https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

User’s Guide for T. E. S. T. (Toxicity Estimation Software Tool) Version 5.1

A Java Application to Estimate Toxicities and Physical Properties from Molecular Structure

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

Todd M. Martin
U.S. Environmental Protection Agency
Center for Computational Toxicology and Exposure
Chemical Characterization and Exposure Division
Cincinnati, OH 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 G-STD-0013882-QP-1-3). 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 Center for Computational Toxicology & Exposure (CCTE) is a scientific organization working to support Agency decisions by providing solutions-driven research to rapidly evaluate the potential human health and environmental risks due to exposures to environmental chemicals and ensure the integrity of the freshwater environment and its capacity to support human well-being. To do this, CCTE research strives to:

- Reduce the time required to thoroughly test chemicals and other emerging materials for human health and ecological toxicity from years to months.
- Expand our understanding of quantitative human and ecological exposures for thousands of chemical substances and emerging materials.
- Develop a comprehensive information system that contains relevant actionable chemical safety and ecological data with the software tools to integrate them for a range of human health and environmental decisions.
- Reduce the time required to characterize freshwater ecosystems and project the future state of ecological condition and ecosystem services from decades to years.
- Demonstrate translation of CCTE data, models, and tools into regulatory decisions by EPA Program Offices, EPA Regions, and States to protect human health and the environment.

Using the knowledge and tools developed from this research, CCTE performs rapid chemical screening and evaluation that allows thousands of chemicals to be evaluated for potential risk in a very short amount of time. The data and tools produced by CCTE researchers can then be leveraged to help Region and Program Offices, states, tribes, and communities make decisions to sustain a healthy society and environment.

Abstract

This guide provides an introduction to QSAR (Quantitative Structure Activity Relationship) models, a detailed description of the QSAR methodologies in T.E.S.T. (Toxicity Estimation Software Tool), 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. USING THE SOFTWARE

1.1. Importing chemicals in single chemical mode

A compound can be imported into the software using the following methods:

- Using the provided molecular structure drawing tool
- Importing from an MDL molfile
- Searching by CAS number, SMILES string, or name

1.1.1. Drawing a molecule using the structure drawing tool

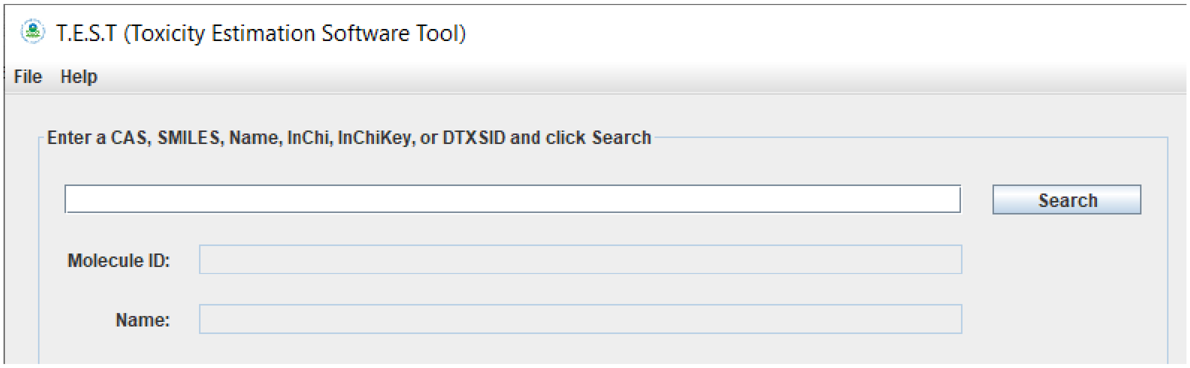
- First, add any rings present in the molecule using the ring template buttons △ □ ⬠ ⬡ ○ ○ ⌬ (click on a button and then click somewhere in the document).
- You can undo any unwanted drawing by clicking Ctrl z.
- Next, add any chains using the / button.
- Next, add double or triple bonds by using / again 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.
- Next, any hetero atoms (non-carbon atoms) need to be set. Use one of the element symbol buttons and then click on an atom to change it to this symbol. Alternatively, you can use the periodic table to choose an element and then click on an atom in the drawing box to change it to that element.
- Finally, the charge can changed by right clicking on an atom and selecting Charge from the pop up menu.

1.1.2. Importing a molecule from an MDL molfile

The structure for a test compound can be imported from an MDL molfile (V2000)[1]. To import a structure using a MDL molfile, click the toolbar button. Click on the location of the file, select the file name, and click open.

1.1.3. Importing by identifier

To import a structure by identifier (either CAS number, SMILES string, chemical name, InChi, InChiKey, or DTXSID) in the search field and click **Search** (or press **Enter**):


T.E.S.T (Toxicity Estimation Software Tool)
File
Help
Enter a CAS, SMILES, Name, InChi, InChiKey, or DTXSID and click Search
Search
Molecule ID:
Name:

The molecule will load in the structure window on the right.

1.2. Importing multiple compounds (batch import)

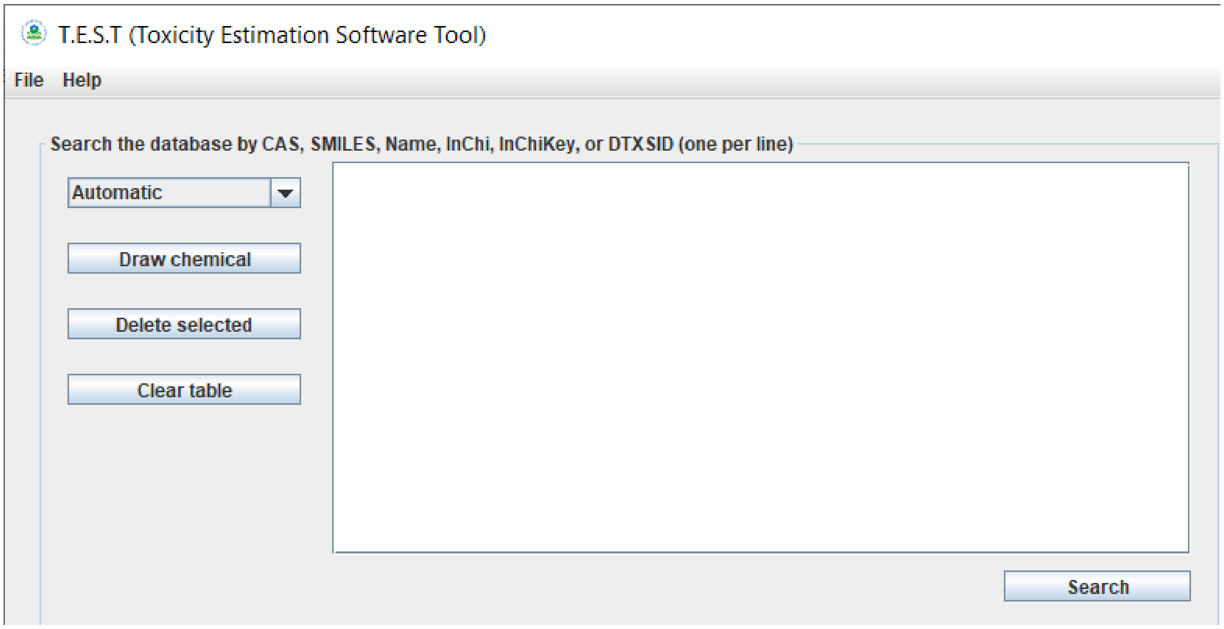
To switch to batch mode, click on the **Switch to Batch Mode** button in the bottom right hand corner or select **Switch to Batch Mode** from the **File** menu.

Multiple compounds can be imported simultaneously several different ways:

- Using the batch search box
- Importing from a MDL SDfile
- Importing from a list of CAS numbers
- Importing from a list of SMILES strings
- Importing one of the training or prediction sets

1.2.1. Using the batch search box

To import structures from the structure database, enter a series of identifiers (either CAS number, SMILES string, chemical name, InChI, InChiKey, or DTXSID) with one identifier on each line in the **Search** field and click **Search**:


T.E.S.T (Toxicity Estimation Software Tool)
File
Help
Search the database by CAS, SMILES, Name, InChi, InChiKey, or DTXSID (one per line)
Automatic
Draw chemical
Delete selected
Clear table
Search

Note: If desired, a custom ID can also be added on each line (separated by a tab character). For example, if searching by smiles:

c1ccccc1	Chemical1

1.2.2. Importing from a MDL SDfile

To import multiple structures from an MDL SDfile select **Batch import from MDL .mol/.sdf file** from the **File** menu.

For best results, one should use SDfiles with a "CAS" field included to uniquely identify each chemical in the file. If the CAS field is not present, the software will attempt to retrieve a match in the database using the molecular structure.

For example, a sample from an SDfile including formaldehyde would be as follows:

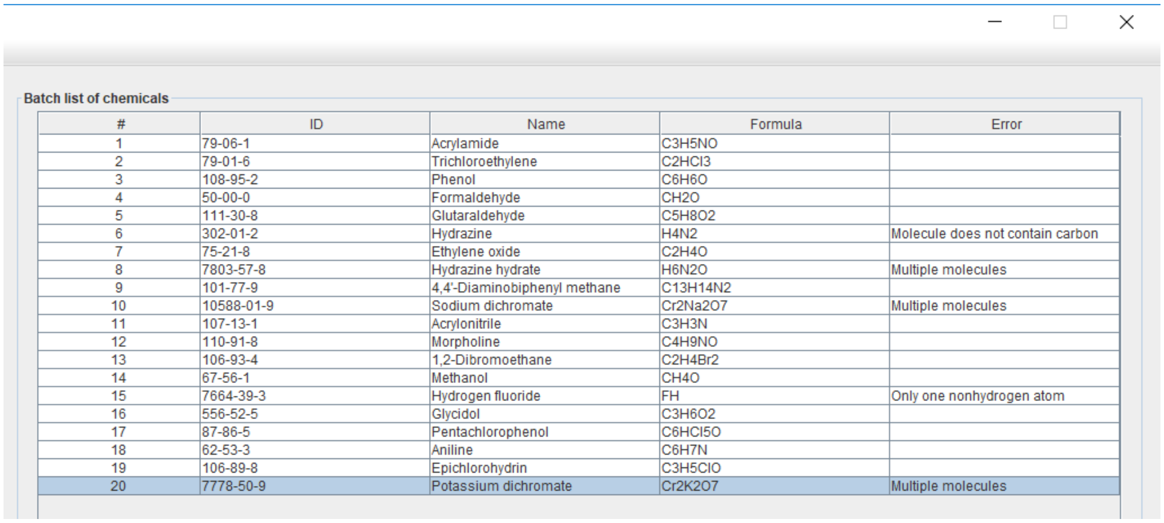
```
Formaldehyde
csChFnd80/07260508122D

 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

$$$$
```

After importing the desired set of chemicals, you can edit an individual chemical in the list by double clicking on its row in the list, which will bring you to the structure drawing tool where you can edit the structure. An example of an imported batch list is as follows:



Batch list of chemicals

#	ID	Name	Formula	Error
1	79-06-1	Acrylamide	C3H5NO	
2	79-01-6	Trichloroethylene	C2HCl3	
3	108-95-2	Phenol	C6H6O	
4	50-00-0	Formaldehyde	CH2O	
5	111-30-8	Glutaraldehyde	C5H8O2	
6	302-01-2	Hydrazine	H4N2	Molecule does not contain carbon
7	75-21-8	Ethylene oxide	C2H4O	
8	7803-57-8	Hydrazine hydrate	H6N2O	Multiple molecules
9	101-77-9	4,4'-Diaminobiphenyl methane	C13H14N2	
10	10588-01-9	Sodium dichromate	Cr2Na2O7	Multiple molecules
11	107-13-1	Acrylonitrile	C3H3N	
12	110-91-8	Morpholine	C4H9NO	
13	106-93-4	1,2-Dibromoethane	C2H4Br2	
14	67-56-1	Methanol	CH4O	
15	7664-39-3	Hydrogen fluoride	FH	Only one nonhydrogen atom
16	556-52-5	Glycidol	C3H6O2	
17	87-86-5	Pentachlorophenol	C6HCl5O	
18	62-53-3	Aniline	C6H7N	
19	106-89-8	Epichlorohydrin	C3H5ClO	
20	7778-50-9	Potassium dichromate	Cr2K2O7	Multiple molecules

1.2.3. Importing from a file containing list of CAS numbers

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

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

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

1.2.4. Importing from a file containing list of SMILES strings

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

The text file should contain the SMILES string and, if desired, a unique identifier on each line. A tab should separate the SMILES string and the identifier. The text file should not container a header line.

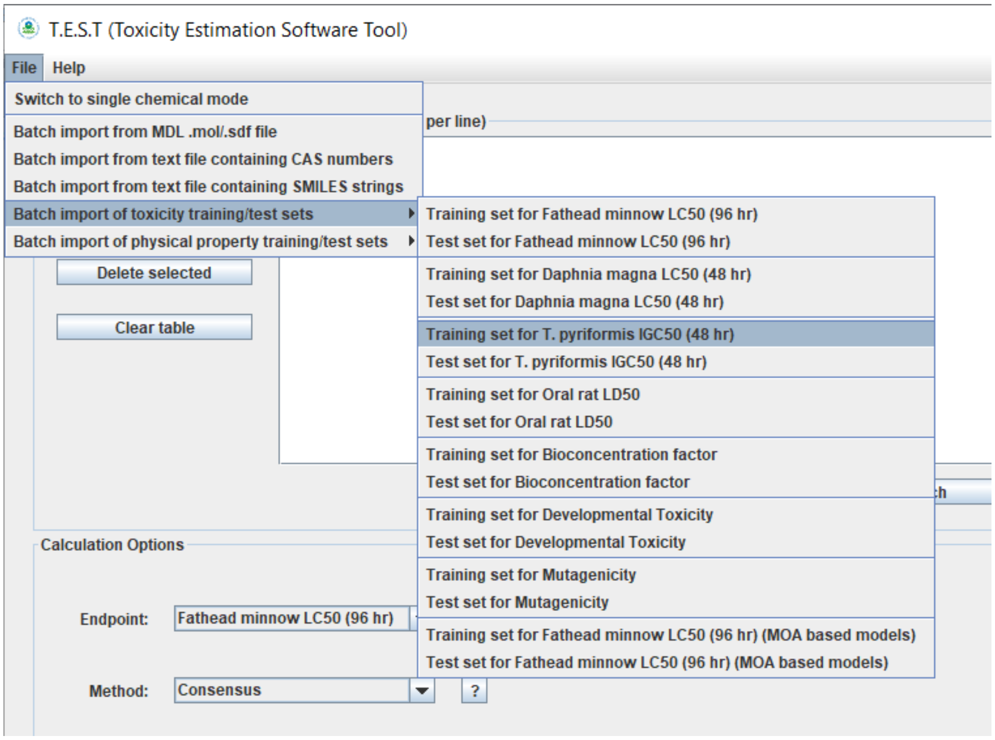
For example, to import benzene and formaldehyde, the contents of the text file should be as follows:

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

If no identifier is present, the software will attempt to retrieve the CAS number based on the molecular structure given by the smiles string.

1.2.5. Importing from training and prediction sets

The training and prediction sets for each endpoint can be loaded in batch mode by going to the File menu and selecting a set from **Batch import of toxicity training/test sets** or **Batch import of physical property training/test sets**:


T.E.S.T (Toxicity Estimation Software Tool)
File
Help
Switch to single chemical mode
Batch import from MDL .mol/.sdf file
Batch import from text file containing CAS numbers
Batch import from text file containing SMILES strings
Batch import of toxicity training/test sets
Batch import of physical property training/test sets
per line)
Delete selected
Clear table
Training set for Fathead minnow LC50 (96 hr)
Test set for Fathead minnow LC50 (96 hr)
Training set for Daphnia magna LC50 (48 hr)
Test set for Daphnia magna LC50 (48 hr)
Training set for T. pyriformis IGC50 (48 hr)
Test set for T. pyriformis IGC50 (48 hr)
Training set for Oral rat LD50
Test set for Oral rat LD50
Training set for Bioconcentration factor
Test set for Bioconcentration factor
Training set for Developmental Toxicity
Test set for Developmental Toxicity
Training set for Mutagenicity
Test set for Mutagenicity
Training set for Fathead minnow LC50 (96 hr) (MOA based models)
Test set for Fathead minnow LC50 (96 hr) (MOA based models)
Calculation Options
?

Endpoint: Fathead minnow LC50 (96 hr)

Method: Consensus

1.2.6. *Adding chemicals to the batch list*

To add chemicals to the list, search using the batch search box or add a chemical by clicking on the **Draw chemical** button.

1.2.7. *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 selected** button (or press the Delete key on the keyboard).

1.2.8. *Returning to Single Chemical Mode*

To return to the single chemical mode, click on the blue **Switch to Single Mode** button.

1.3. Performing toxicity predictions

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. This constraint can be relaxed by checking the **Relax fragment constraint** checkbox. The fragment constraint is described in the Theory section.

Select the output folder by clicking the **Browse…** button in the bottom left hand corner. Note: if "MyToxicity" is not present in the folder name, a "MyToxicity" folder will be appended to the path.

To generate detailed reports in single chemical mode, make sure the **Create detailed reports** checkbox is checked.

To able to save reports in batch chemical mode, make sure the **Create reports** checkbox is checked.

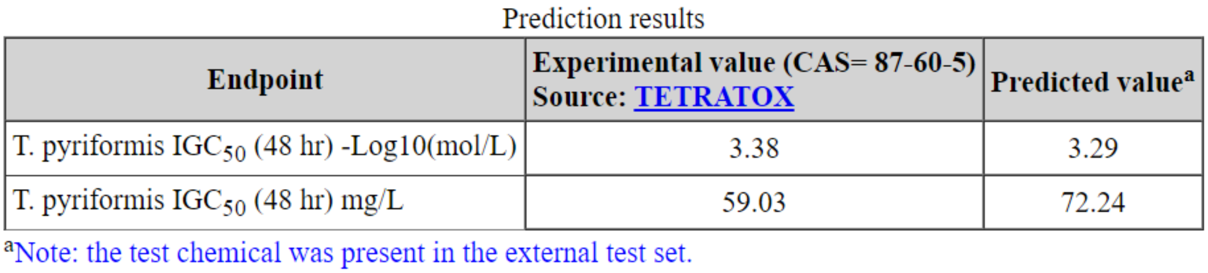
Once the desired options have been selected, the toxicity estimation calculations can be started by clicking the green **Calculate!** button in the bottom right hand corner**.** This button will change to a red **Stop** button while the calculations are processing.

To abort the currently running calculations, click on the red **Stop** button.

1.4. Interpretation of QSAR prediction report

After performing the toxicity estimation calculations using the **Calculate!** button, a QSAR prediction report is generated, which displays the results in the default web browser. The results for 87-60-5 (for the *Tetrahymena pyriformis* IGC_{50} endpoint and the Consensus method) are provided in Table 1.4.1. The predicted toxicity is 72.24 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 predictions are displayed below the table for the consensus method results. The average of the values from all the different QSAR methods is 3.29, which is close to the experimental value of 3.38 (in units of -Log(mol/L)).

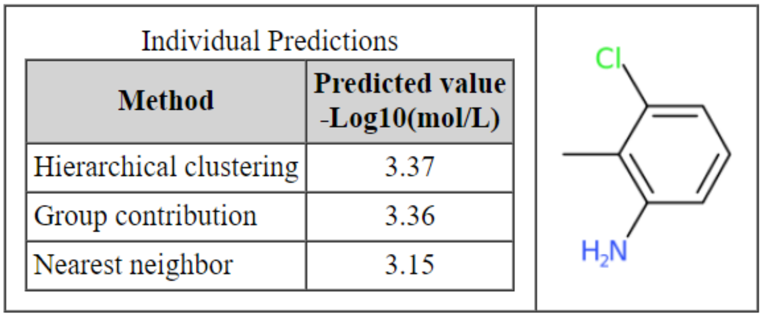
Table 1.4.1. Prediction results from the consensus method for 87-60-5



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	3.29
T. pyriformis IGC_{50} (48 hr) mg/L	59.03	72.24

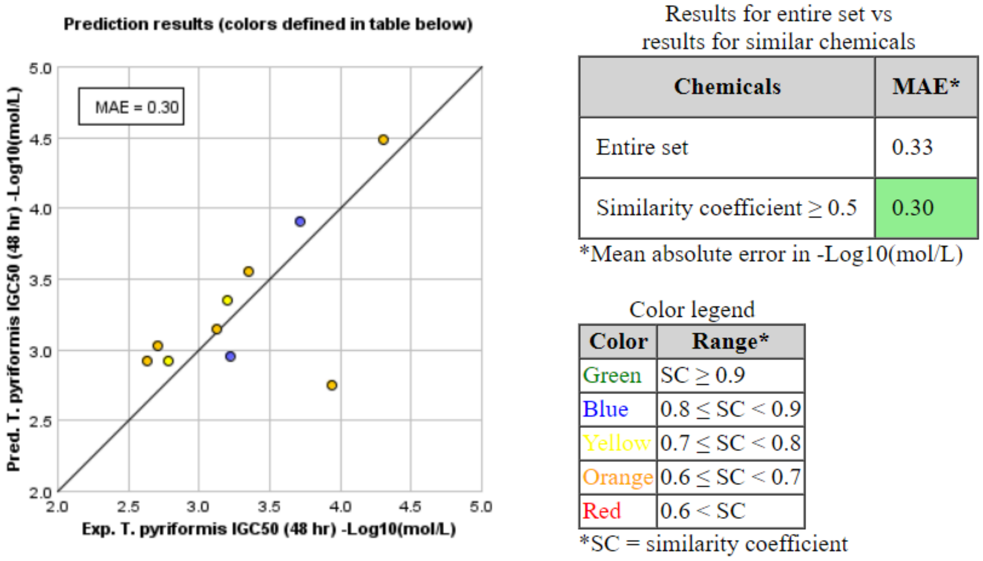
[a]Note: the test chemical was present in the external test set.



Individual Predictions

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

The software provides predictions for similar chemicals from the test set (see Figure 1.4.1). The MAE (mean absolute error) for similar chemicals (0.30) was slightly lower than the value for the entire test set (0.33), indicating increased confidence in the predicted value. The structures for the similar chemicals in the test set are provided by the software and shown in Table 1.4.2.


Prediction results (colors defined in table below)
MAE = 0.30
Pred. T. pyriformis IGC50 (48 hr) -Log10(mol/L)
Exp. T. pyriformis IGC50 (48 hr) -Log10(mol/L)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Results for entire set vs results for similar chemicals

Chemicals	MAE*
Entire set	0.33
Similarity coefficient ≥ 0.5	0.30

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

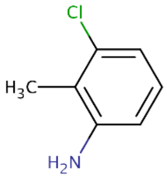
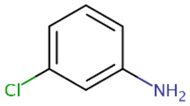
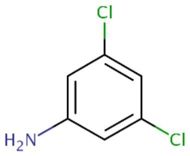
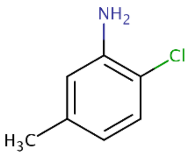
Color legend

Color	Range*
Green	SC ≥ 0.9
Blue	0.8 ≤ SC < 0.9
Yellow	0.7 ≤ SC < 0.8
Orange	0.6 ≤ SC < 0.7
Red	0.6 < SC

*SC = similarity coefficient

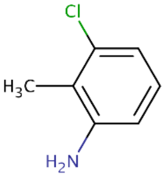
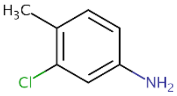
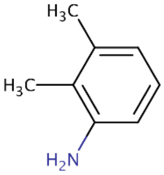
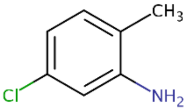
Figure 1.4.1. Predictions for similar chemicals from the test set

Table 1.4.2. 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	3.29
108-42-9		0.84	3.22	2.95
626-43-7		0.80	3.71	3.90
95-81-8		0.77	3.20	3.35
…	…	…	…	…

The most similar chemicals are very similar to the test chemical (benzenes substituted with chloro and amino groups) and were accurately predicted, indicating increased confidence in the predicted value. The program lists the similar chemicals in the training set (see Table 1.4.3). 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 indicates increased confidence in the predicted value because similar chemicals were used to build the QSAR models.

Table 1.4.3. Structures for the similar chemicals in the training set

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
87-60-5 (test chemical)			3.38	3.29
95-74-9		0.89	3.39	3.47
87-59-2		0.85	2.57	2.77
95-79-4		0.84	3.50	3.38
…	…	…	…	…

The details of the predictions for the different QSAR methods can be viewed by clicking on the predicted value for each method. Note: in order to view the details for each QSAR method used in the consensus prediction, the **Create detailed reports checkbox** must be checked.

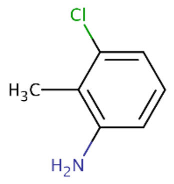
For example, for the Hierarchical clustering method for *T. pyriformis* IGC50 for 87-60-5, the main prediction table is shown in Table 1.4.4. The prediction interval (90% confidence interval) is 48.78 ≤ Tox ≤ 75.30. The experimental value falls within the prediction interval.

Table 1.4.4. 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

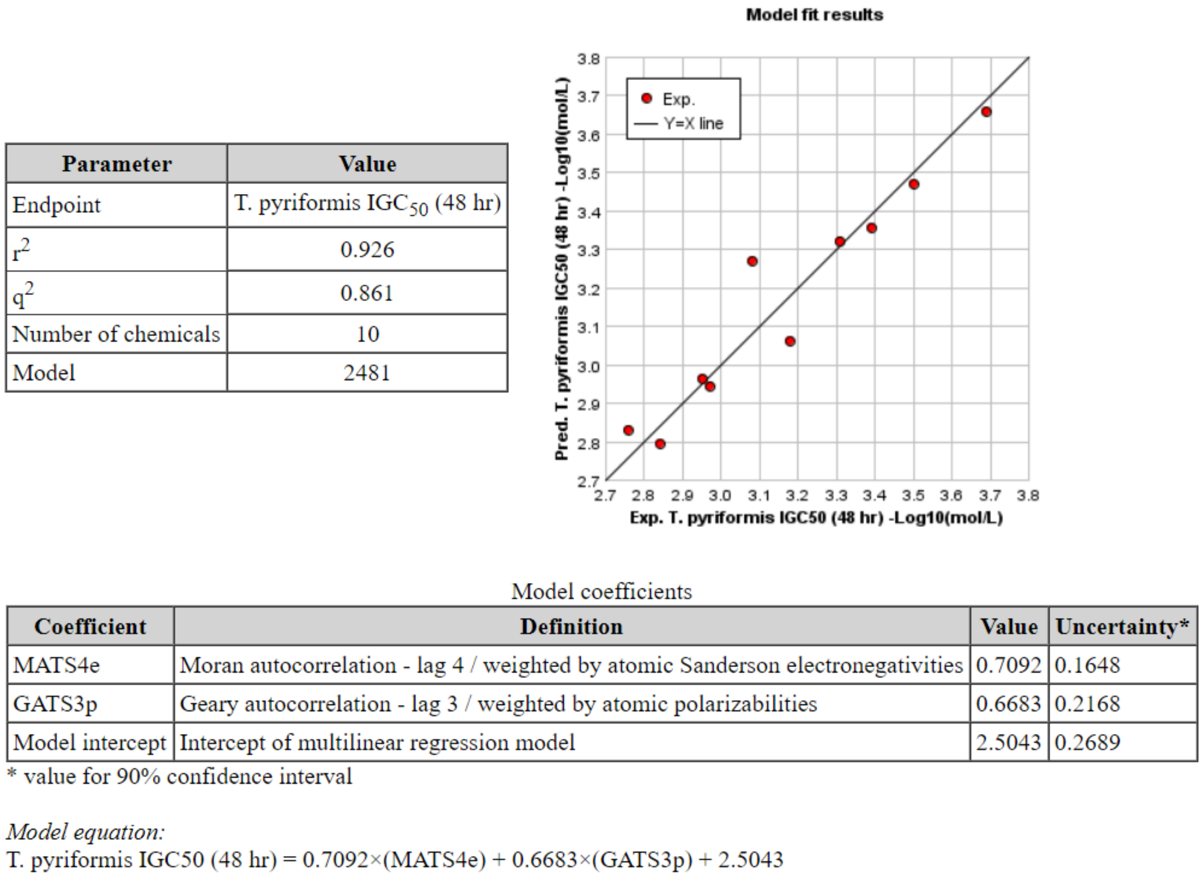
[a]Note: the test chemical was present in the external test set.

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



The predictions from the different clusters were all very similar. 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 details are given in Figure 1.4.2.



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

Model fit results
Exp.
Y=X line
Pred. T. pyriformis IGC50 (48 hr) -Log10(mol/L)
Exp. T. pyriformis IGC50 (48 hr) -Log10(mol/L)
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8

Model coefficients

Coefficient	Definition	Value	Uncertainty*
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
Model intercept	Intercept of multilinear regression model	2.5043	0.2689

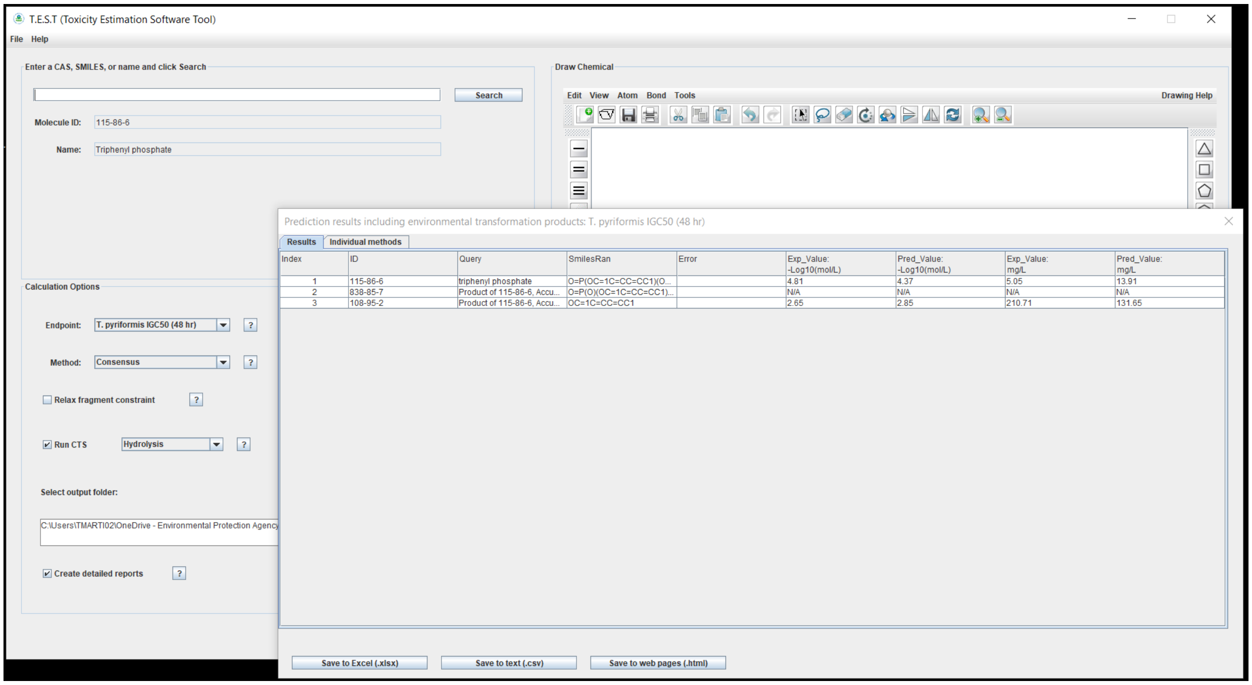
* value for 90% confidence interval

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

Figure 1.4.2. Details for model # 2481

1.5. Generation of environmental transformation products via CTS

In single chemical mode, one can generate predicted environmental transformation products by checking the "Run CTS" checkbox and selecting the desired transformation library (hydrolysis, abiotic reduction, and human metabolism). The software will make a call to the CTS webservice. If likely transformation products are generated, the toxicity (or physical properties) will be displayed as a table in the output. If no transformation products are generated, the output for the only the selected chemical will be displayed.



Molecule ID: 115-86-6

Name: Triphenyl phosphate

Endpoint: T. pyriformis IGC50 (48 hr)

Method: Consensus

☐ Relax fragment constraint

☑ Run CTS Hydrolysis

Select output folder:

C:\Users\TMARTI02\OneDrive - Environmental Protection Agency

☑ Create detailed reports

Prediction results including environmental transformation products: T. pyriformis IGC50 (48 hr)

Index	ID	Query	SmilesRan	Error	Exp_Value: -Log10(mol/L)	Pred_Value: -Log10(mol/L)	Exp_Value: mg/L	Pred_Value: mg/L
1	115-86-6	triphenyl phosphate	O=P(OC=1C=CC=CC1)(O...		4.81	4.37	5.05	13.91
2	838-85-7	Product of 115-86-6, Accu...	O=P(O)(OC=1C=CC=CC1)...		N/A	N/A	N/A	N/A
3	108-95-2	Product of 115-86-6, Accu...	OC=1C=CC=CC1		2.65	2.85	210.71	131.65

1.6. Batch prediction from the command line

To run batch calculations from the command line, utilize the following format:

```
➢ java -cp WebTEST.jar
  ToxPredictor.Application.Calculations.RunFromCommandLine -i inputFilePath -o
  outputFilePath -e endpointAbbreviation -m methodAbbreviation
```

The input file format can be MDL V2000 (use .sdf or .mol extension), a list of SMILES strings (use .smi extension), or a list of CAS numbers (use .txt extension).

For example, to estimate 96 hour fathead minnow LC50 using the hierarchical clustering method use the following text in a Windows .bat file:

```
set "TEST_Installation=C:/Users/UserName/AppData/Local/Programs/TEST 5.1.0.0/"
set Java_Path="%TEST_Installation%jre/bin/java.exe"
set "folder=C:/Users/UserName/Documents/InputFolderName/"
set input="%folder%caslist.txt"
set output="%folder%caslist.xlsx"
set endpoint="LC50"
set method="consensus"
set class=ToxPredictor.Application.Calculations.RunFromCommandLine
cd %TEST_Installation%
%Java_Path% -cp WebTEST.jar %class% -i %input% -o %output% -e %endpoint% -m
%method%
pause
```

There are three types of input formats accepted:

Input format	File extension
List of CAS numbers	.txt
List of SMILES strings	.smi
MDL structure data format	.sdf

The type of output returned depends on the file extension for the output file:

Report type	File extension
Comma separated file	.csv
Excel	.xlsx
Web pages	.html

The endpoint abbreviations are as follows:

Endpoint	Abbreviation
Fathead minnow LC50 (96 hr)	LC50
Daphnia magna LC50 (48 hr)	LC50DM
T. pyriformis IGC50 (48 hr)	IGC50
Oral rat LD50	LD50
Bioconcentration factor	BCF
Developmental Toxicity	DevTox
Mutagenicity	Mutagenicity
Normal boiling point	BP
Vapor pressure at 25°C	VP
Melting point	MP
Flash point	FP
Density	Density
Surface tension at 25°C	ST
Thermal conductivity at 25°C	TC
Viscosity at 25°C	Viscosity
Water solubility at 25°C	WS
Molecular descriptors	Descriptors

The method abbreviations are as follows:

Method	Abbreviation
Hierarchical clustering	hc
Single model	sm
Nearest neighbor	nn
Group contribution	gc
Consensus	consensus

2. 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 toxicity levels (such as the concentration at which 50% of exposed fish 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 to establish priorities for expensive and time-consuming traditional bioassays designed to determine toxicity levels. When conditions do not permit traditional bioassays, QSARs provide a method for estimating toxicity. Additionally, QSAR models are useful for estimating toxicities needed for green process design algorithms such as the Waste Reduction Algorithm [2].

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, without requiring any external programs. Users can input a chemical to be evaluated by drawing it in an included chemical sketcher window, entering a CAS, SMILES, or name, 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.).

2.1. Toxicity Endpoints

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

1. 96 hour fathead minnow LC_{50} (concentration of the test chemical in water in mg/L that is lethal to 50% of exposed fathead minnows after 96 hours)
2. 48 hour *Daphnia magna* LC_{50} (concentration of the test chemical in water in mg/L that is lethal to 50% of exposed *Daphnia magna* after 48 hours)
3. 48 hour *Tetrahymena pyriformis* IGC_{50} (concentration of the test chemical in water in mg/L that results in 50% growth inhibition to *Tetrahymena pyriformis* after 48 hours)
4. Oral rat LD_{50} (amount of chemical in mg/kg body weight that is lethal to 50% of rats after oral ingestion)
5. Bioaccumulation factor (ratio of the chemical concentration in fish to that in water at steady state)
6. Developmental toxicity (binary indication of whether or not a chemical can interfere with

normal development of 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 a chemical 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 a solid state changes to a liquid state)

2.2. QSAR Methodologies

T.E.S.T allows you to estimate toxicity values using several different advanced QSAR methodologies [3]:

- **Hierarchical clustering 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.
- **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).

T.E.S.T provides multiple prediction methodologies so users can have greater confidence in the predicted toxicities if the predictions from different methods are similar. 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 presented in Table 2.2.1.

Table 2.2.1. Advantages and disadvantages of the QSAR methods in T.E.S.T.

Method	**Advantages**	**Disadvantages**
Hierarchical clustering	• 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
Nearest neighbor	• Provides a quick estimate of toxicity • Allows one to determine structural analogs for a given test compound • Always provides 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
Consensus	• Was shown to achieve the best prediction results during external validation	• Cannot provide external estimates of toxicity for compounds in the training set

3. THEORY

3.1. 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 (CDK) [4]. The 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).

3.2. QSAR Methodologies

3.2.1. 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 provided in Figure 3.2.1.

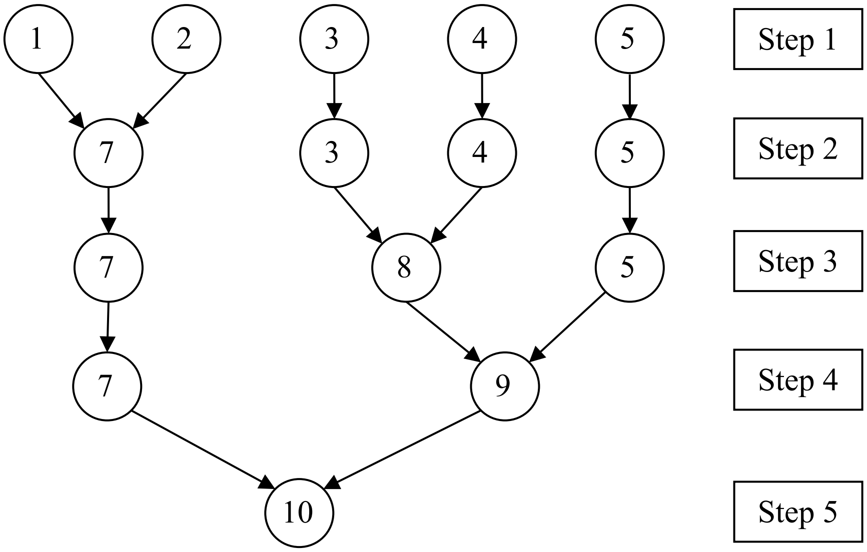


Figure 3.2.1. Hierarchical clustering with five chemicals

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:

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

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

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

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*:

$$C_j = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij} \tag{3}$$

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:

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

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 whether 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 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, or 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. The outlier statistical tests described above may not be applicable 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, then 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, then 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 the 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, then the model is considered to be valid. Additionally, both the LOO sensitivity and specificity must be at least 0.5 to avoid models that 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 the prediction and weight, respectively, 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

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

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

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

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

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

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

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

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

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

where *t* is the t-statistic, α = 0.1 (significance level for a 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:

$$\hat{y} - \hat{u} \le Toxicity \le y + \hat{u} \qquad (13)$$

The prediction interval indicates 90% confidence 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]

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

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 whether the test chemical is within the multidimensional ellipsoid defined by the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model). The model ellipsoid constraint is satisfied if the leverage of the test compound (h_{00}) is less than the maximum leverage value for all the compounds used in the model [17]. The second constraint, the Rmax constraint, checks whether the distance from the test chemical to the centroid of the cluster is less than the maximum distance of 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):

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

where $distance_i$ is the distance of chemical *i* to the centroid of the cluster.

The last constraint, the fragment constraint, requires that the compounds in the cluster have at least one example of each of the fragments contained in the test chemical. For example, if 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 in which 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 used because 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 the best model is selected from each step of the hierarchical clustering process. 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.

3.2.2. 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 clustering 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 [18]. 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.

3.2.3. Group contribution

The group contribution method is based on the group contribution approach of Martin and Young [19]. 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 clustering methodology, predictions are made using the model ellipse and fragment constraints.

The advantage of this approach is that 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.

3.2.4. 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. To make a prediction, each of the structural analogs must exceed a certain minimum cosine similarity coefficient (SCmin):

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

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*. SCmin was set at 0.5 so that the prediction coverage was similar to the other QSAR methods [3]. 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.

3.2.5. Mode of action

In the mode of action (MOA) method, the toxicity is predicted using a two-step process [20, 21]. 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. **Note: for the mode of action method, the training and prediction sets for 96 hour fathead minnow toxicity do not match those for the other QSAR methods.**

3.2.6. 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)[22]. If only a single QSAR methodology can make a prediction, the predicted value is deemed unreliable and not used. The consensus method typically provides the highest prediction accuracy because errant predictions are dampened by the predictions from the other methods.

Additionally, this method provides the highest prediction coverage because several methods with slightly different applicability domains are used to make a prediction.

3.3. Validation Methods

3.3.1. Statistical external validation

The predictive ability of each of the QSAR methodologies was evaluated using statistical external validation [23]. 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 [24-27]. Starting in version 3.0, random selection was used to develop the training and test sets because the Kennard-Stone method may yield 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* [28]. This was done for comparison purposes.

A QSAR model is considered to have 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 the correlation coefficient between the observed and predicted toxicities for the test set, R_o^2 is the 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 is evaluated in terms of equations 18 and 19. Additionally, the accuracy is 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 LOO 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.

3.4. Prediction of activity and endpoint values

If the endpoint is binary (e.g., mutagenicity or developmental toxicity), the calculated activity is defined as follows:

if calculated score < 0.5, then activity = negative
else if calculated score ≥ 0.5, then activity = positive.

For the continuous endpoints, the QSAR models were fit to experimental values with the following units:

Endpoint	Units
Fathead minnow LC50 (96 hr) *Daphnia magna* LC50 (48 hr) *T. pyriformis* IGC50 (48 hr) Water solubility at 25°C	$-\text{Log}_{10}$ (mol/L)
Oral rat LD50	$-\text{Log}_{10}$ (mol/kg)
Bioconcentration factor	Log_{10}
Viscosity at 25°C	Log_{10} (cP)
Vapor pressure at 25°C	Log_{10} (mmHg)
Normal boiling point Melting point Flash point	°C
Density	g/cm^3
Surface tension	dyn/cm
Thermal conductivity	mW/mK

3.5. Generation of transformation products using CTS

T.E.S.T. has the capability to generate transformation products using the CTS (Chemical Transformation Simulator)[31]. CTS is a web-based tool for predicting environmental and biological transformation pathways and physicochemical properties of organic chemicals. CTS can determine transformation products via hydrolysis, abiotic reduction, and human metabolism reaction pathways. T.E.S.T. will estimate toxicity values (or physical properties) for the likely transformation products for the selected chemical.

4. EXPERIMENTAL DATA SETS

4.1. 96 hour fathead minnow LC_{50} data set

The fathead minnow LC_{50} endpoint represents the concentration in water that is lethal to half of exposed fathead minnows (*Pimephales promelas*) in 4 days (96 hours). The data set for this endpoint was obtained by downloading the ECOTOX aquatic toxicity database[32].

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 (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 onwards, all isomers were kept because 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 Clustering, Single Model, Group Contribution, Nearest Neighbor, and Consensus methods, the data set was divided randomly into a training set (80% of the overall set) and a test set (20% of the overall set).

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

The *Daphnia magna* LC_{50} endpoint represents the concentration in water that is lethal to half of exposed *D. magna* (a water flea) in 48 hours. The data set for this endpoint was obtained from the ECOTOX aquatic toxicity database[32]. 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 541 chemicals. The modeled endpoint was $-Log_{10}$ (LC_{50} mol/L).

4.3. 40 hour *Tetrahymena pyriformis* IGC_{50} data set

The *Tetrahymena pyriformis* IGC_{50} endpoint represents the 50% growth inhibitory concentration for *T. pyriformis* (a protozoan ciliate) after 40 hours. The IGC_{50} training set was obtained from Schultz and coworkers [22, 33-70]. The final *T. pyriformis* IGC_{50} data set contained 1792 chemicals. The modeled endpoint was $-Log_{10}$ (IGC_{50} mol/L).

4.4. Oral rat LD_{50} data set

The oral rat LD_{50} endpoint represents the amount of the chemical (mass of the chemical in mg per body weight of the rat in kg) which when orally ingested is lethal to half of the rats. The dataset for this endpoint was obtained by downloading records from the ChemID*plus* database [71], from which 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 (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 $–Log_{10}$ (LD_{50} mol/kg).

4.5. 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 [72]. Data were compiled from several different databases [73-76]. The final dataset consists of 676 chemicals (after removing salts, mixtures, and ambiguous compounds). The modeled endpoint was the Log_{10}(BCF).

4.6. Developmental toxicity data set

The developmental toxicity endpoint is defined by whether a chemical is associated with developmental toxicity outcomes in humans and/or animals. Developmental toxicity includes any interference with normal development, both before and after birth. A dataset of 293 chemicals was created by Arena and Coworkers [77, 78] by combining data from the Teratogen Information System (TERIS) [79] and FDA guidelines [80]. The developmental toxicity values were taken from the revised binary toxicity values developed for the CAESAR project [28]. 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).

4.7. 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 (revertant colonies). A compound is classified as 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 [81, 82]. The final dataset consists of 5743 chemicals (after removing salts, mixtures, ambiguous compounds, and compounds without CAS numbers).

4.8. 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 [83]. Forty-one chemicals were removed from the data set because they were previously shown to be poorly predicted and had experimental values that were significantly different (>50K) from other sources such as NIST[84] and LookChem [85]. The final data set contained 5759 chemicals. The modeled property was the boiling point in °C.

4.9. 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 [85]. 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 chemicals 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.

4.10. 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 [85]. Chemicals with flash points greater than 1000°C were omitted from the data set. The modeled property was the flash point in °C.

4.11. Thermal conductivity

Thermal conductivity is defined as the ability of a material to conduct heat. The thermal conductivity values at 25°C for 442 chemicals were obtained from Jamieson and Vargaftik [86, 87] as follows:

- If a value was available at 25°C, then this value was used.
- If an experimental value was not available, then a value was extrapolated to 25°C (as long as the closest data point was within 10°C of 25°C).
- If the temperature coefficient was not available (or only a single data point was available), then the thermal conductivity of the nearest data point was used (as long as the closest data point was 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.

4.12. 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). Viscosity data at 25°C for 557 chemicals were obtained from Viswanath and Riddick [88, 89] as follows:

1. If a value was available at 25°C, then this value was used.
2. If an experimental value was not available, then a value was 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 to expand the size of the overall dataset. The modeled property was $\log_{10}$(viscosity cP).

4.13. 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 [90]. 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.

4.14. 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 [83]. Chemicals with water solubilities exceeding 1,000,000 mg/L were omitted from the overall dataset. Additionally, data were limited to data points that are within 10°C of 25°C. 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).

4.15. 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 [83]. The modeled property was Log10(vapor pressure mmHg).

4.16. Melting point

Melting point is the temperature, in °C, at which a chemical in a solid state changes to a liquid state. The melting points for 9385 chemicals were obtained from the database in EPI Suite [83].

The modeled property was Log10(vapor pressure mmHg).

5. VALIDATION RESULTS

5.1. 96 hour fathead minnow LC_{50}

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 for which a condition in equation 18 or 19 was not met. Models that 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 shown in Figure 5.1.1.

Table 5.1.1. 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 clustering	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
Group contribution	0.686	0.123	0.949	0.811	0.579	0.872
Nearest neighbor	0.667	0.080	1.000	0.877	0.649	0.939
Consensus	0.729	0.115	0.966	0.767	0.551	0.951

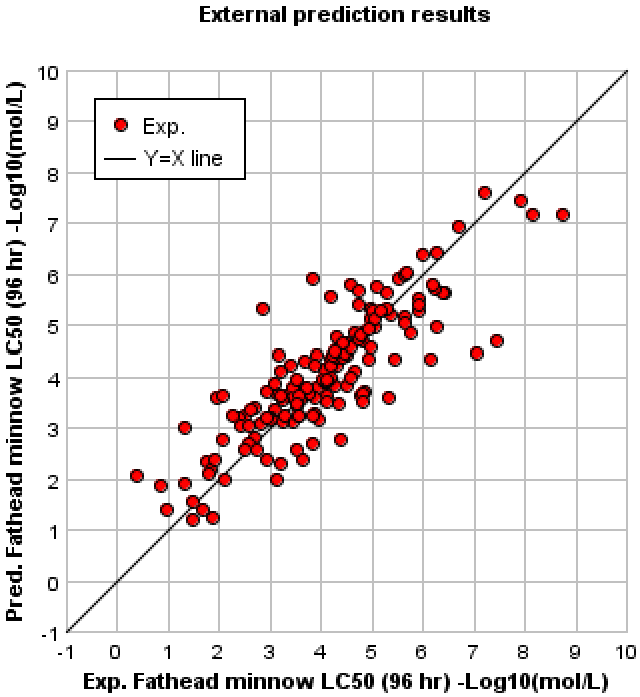


Figure 5.1.1. Experimental vs predicted values for the fathead minnow LC_{50} test set

5.2. 48 hour *Daphnia magna* LC_{50}

The consensus method yielded comparable results to the hierarchical clustering method (see Table 5.2.1). The prediction results for the consensus method are provided in Figure 5.2.1.

Table 5.2.1. 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 clustering	0.630	0.232	0.954	1.018	0.759	0.982
Single Model	0.530	0.250	0.979	1.191	0.913	0.982
Group contribution	0.476	0.413	0.963	1.152	0.879	0.853
Nearest neighbor	0.642	0.122	0.966	1.010	0.724	0.899
Consensus	0.616	0.233	0.969	1.042	0.786	0.982

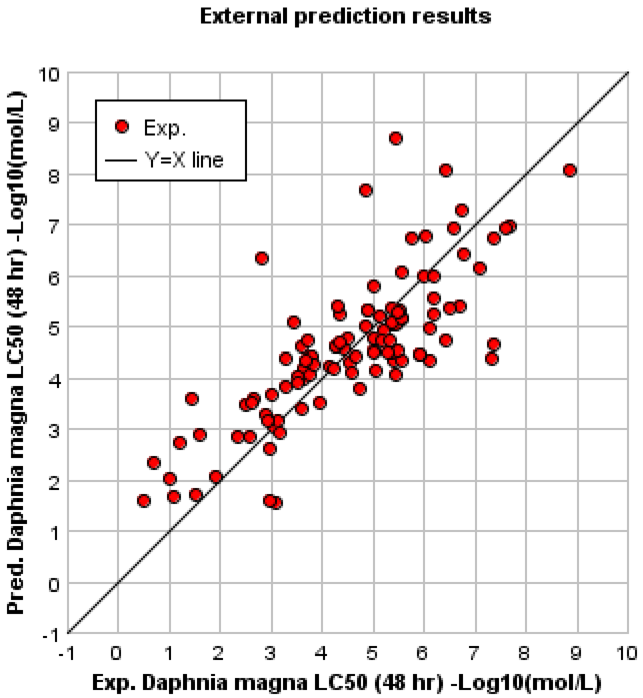


Figure 5.2.1. Experimental vs predicted values for the *Daphnia magna* LC_{50} test set

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

Again, the consensus method achieved the best results (see Table 4.3.1). The prediction results for the consensus method are shown in Figure 5.3.1.

Table 5.3.1. 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 clustering	0.718	0.023	0.978	0.540	0.358	0.933
Group contribution	0.682	0.066	0.994	0.576	0.411	0.955
Nearest neighbor	0.600	0.170	0.976	0.638	0.451	0.986
Consensus	0.739	0.070	0.983	0.505	0.355	0.966

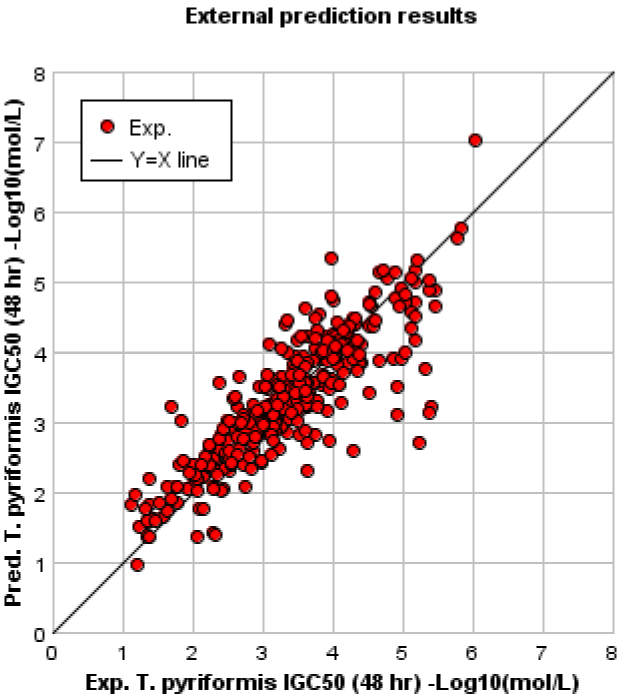


Figure 5.3.1. Experimental vs predicted values for the *T. pyriformis* IGC_{50} test set

5.4. Oral rat LD_{50} dataset

It was not possible to develop a single model or a group contribution model that fit the entire training set (see Table 5.4.1). 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 finding 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 [91]. The prediction results for the consensus method are shown in Figure 5.4.1.

Table 5.4.1. 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 clustering	0.578	0.184	0.969	0.650	0.460	0.875
Nearest neighbor	0.557	0.243	0.961	0.656	0.477	0.993
Consensus	0.633	0.188	0.968	0.595	0.436	0.875

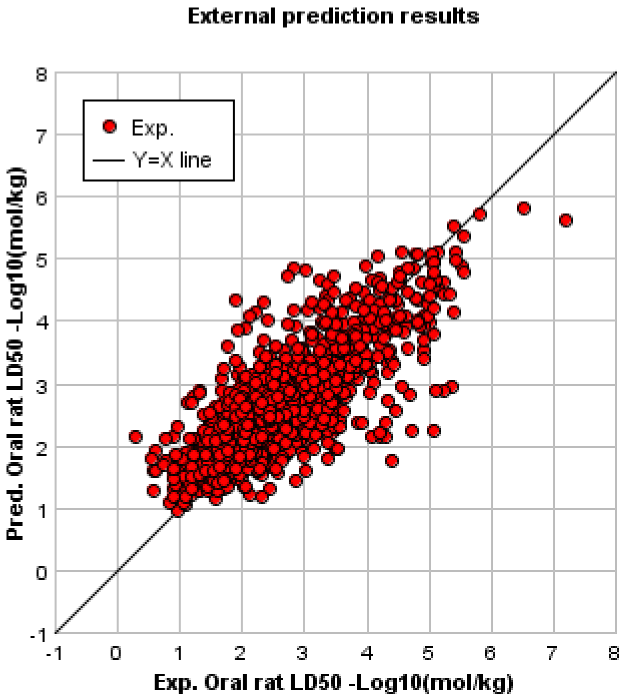


Figure 5.4.1. Experimental vs predicted values for the oral rat LD_{50} test set

5.5. Bioconcentration factor (BCF)

Again, the consensus method yielded the best statistics if one considers both prediction accuracy and coverage (see Table 5.5.1.). The prediction results for the consensus method are shown in Figure 5.5.1.

Table 5.5.1. Prediction results for the BCF test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.735	0.019	0.888	0.712	0.541	0.926
Single Model	0.742	0.082	0.901	0.684	0.542	0.926
Group contribution	0.675	0.187	0.888	0.761	0.623	0.874
Nearest neighbor	0.609	0.099	0.931	0.884	0.604	0.948
Consensus	0.754	0.076	0.898	0.670	0.523	0.926

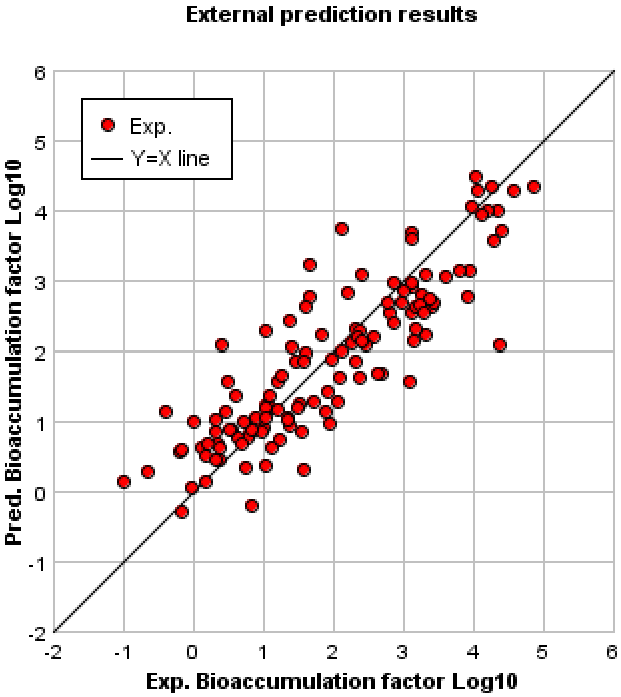


Figure 5.5.1. Experimental vs predicted values for the BCF test set

The BCFBAF (bioconcentration factor bioaccumulation factor) module (v. 3.00) of US EPA's EPI Suite software package [83] 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.

5.6. Developmental toxicity

The consensus method achieved the best results for the EPA-developed QSAR methods (in terms of prediction accuracy and coverage) (see Table 5.6.1). All of the methods achieved appreciably higher prediction sensitivities than specificities. This is acceptable for regulatory applications because reducing the proportion of false negatives is more crucial than reducing the proportion of false positives.

Table 5.6.1. Prediction results for the reproductive toxicity test set

Method	Concordance	Sensitivity	Specificity	Coverage
Hierarchical clustering	0.724	0.829	0.471	1.000
Single Model	0.732	0.850	0.438	0.966
Nearest neighbor	0.795	0.844	0.667	0.759
Consensus	0.772	0.900	0.471	0.983

5.7. Ames mutagenicity

Again, the consensus method achieved the best prediction accuracy (concordance) and prediction coverage (see Table 5.7.1). 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 5.7.1. Prediction results for the Ames mutagenicity test set

Method	Concordance	Sensitivity	Specificity	Coverage
Hierarchical clustering	0.763	0.776	0.746	0.956
Nearest neighbor	0.770	0.783	0.753	0.990
Consensus	0.777	0.794	0.755	0.948

5.8. Normal boiling point

The consensus method achieved the best statistics in terms of both prediction accuracy and coverage (see Table 5.8.1. In general, the prediction statistics for the physical properties were excellent. The prediction results for the consensus method are shown in Figure 5.8.1.

Table 5.8.1. 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 clustering	0.950	0.001	0.991	18.690	10.592	0.935
Group contribution	0.897	0.002	0.997	27.554	17.001	0.977
Nearest neighbor	0.877	0.005	0.968	29.967	19.754	0.988
Consensus	0.940	0.003	0.986	20.547	12.488	0.977

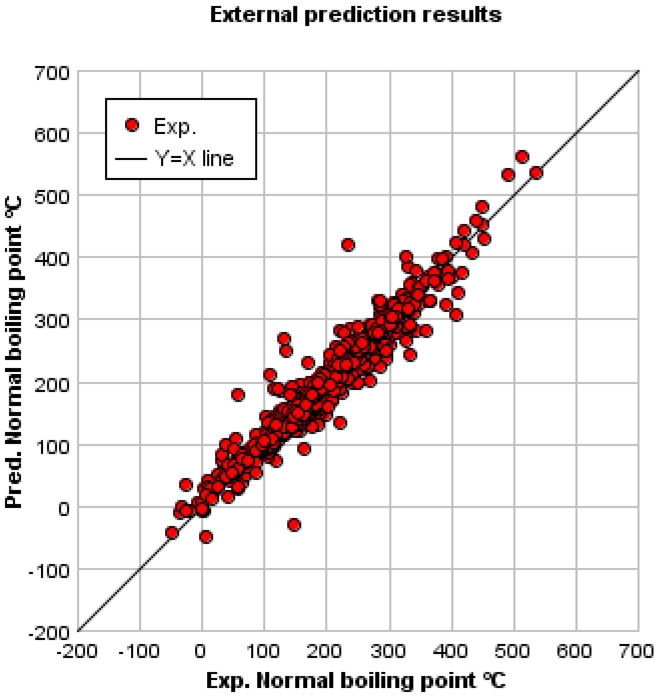


Figure 5.8.1. Experimental vs predicted values for the normal boiling point test set

5.9. Density

For this property, the hierarchical clustering and FDA methods gave a slightly higher R^2 value than the consensus method (see Table 5.9.1). However, the consensus method yielded a near 100% prediction coverage. The prediction results for the consensus method are shown in Figure 5.9.1.

Table 5.91 Prediction results for the density test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.972	0.001	0.997	0.053	0.026	0.942
Group contribution	0.872	0.005	0.997	0.116	0.071	0.992
Nearest neighbor	0.858	0.021	0.979	0.121	0.073	0.997
Consensus	0.938	0.006	0.990	0.080	0.046	0.992

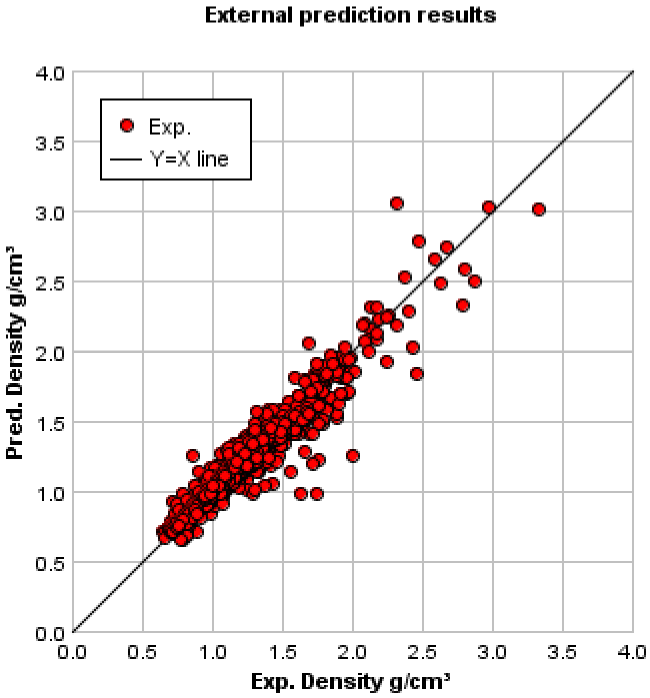


Figure 5.9.1. Experimental vs predicted values for the density test set

5.10. Flash point

For this property, the consensus method produces the best results in terms of prediction accuracy and coverage (see Table 5.10.1). The prediction results for the consensus method are provided in Figure 5.10.1.

Table 5.10.1. Prediction results for the flash point test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.870	0.008	0.961	28.911	16.753	0.924
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	0.873	0.010	0.953	29.064	17.571	0.987

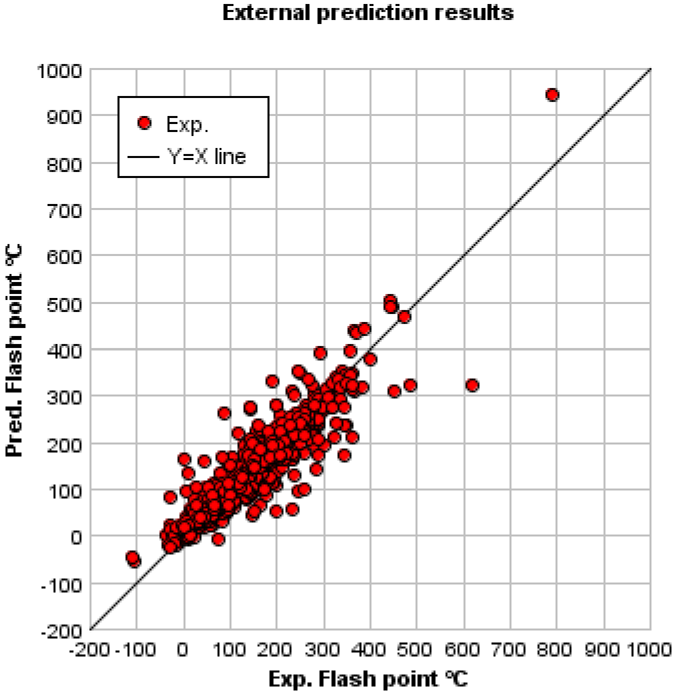


Figure 5.10.1. Experimental vs predicted values for the flash point test set

5.11. Thermal conductivity

For this property, the hierarchical clustering method produces similar results to the consensus method (see Table 5.11.1). The prediction results for the consensus method are provided in Figure 5.11.1.

Table 5.11.1. Prediction results for the thermal conductivity test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.905	0.025	0.996	11.062	6.771	0.956
Single Model	0.890	0.031	0.992	11.864	8.524	0.956
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	0.913	0.042	0.993	10.936	6.802	0.956

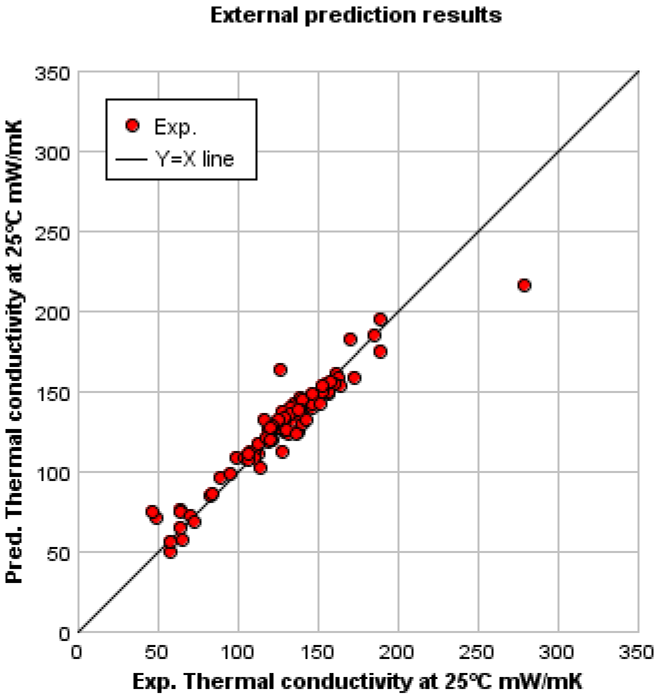


Figure 5.11.1. Experimental vs predicted values for the thermal conductivity test set

5.12. Viscosity

For this property, the consensus method produces the best results if you consider both prediction accuracy and coverage (see Table 5.12.1). 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 shown in Figure 5.12.1.

Table 5.12.1. Prediction results for the viscosity test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	*k*	RMSE	MAE	Coverage
Hierarchical clustering	0.867	0.001	0.808	0.215	0.131	0.929
Single Model	0.642	0.011	0.624	0.347	0.218	0.929
Group contribution	0.888	0.002	0.830	0.200	0.113	0.814
Nearest neighbor	0.757	0.009	0.725	0.289	0.194	0.920
Consensus	0.864	0.005	0.751	0.228	0.133	0.929

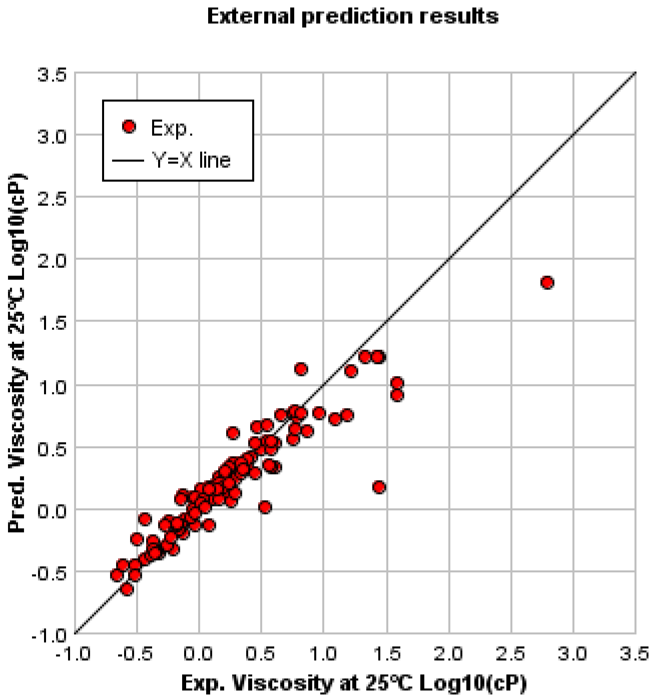


Figure 5.12.1. Experimental vs predicted values for the viscosity test set

5.13. Surface tension

For this property, the consensus method produces the best results in terms of prediction accuracy and coverage (see Table 5.13.1). The prediction results for the consensus method are shown in Figure 5.13.1.

Table 5.13.1. Prediction results for the surface tension test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.929	0.016	0.989	1.792	1.038	0.919
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	0.889	0.033	0.985	2.245	1.414	0.926

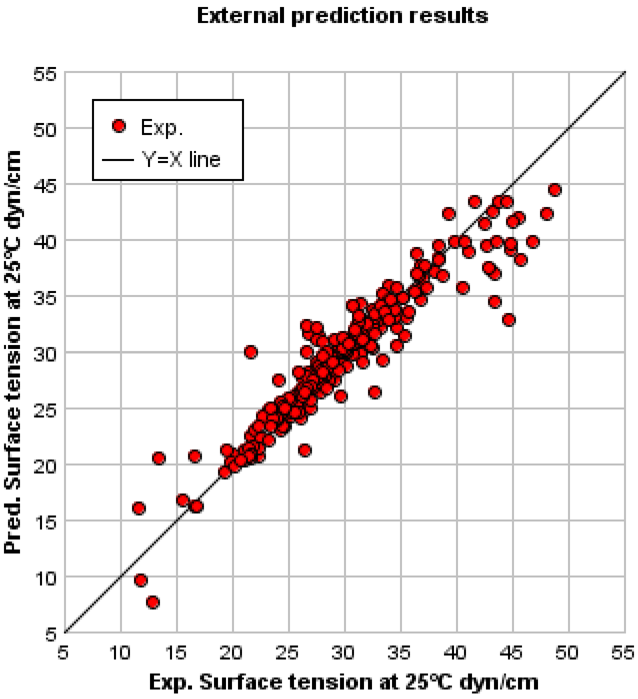


Figure 5.13.1. Experimental vs predicted values for the surface tension test set

5.14. Water solubility

For this property, the consensus method produces the best statistics in terms of prediction accuracy and coverage (see Table 5.14.1). The prediction results for the consensus method are shown in Figure 5.14.1.

Table 5.14.1. Prediction results for the water solubility test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.835	0.015	0.943	0.900	0.600	0.934
Group contribution	0.766	0.039	0.933	1.074	0.798	0.982
Nearest neighbor	0.791	0.022	0.950	1.024	0.735	0.985
Consensus	0.844	0.025	0.941	0.872	0.617	0.980

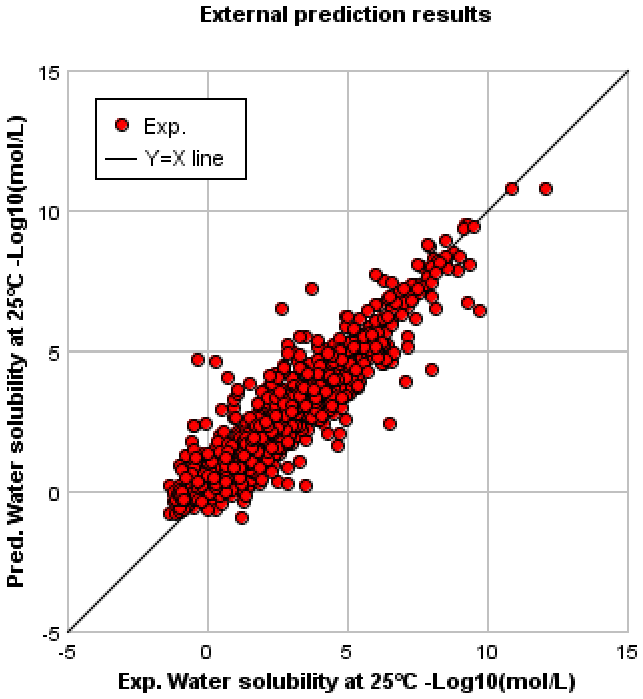


Figure 5.14.1. Experimental vs predicted values for the water solubility test set

5.15. Vapor pressure

The prediction statistics were excellent and again the consensus method achieved the best results (see Table 5.15.1). The prediction results for the consensus method are provided in Figure 5.15.1.

Table 5.15.1. Prediction results for the vapor pressure test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.955	0.001	0.976	0.754	0.460	0.940
Group contribution	0.929	0.001	1.020	0.999	0.608	0.968
Nearest neighbor	0.878	0.001	0.937	1.251	0.824	0.980
Consensus	0.948	0.001	0.978	0.818	0.500	0.970

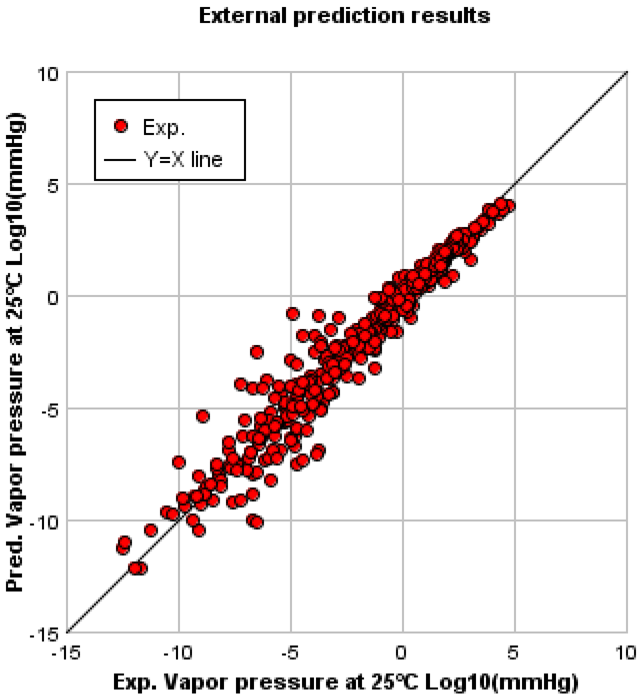


Figure 5.15.1. Experimental vs predicted values for the vapor pressure test set

5.16. Melting point

The prediction statistics were very good and the again the consensus method achieved the best results (see Table 5.16.1.). The prediction results for the consensus method are shown in Figure 5.16.1.

Table 5.16.1. Prediction results for the water solubility test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.809	0.011	0.891	44.509	31.480	0.932
Group contribution	0.704	0.065	0.837	54.947	41.274	0.997
Nearest neighbor	0.738	0.017	0.850	52.092	37.832	0.998
Consensus	0.813	0.026	0.858	43.771	31.888	0.998

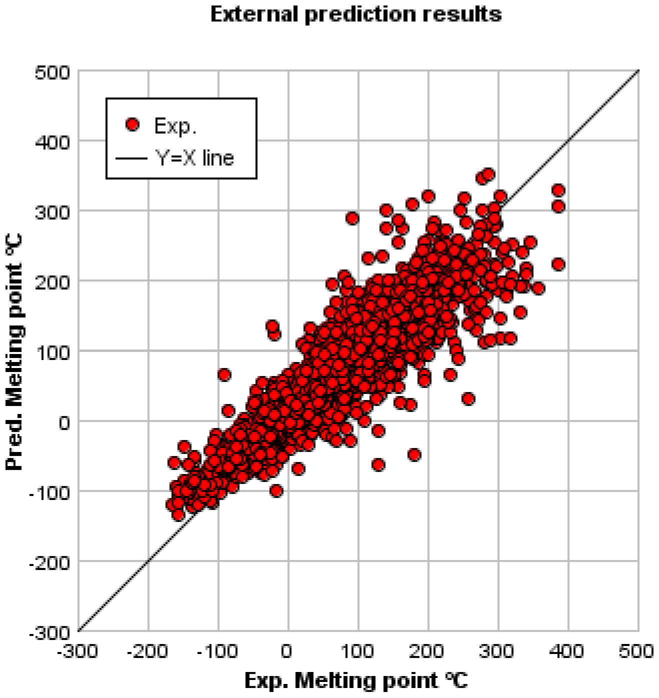


Figure 5.16.1. Experimental vs predicted values for the melting point test set

6. Bibliography

1. Wikipedia. *Molfile*. 2019 8/10/19]; Available from: https://en.wikipedia.org/wiki/Chemical_table_file#Molfile.
2. US EPA. *Environmental Optimization Using the Waste Reduction Algorithm*. 2011 4/18/16]; Available from: nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100DZKT.TXT.
3. Martin, T.M., et al., *A Hierarchical Clustering Methodology for the Estimation of Toxicity.* Toxicology Mechanisms and Methods, 2008. **18**: p. 251–266.
4. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics.* Journal of Chemical Information and Computer Sciences, 2003. **43**: p. 493-500.
5. Sourceforge.net. *Chemistry Development Kit (CDK)*. 2016 4/14/2016]; Available from: https://sourceforge.net/projects/cdk/.
6. Elsevier MDL. *MDL QSAR Version 2.2*. 2006 8/17/2006]; Available from: http://www.mdl.com/products/predictive/qsar/index.jsp.
7. Talete. *Dragon Version 5.4*. 2006 5/26/09]; Available from: http://www.talete.mi.it/.
8. Edusoft-LC. *Molconn-z Version 4.0*. 2006 5/26/09]; Available from: http://www.edusoft-lc.com/molconn/.
9. Romesburg, H.C., *Cluster Analysis for Researchers*. 1984, Belmont, CA: Lifetime Learning Publications.
10. Eriksson, L., et al., *Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.* Environmental Health Perspectives, 2003. **111**(10): p. 1361-1375.
11. Topliss, J.G. and R.P. Edwards, *Chance factors in Studies of Quantitative Structure-Activity Relationships.* Journal of Medicinal Chemistry, 1979. **22**(10): p. 1238-1244.
12. The University of Waikato. *WEKA - The Waikato Environment for Knowledge Analysis*. 2007 5/26/09]; Available from: http://www.cs.waikato.ac.nz/~ml/weka/.
13. Witten, I.H., *Data Mining: Practical machine learning tools and techniques*. 2005, San Francisco: Morgan Kaufmann.
14. Kutner, M.H., Nachtsheim, C. J., Neter, J., and Li, W. , *Applied Linear Statistical Models*. 2004, New York: McGraw-Hill.
15. Eriksson, L., et al., *Multi- and Megavariate Data Analysis - Principles and Applications*. 2001, Umea, Sweden: Umetrics AB.
16. Wikipedia.org. *Weighted mean*. 2016 [cited 2016 4/14/16]; Available from: http://en.wikipedia.org/wiki/Weighted_mean.
17. Montgomery, D.C., *Introduction to linear regression analysis*. 1982, John Wiley and Sons: New York. p. 141.
18. Benigni, R. and A.M. Richard, *QSARS of mutagens and carcinogens: Two case studies illustrating problems in the construction of models for noncongeneric chemicals.* Mutation Research, 1996. **371**: p. 29-46.
19. Martin, T.M. and D.M. Young, *Prediction of the Acute Toxicity (96-h LC_{50}) of Organic Compounds ti the Fathead Minnow (Pimephales promelas) Using a Group Contribution Method.* Chemical Research in Toxicology, 2001. **14**: p. 1378-1385.

20. Martin, T.M., et al., *Prediction of Aquatic Toxicity Mode of Action Using Linear Discriminant and Random Forest Models.* Journal of Chemical Information and Modeling, 2013. **53**(9): p. 2229-2239.
21. Martin, T.M., et al., *Comparison of global and mode of action-based models for aquatic toxicity.* SAR and QSAR in Environmental Research, 2015. **26**(3): p. 245-262.
22. Zhu, H., et al., *Combinational QSAR Model of Chemical Toxicants Tested against Tetrahymena pyriformis.* Journal of Chemical Information and Modeling, 2008. **48**: p. 766 - 784.
23. Gramatica, P. and P. Pilutti, *Evaluation of different statistical approaches for the validation of quantitative structure-activity relationships*. 2004, The European Commission - Joint Research Centre, Institute for Health & Consumer Protection - ECVAM: Ispra, Italy.
24. Bourguignon, B., et al., *Optimization in Irregularly Shaped Regions: pH and Solvent Strength in Reversed-Phase High-Performance Liquid Chromatography Separations.* Analytical Chemistry, 1994. **66**: p. 893-904.
25. Bourguignon, B., et al., Journal of Chromatography Science, 1994. **32**: p. 144-152.
26. Kennard, R.W. and L.A. Stone, Technometrics, 1969. **11**: p. 137-148.
27. Snarey, M., et al., *Comparison of Algorithms for Dissimilarity-Based Compound Selection.* Journal of Molecular Graphics and Modeling, 1997. **15**: p. 372-385.
28. CAESAR. *Developmental Toxicity Model*. 2009 9/21/09]; Available from: http://www.caesar-project.eu/index.php?page=results&section=endpoint&ne=5.
29. Golbraikh, A., et al., *Rational Selection of Training and Test sets for the Development of Validated QSAR Models.* Journal of Computer-Aided Molecular Design, 2003. **17**: p. 241-253.
30. Golbraikh, A. and A. Tropsha, *Beware of q^2!* Journal of Molecular Graphics and Modeling, 2002. **20**: p. 269-276.
31. US EPA. *CTS: Chemical Transformation Simultor*. 2019 08/27/2019]; Available from: https://qed.epacdx.net/cts/.
32. US EPA. *ECOTOX Database*. 2016 4/14/2016]; Available from: http://cfpub.epa.gov/ecotox/.
33. Akers, K.S., G.D. Sinks, and T.W. Schultz, *Structure–toxicity relationships for selected halogenated aliphatic chemicals.* Environmental Toxicology and Pharmacology, 1999. **7**: p. 33–39.
34. Aptula, A.O., et al., *Chemistry-Toxicity Relationships for the Effects of Di- and Trihydroxybenzenes to Tetrahymena pyriformis.* Chemical Research in Toxicology, 2005. **18**(5): p. 844-854.
35. Bearden, A.P. and T.W. Schultz, *Structure–Activity Relationships For Pimephales And Tetrahymena: A Mechanism Of Action Approach.* Environmental Toxicology and Chemistry, 1997. **16**(6): p. 1311–1317.
36. Bohme, A., et al., *Thiol Reactivity and Its Impact on the Ciliate Toxicity of Unsaturated Aldehydes, Ketones, and Esters.* Chemical Research in Toxicology, 2010. **23**: p. 1905-1912.

37. Cottrell, M.B. and T.W. Schultz, *Structure–Toxicity Relationships for Methyl Esters of Cyanoacetic Acids to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 2003. **70**: p. 549–556.

38. Cronin, M.T.D., et al., *Structure-Toxicity Relationships for Aliphatic Compounds Encompassing a Variety of Mechanisms of Toxic Action to Vibrio fischeri.* SAR and QSAR in Environmental Research, 2000. **11**(3-4): p. 301-312.

39. Cronin, M.T.D., et al., *Parametrization of Electrophilicity for the Prediction of the Toxicity of Aromatic Compounds.* Chem. Res. Toxicol., 2001. **14**: p. 1498-1505.

40. Cronin, M.T.D., et al., *Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis.* Chemosphere, 2002. **49**: p. 1201–1221.

41. DeWeese, A.D. and T.W. Schultz, *Structure–Activity Relationships for Aquatic Toxicity to Tetrahymena: Halogen-Substituted Aliphatic Esters.* Environmental Toxicology, 2001. **16**(1): p. 54–60.

42. Dimitrov, S., et al., *Interspecies Quantitative Structure–Activity Relationship Model For Aldehydes: Aquatic Toxicity.* Environmental Toxicology and Chemistry, 2004. **23**(2): p. 463-470.

43. Ellison, C.M., et al., *Definition of the structural domain of the baseline non-polar narcosis model for Tetrahymena pyriformis.* SAR and QSAR in Environmental Research, 2008. **19**(7–8): p. 751–783.

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

45. Muccini, M., et al., *Aquatic Toxicities of Halogenated Benzoic Acids to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 1999. **62**: p. 616-622.

46. Netzeva, T.I., et al., *Partial least squares modelling of the acute toxicity of aliphatic compounds to Tetrahymena pyriformis.* SAR and QSAR in Environmental Research, 2003. **14**(4): p. 265-83.

47. Netzeva, T.I. and T.W. Schultz, *QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data.* Chemosphere, 2005. **61**(11): p. 1632-1643.

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

49. Roberts, D.W., et al., *Experimental Reactivity Parameters for Toxicity Modeling: Application to the Acute Aquatic Toxicity of SN2 Electrophiles to Tetrahymena pyriformis.* Chemical Research in Toxicology, 2010. **23**: p. 228–234.

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

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

52. Schultz, T.W. and M. Tichy, *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**: p. 681-688.
53. Schultz, T.W. and J.L. Comeaux, *Structure-Toxicity Relationships for Aliphatic Isothiocyanates to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 1996. **56**: p. 638-642.
54. Schultz, T.W. and A.P. Bearden, *Structure-Toxicity Relationships for Selected Naphthoquinones to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 1998. **61**: p. 405-410.
55. Schultz, T.W., *Structure-Toxicity Relationships for Benzenes Evaluated with Tetrahymena pyriformis.* Chemical Research in Toxicology, 1999. **12**: p. 1262-1267.
56. Schultz, T.W., G.D. Sinks, and L.A. Miller, *Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis.* Environmental Toxicology, 2001. **16**(6): p. 543-549.
57. Schultz, T.W., T.I. Netzeva, and M.T.D. Cronin, *Selection of data sets for qsars: Analyses of tetrahymena toxicity from aromatic compounds.* SAR and QSAR in Environmental Research, 2003. **Vol. 14**(1): p. pp. 59–81.
58. Schultz, T.W. and V.A. Tucker, *Structure-Toxicity Relationships for the Effects of N- and N,N-Alkyl Thioureas to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 2003. **70**: p. 1251-1258.
59. Schultz, T.W. and J.T. Burgan, *pH-Stress and Toxicity of Nitrophenols to Tetrahymena pyriformis.* Bull. Environ. Contam. Toxicol., 2003. **71**: p. 1069-1076.
60. Schultz, T.W., et al., *Population Growth Impairment of Aliphatic Alcohols to Tetrahymena.* Environmental Toxicology, 2004. **19**(1): p. 1-10.
61. Schultz, T.W., J.W. Yarbrough, and M. Woldemeskel, *Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates.* Cell Biology and Toxicology, 2005. **21**(3-4): p. 181-189.
62. Schultz, T.W., et al., *Structure-Toxicity Relationships for the Effects to Tetrahymena pyriformis of Aliphatic, Carbonyl-Containing, α,β-Unsaturated Chemicals.* Chemical Research in Toxicology, 2005. **18**: p. 330-341.
63. Schultz, T.W., J.W. Yarbrough, and S.K. Koss, *Identification of reactive toxicants: Structure–activity relationships for amides.* Cell Biol Toxicol, 2006. **22**: p. 339–349.
64. Schultz, T.W. *Tetratox*. 2007 5/26/09]; Available from: http://www.vet.utk.edu/TETRATOX/.
65. Schultz, T.W., et al., *Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action.* QSAR Comb. Sci., 2007. **26**(2): p. 238-254.
66. Schultz, T.W., et al., *Structure-activity relationships for abiotic thiol reactivity and aquatic toxicity of halo-substituted carbonyl compounds.* SAR and QSAR in Environmental Research, 2007. **18**(1-2): p. 21-29.
67. Schultz, T.W., C.L. Sparfkin, and A.O. Aptula, *Reactivity-based toxicity modelling of five-membered heterocyclic compounds: Application to Tetrahymena pyriformis.* SAR and QSAR in Environmental Research, 2010. **21**(7-8): p. 681-691.

68. Schwöbel, J.A.H., J.C. Madden, and M.T.D. Cronin, *Application of a computational model for Michael addition reactivity in the prediction of toxicity to Tetrahymena pyriformis.* Chemosphere, 2011. **85**(6): p. 1066-1074.

69. Seward, J.R., E.L. Hamblen, and T.W. Schultz, *Regression comparisons of Tetrahymena pyriformis and Poecilia reticulata toxicity.* Chemosphere, 2002. **47**: p. 93–101.

70. Sinks, G.D. and T.W. Schultz, *Correlation Of Tetrahymena And Pimephales Toxicity: Evaluation Of 100 Additional Compounds.* Environmental Toxicology and Chemistry, 2001. **20**(4): p. 917–921.

71. U.S. National Library of Medicine. *ChemIDplus*. 2016 4/14/16]; Available from: http://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsp.

72. Hamelink, J.L., *Current bioconcentration test methods and theory*, in *Aquatic Toxicology and Hazard Evaluation*, F.L. Mayer and J.L. Hamelink, Editors. 1977, ASTM STP: West Conshohocken, PA p. 149-161.

73. Dimitrov, S., et al., *Base-line model for identifying the bioaccumulation potential of chemicals.* SAR and QSAR in Environmental Research, 2005. **16**: p. 531-554

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

75. EURAS. *EURAS bioconcentration factor (BCF) Gold Standard Database*. 3/30/18]; Available from: http://ambit.sourceforge.net/euras/.

76. 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**: p. 1701-1707.

77. Arena, V.C., et al., *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 and QSAR in Environmental Research, 2004. **15**(1): p. 1-18.

78. Sussman, N.B., et al., *Decision Tree SAR Models for Developmental Toxicity Based on an FDA/TERIS Database.* SAR and QSAR in Environmental Research, 2003. **14**(2): p. 83-96.

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

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

81. Hansen, K., et al., *Benchmark Data Set for in Silico Prediction of Ames Mutagenicity.* Journal of Chemical Information and Modeling, 2009. **49**(9): p. 2077-2081.

82. Benchmark, T. 4/30/10]; Available from: http://ml.cs.tu-berlin.de/toxbenchmark/.

83. US EPA. *EPI Suite, Version 4.0*. 2009 5/21/09]; Available from: http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm.

84. NIST. *NIST Chemistry WebBook*. 2010; Available from: http://webbook.nist.gov/chemistry/.

85. Lookchem.com. 2011; Available from: http://www.lookchem.com.

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

87. Vargaftik, N.B., et al., *Handbook of thermal conductivity of liquids and gases*. 1994, Boca Raton: CRC Press. 358.

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

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

90. Jasper, J.J., *The Surface Tension of Pure Liquid Compounds.* J. Phys. Chem. Ref. Data, 1972. **1**: p. 841-1009.

91. Zhu, H., et al., *Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure.* Chem. Res. Toxicol., 2009. **22**: p. 1913–1921.

EPA
United States
Environmental Protection
Agency

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