

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

*A Program to Estimate Toxicity from Molecular Structure*

Version 3.0



©2008 U.S. Environmental Protection Agency

# Table of Contents

	Page
1. INTRODUCTION	3
2. THEORY	7
3. EXPERIMENTAL DATA SETS	18
4. VALIDATION OF THE QSAR METHODOLOGIES	22
5. USING THE SOFTWARE	35
6. REFERENCES	50

## 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:

$$\text{Toxicity} = ax_1 + bx_2 + c$$

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 ([http://www.epa.gov/nrmrl/std/cppb/war/sim\\_war.htm](http://www.epa.gov/nrmrl/std/cppb/war/sim_war.htm)).

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:

- 96 hour fathead minnow LC<sub>50</sub> (concentration of the test chemical in water in mg/L that causes 50% of fathead minnow to die after 96 hours)
- 48 hour *Tetrahymena pyriformis* IGC<sub>50</sub> (concentration of the test chemical in water in mg/L that causes 50% growth inhibition to *Tetrahymena pyriformis* after 40 hours)
- Oral rat LD<sub>50</sub> (amount of chemical in mg/kg body weight that causes 50% of rats to die after oral ingestion)
- 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)

## 1.2. QSAR Methodologies

T.E.S.T allows you to estimate toxicity values using several different advanced Quantitative Structure Activity Relationship (QSAR) methodologies (Martin et al. 2008):

- **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 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 other QSAR methods (provided the predictions are within the respective applicability domains)

T.E.S.T provides multiple prediction methodologies so that one can have greater confidence in the predicted toxicities (assuming the predicted toxicities are fairly 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 different QSAR methods have different advantages and disadvantages:

Method	Advantages	Disadvantages
Hierarchical	<ul style="list-style-type: none"> <li>• Can produce more reliable predictions since predictions are made from multiple models</li> </ul>	<ul style="list-style-type: none"> <li>• Cannot provide external estimates of toxicity for compounds in the training set</li> </ul>
Single model	<ul style="list-style-type: none"> <li>• Single transparent model can be easily viewed/exported</li> <li>• The model does not need to rely on clustering the chemicals correctly</li> </ul>	<ul style="list-style-type: none"> <li>• Since the model is fit to the entire dataset it may incorrectly predict the trends in toxicity for certain chemical classes</li> <li>• Cannot provide external estimates of toxicity for compounds in the training set</li> </ul>
Group contribution	<ul style="list-style-type: none"> <li>• Single transparent model can be easily viewed/exported</li> <li>• Estimates of toxicity can be made without using a computer program</li> </ul>	<ul style="list-style-type: none"> <li>• The model doesn't correct for the interactions of adjacent fragments</li> <li>• Since the model is fit to the entire dataset it may incorrectly predict the trends in toxicity for certain chemical classes</li> <li>• Cannot provide external estimates of toxicity for compounds in the training set</li> </ul>
FDA	<ul style="list-style-type: none"> <li>• Can generate a new model based the closest analogs to the test compound</li> <li>• Always provides an external prediction of toxicity</li> </ul>	<ul style="list-style-type: none"> <li>• Predictions sometimes take longer since it has to generate a new model each time</li> </ul>
Nearest neighbor	<ul style="list-style-type: none"> <li>• Provides a quick estimate of toxicity</li> <li>• Allows one to determine structural analogs for a given test compound</li> <li>• Always provide an external prediction of toxicity</li> </ul>	<ul style="list-style-type: none"> <li>• It does not use a QSAR model to correlate the differences between the test compound and the nearest neighbors</li> <li>• Was shown to achieve the worst prediction results during external validation</li> </ul>
Consensus	<ul style="list-style-type: none"> <li>• Was shown to achieve the best prediction results during external validation</li> </ul>	<ul style="list-style-type: none"> <li>• Cannot provide external estimates of toxicity for compounds in the training set</li> </ul>

## 2. THEORY

### 2.1 Molecular Descriptors

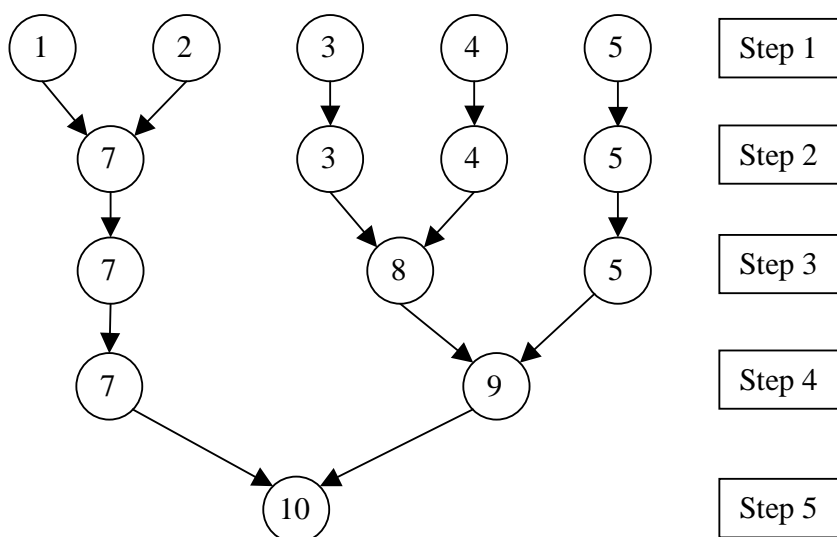
Molecular descriptors are physical characteristics of the structure of chemicals such as the molecular weight or the number of benzene rings. The overall pool of descriptors in the software contains 790 2-dimensional descriptors. The descriptors include the following classes of descriptors: E-state values 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 literatures sources 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 (Steinbeck et al. 2003). The Chemistry Development Kit (CDK) is a Java library for structural chemo- and bioinformatics which is available at the following link: <http://sourceforge.net/projects/cdk>. The descriptor values were validated using MDL QSAR (Elsevier MDL 2006), Dragon (Talete 2006), and Molconn-z (Edusoft-LC 2006). 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 (Romesburg 1984) 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 as follows:



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:

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

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

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

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  $k$ :

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

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:

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

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  $n_k/5$  since the recommended ratio of compounds to variables should be at least 5 (Eriksson et al. 2003; Topliss and Edwards 1979) 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 (The University of Waikato 2007; Witten 2005).

The genetic algorithm is used to maximize the adjusted 5 fold leave many out cross validation coefficient ( $q_{adj, LMO}^2$ ):

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

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  $q_{adj,LMO}^2$ ).

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 good (see pg 67 of (Eriksson et al. 2001)). If the  $q^2$  LOO is less than 0.5, the model from the cluster is not used to make predictions for test compounds.

The predicted toxicity ( $\hat{y}$ ) for a test chemical is given by the weighted average for all the valid predictions (Wikipedia.org 2008):

$$\hat{y} = \frac{\sum_{j=1}^{nvc} w_j \hat{y}_j}{\sum_{j=1}^{nvc} w_j} \quad (6)$$

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 (Wikipedia.org 2008)

$$w_j = \frac{1}{se_j^2} \quad (7)$$

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

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

where  $\sigma_j^2$  is given by

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

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

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

where  $X_o$  is the vector of model descriptor values for the test compound.

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

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

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

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

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

$$\hat{y} - \hat{u} \leq Toxicity \leq \hat{y} + \hat{u} \quad (13)$$

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 (Montgomery 1982)

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

The uncertainty is a function of the quality of the regression model (from the  $\sigma^2$  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 (Montgomery 1982). 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):

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

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<sub>3</sub> [aliphatic attach]), one compound with a methylene fragment (-CH<sub>2</sub> [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 wouldn't 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.

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 FDA (Food and Drug Administration) method is based on the work of Contrera and coworkers (Contrera, Matthews, and Benz 2003). 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

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

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 thirty 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 (Martin et al. 2008). 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. 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 (Benigni and Richard 1996). 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, ...) which does not increase linearly.

#### 2.2.4. *Group contribution*

The group contribution approach is based on the group contribution approach of Martin and Young (Martin and Young 2001). 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 since 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 any 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 ( $SC_{\min}$ ).  $SC_{\min}$  was set at 0.5 so that the prediction coverage was similar to the other QSAR methods (Martin et al. 2008). 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 the structural analogs are not accounted for.

### 2.2.6 *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)(Zhu et al. 2008). 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.

### 3. EXPERIMENTAL DATA SETS

#### 3.1. 96 hour fathead minnow LC<sub>50</sub> data set

The fathead minnow LC<sub>50</sub> endpoint represents the concentration in water which kills half of fathead minnow in 4 days (96 hours). The data set for this endpoint was obtained by downloading the ECOTOX aquatic toxicity database at the following link:

<http://cfpub.epa.gov/ecotox/>

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<sub>50</sub> 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, Hg, or Sn
- Compounds must represent a single pure component (i.e. salts, undefined isomeric mixtures, polymers, or mixtures were removed)

The LC<sub>50</sub> values were taken from the “Conc 1 (ug/L)” field in ECOTOX.

For chemicals with multiple LC<sub>50</sub> values, the median value was used.

In version 2.0 of T.E.S.T., the final dataset consisted of 819 chemicals. 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. The final fathead minnow LC<sub>50</sub> data set contained **809** chemicals. For use in QSAR modeling, the experimental values in µg/L were converted to -Log (LC<sub>50</sub> mol/L).

### 3.2. 40 hour *Tetrahymena pyriformis* IGC<sub>50</sub> data set

The *Tetrahymena pyriformis* IGC<sub>50</sub> endpoint represents the 50% growth inhibitory concentration of the *T. pyriformis* organism (a protozoan ciliate) after 40 hours. The IGC<sub>50</sub> training set was obtained from Zhu and coworkers (Zhu et al. 2008). Zhu and coworkers developed a training set containing 644 chemicals and two different prediction sets (one containing 339 chemicals and one containing 110 chemicals). Zhu and coworkers compiled the IGC<sub>50</sub> values from several publications of the Schultz group (Schultz 2007; Schultz and Netzeva 2004; Schultz et al. 2007).

The data sets developed by Zhu and coworkers were pooled to form a data set consisting of 1093 chemicals. The final *Tetrahymena pyriformis* IGC<sub>50</sub> data set consisted of 1085 chemicals (8 salts were omitted). The modeled endpoint was  $-\log(\text{IGC}_{50} \text{ mol} / \text{L})$ .

### 3.3. Oral rat LD<sub>50</sub> data set

The oral rat LD<sub>50</sub> 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 at the following link: <http://chem.sis.nlm.nih.gov/chemidplus/>

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<sub>50</sub> values were used (i.e. chemicals with LD<sub>50</sub> 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 a total of 106 2d isomers. In version 3.0, only one isomer was kept, using the average toxicity value. The final oral rat LD<sub>50</sub> data set contained **7286** chemicals. The modeled endpoint was the  $-\text{Log}(\text{LD}_{50} \text{ mol/kg})$ .

### **3.4. 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 (Hamelink 1977). A dataset of 643 chemicals was compiled from several different databases (Dimitrov et al. 2005; Arnot and Gobas 2006; EURAS; Zhao 2008). The final dataset consists of 610 chemicals (after removing salts, mixtures, and ambiguous compounds). The modeled endpoint was the Log (BCF).

## 4. VALIDATION OF THE QSAR METHODOLOGIES

### 4.1 Validation Methods

#### 4.1.1 Statistical external validation

The predictive ability of each of the QSAR methodologies was evaluated using statistical external validation (Gramatica and Pilutti 2004). In version 3.0 each overall data set was randomly divided into a training set (80% of the overall set) and a test set (20% of the overall set). 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 (Bourguignon, Deaguiar, Khots et al. 1994; Bourguignon, Deaguiar, Thorre et al. 1994; Kennard and Stone 1969; Snarey et al. 1997). 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).

A QSAR model has acceptable predictive power if the following conditions are satisfied (Golbraikh et al. 2003):

$$q^2 > 0.5; \quad (17)$$

$$R^2 > 0.6; \quad (18)$$

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

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 (Golbraikh and Tropsha 2002). The prediction coverage (fraction of chemicals predicted) must also be considered because the prediction accuracy (in terms of  $R^2_{\text{abs}}$  and RMSE) can sometimes be improved at the sacrifice of the prediction coverage.

#### 4.1.2 External validation

The predictive ability of the QSAR methodologies for the fathead minnow  $LC_{50}$  endpoint were also validated using external validation (using experimental data that was never part of the training set). External validation is the most demanding way to predictively validate a model (Eriksson et al. 2003). The external toxicity data set consisted of *rainbow trout*  $LC_{50}$  values for chemicals not present in the overall fathead minnow  $LC_{50}$  training set (the  $LC_{50}$  for the fathead minnow and the rainbow trout are highly correlated).

## 4.2 96 Hour Fathead minnow lethal concentration (LC<sub>50</sub>)

### 4.2.1 Statistical External Validation

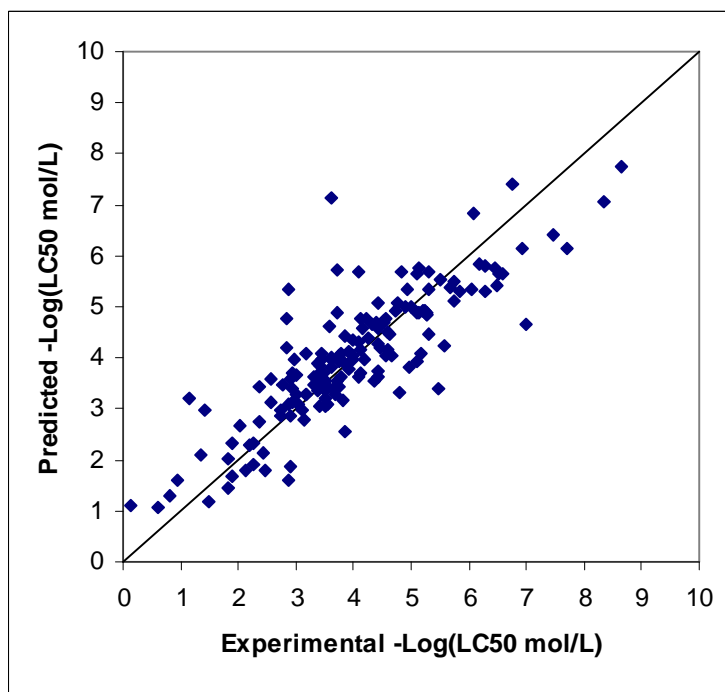
The prediction results for the fathead minnow LC<sub>50</sub> test set were as follows:

<b>Method</b>	<b>R<sup>2</sup></b>	<b><math>\frac{(R^2 - R_0^2)}{R^2}</math></b>	<b>k</b>	<b>RMSE</b>	<b>MAE</b>	<b>Coverage</b>
Hierarchical	0.693	0.105	0.960	0.821	0.599	0.988
Group contribution	0.684	0.072	0.980	0.822	0.600	0.883
Single Model	0.684	0.159	0.973	0.832	0.588	0.988
Nearest neighbor	0.602	0.108	0.980	0.952	0.700	0.938
FDA	0.582	0.146	0.970	0.990	0.735	0.981
Consensus	0.717	0.120	0.975	0.785	0.572	0.988

The consensus approach achieved the best results in terms of all the prediction statistics. The hierarchical and single model methods achieved slightly worse prediction statistics (the MAEs were about 0.02 log units higher than the consensus method). The group contribution method achieved R<sup>2</sup> and MAE values that were similar to the hierarchical and single model methods but the prediction coverage was lower (88% vs. 99%). The nearest neighbor and FDA methods achieved the worst prediction statistics (the MAEs were about 0.1 log units higher than those for the hierarchical method).

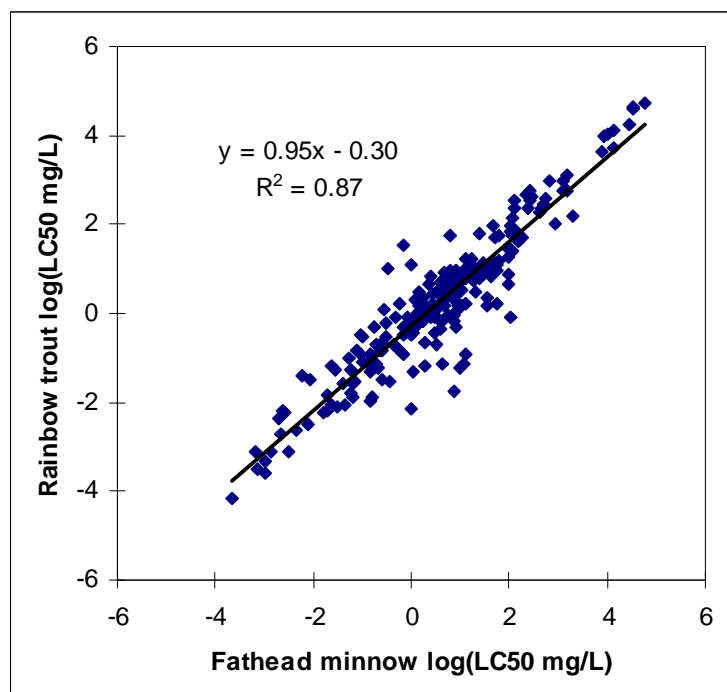
All of the QSAR methods (except for the FDA method) achieved an R<sup>2</sup> value (correlation coefficient for the test set) that met the constraint in equation 18. All of the QSAR methods achieved a k value (slope of the line Y=kX for the test set) which met the constraint in equation 19. All of the methods (except for the group contribution method) slightly violated the constraint for  $\frac{(R^2 - R_0^2)}{R^2}$  in equation 19.

The predicted values for the test set for the fathead minnow  $LC_{50}$  endpoint for the consensus method are as follows:



#### 4.2.2 External Validation

The predictive ability of the QSAR methodologies for the fathead minnow  $LC_{50}$  endpoint were also validated using true external validation (using experimental data that was never part of any training set). The external toxicity data was obtained by querying the ECOTOX database for *rainbow trout*  $LC_{50}$  values for chemicals not present in the overall fathead minnow  $LC_{50}$  training set of 819 chemicals.  $LC_{50}$  data for the rainbow trout were shown to be highly correlated with fathead minnow  $LC_{50}$  data (Sulaiman 1993; Thurston et al. 1985). There are 231 chemicals in the ECOTOX database (omitting salts and inorganic chemicals) which possess  $LC_{50}$  values for both the fathead minnow and the rainbow trout. The experimental  $LC_{50}$  values were again highly correlated:



The root mean square difference between the LC<sub>50</sub> values for the two species was 0.66 log units. For the 10 chemicals studied by Sulaiman (Sulaiman 1993), the root mean squared difference was 0.42 log units.

The prediction results for the different QSAR methodologies (hierarchical, single model, group contribution, FDA, and nearest neighbor) will be compared to those obtained using ECOSAR (v. 1.0)(USEPA 2009). ECOSAR is a computer program developed by the US EPA to predict aquatic toxicity to fish, aquatic invertebrates, and green algae. ECOSAR predicts acute fish toxicity using QSAR models for many different chemical classes. The models in ECOSAR are linear regressions between the toxicity and the octanol water partition coefficient. The fish toxicity values generated by ECOSAR are not species specific. Sometimes the ECOSAR software produces multiple predictions for the 96 hour fish toxicity for a given chemical (if a chemical can be assigned to more than one chemical class or if predictions are made using two different octanol water partition coefficient values). It was determined that slightly better

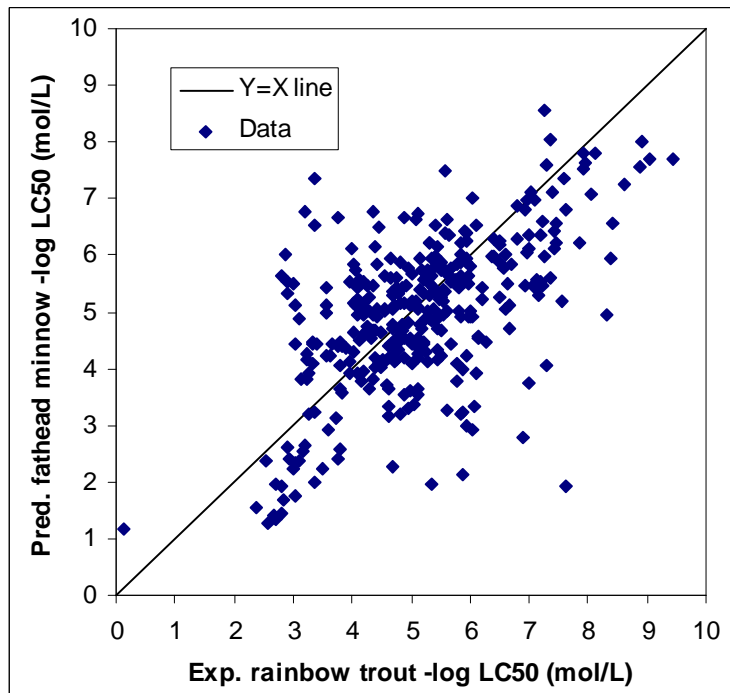
prediction results are achieved if the most conservative (i.e. smallest) toxicity value is used each time (i.e. instead of taking a geometric average). In some cases, ECOSAR indicated that the octanol water partition coefficient was too high for an accurate prediction or the predicted LC<sub>50</sub> concentration was greater than the chemical's water solubility. These predictions were considered to be outside the prediction space of the respective QSAR model.

The prediction results for the external rainbow trout dataset (n = 361) are as follows:

Method	RMSE	Coverage
Hierarchical	1.19	0.86
Single model	1.23	0.84
Group contribution	1.29	0.52
FDA	1.33	0.77
Nearest neighbor	1.29	0.70
Consensus	1.23	0.97
ECOSAR	1.57	0.78

As was the case in the statistical external validation, the consensus method achieved the best results in terms of both the RMSE and the prediction coverage. The consensus method achieved an RMSE that was 0.38 log units lower than the value obtained using ECOSAR. In addition the prediction coverage was appreciably higher (97% versus 78%). If one includes the ECOSAR predictions for chemicals where the predicted toxicity is greater than the water solubility, the RMSE remains the same at 1.57 log units and the coverage increases to 100%.

The predictions for the consensus method for the rainbow trout external set are as follows:



The  $R^2$  value (0.35) and the RMSE (1.23 log units) for the consensus method for the rainbow trout external set were poorer than the values achieved for the statistical validation exercise using fathead minnow experimental (0.72 and 0.79, respectively). A large portion of the increased RMSE can be attributed to the species difference between the fathead minnow and the rainbow trout (the root mean squared difference between the two species was previously shown to be ~0.42-0.66 log units). While the results for the rainbow trout test set were not excellent, they do indicate that the software can achieve superior prediction results to the ECOSAR program.

### 4.3 Tetrahymena pyriformis 50% growth inhibitory concentration (IGC<sub>50</sub>)

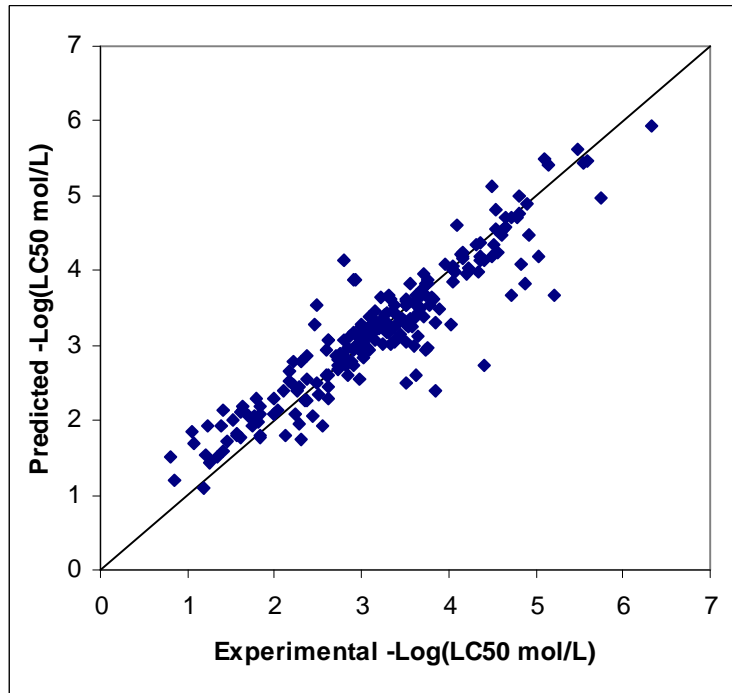
#### 4.3.1 Statistical External Validation

The prediction results for the IGC<sub>50</sub> test set were as follows:

<b>Method</b>	<b><math>R^2</math></b>	<b><math>\frac{(R^2 - R_0^2)}{R^2}</math></b>	<b><math>k</math></b>	<b>RMSE</b>	<b>MAE</b>	<b>Coverage</b>
Hierarchical	0.849	0.037	0.980	0.403	0.277	0.977
Group contribution	0.827	0.025	0.993	0.431	0.318	0.940
Single Model	0.784	0.042	0.988	0.484	0.338	0.977
Nearest neighbor	0.751	0.093	0.983	0.523	0.388	0.963
FDA	0.817	0.033	0.983	0.445	0.303	0.977
Consensus	0.851	0.049	0.983	0.406	0.283	1.000

Again the consensus method achieved the best results if one takes into both prediction accuracy and coverage. The nearest neighbor method achieved the worst results (the MAE was about 0.11 log units higher than the consensus method). The group contribution method and the FDA method produced results that were slightly worse than the hierarchical method. All of the methods achieved a prediction coverage (fraction of chemicals predicted) of at least 94%. All of the QSAR methods met the constraints in equation 18 and 19.

The prediction results for the consensus method are given by:



## 4.4 Oral rat LD<sub>50</sub> dataset

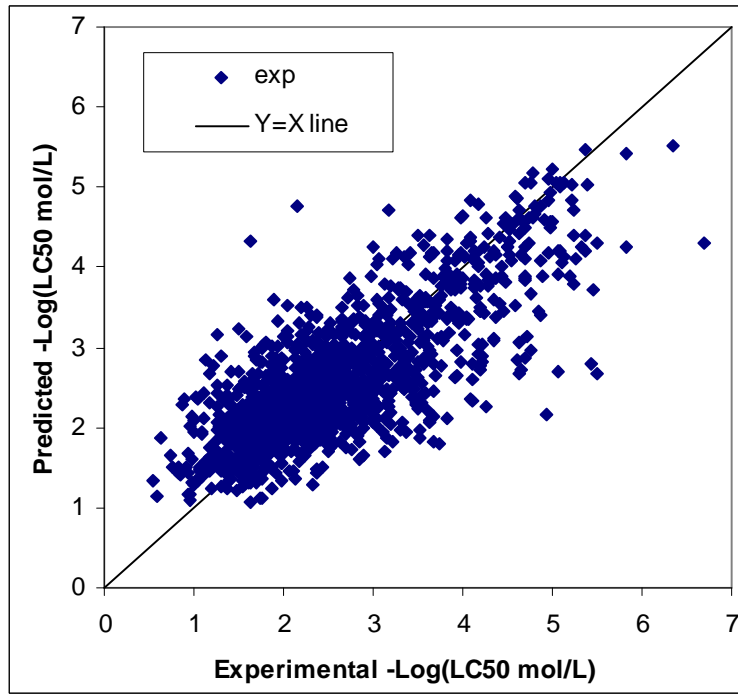
### 4.4.1 Statistical External Validation

The prediction results for the LD<sub>50</sub> endpoint were as follows:

<b>Method</b>	<b><math>R^2</math></b>	<b><math>\frac{(R^2 - R_0^2)}{R^2}</math></b>	<b><math>k</math></b>	<b>RMSE</b>	<b>MAE</b>	<b>Coverage</b>
Hierarchical	0.573	0.194	0.951	0.654	0.473	0.847
Nearest neighbor	0.546	0.293	0.957	0.662	0.479	0.995
FDA	0.555	0.236	0.959	0.658	0.481	0.987
Consensus	0.620	0.265	0.954	0.596	0.437	1.000

It was not possible to develop a single model or a group contribution model that fit the entire training set. The consensus method achieved the best results in terms of both prediction accuracy and prediction coverage. The nearest neighbor and FDA methods achieved MAE values that were slightly higher (0.04 log units) at about 99% coverage. The hierarchical method achieved prediction accuracy similar to the nearest neighbor and FDA methods but the prediction coverage was lower (85%). All of the methods (with the exception of the consensus method) barely missed meeting the constraint for  $R^2$  in equation 18. None of the methods met the constraint in equation 19 for  $\frac{(R^2 - R_0^2)}{R^2}$ . This indicates that while the prediction error for the oral rat LD<sub>50</sub> endpoint is fairly low, the methods do not correlate the experimental data as well as they do for the fathead minnow LC<sub>50</sub> and the *T. pyriformis* IGC<sub>50</sub> endpoints.

The prediction results for the consensus method are given by:



## 4.5 Bioaccumulation factor (BCF)

### 4.5.1 Statistical External Validation

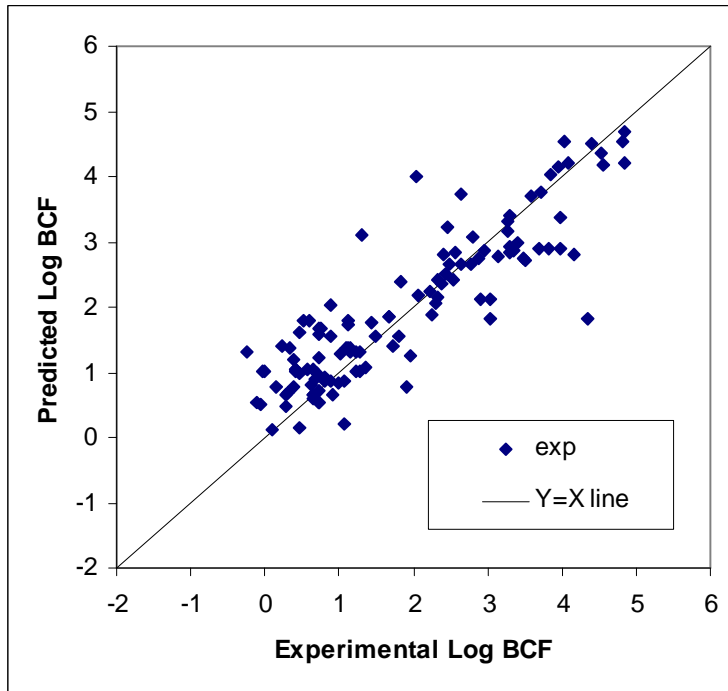
The prediction results for the BCF endpoint were as follows:

<b>Method</b>	<b><math>R^2</math></b>	<b><math>\frac{(R^2 - R_0^2)}{R^2}</math></b>	<b><math>k</math></b>	<b>RMSE</b>	<b>MAE</b>	<b>Coverage</b>
Hierarchical	0.755	0.103	0.961	0.691	0.507	0.952
Single Model	0.740	0.114	0.934	0.704	0.545	0.952
Nearest neighbor	0.704	0.094	0.960	0.776	0.536	0.871
FDA	0.731	0.063	0.938	0.716	0.548	0.927
Consensus	0.781	0.115	0.945	0.656	0.480	0.968

Again the consensus method yielded the best results in terms of prediction accuracy and coverage. The other methods yielded lower  $R^2$  values and prediction coverage.

The BCFBAF module (v. 3.0) of US EPA's EPI Suite software package (USEPA 2009) yielded an  $R^2$  value of 0.772 and MAE of 0.492 (coverage =100%). Thus the predictions for the consensus method are comparable to those from EPI Suite. However, this may not be a fair comparison since some of the chemicals in the prediction set may have appeared in the training set for the BCF model in EPI Suite.

The prediction results for the consensus method are given by:


















## 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 also use the periodic table  to choose an element. Finally with  you can go through some common elements by clicking on an atom repeatedly. With  and  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* ([http://www.mdl.com/solutions/white\\_papers/ctfile\\_formats.jsp](http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp))

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.

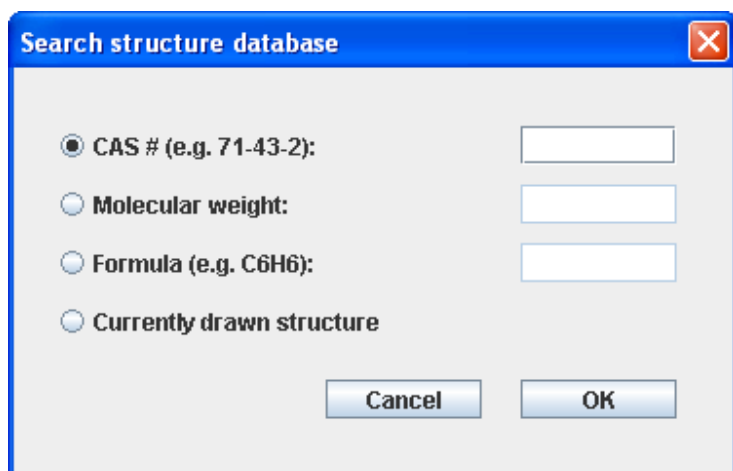


For example, to import benzene enter c1ccccc1 as the SMILES string. A SMILES string can be pasted from the clipboard by selecting **Generate from SMILES on clipboard**.

#### 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:



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 also 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:  
<http://www.epa.gov/nrmrl/std/cppb/qsar/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  0  0999 v2000
    0.0000  0.0000  0.0000 C  0  0  0  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  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 a unique identifier on each line. The SMILES string and the identifier can be separated by a comma, tab, or a space. The text file should not contain 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
```

### 3.2.4. 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 as follows:

**T.E.S.T (Toxicity Estimation Software Tool)**

#	ID	Formula	Error
1	50-30-6	C7H4O2Cl2	
2	51-03-6	C19H30O5	
3	52-51-7	C3H6NO4Br	
4	52-60-8	C13H21O3S2P	
5	54-11-5	C10H14N2	
6	56-72-4	C14H16O5SPCl	
7	56-81-5	C3H8O3	
8	57-24-9	C21H22N2O2	
9	58-36-6	C24H16O3As2	
10	60-51-5	C5H12NO3S2P	
11	61-82-5	C2H4N4	
12	62-38-4	C8H8O2Hg	
13	64-00-6	C11H15NO2	
14	72-55-9	C14H8Cl4	
15	72-56-0	C18H20Cl2	
16	75-64-9	C4H11N	
17	76-06-2	CNO2Cl3	
18	77-48-5	C5H6N2O2Br2	
19	78-48-8	C12H27OS3P	
20	79-09-4	C3H6O2	
21	79-33-4	C3H6O3	

Note: double click to edit a chemical

Endpoint: 
 Method:

Relax fragment constraint

#### 5.2.4. *Deleting chemicals from 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.

#### 5.2.5. *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).

#### 5.2.6. *Saving the batch list*

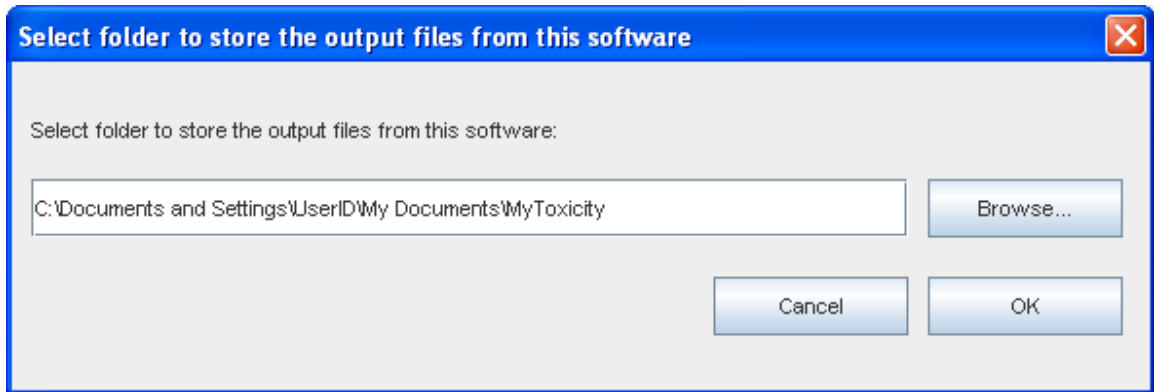
To save the batch list as an MDL SD file, click on the **Save list as SDF** button. This feature allows you to save changes to your list.

#### 5.2.7. *Closing the batch list*

To close the batch list click on the **Close batch list** button. One can also 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.
- Select a toxicity endpoint using the drop down list provided (the fathead minnow LC<sub>50</sub> 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. 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:



- 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 fenthion (for the Fathead minnow LC<sub>50</sub> endpoint and the *Consensus method*) are as follows:

Prediction results		
Endpoint	Experimental value CAS: 55-38-9 Source: <a href="#">ECOTOX</a>	Predicted value <sup>a</sup>
Fathead minnow LC <sub>50</sub> -Log(mol/L)	5.00	4.99
Fathead minnow LC <sub>50</sub> mg/L	2.81	2.84

<sup>a</sup>Note: the test chemical was present in the external test set.

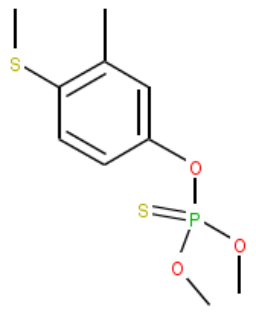

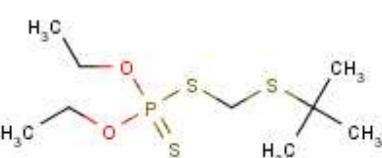
The predicted toxicity is 2.84 mg/L and the experimental value is 2.81 mg/L. The prediction is flagged in this example because the chemical was part of the external test set.

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:

Individual Predictions	
Method	Predicted value -Log(mol/L)
Hierarchical clustering	<a href="#">4.64</a>
Single model	<a href="#">4.97</a>
Group contribution	<a href="#">N/A</a>
FDA	<a href="#">4.59</a>
Nearest neighbor	<a href="#">5.76</a>

The predicted values from the different QSAR methods are distributed about the experimental value of 4.99 (in units of -Log(mol/L)). A prediction for the *Group contribution method* is not available because it failed the fragment constraint (there were no chemicals in the training set which possessed a sulfur atom attached to an aromatic atom).

The software also provides predictions for chemicals in the test set which are similar to the test chemical. The predictions for the most similar chemicals in the test set are as follows:

CAS	Structure	Similarity Coefficient	Experimental value -Log(mol/L)	Predicted value -Log(mol/L)
55-38-9 (test chemical)			5.00	4.99
78-34-2		0.60	5.14	5.77
13071-79-9		0.51	6.50	5.41

These results illustrate that similar chemicals in the test set are fairly toxic and are fairly accurately predicted.

One can view the details of the predictions for the different QSAR methods by clicking on the predicted value for each method. For example, the details for the *Hierarchical clustering method* are as follows:

Prediction results			
Endpoint	Experimental value CAS: 55-38-9 Source: <a href="#">ECOTOX</a>	Predicted value <sup>a</sup>	Prediction interval
Fathead minnow LC <sub>50</sub> - Log(mol/L)	5.00	4.64	4.18 ≤ Tox ≤ 5.11
Fathead minnow LC <sub>50</sub> mg/L	2.81	6.32	2.15 ≤ Tox ≤ 18.58

<sup>a</sup>Note: the test chemical was present in the external test set.

Cluster model predictions and statistics					
Cluster model	Test chemical descriptor values	Prediction interval -Log(mol/L)	r <sup>2</sup>	q <sup>2</sup>	#chemicals
<a href="#">1245</a>	<a href="#">Descriptors</a>	4.813 ± 0.638	0.872	0.800	26
<a href="#">1256</a>	<a href="#">Descriptors</a>	3.081 ± 0.944	0.836	0.662	30
<a href="#">1282</a>	<a href="#">Descriptors</a>	5.044 ± 0.976	0.773	0.692	156
<a href="#">1288</a>	<a href="#">Descriptors</a>	5.457 ± 1.156	0.681	0.626	235
<a href="#">1290</a>	<a href="#">Descriptors</a>	5.059 ± 1.397	0.672	0.611	264
<a href="#">1292</a>	<a href="#">Descriptors</a>	4.968 ± 1.301	0.730	0.691	644

Cluster models with violated constraints				
Cluster Model	r <sup>2</sup>	q <sup>2</sup>	# chemicals	Message
<a href="#">955</a>	0.972	0.848	5	Rmax constraint not met
<a href="#">1134</a>	0.851	0.634	11	Rmax constraint not met
<a href="#">1162</a>	0.902	0.778	11	Fragment constraint not met
<a href="#">1168</a>	0.854	0.757	17	Fragment constraint not met
<a href="#">1198</a>	0.860	0.719	18	Model ellipsoid constraint not met
<a href="#">1234</a>	0.834	0.770	19	Fragment constraint not met
<a href="#">1238</a>	0.829	0.693	22	Model ellipsoid constraint not met

[Descriptor values for test chemical](#)

The predicted toxicity is 6.32 mg/L and the experimental value is 2.81 mg/L. The prediction interval is  $2.15 \leq \text{Tox} \leq 18.58$  mg/L (one is 90% confident that the predicted value is between 2.15 and 18.58 mg/L).

For the hierarchical method, the web page also includes a summary of all the predictions from the models used to estimate the toxicity. For the example compound, seven hierarchical cluster models were used to generate the estimated toxicity value. Most of the models generated toxicity values of about 5 (in terms of  $-\text{Log}(\text{mol/L})$ ). The web page also includes the models which were not used to make a prediction since the test chemical violated one or more of the models' constraints. For example the model for cluster #1162 could not be used because fragment constraint was violated (since none of the chemicals in this cluster contained a sulfur atom that was attached to an aromatic ring).

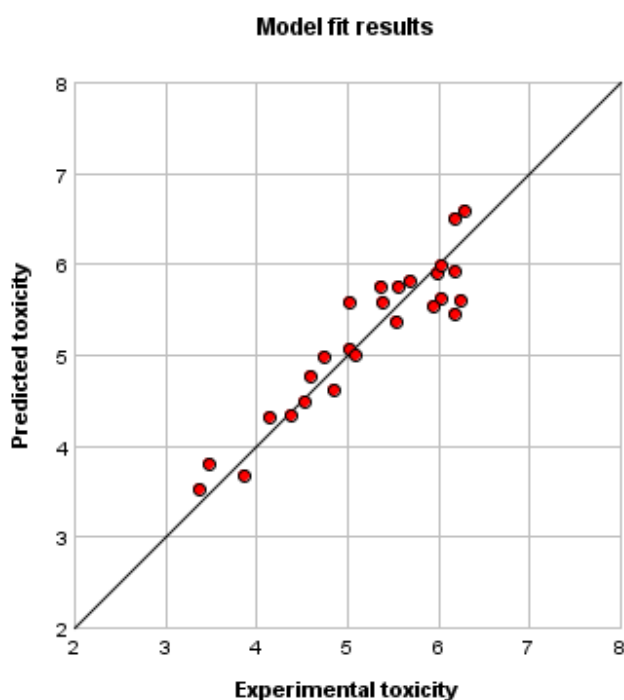
One can click on the link for each model (in the **Cluster model** column) to display its statistics, regression plot, parameters, and chemical descriptor values.

For example for model #1245, the model statistics are as follows:

Parameter	Value
Endpoint	Fathead minnow LC <sub>50</sub>
r <sup>2</sup>	0.872
q <sup>2</sup>	0.800
#chemicals	26
Model	1245

The r<sup>2</sup> value is the correlation coefficient and the q<sup>2</sup> value is the leave one out correlation coefficient.

The model regression plot is as follows:



The model parameters are as follows:

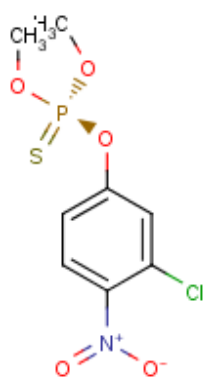
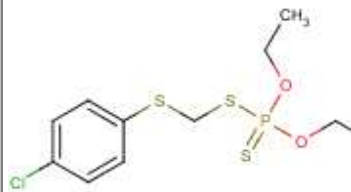
Model coefficients			
Coefficient	Definition	Value	Uncertainty*
Intercept	Model intercept	5.6003	1.6837
SdssS	Sum of ( = S < ) E-States (SdssS)	1.1890	0.4787
Hmin	Minimum hydrogen E-State value in molecule.	-1.8248	0.6767
BEHm7	Highest eigenvalue n. 7 of Burden matrix / weighted by atomic masses	1.1915	0.5755
BELp5	Lowest eigenvalue n. 5 of Burden matrix / weighted by atomic polarizabilities	-1.6559	0.9128
MATS7v	Moran autocorrelation - lag 7 / weighted by atomic van der Waals volumes	3.3274	1.0675

\* value for 90% confidence interval

The above table indicates that the equation for the model is as follows:

$$\text{Toxicity} = 1.1890 \times (\text{SdssS}) - 1.8248 \times (\text{Hmin}) + 1.1915 \times (\text{BEHm7}) - 1.6559 \times (\text{BELp5}) + 3.3274 \times (\text{MATS7v}) + 5.6003$$

The descriptors for the model chemicals are as follows:

Number	Chemical	Structure	Exp. Toxicity -Log(mol/L)	Fit Toxicity -Log(mol/L)	SdssS	...	MATS7v
1	500-28-7		5.0270	5.0656	0.0000	...	-0.1315
...	...	...	...	...	...	...	...
26	786-19-6		6.1770	5.9257	0.0000	...	-0.0802

One can click on the links in the **Test chemical descriptor values** column to view the descriptor values for the test chemical for the given cluster model. For cluster#1245, the model descriptors for the test chemical are as follows:

Descriptor Values	
Descriptor	Value
SdssS	0.0000
Hmin	0.5962
BEHm7	2.5664
BELp5	1.2518
MATS7v	-0.2057

## 6. REFERENCES

- Arnot, J.A., and Gobas, F.A.P.C. 2006. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environ. Rev.* 14:257-297.
- Benigni, R., and Richard, A.M. 1996. QSARS of mutagens and carcinogens: Two case studies illustrating problems in the construction of models for noncongeneric chemicals. *Mutat. Res.* 371:29-46.
- Bourguignon, B., Deaguiar, P.F., Khots, M.S., and Massart, D.L. 1994. Optimization in Irregularly Shaped Regions: pH and Solvent Strength in Reversed-Phase High-Performance Liquid Chromatography Separations. *Analytical Chemistry* 66:893-904.
- Bourguignon, B., Deaguiar, P.F., Thorre, K., and Massart, D.L. 1994. *Journal of Chromatography Science* 32:144-152.
- Contrera, J.F., Matthews, E.J., and Benz, R.D. 2003. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharm.* 38:243-259.
- Dimitrov, S., Dimitrova, N., Parkerton, T., Combers, M., Bonnell, M., and Mekenyan, O. 2005. Base-line model for identifying the bioaccumulation potential of chemicals. *SAR QSAR Environ. Res.* 16:531-554
- Edusoft-LC, Molconn-z Version 4.0. <http://www.edusoft-lc.com/molconn/> (Accessed on 5/26/09).
- Elsevier MDL, MDL QSAR Version 2.2. <http://www.mdl.com/products/predictive/qsar/index.jsp> (Accessed on 8/17/2006).
- Eriksson, L., Jaworska, J.S., Worth, A.P., Cronin, M.T.D., McDowell, R.M., and Gramatica, P. 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Persp.* 111 (10):1361-1375.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S. 2001. *Multi- and Megavariate Data Analysis - Principles and Applications*. Umea, Sweden: Umetrics AB.
- EURAS, Establishing a bioconcentration factor (BCF) Gold Standard Database. <http://www.euras.be/eng/project.asp?ProjectId=92> (Accessed on 5/20/09).
- Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., and Tropsha, A. 2003. Rational Selection of Training and Test sets for the Development of Validated QSAR Models. *J. Comput. Aid. Mol. Des.* 17:241-253.
- Golbraikh, A., and Tropsha, A. 2002. Beware of  $q^2$ ! *J. Mol. Graph. Model.* 20:269-276.
- Gramatica, P., and Pilutti, P. 2004. 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.
- Hamelink, J.L. 1977. Current bioconcentration test methods and theory. In *Aquatic Toxicology and Hazard Evaluation*, edited by F. L. Mayer and J. L. Hamelink. West Conshohocken, PA ASTM STP.
- Kennard, R.W., and Stone, L.A. 1969. *Technometrics* 11:137-148.
- Martin, T.M., Harten, P., Venkatapathy, R., Das, S., and Young, D.M. 2008. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Method.* 18:251-266.

- Martin, T.M., and Young, D.M. 2001. Prediction of the Acute Toxicity (96-h LC<sub>50</sub>) of Organic Compounds to the Fathead Minnow (*Pimephales promelas*) Using a Group Contribution Method. *Chem. Res. Toxicol.* 14:1378-1385.
- Montgomery, D.C. 1982. Introduction to linear regression analysis. New York: John Wiley and Sons.
- Romesburg, H.C. 1984. *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications.
- Schultz, T. W., Tetratox. <http://www.vet.utk.edu/TETRATOX/> (Accessed on 5/26/09).
- Schultz, T.W., Hewitt, M., Netzeva, T.I., and Cronin, M.T.D. 2007. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR Comb. Sci.* 26 (2):234-254.
- Schultz, T.W., and Netzeva, T.I. 2004. Development and Evaluation of QSARs for Ecotoxic Endpoints: The Benzene Response-Surface Model for Tetrahymena Toxicity. In *Modeling Environmental Fate and Toxicity*, edited by M. T. D. Cronin and D. J. Livingstone. Boca Raton, FL: CRC Press.
- Snarey, M., Terrett, N.K., Willet, P., and Wilton, D.J. 1997. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graph. Model.* 15:372-385.
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. 2003. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comp. Sci.* 43:493-500.
- Sulaiman, A.H. 1993. Acute toxicity relationships for two species of fish using a simultaneous testing method. *Sci. Total Environ.* 134, Supplement 2:1001-1009.
- Talete, Dragon Version 5.4. <http://www.talete.mi.it/> (Accessed on 5/26/09).
- The University of Waikato, WEKA - The Waikato Environment for Knowledge Analysis. <http://www.cs.waikato.ac.nz/~ml/weka/> (Accessed on 5/26/09).
- Thurston, R.V., Gilfoil, T.A., Meyn, E.L., Zajdel, R.K., Aoki, T.I., and Veith, G. 1985. Comparative toxicity of ten organic chemicals to ten common aquatic species. *Water Res.* 19 (9):1145-1156.
- Topliss, J.G., and Edwards, R.P. 1979. Chance factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* 22 (10):1238-1244.
- US EPA, EPI SUITE, Version 4.0. <http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm> (Accessed on 5/21/09).
- Wikipedia.org, Weighted mean. [http://en.wikipedia.org/wiki/Weighted\\_mean](http://en.wikipedia.org/wiki/Weighted_mean) (Accessed on 5/26/09).
- Witten, I.H. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
- Zhao, C.B., E.; Chana, A.; Roncaglioni, A.; Benfenati, E. 2008. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73:1701-1707.
- Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Öberg, T., Dao, P., Cherkasov, A., and Tetko, I.V. 2008. Combinational QSAR Model of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* 48:766 - 784.