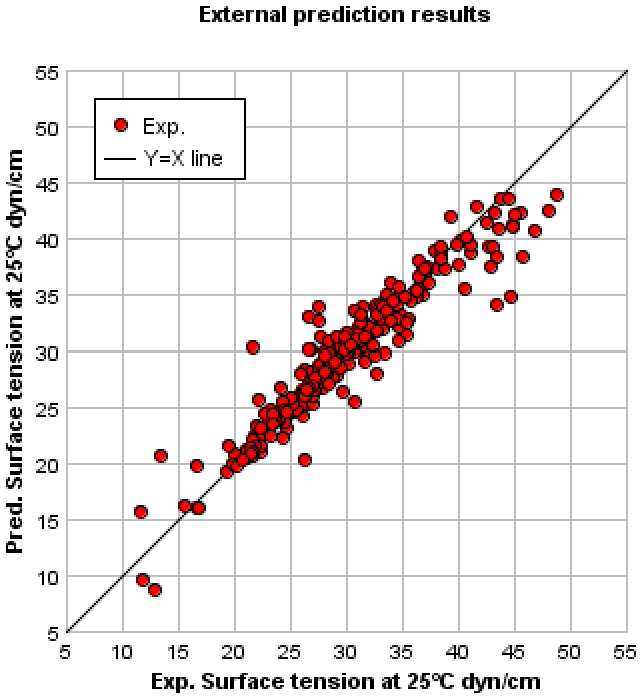


Figure 4.13.1. Experimental vs predicted values for the surface tension test set

4.14. Water solubility

4.14.1. Statistical External Validation

For this property, the consensus method gives the best statistics in terms of prediction accuracy and coverage (see Table 4.14.1). The prediction results for the consensus method are given in Figure 4.14.1.

Table 4.14.1. Prediction results for the water solubility test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical	0.834	0.015	0.943	0.903	0.601	0.935
FDA	0.809	0.014	0.950	0.953	0.639	0.984
Group contribution	0.766	0.039	0.933	1.074	0.798	0.982
Nearest neighbor	0.791	0.022	0.950	1.023	0.735	0.985
Consensus	0.857	0.021	0.943	0.835	0.578	0.987

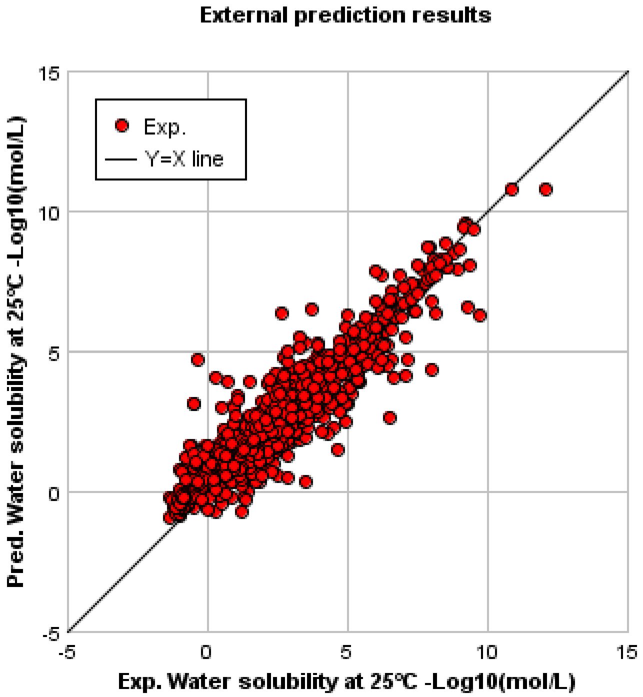


Figure 4.14.1. Experimental vs predicted values for the water solubility test set

4.15. Vapor pressure

4.15.1. Statistical External Validation

The prediction statistics were excellent and again the consensus method achieved the best results (see Table 4.15.1). The prediction results for the consensus method are given in Table 4.15.1.

Table 4.15.1. Prediction results for the vapor pressure test set

Method	***R*²**	$\frac{R^2 - R_0^2}{R^2}$	***k***	**RMSE**	**MAE**	**Coverage**
Hierarchical	0.956	0.001	0.977	0.745	0.455	0.940
FDA	0.946	0.001	0.985	0.827	0.494	0.982
Group contribution	0.929	0.001	1.020	0.998	0.608	0.968
Nearest neighbor	0.878	0.001	0.937	1.251	0.823	0.980
Consensus	0.954	0.001	0.980	0.769	0.466	0.980

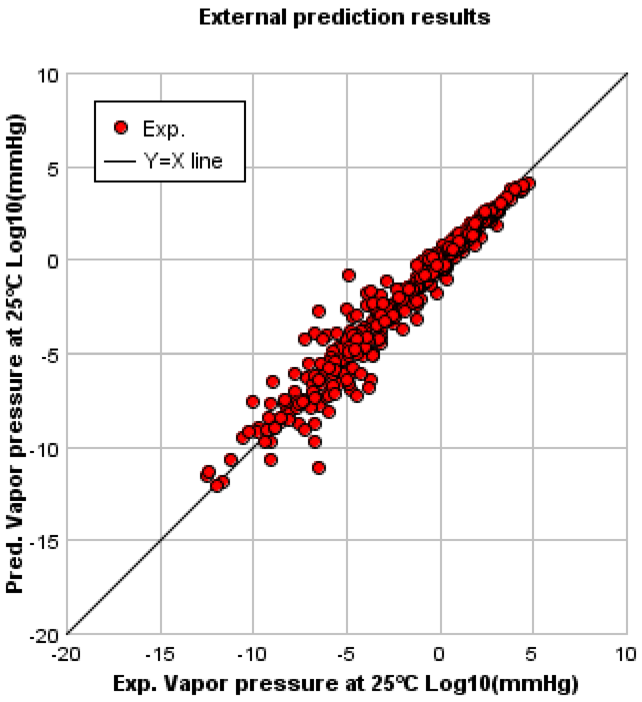


Figure 4.15.1. Experimental vs predicted values for the vapor pressure test set

4.16. Melting point

4.16.1. Statistical External Validation

The prediction statistics were very good and the again the consensus method achieved the best results (see Table 4.16.1.). The prediction results for the consensus method are given in Figure 4.16.1.

Table 4.16.1. Prediction results for the water solubility test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical	0.811	0.011	0.892	44.355	31.433	0.932
FDA	0.801	0.011	0.879	45.095	32.920	0.993
Group contribution	0.704	0.065	0.837	54.947	41.274	0.997
Nearest neighbor	0.738	0.017	0.850	52.095	37.837	0.998
Consensus	0.834	0.021	0.863	41.464	30.207	0.998

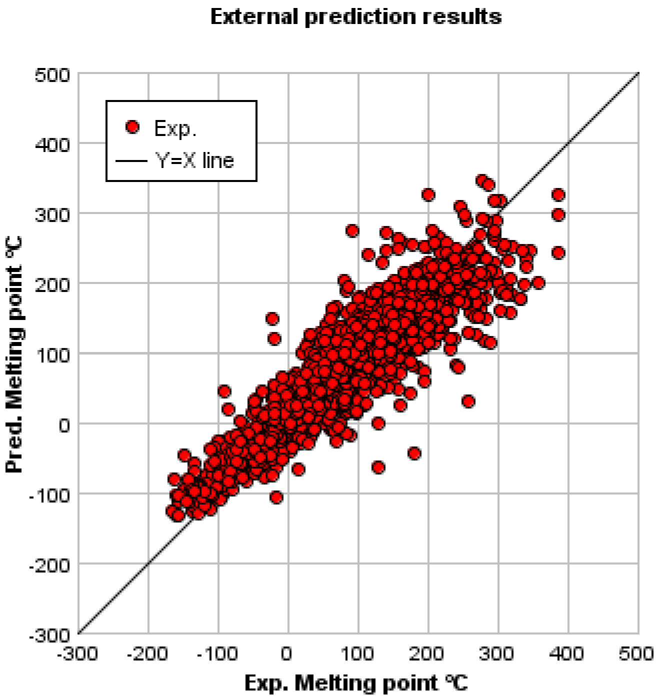


Figure 4.16.1. Experimental vs predicted values for the melting point test set

5. USING THE SOFTWARE

5.1. Importing a single compound

A compound can be imported into the software several different ways:

- Drawn using the provided molecular structure drawing tool
- Imported from an MDL molfile
- Imported from a SMILES string
- Imported from the included structure data base

5.1.1. Drawing a molecule using the structure drawing tool

- First, add any rings present in the molecule using the ring template buttons △ □ ⬠ ⬡ ○ ○ ⌬ (click on a button and then click somewhere in the document).
- Next, step add any chains using the / button.
- Next, add double or triple bonds by using / again and clicking on the bonds to make them double or triple bonds. You can use ▶ and ||| to make existing bonds wedge bonds or you can draw wedge bonds directly.
- Finally, any hetero atoms (non carbon atoms) need to be set. Either use one of the element symbol buttons and click on an atom to change it to this symbol. You can use the periodic table H LiBE to choose an element.
- Finally, with C→O H←N you can go through some common elements by clicking on an atom repeatedly. With +1 and -1 you can change the charge.

5.1.2. Importing a molecule from an MDL molfile

The structure for a test compound can be imported from an MDL molfile (https://en.wikipedia.org/wiki/Chemical_table_file)

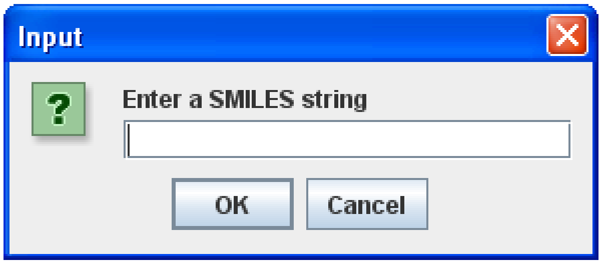
To import a structure using a MDL molfile, select Import from MDL molfile from the File menu.

5.1.3. Import a molecule from a SMILES string

The structure for a test compound can be imported from a *SMILES string* (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html).

To import a structure using a SMILES string, select **Generate from SMILES string** from the **File** menu**.**

Enter the desired SMILES string in the dialog box provided and press OK.

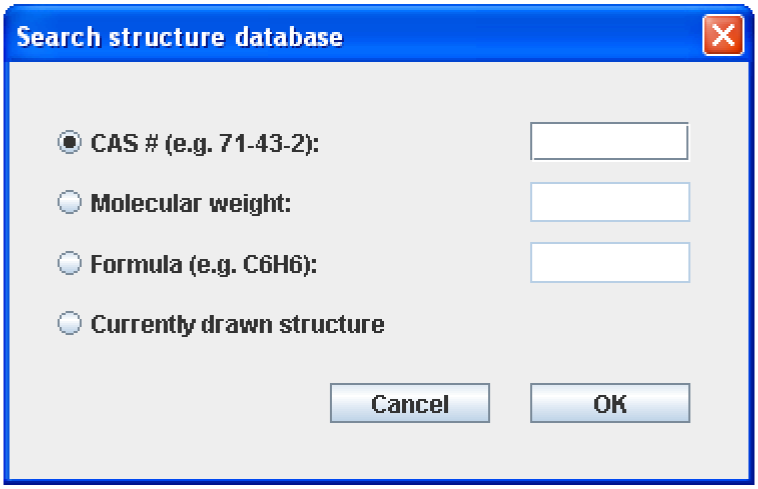


For example, to import benzene enter c1ccccc1 as the SMILES string. A SMILES string can be pasted from the clipboard by selecting **Generate from SMILES on clipboard**.

5.1.4. Import from the structure database

To import a structure from the structure database, first select **Import from structure database** from the **File** menu**.**

One can then import a structure from the CAS number, molecular weight, or formula:



● CAS # (e.g. 71-43-2):
○ Molecular weight:
○ Formula (e.g. C6H6):
○ Currently drawn structure

One can enter the CAS number with or without dashes (i.e. 71-43-2 or 71432). The **Currently drawn structure** option allows you to retrieve the CAS number for a given drawn structure (assuming it is available in the database included with the software).

You can import a chemical by its CAS number by entering a CAS number in the **Molecule ID** field and pressing enter.

5.2. Importing multiple compounds (batch import)

Multiple compounds can be imported simultaneously several different ways:

- Importing from a MDL SDfile

- Importing from a list of CAS numbers
- Importing from a list of SMILES strings

Sample files in each of these formats are available in a zip file at the following link: https://www.epa.gov/sites/production/files/2015-07/samplefiles.zip

5.2.1. Importing from a MDL SDfile

To import multiple structures from an MDL SDfile select **Batch import from MDL SDfile** from the **Import Chemical** menu option.

For best results, one should use SDfiles with either a “CAS” or a “Name” field included to uniquely identify each chemical in the file. The program first looks for a “CAS” field and then looks for “Name” field when assigning identifiers. For example, a sample from an SDfile including formaldehyde would be as follows:

```
Formaldehyde
csChFnd80/07260508122D

 2 1 0 0 0 0 0 0 0 0999 V2000
 0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
 1.4000 0.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
 1 2 2 0 0 0 0
M END

> <CAS>
50-00-0

> <Name>
Formaldehyde

$$$$
```

5.2.2. Importing from a list of CAS numbers

To import multiple structures from a list of CAS numbers (in a text file), select **Batch import from list of CAS numbers** from the **Import Chemical** menu option.

For example to import benzene and formaldehyde, the contents of the text file should be as follows:

```
71-43-2
50-00-0
```

5.2.3. Importing from a list of SMILES strings

To import multiple structures from a list of SMILES strings (in a text file), select **Batch import from list of SMILES strings** from the **Import Chemical** menu option.

The text file should contain the SMILES string and an unique identifier on each line. A comma, tab, or a space can separate the SMILES string and the identifier. The text file should not container a header line.

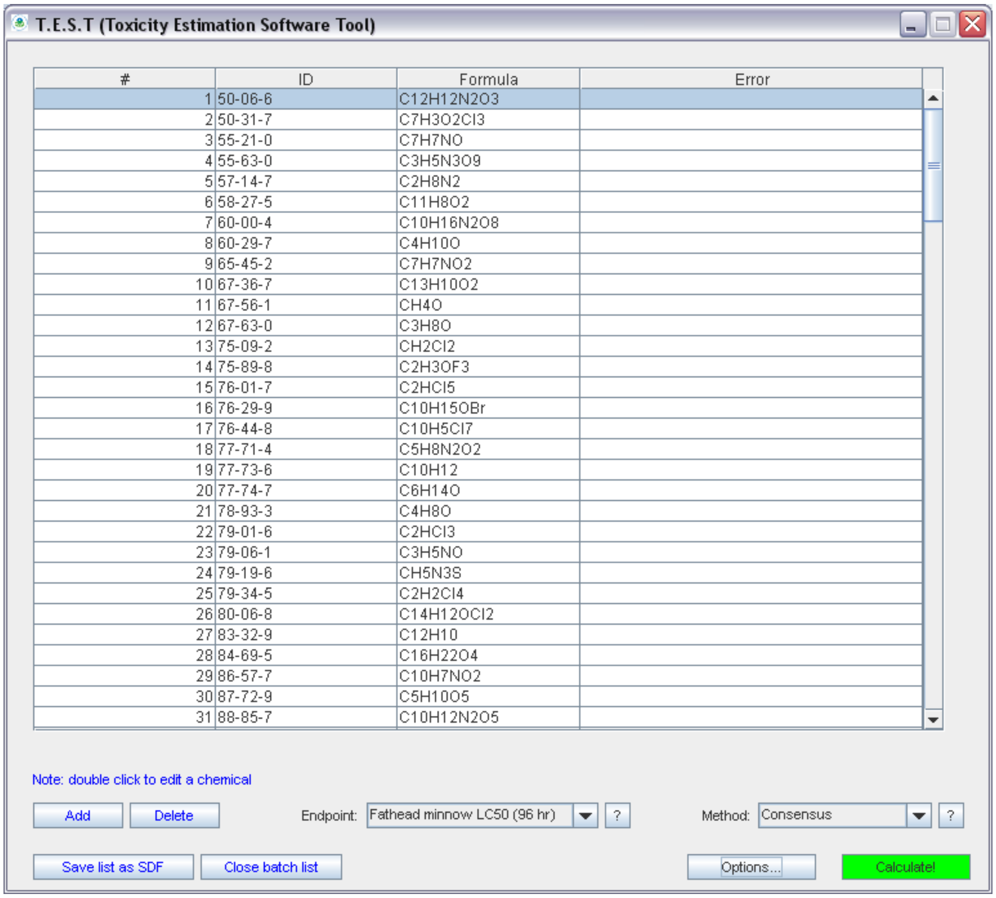
For example to import benzene and formaldehyde, the contents of the text file should be as follows:

```
c1ccccc1                        71-43-2
C=O                             50-00-0
```

5.2.4. Editing a chemical in the batch list

After importing the desired set of chemicals, you can edit individual chemicals in the list by double clicking on its row in the list. An example of an imported batch list is given in Figure 5.2.4.

Figure 5.2.4. Batch mode screen in T.E.S.T.



#	ID	Formula	Error
1	50-06-6	C12H12N2O3	
2	50-31-7	C7H3O2Cl3	
3	55-21-0	C7H7NO	
4	55-63-0	C3H5N3O9	
5	57-14-7	C2H8N2	
6	58-27-5	C11H8O2	
7	60-00-4	C10H16N2O8	
8	60-29-7	C4H10O	
9	65-45-2	C7H7NO2	
10	67-36-7	C13H10O2	
11	67-56-1	CH4O	
12	67-63-0	C3H8O	
13	75-09-2	CH2Cl2	
14	75-89-8	C2H3OF3	
15	76-01-7	C2HCl5	
16	76-29-9	C10H15OBr	
17	76-44-8	C10H5Cl7	
18	77-71-4	C5H8N2O2	
19	77-73-6	C10H12	
20	77-74-7	C6H14O	
21	78-93-3	C4H8O	
22	79-01-6	C2HCl3	
23	79-06-1	C3H5NO	
24	79-19-6	CH5N3S	
25	79-34-5	C2H2Cl4	
26	80-06-8	C14H12OCl2	
27	83-32-9	C12H10	
28	84-69-5	C16H22O4	
29	86-57-7	C10H7NO2	
30	87-72-9	C5H10O5	
31	88-85-7	C10H12N2O5	

Note: double click to edit a chemical

Endpoint: Fathead minnow LC50 (96 hr)

Method: Consensus

5.2.5. Adding chemicals to the batch list

To add chemicals to the list, click the **Add** button. Double click on the new chemical to add the molecular structure for the new chemical.

5.2.6. Deleting chemicals from the batch list

To delete chemicals from the list, select one or more rows in the batch list and click the **Delete** button (or press the Delete key on the keyboard).

5.2.7. Saving the batch list

To save the batch list as an MDL SD file, click on the **Save list as SDF** button. This feature allows you to save changes to your list.

5.2.8. Closing the batch list

To close the batch list click on the Close batch list button. One can close the batch list by deleting all the chemicals in the list.

5.3. Performing toxicity predictions

If the **Molecule ID** is blank, enter a unique identifier for the compound. It is recommended that the CAS number be used for the **Molecule ID** but the name can be used as well. The software needs the **Molecule ID** in order to generate the output web pages. *Warning: if two molecules have the same Molecule ID, the results files will get overwritten.*

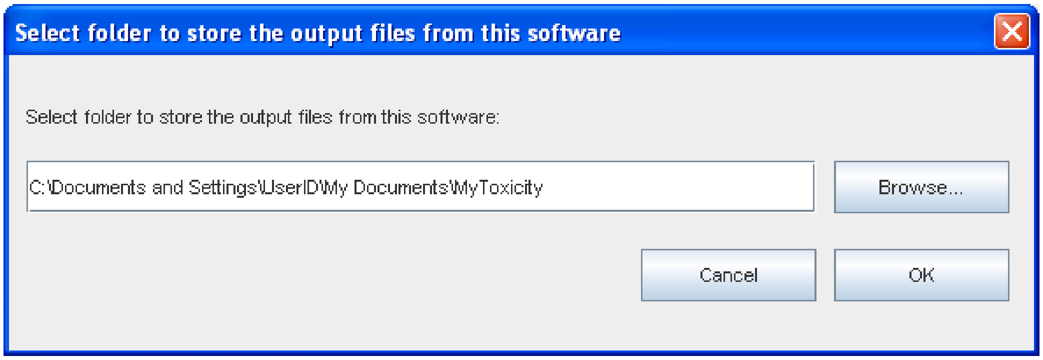
Select a toxicity endpoint using the drop down list provided (the fathead minnow LC_{50} is selected by default).

Select a QSAR toxicity estimation method using the drop down list provided (the hierarchical clustering method is chosen by default). The methodologies are described in detail in the Theory section.

Sometimes predictions for a given chemical cannot be made because the model(s) violate the fragment constraint. The fragment constraint says that in order for a prediction to be made using a given model, the chemicals used in the construction of the model must possess at least one example of each molecular fragment present in the test compound. One can relax this constraint by checking the **Relax fragment constraint** checkbox (now accessed by clicking the **Options** button on the bottom of the screen). The fragment constraint is described in the Theory section.

Once the desired options have been selected, one can start the toxicity estimation calculations by clicking **Calculate!.**

Before the calculations can proceed, one must first select the location where the output files will be stored:


Select folder to store the output files from this software

Select folder to store the output files from this software:

C:\Documents and Settings\UserID\My Documents\MyToxicity

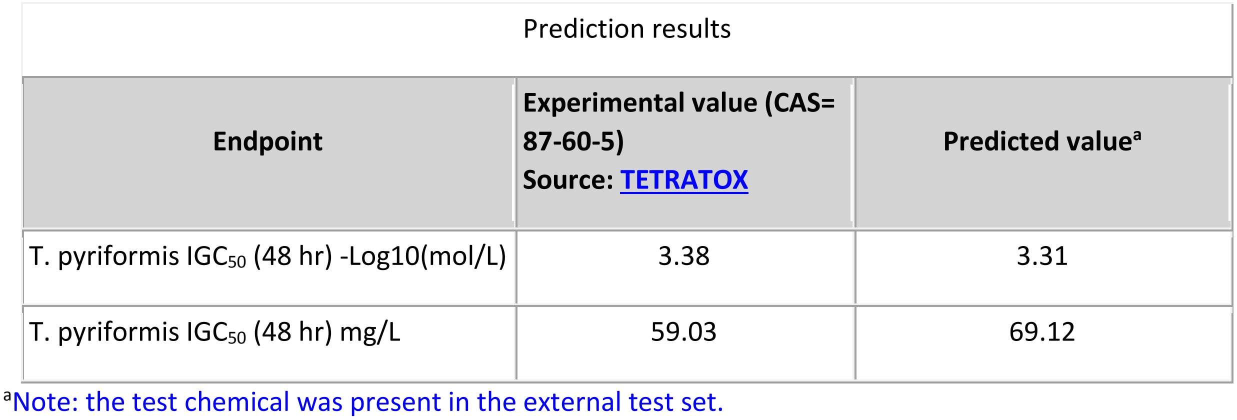
The output folder can be changed at any time by choosing **Select output folder** from the **Options** screen. The software will remember the selected output folder the next time the software is loaded.

If one wishes to abort the currently running calculations, click on the red **Stop** button.

5.4. Interpretation of results

After performing the toxicity estimation calculations, a web page is generated which displays the results. The results for 87-60-5 (for the Tetrahymena pyriformis IGC_{50} endpoint and the *Consensus method*) are given in Table 5.4.1. The predicted toxicity is 69.12 mg/L and the experimental value is 59.03 mg/L. The prediction is flagged in this example because the chemical was part of the external test set. The predicted toxicity from the consensus method represents the average of the predicted toxicities from all the different QSAR methods incorporated into the TEST software. The individual prediction are given in Table 5.4.2. The average of the values from all the different QSAR methods is 3.31 which is close to the experimental value of 3.38 (in units of -Log(mol/L)).

Table 5.4.1. Prediction results from the consensus method for 87-60-5

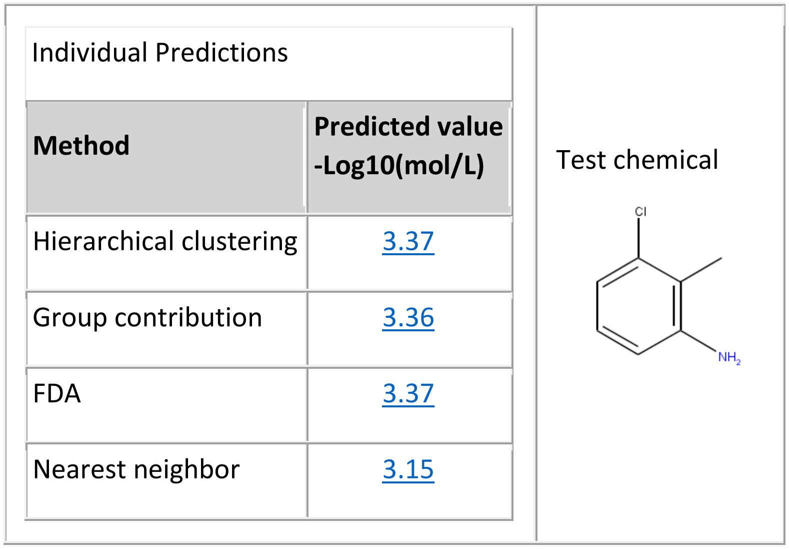


Prediction results

Endpoint	**Experimental value (CAS= 87-60-5)** **Source: TETRATOX**	**Predicted value[a]**
T. pyriformis IGC_{50} (48 hr) -Log10(mol/L)	3.38	3.31
T. pyriformis IGC_{50} (48 hr) mg/L	59.03	69.12

[a]Note: the test chemical was present in the external test set.

Table 5.4.2 Individual predictions for 87-60-5

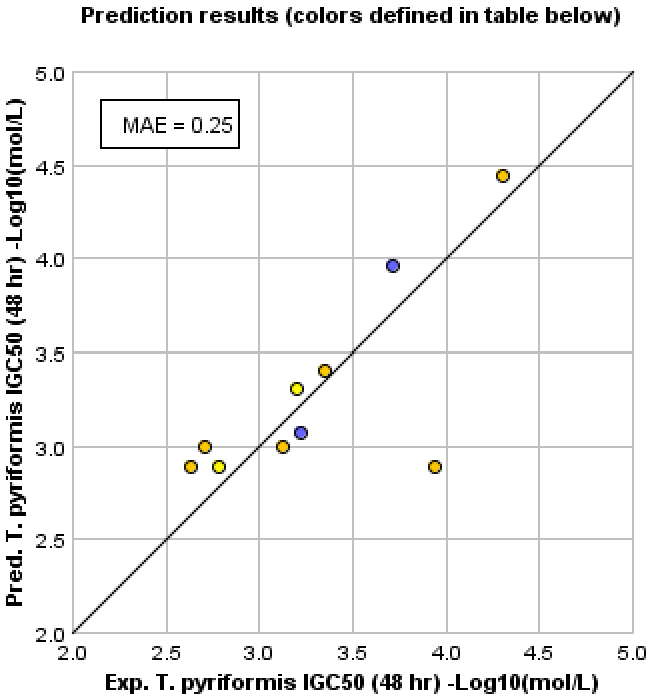


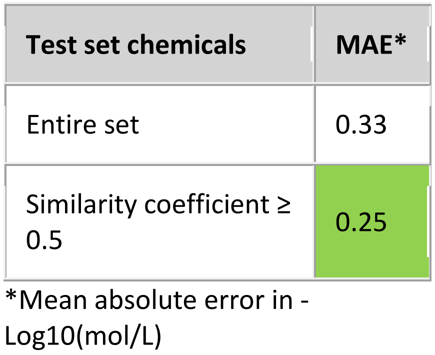
Individual Predictions

Method	**Predicted value -Log10(mol/L)**
Hierarchical clustering	3.37
Group contribution	3.36
FDA	3.37
Nearest neighbor	3.15

Test chemical

The software provides predictions for similar chemicals from the test set (see Figure 5.4.1). The colors of the data points are defined in Table 5.4.3. The MAE (mean absolute error) for similar chemicals (0.25) was slightly lower than the value for the entire test set (0.33). This increases ones confidence in the predicted value. The structures for the similar chemicals in the test set are given in Table 5.4.3.



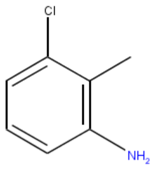
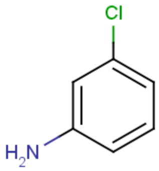
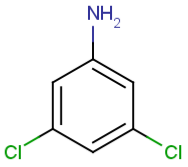
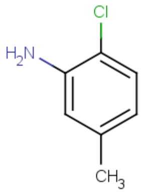


Test set chemicals	MAE*
Entire set	0.33
Similarity coefficient ≥ 0.5	0.25

*Mean absolute error in -Log10(mol/L)

Figure 5.4.1. Predictions for similar chemicals from the test set

Table 5.4.3. Structures for the similar chemicals in the test set

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
87-60-5 (test chemical)			3.38	3.31
108-42-9		0.84	3.22	3.07
626-43-7		0.80	3.71	3.96
95-81-8		0.77	3.20	3.31
…	…	…	…	…

The most similar chemicals are very similar to the test chemical (benzenes substituted with chloro and amino groups) and were accurately predicted. This increases ones confidence in the predicted value. The program lists the similar chemicals in the training set (see Table 5.4.4). As shown by the fairly large similarity coefficients, there are very similar chemicals in the training set (the only difference is the substitution pattern). This increases ones confidence in the predicted value because similar chemicals were used to build the QSAR models.

One can view the details of the predictions for the different QSAR methods by clicking on the predicted value for each method. For example, for the *Hierarchical clustering method* the main prediction table is given in Table 5.4.5. The prediction interval is 48.78 ≤ Tox ≤ 75.30 (one is 90% confident that the predicted value is between 48.78 and 75.30). The experimental value falls within the prediction interval.

Table 5.4.4. Structures for the similar chemicals in the training set

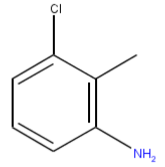
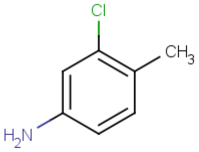
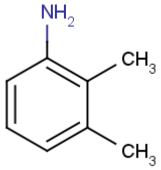
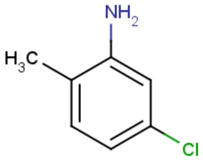
CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)
87-60-5 (test chemical)			3.38
95-74-9		0.89	3.39
87-59-2		0.85	2.57
95-79-4		0.84	3.50
…	…	…	…

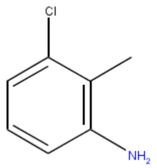
Table 5.4.5. Prediction from the hierarchical clustering method.

Prediction results			
Endpoint	**Experimental value (CAS= 87-60-5)** **Source: TETRATOX**	**Predicted value[a]**	**Prediction interval**
T. pyriformis IGC_{50} (48 hr) -Log10(mol/L)	3.38	3.37	3.27 ≤ Tox ≤ 3.46
T. pyriformis IGC_{50} (48 hr) mg/L	59.03	60.61	48.78 ≤ Tox ≤ 75.30

[a]Note: the test chemical was present in the external test set.

Cluster model predictions and statistics					
Cluster model	**Test chemical descriptor values**	**Prediction interval -Log10(mol/L)**	**r^2**	**q^2**	**#chemicals**
2362	Descriptors	3.31 ± 0.25	0.909	0.834	7
2481	Descriptors	3.48 ± 0.21	0.926	0.861	10
2562	Descriptors	3.40 ± 0.23	0.911	0.834	17
2621	Descriptors	3.24 ± 0.28	0.884	0.796	28
…	…	…	…	…	…

Test chemical



The predictions from the different clusters were all very similar. One can click on the link for each model (in the Cluster model column) to display its statistics, regression plot, parameters, and chemical descriptor values. For example for model #2481, the model statistics are given in Table 5.4.6 and the model regression plot is given in Figure 5.4.2.

Table 5.4.6. Regression statistics for model 2481

Parameter	Value
Endpoint	T. pyriformis IGC_{50} (48 hr)
r^2	0.926
q^2	0.861
#chemicals	10
Model	Model # 2481

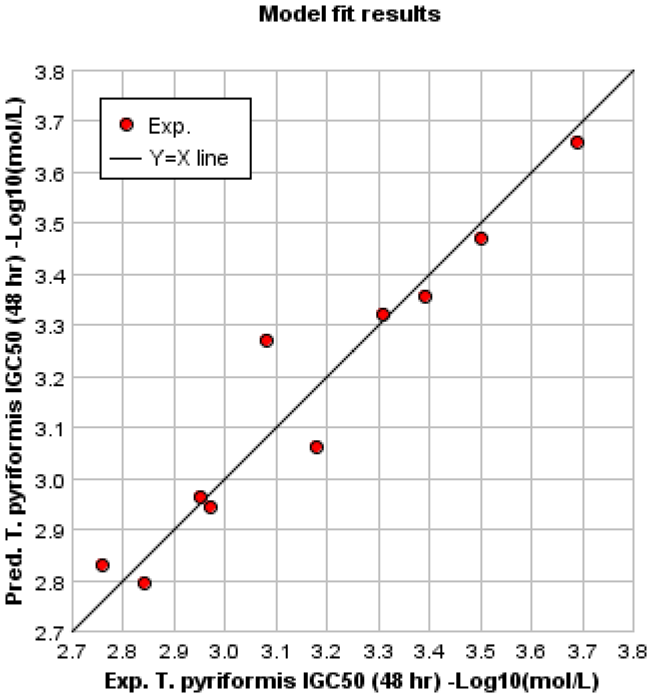


Figure 5.4.2. Model regression plot for model 2481

Table 5.4.6. Model parameters for model 2481

Model coefficients			
Coefficient	**Definition**	**Value**	**Uncertainty***
Intercept	Model intercept	2.5043	0.2689
MATS4e	Moran autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities	0.7092	0.1648
GATS3p	Geary autocorrelation - lag 3 / weighted by atomic polarizabilities	0.6683	0.2168

* value for 90% confidence interval

Table 5.4.6. indicates that the equation for the model is as follows:

Model equation:

T. pyriformis IGC50 (48 hr) = 0.7092×(MATS4e) + 0.6683×(GATS3p) + 2.5043

The fit results (and structures) for each chemical in the model's training set can be obtained by clicking on Model 2481 fit results by chemical.

The descriptor values (in a "|" delimited text file) can be obtained by clicking on Model 2481 training set descriptors.

Bibliography

(1) US EPA. Environmental Optimization Using the Waste Reduction Algorithm. nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100DZKT.TXT (accessed 4/18/16).

(2) Martin, T. M.; Harten, P.; Venkatapathy, R.; Das, S.; Young, D. M. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Method.* **2008,** *18*, 251–266.

(3) CAESAR. Developmental Toxicity Model. http://www.caesar-project.eu/index.php?page=results&section=endpoint&ne=5 (accessed 9/21/09).

(4) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comp. Sci.* **2003,** *43*, 493-500.

(5) Sourceforge.net. Chemistry Development Kit (CDK). https://sourceforge.net/projects/cdk/ (accessed 4/14/2016).

(6) Elsevier MDL. MDL QSAR Version 2.2. http://www.mdl.com/products/predictive/qsar/index.jsp (accessed 8/17/2006).

(7) Talete. Dragon Version 5.4. http://www.talete.mi.it/ (accessed 5/26/09).

(8) Edusoft-LC. Molconn-z Version 4.0. http://www.edusoft-lc.com/molconn/ (accessed 5/26/09).

(9) Romesburg, H. C., *Cluster Analysis for Researchers*. Lifetime Learning Publications: Belmont, CA, 1984.

(10) Eriksson, L.; Jaworska, J. S.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Persp.* **2003,** *111*, 1361-1375.

(11) Topliss, J. G.; Edwards, R. P. Chance factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979,** *22*, 1238-1244.

(12) The University of Waikato. WEKA - The Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/~ml/weka/ (accessed 5/26/09).

(13) Witten, I. H., *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann: San Francisco, 2005.

(14) Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. , *Applied Linear Statistical Models*. McGraw-Hill: New York, 2004.

(15) Eriksson, L.; Johannson, E.; Kettaneh-Wold, N.; Wold, S., *Multi- and Megavariate Data Analysis - Principles and Applications*. Umetrics AB: Umea, Sweden, 2001.

(16) Wikipedia.org. Weighted mean. http://en.wikipedia.org/wiki/Weighted_mean (accessed 4/14/16).

(17) Montgomery, D. C., Introduction to linear regression analysis. In John Wiley and Sons: New York, 1982; p 141.

(18) Contrera, J. F.; Matthews, E. J.; Benz, R. D. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharm.* **2003,** *38*, 243-259.

(19) Benigni, R.; Richard, A. M. QSARS of mutagens and carcinogens: Two case studies illustrating problems in the construction of models for noncongeneric chemicals. *Mutat. Res.* **1996,** *371*, 29-46.

(20) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC_{50}) of Organic Compounds ti the Fathead Minnow (*Pimephales promelas*) Using a Group Contribution Method. *Chem. Res. Toxicol.* **2001,** *14*, 1378-1385.

(21) Martin, T. M.; Grulke, C. M.; Young, D. M.; Russom, C. L.; Wang, N. Y.; Jackson, C. R.; Barron, M. G. Prediction of Aquatic Toxicity Mode of Action Using Linear Discriminant and Random Forest Models. *J. Chem. Inf. Model.* **2013,** *53*, 2229-2239.

(22) Martin, T. M.; Young, D. M.; Lilavois, C. R.; Barron, M. G. Comparison of global and mode of action-based models for aquatic toxicity. *SAR QSAR Environ. Res.* **2015,** *26*, 245-262.

(23) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinational QSAR Model of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008,** *48*, 766 - 784.

(24) Gramatica, P.; Pilutti, P. *Evaluation of different statistical approaches for the validation of quantitative structure-activity relationships*; The European Commission - Joint Research Centre, Institute for Health & Consumer Protection - ECVAM: Ispra, Italy, 2004.

(25) Bourguignon, B.; Deaguiar, P. F.; Khots, M. S.; Massart, D. L. Optimization in Irregularly Shaped Regions: pH and Solvent Strength in Reversed-Phase High-Performance Liquid Chromatography Separations. *Analytical Chemistry* **1994,** *66*, 893-904.

(26) Bourguignon, B.; Deaguiar, P. F.; Thorre, K.; Massart, D. L. *Journal of Chromatography Science* **1994,** *32*, 144-152.

(27) Kennard, R. W.; Stone, L. A. *Technometrics* **1969,** *11*, 137-148.

(28) Snarey, M.; Terrett, N. K.; Willet, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graph. Model.* **1997,** *15*, 372-385.

(29) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational Selection of Training and Test sets for the Development of Validated QSAR Models. *J. Comput. Aid. Mol. Des.* **2003,** *17*, 241-253.

(30) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002,** *20*, 269-276.

(31) US EPA. ECOTOX Database. http://cfpub.epa.gov/ecotox/. (accessed 4/14/2016).

(32) Akers, K. S.; Sinks, G. D.; Schultz, T. W. Structure–toxicity relationships for selected halogenated aliphatic chemicals. *Environmental Toxicology and Pharmacology* **1999,** *7*, 33–39.

(33) Aptula, A. O.; Roberts, D. W.; Cronin, M. T. D.; Schultz, T. W. Chemistry-Toxicity Relationships for the Effects of Di- and Trihydroxybenzenes to Tetrahymena pyriformis. *Chem. Res. Toxicol.* **2005,** *18*, 844-854.

(34) Bearden, A. P.; Schultz, T. W. Structure–Activity Relationships For Pimephales And Tetrahymena: A Mechanism Of Action Approach. *Environmental Toxicology and Chemistry* **1997,** *16*, 1311–1317.

(35) Bohme, A.; Thaens, D.; Schramm, F.; Paschke, A.; Schuurmann, G. Thiol Reactivity and Its Impact on the Ciliate Toxicity of Unsaturated Aldehydes, Ketones, and Esters. *Chem. Res. Toxicol.* **2010,** *23*, 1905-1912.

(36) Cottrell, M. B.; Schultz, T. W. Structure–Toxicity Relationships for Methyl Esters of Cyanoacetic Acids to Tetrahymena pyriformis. *Bull. Environ. Contam. Toxicol.* **2003,** *70*, 549–556.

(37) Cronin, M. T. D.; Bowers, G. S.; Sinks, G. D.; Schultz, T. W. Structure-Toxicity Relationships for Aliphatic Compounds Encompassing a Variety of Mechanisms of Toxic Action to Vibrio fischeri. *SAR QSAR Environ. Res.* **2000,** *11*, 301-312.

(38) Cronin, M. T. D.; Manga, N.; Seward, J. R.; Sinks, G. D.; Schultz, T. W. Parametrization of Electrophilicity for the Prediction of the Toxicity of Aromatic Compounds. *Chem. Res. Toxicol.* **2001,** *14*, 1498-1505.

(39) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. *Chemosphere* **2002,** *49*, 1201–1221.

(40) DeWeese, A. D.; Schultz, T. W. Structure–Activity Relationships for Aquatic Toxicity to Tetrahymena: Halogen-Substituted Aliphatic Esters. *Environ. Toxicol.* **2001,** *16*, 54–60.

(41) Dimitrov, S.; Koleva, Y.; Schultz, T. W.; Walker, J. D.; Mekenyan, O. Interspecies Quantitative Structure–Activity Relationship Model For Aldehydes: Aquatic Toxicity. *Environmental Toxicology and Chemistry* **2004,** *23*, 463-470.

(42) Ellison, C. M.; Cronin, M. T. D.; Madden, J. C.; Schultz, T. W. Definition of the structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2008,** *19*, 751–783.

(43) Gagliardi, S. R.; Schultz, T. W. Regression Comparisons of Aquatic Toxicity of Benzene Derivatives: Tetrahymena pyriformis and Rana japonica. *Bull. Environ. Contam. Toxicol.* **2005,** *74*, 256–262.

(44) Muccini, M.; Layton, A. C.; Sayler, G. S.; Schultz, T. W. Aquatic Toxicities of Halogenated Benzoic Acids to Tetrahymena pyriformis. *Bull. Environ. Contam. Toxicol.* **1999,** *62*, 616-622.

(45) Netzeva, T. I.; Schultz, T. W.; Aptula, A. O.; Cronin, M. T. D. Partial least squares modelling of the acute toxicity of aliphatic compounds to Tetrahymena pyriformis. *SAR QSAR Environ. Res.* **2003,** *14*, 265-83.

(46) Netzeva, T. I.; Schultz, T. W. QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. *Chemosphere* **2005,** *61*, 1632-1643.

(47) Ren, S.; Frymier, P. D.; Schultz, T. W. An exploratory study of the use of multivariate techniques to determine mechanisms of toxic action. *Ecotoxicology and Environmental Safety* **2003,** *55*, 86-97.

(48) Roberts, D. W.; Schultz, T. W.; Wolf, E. M.; Aptula, A. O. Experimental Reactivity Parameters for Toxicity Modeling: Application to the Acute Aquatic Toxicity of SN2 Electrophiles to Tetrahymena pyriformis. *Chem. Res. Toxicol.* **2010,** *23*, 228–234.

(49) Schultz, T. W.; Kier, L. B.; Hall, L. H. Structure-Toxicity Relationships of Selected Nitrogenous Heterocyclic Compounds. III. Relations Using Molecular Connectivity. *Bull. Environ. Contam. Toxicol.* **1982,** *28*, 373-378.

(50) Schultz, T. W.; Wesley, S. K.; Baker, L. L. Structure-Activity Relationships for Di and Tri Alkyl and/or Halogen Substituted Phenol. *Bull. Environ. Contam. Toxicol.* **1989,** *43*, 192-198.

(51) Schultz, T. W.; Tichy, M. Structure-Toxicity Relationships for Unsaturated Alcohols to *Tetrahymena pyriformis*: C_5 and C_6 analogs and Primary Propargylic Alcohols. *Bull. Environ. Contam. Toxicol.* **1993,** *51*, 681-688.

(52) Schultz, T. W.; Comeaux, J. L. Structure-Toxicity Relationships for Aliphatic Isothiocyanates to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **1996,** *56*, 638-642.

(53) Schultz, T. W.; Bearden, A. P. Structure-Toxicity Relationships for Selected Naphthoquinones to Tetrahymena pyriformis. *Bull. Environ. Contam. Toxicol.* **1998,** *61*, 405-410.

(54) Schultz, T. W. Structure-Toxicity Relationships for Benzenes Evaluated with Tetrahymena pyriformis. *Chem. Res. Toxicol.* **1999,** *12*, 1262-1267.

(55) Schultz, T. W.; Sinks, G. D.; Miller, L. A. Population growth impairment of sulfur-containing compounds to *Tetrahymena pyriformis*. *Environ. Toxicol.* **2001,** *16*, 543-549.

(56) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Selection of data sets for qsars: Analyses of tetrahymena toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003,** *Vol. 14*, pp. 59–81.

(57) Schultz, T. W.; Tucker, V. A. Structure-Toxicity Relationships for the Effects of N- and N,N-Alkyl Thioureas to Tetrahymena pyriformis. *Bull. Environ. Contam. Toxicol.* **2003,** *70*, 1251-1258.

(58) Schultz, T. W.; Burgan, J. T. pH-Stress and Toxicity of Nitrophenols to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **2003,** *71*, 1069-1076.

(59) Schultz, T. W.; Seward-Nagel, J.; Foster, K. A.; Tucker, V. A. Population Growth Impairment of Aliphatic Alcohols to Tetrahymena. *Environ. Toxicol.* **2004,** *19*, 1-10.

(60) Schultz, T. W.; Yarbrough, J. W.; Woldemeskel, M. Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. *Cell Biol. Toxicol.* **2005,** *21*, 181-189.

(61) Schultz, T. W.; Netzeva, T. I.; Roberts, D. W.; Cronin, M. T. D. Structure-Toxicity Relationships for the Effects to Tetrahymena pyriformis of Aliphatic, Carbonyl-Containing, α,β-Unsaturated Chemicals. *Chem. Res. Toxicol.* **2005,** *18*, 330-341.

(62) Schultz, T. W.; Yarbrough, J. W.; Koss, S. K. Identification of reactive toxicants: Structure–activity relationships for amides. *Cell Biol Toxicol* **2006,** *22*, 339–349.

(63) Schultz, T. W. Tetratox. http://www.vet.utk.edu/TETRATOX/ (accessed 5/26/09).

(64) Schultz, T. W.; Hewitt, M.; Netzeva, T. I.; Cronin, M. T. D. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR Comb. Sci.* **2007,** *26*, 238-254.

(65) Schultz, T. W.; Ralston, K. E.; Roberts, D. W.; Veith, G. D.; Aptula, A. O. Structure-activity relationships for abiotic thiol reactivity and aquatic toxicity of halo-substituted carbonyl compounds. *SAR QSAR Environ. Res.* **2007,** *18*, 21-29.

(66) Schultz, T. W.; Sparfkin, C. L.; Aptula, A. O. Reactivity-based toxicity modelling of five-membered heterocyclic compounds: Application to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2010,** *21*, 681-691.

(67) Schwöbel, J. A. H.; Madden, J. C.; Cronin, M. T. D. Application of a computational model for Michael addition reactivity in the prediction of toxicity to *Tetrahymena pyriformis*. *Chemosphere* **2011,** *85*, 1066-1074.

(68) Seward, J. R.; Hamblen, E. L.; Schultz, T. W. Regression comparisons of Tetrahymena pyriformis and Poecilia reticulata toxicity. *Chemosphere* **2002,** *47*, 93–101.

(69) Sinks, G. D.; Schultz, T. W. Correlation Of Tetrahymena And Pimephales Toxicity: Evaluation Of 100 Additional Compounds. *Environmental Toxicology and Chemistry* **2001,** *20*, 917–921.

(70) U.S. National Library of Medicine. ChemIDplus. http://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsp (accessed 4/14/16).

(71) Hamelink, J. L., Current bioconcentration test methods and theory. In *Aquatic Toxicology and Hazard Evaluation*, Mayer, F. L.; Hamelink, J. L., Eds.; ASTM STP: West Conshohocken, PA 1977; Vol. 634, pp 149-161.

(72) Dimitrov, S.; Dimitrova, N.; Parkerton, T.; Combers, M.; Bonnell, M.; Mekenyan, O. Base-line model for identifying the bioaccumulation potential of chemicals. *SAR QSAR Environ. Res.* **2005,** *16*, 531-554

(73) Arnot, J. A.; Gobas, F. A. P. C. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environ. Rev.* **2006,** *14*, 257-297.

(74) EURAS. Establishing a bioconcentration factor (BCF) Gold Standard Database. http://www.euras.be/eng/project.asp?ProjectId=92 (accessed 5/20/09).

(75) Zhao, C. B., E.; Chana, A.; Roncaglioni, A.; Benfenati, E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* **2008,** *73*, 1701-1707.

(76) Arena, V. C.; Sussman, N. B.; Mazumdar, S.; Yu, S.; Macina, O. T. The Utility of Structure-Activity Relationship (SAR) Models for Prediction and Covariate Selection in Developmental Toxicity: Comparative Analysis of Logistic Regression and Decision Tree Models. *SAR QSAR Environ. Res.* **2004,** *15*, 1-18.

(77) Sussman, N. B.; Arena, V. C.; Yu, S.; Mazumdar, S.; Thampatty, B. P. Decision Tree SAR Models for Developmental Toxicity Based on an FDA/TERIS Database. *SAR QSAR Environ. Res.* **2003,** *14*, 83-96.

(78) Briggs, G. G.; Freeman, R. K.; Yaffe, S. J., *Drugs in Pregnancy and Lactation, 3rd ed.* Williams and Wilkens: Baltimore, MD, 1990.

(79) Shepard, T. H., *Catalog of Teratologic Agents, 5th ed.* Johns Hopkins University Press: Baltimore, MD, 1992.

(80) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009,** *49*, 2077-2081.

(81) Benchmark, T. http://ml.cs.tu-berlin.de/toxbenchmark/ (accessed 4/30/10).

(82) US EPA. EPI Suite, Version 4.0. http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm (accessed 5/21/09).

(83) NIST. NIST Chemistry WebBook. http://webbook.nist.gov/chemistry/ (accessed

(84) Lookchem.com. http://www.lookchem.com (accessed

(85) Jamieson, D. T. I., J.B; Tudhope, J.S. , *Liquid Thermal Conductivity. A Data Survey to 1973*. H. M. Stationary Office: Edinburgh, 1975.

(86) Vargaftik, N. B.; Filippov, L. P.; Tarzimanov, A. A.; Totskii, E. E., *Handbook of thermal conductivity of liquids and gases*. CRC Press: Boca Raton, 1994; p 358.

(87) Viswanath, D. S. N., G., *Data Book on the Viscosity of Liquids*. Hemisphere Pub. Co.: New York, 1989.

(88) Riddick, J. A.; Bunger, W. B.; Sakano, T. K., *Organic Solvents Physical Properties and Methods of Purification, 4th ed.* Wiley: New York, 1986.

(89) Jasper, J. J. The Surface Tension of Pure Liquid Compounds. *J. Phys. Chem. Ref. Data* **1972,** *1*, 841-1009.

(90) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chem. Res. Toxicol.* **2009,** *22*, 1913–1921.

EPA
United States
Environmental Protection
Agency

Office of Research
and Development
(8101R)
Washington, DC
20460