EPA
United States
Environmental Protection
Agency

Overview of T.E.S.T.
(Toxicity Estimation Software Tool)

Todd Martin, US EPA, Cincinnati, OH, USA

Office of Research and Development
National Risk Management Research Laboratory / Sustainable Technology Division / Clean Processes Branch

April 26, 2018

United States
Environmental Protection
Agency

Goal

- Our goal was to develop user friendly software that can estimate toxicity and physical properties from molecular structure
  - Experimental data such as critical properties or biological assays are not used
  - Values can be used for alternatives assessment

United States
Environmental Protection
Agency

OECD* Principles

- An unambiguous algorithm
- A defined endpoint
- A defined domain of applicability
- Appropriate measures of goodness-of-fit, robustness and predictivity
- A mechanistic interpretation, if possible

*Organisation for Economic Co-operation and Development: http://bit.ly/2r8bVAs

EPA
United States
Environmental Protection
Agency

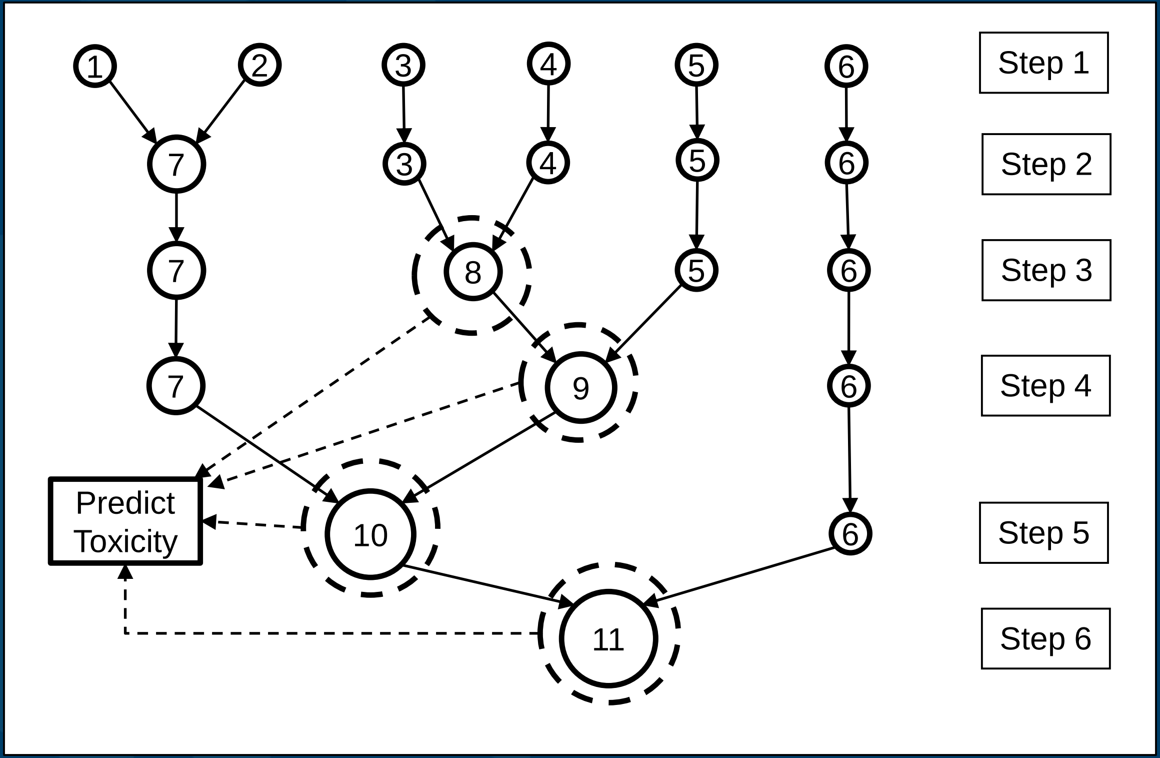
Methods

- There are several quantitative structure activity relationship (QSAR) methods available in TEST:
  - Hierarchical clustering
  - Single Model
  - Group contribution
  - FDA (Food and Drug Administration)
  - Nearest neighbor
  - Consensus
- See the TEST User’s guide for detailed information

EPA United States Environmental Protection Agency

Hierarchical clustering

- Similar chemicals are grouped together but not necessarily on expert defined chemical classes
- Uses structural information from entire dataset instead of just from chemicals in SAR



- Clustering is based on Ward's method (which aims to minimize the variance of the clusters)
- A prediction is made using the closest cluster from each step in the clustering

Hierarchical clustering, cont.

- Predictions made using weighted average of several different models:

$$Tox = \sum_{i=1}^{k} w_i \times Tox_i \bigg/ \sum_{i=1}^{k} w_i$$

- The weights are based on the standard error for each prediction:

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

  - For binary endpoints (i.e. mutagenicity) the predictions are equally weighted (w_j=1)

Hierarchical Clustering, cont.

- Advantages
  - Most accurate single method since prediction represents prediction from multiple models
- Disadvantages
  - Cannot provide external estimates of toxicity for compounds in the training set

EPA United States Environmental Protection Agency

Single model

- Predictions are made using multilinear regression model fit to entire training set:

$$Tox = \sum a_i x_i + a_0$$

- Descriptors, x_i, are 2d molecular descriptors
- Example, 48 hr *Daphnia magna* LC_{50} model:
  - Toxicity = 1.2157 × (xc4) + 0.1341 × (StN) + 0.6974 × (SsSH) - 1.3213 × (SsOH_acnt) + 0.8605 × (Hmax) + 1.4685 × (ssi) - 0.9197 × (MDEN33) + 0.2238 × (BEHm1) + 1.4502 × (BEHp1) + 2.4060 × (Mv) + 1.9085 × (MATS1m) - 2.4036 × (MATS1e) - 0.3463 × (GATS3m) + 0.0255 × (AMR) - 1.4215 × (-C(=S)- [2 nitrogen attach]) - 0.7185 × (AN) - 1.0232 × (-N< [attached to P]) - 1.5228 × (-S(=O)(=O)- [aromatic attach]) - 6.5594

Single model, cont.

- Advantages
  - Single transparent model can be easily viewed/exported
  - The model does not need to rely on clustering the chemicals correctly
- Disadvantages
  - 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

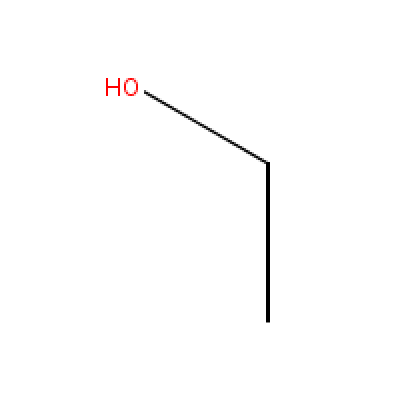
Group contribution

- Predictions are made using multilinear regression model fit to entire training set:

$$Tox = \sum a_i x_i + a_0$$

- Descriptors, x_i, are molecular fragment counts

Descriptor	x_i	a_i	$a_i \times x_i$
-CH3 [aliphatic attach]	1	0.23	0.23
-CH2- [aliphatic attach]	1	0.27	0.27
-OH [aliphatic attach]	1	-0.58	-0.58
Model intercept (a_0)	1	1.96	1.96
Tox (-Log10(LC_{50} mol/L))			1.88



Group contribution, cont.

- Advantages
  - Easy to understand the model and estimates can be made without using a computer program
  - Toxicity estimates are rapid and can be used for molecular design
- Disadvantages
  - 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

FDA

- Predictions are made using a multilinear regression model fit to the 30-75 most similar compounds in the training set:

$$Tox = \sum a_i x_i + a_0$$

- Descriptors, *x_i*, are 2d molecular descriptors
- Example model built for benzene for FHM LC50:
  - Toxicity = 0.4642 × (SsssCH) + 0.3255 × (SdssC) + 0.7706 × (Hmin) + 0.7088 × (iedem) - 1.0033 × (BEHm3) + 0.8268 × (ALOGP) + 2.5756

FDA, cont.

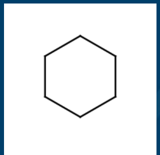
- Advantages
  - Can generate a new model based on the closest analogs to the test compound
  - Always provides an external prediction of toxicity
- Disadvantages
  - Predictions sometimes take longer since it has to generate a new model each time

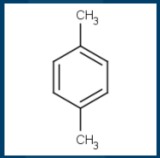
Nearest Neighbor

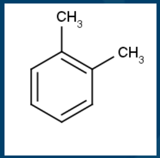
- Predicted toxicity is simply the average of the three nearest neighbors (i.e. read across)
- The neighbors are those with highest similarity coefficient:

$$SC_{i,k} = \frac{\sum_{j=1}^{\#descriptor\,s} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{\#descriptor\,s} x_{ij}^{\;2} \cdot \sum_{j=1}^{\#descriptor\,s} x_{kj}^{\;2}}}$$

- All neighbors must exceed a minimum cosine similarity coefficient
- For example the predicted FHM LC_{50} for benzene is made using average of values for







United States
Environmental Protection
Agency

Nearest neighbor, cont.

- Advantages
  - Provides a quick estimate of toxicity
  - Allows one to determine structural analogs for a given test compound
  - Always provide an external prediction of toxicity
- Disadvantages
  - 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 model

- The consensus prediction is simply the average predicted value for all the models that have predictions inside their applicability domain
- A prediction is made if at least two models have a valid prediction in terms of their respective applicability domain
- Using multiple models minimizes bad predictions and maximizes prediction accuracy
- Using different applicability domains maximizes prediction coverage
- This method is recommended method to use

Consensus, cont.

- Advantages
  - Was shown to achieve the best prediction accuracy and coverage during external validation
- Disadvantages
  - Cannot provide external estimates of toxicity for compounds in the training set
  - Calculations take longer

Applicability Domain

- Model ellipsoid constraint
  - Test chemical must be within ellipsoid of descriptor values for model chemicals (based on descriptors in 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:

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

United States Environmental Protection Agency

Applicability Domain, cont.

- Rmax constraint
  - Distance to the centroid of the cluster must be < the maximum distance for any cluster chemical (based on entire descriptor pool)

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

Applicability Domain, cont.

- Fragment Constraint
  - Compounds in the cluster must have at least one example of each of the fragments contained in the test chemical
    - Note: not used for binary endpoints (i.e. mutagenicity)
- Example:
  - If a cluster contained only primary alcohols, it shouldn’t be used to predict the toxicity for a primary aldehyde (since the cluster doesn’t contain any compounds with an aldehyde group)

Applicability Domain, cont.

Method	AD Measures
Hierarchical clustering	Ellipsoid, Rmax, Fragment
Single model	Ellipsoid, Rmax, Fragment
FDA	Ellipsoid, Fragment
Group contribution	Ellipsoid, Fragment
Nearest neighbor	Must have 3 chemicals with SC > SC_{min}

Molecular descriptors

- TEST generates ~800 descriptors:
  - Estate values and E-state counts
  - Constitutional descriptors
  - Topological descriptors
  - Walk and path counts
  - Connectivity
  - Information content
  - 2d autocorrelation
  - Burden eigenvalue
  - Molecular property (such as Kow)
  - Kappa
  - Hydrogen bond acceptor/donor counts
  - Molecular distance edge
  - Molecular fragment counts
- See Molecular Descriptor Guide in TEST (accessible from Help menu or from link on website)

Required Model Statistics

- Continuous endpoints
  - $q^2 \geq 0.5$
- Binary endpoints
  - LOO Concordance ≥ 0.8
  - LOO Sensitivity ≥ 0.5
  - LOO Specificity ≥ 0.5

Validation Procedure

- The overall datasets are randomly divided into a training set (80%) and a test set (20%) five times
  - Splitting is done in 5 fold fashion and models are fit to a new set of descriptors each time
- The results are reported for the random splitting that provides results closest to the average results
  - Goal is to provide a reasonable estimate of the predictive ability of the models
- Test set results are evaluated in terms of
  - Prediction accuracy (r^2)
  - Prediction coverage (fraction predicted)

United States Environmental Protection Agency

Endpoints

Endpoint	Description
96 hr fathead minnow LC_{50}	Concentration in mg/L that causes 50% of fathead minnows to die after 96 hours
48 hour *Daphnia magna* LC_{50}	Concentration in mg/L that causes 50% of *Daphnia magna* to die after 48 hours
48 hour *Tetrahymena pyriformis* IGC_{50}	Concentration in mg/L that causes 50% growth inhibition to *Tetrahymena pyriformis* after 48 hours
Oral rat LD_{50}	Amount in mg/kg body weight that causes 50% of rats to die after oral ingestion

Endpoints, cont.

Endpoint	Description
Bioaccumulation factor	Ratio of the chemical concentration in fish as a result of absorption via the respiratory surface to that in water at steady state
Developmental toxicity	Whether or not a chemical causes developmental toxicity effects to humans or animals
Ames mutagenicity	A compound is positive for mutagenicity if it induces revertant colony growth in any strain of *Salmonella typhimurium*

United States Environmental Protection Agency

Physical properties in T.E.S.T.

Property	Description
Normal boiling point	Temperature (° C) at which a chemical boils at atmospheric pressure (1 atm)
Vapor pressure	The pressure (mmHg) exerted by a vapor in thermodynamic equilibrium with the liquid phase at 25° C in a closed system
Melting point	The temperature (° C) at which a chemical changes state from solid to liquid
Flash point	The lowest temperature (° C) at which a chemical can vaporize to form an ignitable mixture in air
Density	The mass per unit volume (g/cm^3)

EPA United States Environmental Protection Agency

Physical properties, cont.

Property	Description
Surface tension	A property of the surface of a liquid (dyn/cm) that allows it to resist an external force
Thermal conductivity	The property of a material (mW/mK) reflecting its ability to conduct heat
Viscosity	A measure of the resistance of a fluid to flow (cP) defined as the proportionality constant between shear rate and shear stress
Water solubility	The amount of a chemical (mg/L) that will dissolve in liquid water to form a homogeneous solution

United States
Environmental Protection
Agency

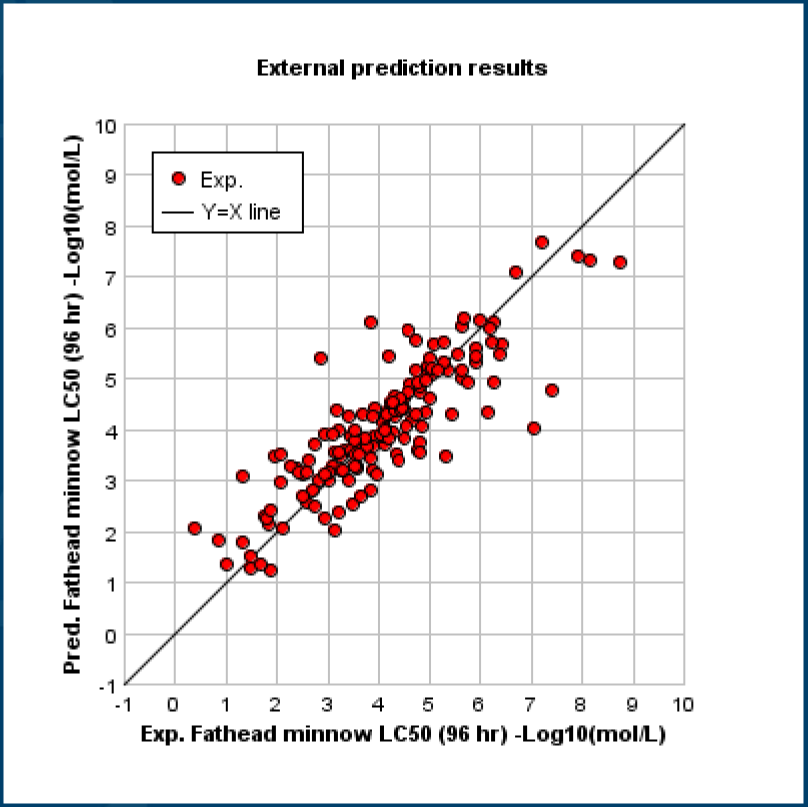
Future Endpoints in T.E.S.T.

- Skin sensitization/irritation/corrosion potential
- Eye irritation potential
- Octanol water partition coefficient
- Requests???

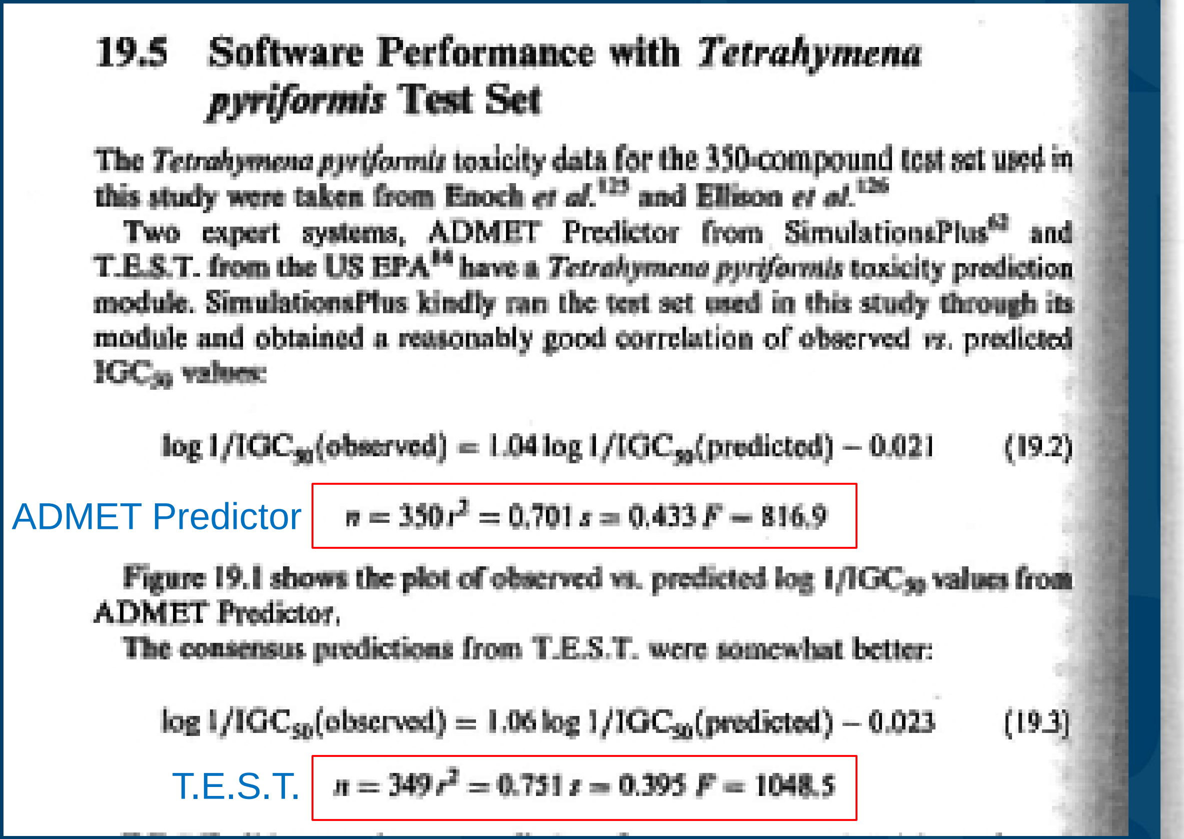
$$\frac{R^2 - R_0^2}{R^2}$$

96 hour fathead minnow LC_{50}

Method	R^2	Coverage
Hierarchical	0.710	0.951
Single Model	0.704	0.945
FDA	0.626	0.945
GC	0.686	0.872
NN	0.667	0.939
Consensus	0.728	0.951
ECOSAR	0.620	0.976



IGC$_{50}$ performance*



19.5 Software Performance with *Tetrahymena pyriformis* Test Set

The *Tetrahymena pyriformis* toxicity data for the 350-compound test set used in this study were taken from Enoch *et al.*[125] and Ellison *et al.*[126]

Two expert systems, ADMET Predictor from SimulationsPlus[62] and T.E.S.T. from the US EPA[84] have a *Tetrahymena pyriformis* toxicity prediction module. SimulationsPlus kindly ran the test set used in this study through its module and obtained a reasonably good correlation of observed *vs.* predicted IGC$_{50}$ values:

$$\log 1/\mathrm{IGC}_{50}(\text{observed}) = 1.04 \log 1/\mathrm{IGC}_{50}(\text{predicted}) - 0.021 \qquad (19.2)$$

ADMET Predictor

$$n = 350 \; r^2 = 0.701 \; s = 0.433 \; F = 816.9$$

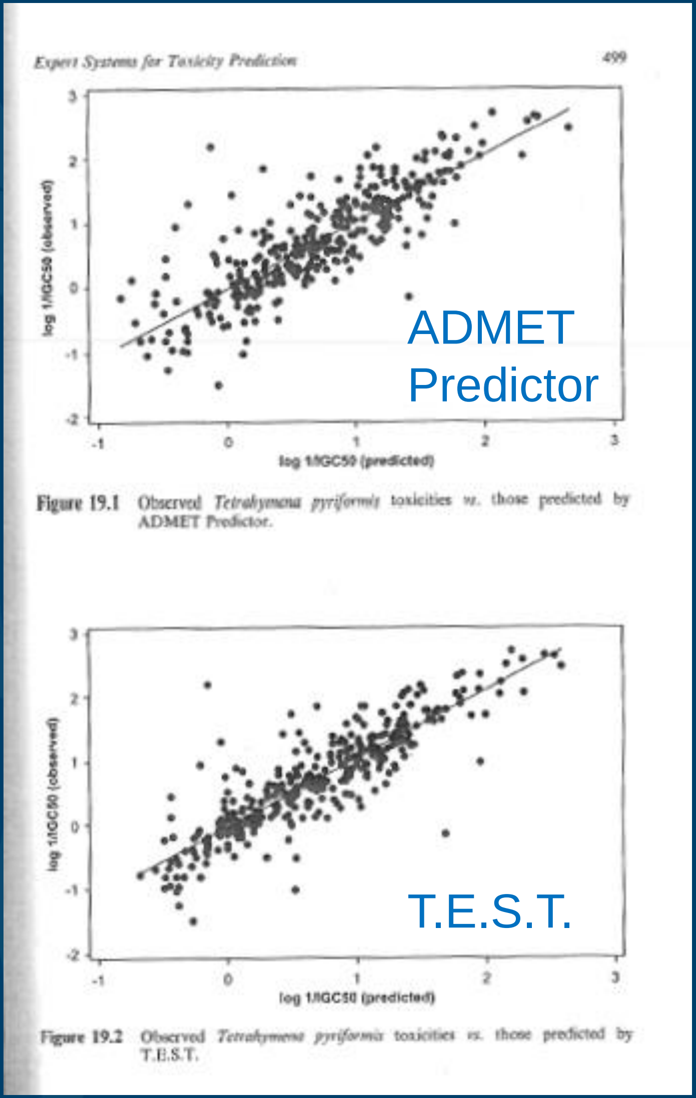
Figure 19.1 shows the plot of observed vs. predicted log 1/IGC$_{50}$ values from ADMET Predictor.

The consensus predictions from T.E.S.T. were somewhat better:

$$\log 1/\mathrm{IGC}_{50}(\text{observed}) = 1.06 \log 1/\mathrm{IGC}_{50}(\text{predicted}) - 0.023 \qquad (19.3)$$

T.E.S.T.

$$n = 349 \; r^2 = 0.751 \; s = 0.395 \; F = 1048.5$$



Expert Systems for Toxicity Prediction 499

Figure 19.1 Observed *Tetrahymena pyriformis* toxicities *vs.* those predicted by ADMET Predictor.

Figure 19.2 Observed *Tetrahymena pyriformis* toxicities *vs.* those predicted by T.E.S.T.

*Dearden, 2010

United States Environmental Protection Agency

Mutagenicity performance*

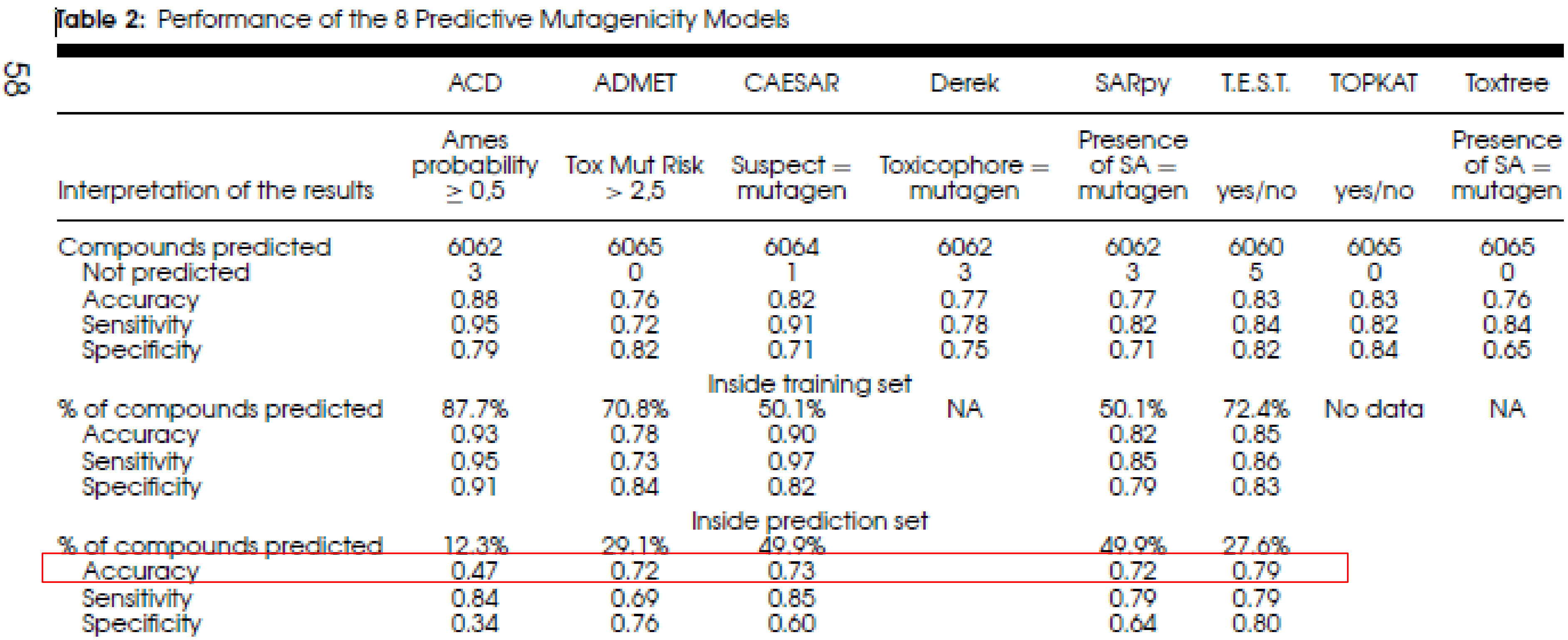


Table 2: Performance of the 8 Predictive Mutagenicity Models

	ACD	ADMET	CAESAR	Derek	SARpy	T.E.S.T.	TOPKAT	Toxtree
Interpretation of the results	Ames probability ≥ 0.5	Tox Mut Risk > 2.5	Suspect = mutagen	Toxicophore = mutagen	Presence of SA = mutagen	yes/no	yes/no	Presence of SA = mutagen
Compounds predicted	6062	6065	6064	6062	6062	6060	6065	6065
Not predicted	3	0	1	3	3	5	0	0
Accuracy	0.88	0.76	0.82	0.77	0.77	0.83	0.83	0.76
Sensitivity	0.95	0.72	0.91	0.78	0.82	0.84	0.82	0.84
Specificity	0.79	0.82	0.71	0.75	0.71	0.82	0.84	0.65
			Inside training set					
% of compounds predicted	87.7%	70.8%	50.1%	NA	50.1%	72.4%	No data	NA
Accuracy	0.93	0.78	0.90		0.82	0.85		
Sensitivity	0.95	0.73	0.97		0.85	0.86		
Specificity	0.91	0.84	0.82		0.79	0.83		
			Inside prediction set					
% of compounds predicted	12.3%	29.1%	49.9%		49.9%	27.6%		
Accuracy	0.47	0.72	0.73		0.72	0.79		
Sensitivity	0.84	0.69	0.85		0.79	0.79		
Specificity	0.34	0.76	0.60		0.64	0.80		

- T.E.S.T. achieved highest prediction accuracy for external set

*Bakhtyari, 2013

EPA
United States
Environmental Protection
Agency

Developmental Toxicity

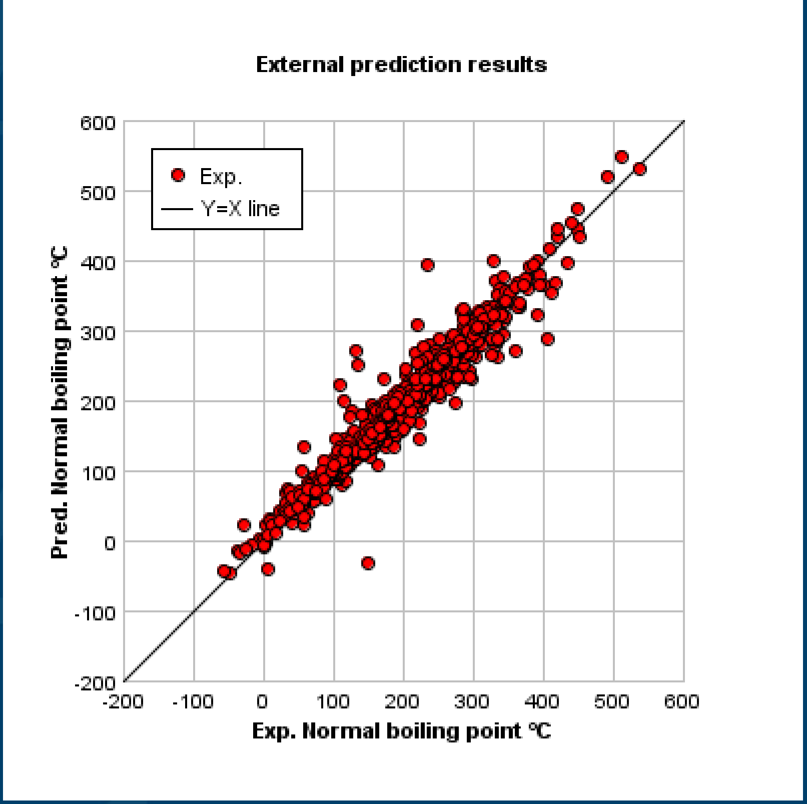
Method	Concor-dance	Sensitivity	Specificity	Coverage
Hierarchical	0.741	0.854	0.471	1.000
Single Model	0.754	0.900	0.412	0.983
FDA	0.672	0.780	0.412	1.000
Nearest neighbor	0.795	0.844	0.667	0.759
Consensus	0.759	0.902	0.412	1.000
Random Forest	0.852	0.949	0.600	0.931

$$\frac{R^2 - R_0^2}{R^2}$$

United States Environmental Protection Agency

Normal boiling point

Method	R^2	Coverage
Hierarchical	0.949	0.935
FDA	0.936	0.988
GC	0.897	0.977
NN	0.877	0.988
Consensus	0.947	0.986

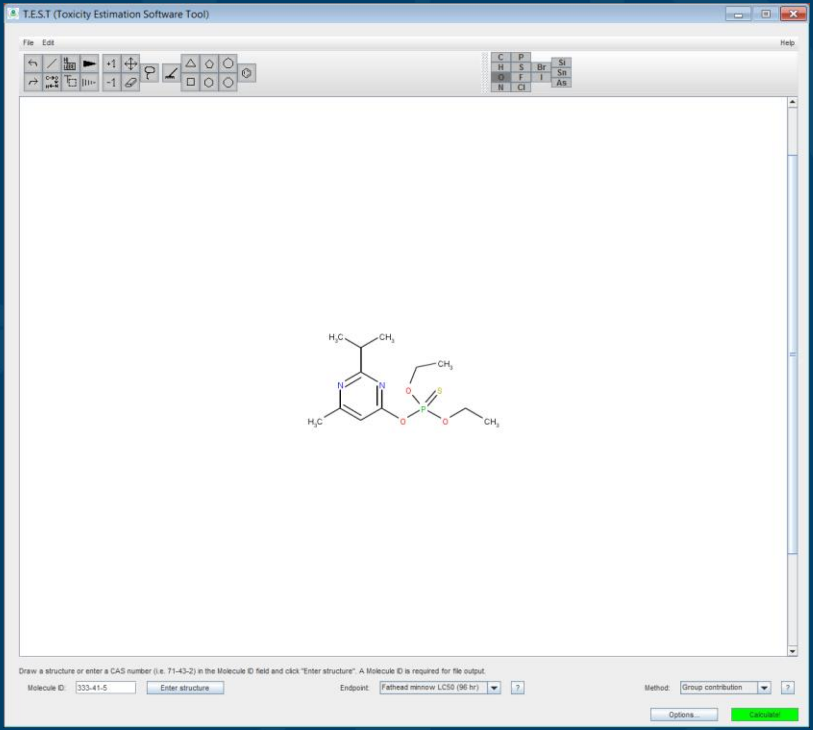


When not to use T.E.S.T.

- Compounds containing elements other than C, H, O, N, F, Cl, Br, I, S, P, Si, As
- Inorganic compounds
- Polymers
- Mixtures (more than one molecule)
- Salts / Ionic species
- Very complicated polycyclic aromatics such as Bucky balls
- When only one model can make a prediction (especially if method is NN method)

Where can I get T.E.S.T.?

http://bit.ly/1suh4kr



Draw a structure or enter a CAS number (i.e. 71-43-2) in the Molecule ID field and click "Enter structure". A Molecule ID is required for file output.

Molecule ID: 333-41-5

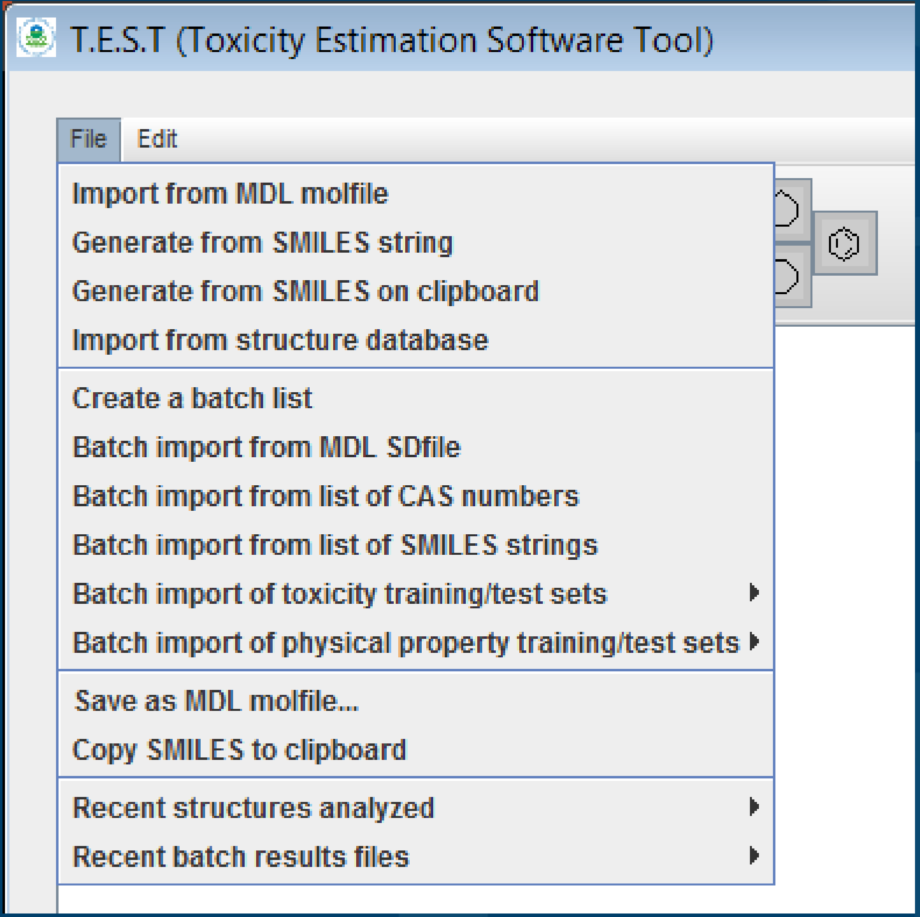
Endpoint: Fathead minnow LC50 (96 hr)

Method: Group contribution

United States
Environmental Protection
Agency

Tutorial

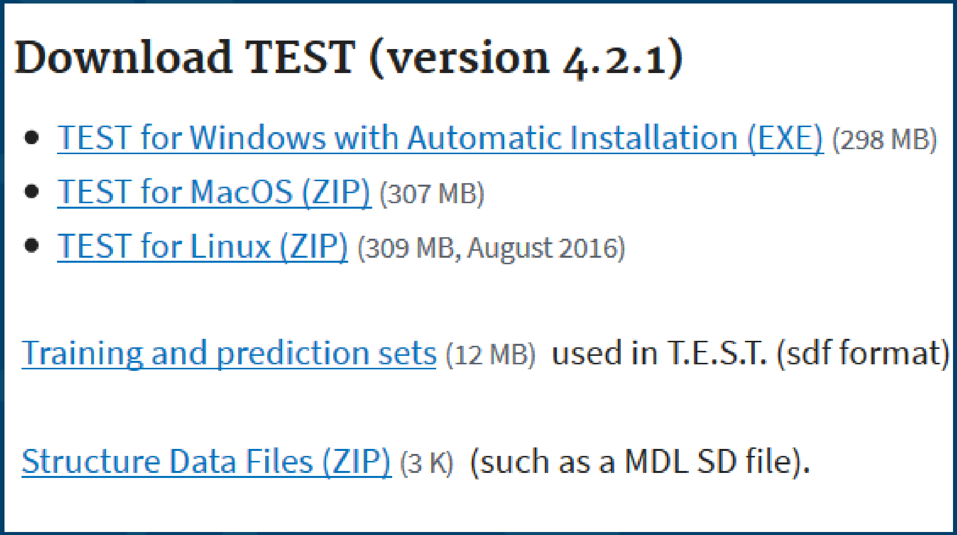
Importing files



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

File Edit

- Import from MDL molfile
- Generate from SMILES string
- Generate from SMILES on clipboard
- Import from structure database
- Create a batch list
- Batch import from MDL SDfile
- Batch import from list of CAS numbers
- Batch import from list of SMILES strings
- Batch import of toxicity training/test sets ▶
- Batch import of physical property training/test sets ▶
- Save as MDL molfile...
- Copy SMILES to clipboard
- Recent structures analyzed ▶
- Recent batch results files ▶



Download TEST (version 4.2.1)

- TEST for Windows with Automatic Installation (EXE) (298 MB)
- TEST for MacOS (ZIP) (307 MB)
- TEST for Linux (ZIP) (309 MB, August 2016)

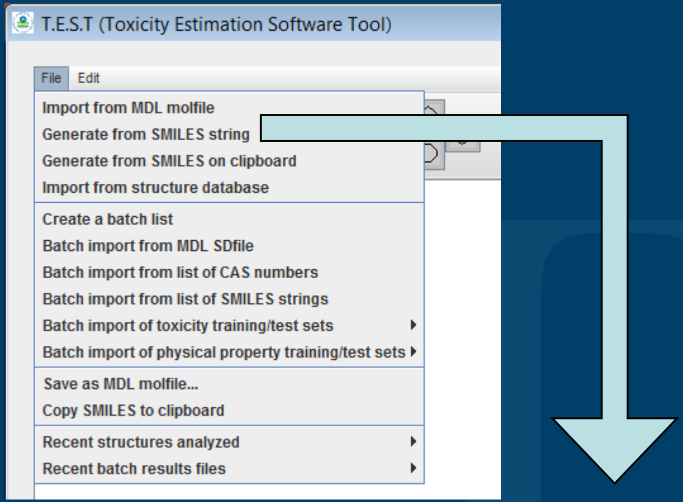
Training and prediction sets (12 MB) used in T.E.S.T. (sdf format)

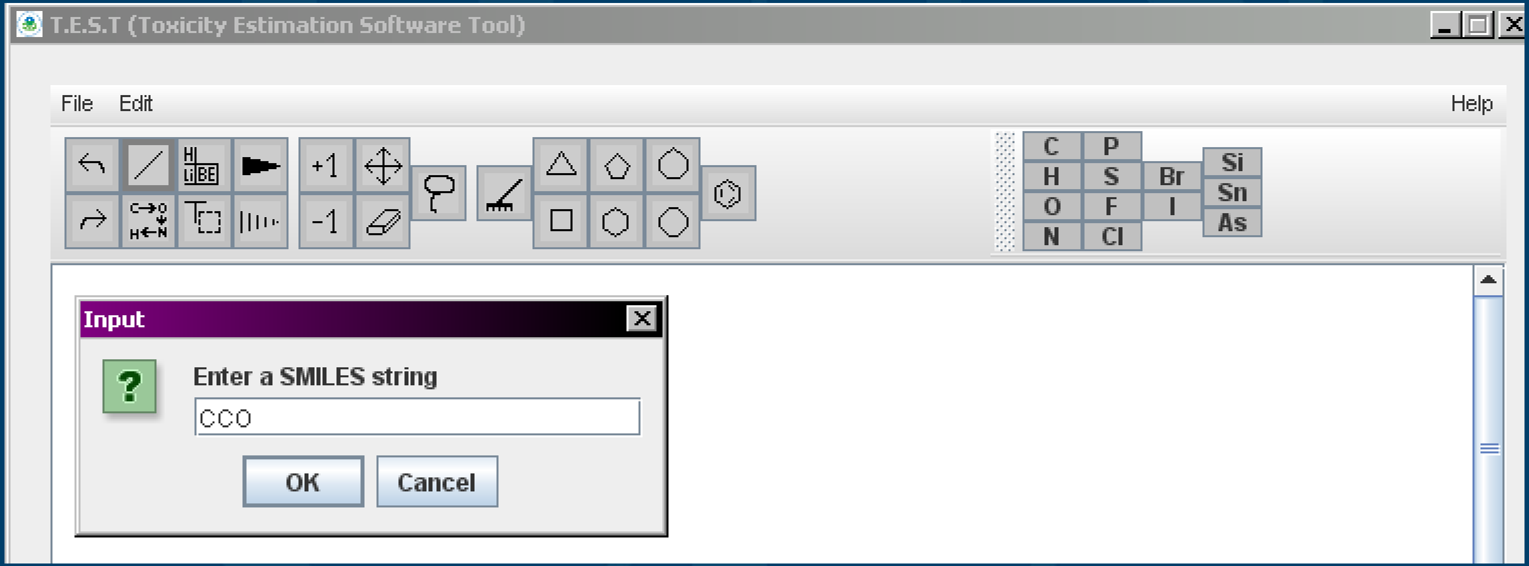
Structure Data Files (ZIP) (3 K) (such as a MDL SD file).

Example of SD File

```
Benzene, ID: C71432
  NIST    04042217093D 1   1.00000     0.00000
NIST Chemistry WebBook
 12 12  0     0  0              1 V2000
    3.2883    3.3891    0.2345 C   0  0  0  0  0  0           0  0  0
    1.9047    3.5333    0.2237 C   0  0  0  0  0  0           0  0  0
    3.8560    2.1213    0.1612 C   0  0  0  0  0  0           0  0  0
    1.0888    2.4099    0.1396 C   0  0  0  0  0  0           0  0  0
    3.0401    0.9977    0.0771 C   0  0  0  0  0  0           0  0  0
    1.6565    1.1421    0.0663 C   0  0  0  0  0  0           0  0  0
    3.9303    4.2734    0.3007 H   0  0  0  0  0  0           0  0  0
    1.4582    4.5312    0.2815 H   0  0  0  0  0  0           0  0  0
    4.9448    2.0077    0.1699 H   0  0  0  0  0  0           0  0  0
    0.0000    2.5234    0.1311 H   0  0  0  0  0  0           0  0  0
    3.4870    0.0000    0.0197 H   0  0  0  0  0  0           0  0  0
    1.0145    0.2578    0.0000 H   0  0  0  0  0  0           0  0  0
  2  1  2  0     0  0
  1  3  1  0     0  0
  1  7  1  0     0  0
  4  2  1  0     0  0
  2  8  1  0     0  0
  3  5  2  0     0  0
  3  9  1  0     0  0
  6  4  2  0     0  0
  4 10  1  0     0  0
  5  6  1  0     0  0
  5 11  1  0     0  0
  6 12  1  0     0  0
M  END
> <CAS>
71-43-2
```

SMILES Example


T.E.S.T (Toxicity Estimation Software Tool)
File
Edit
Import from MDL molfile
Generate from SMILES string
Generate from SMILES on clipboard
Import from structure database
Create a batch list
Batch import from MDL SDfile
Batch import from list of CAS numbers
Batch import from list of SMILES strings
Batch import of toxicity training/test sets
Batch import of physical property training/test sets
Save as MDL molfile...
Copy SMILES to clipboard
Recent structures analyzed
Recent batch results files

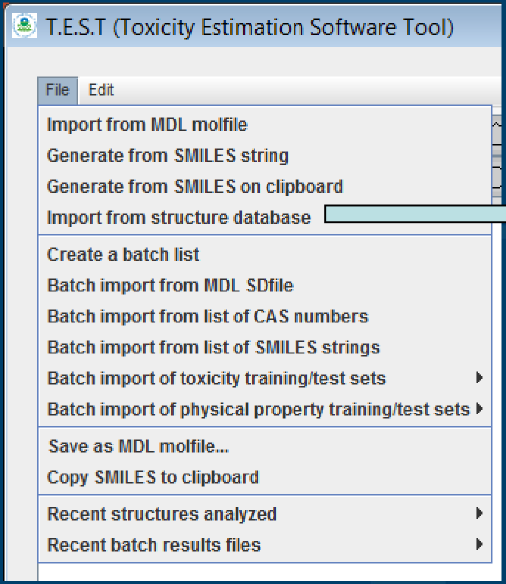


Enter a SMILES string

CCO

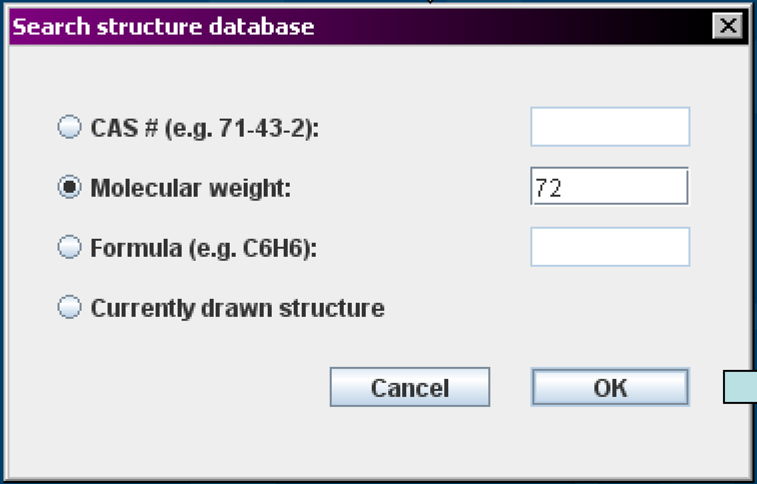
Importing from the database

There are approximately 20,000 compounds in the database



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

File
Edit
Import from MDL molfile
Generate from SMILES string
Generate from SMILES on clipboard
Import from structure database
Create a batch list
Batch import from MDL SDfile
Batch import from list of CAS numbers
Batch import from list of SMILES strings
Batch import of toxicity training/test sets
Batch import of physical property training/test sets
Save as MDL molfile...
Copy SMILES to clipboard
Recent structures analyzed
Recent batch results files



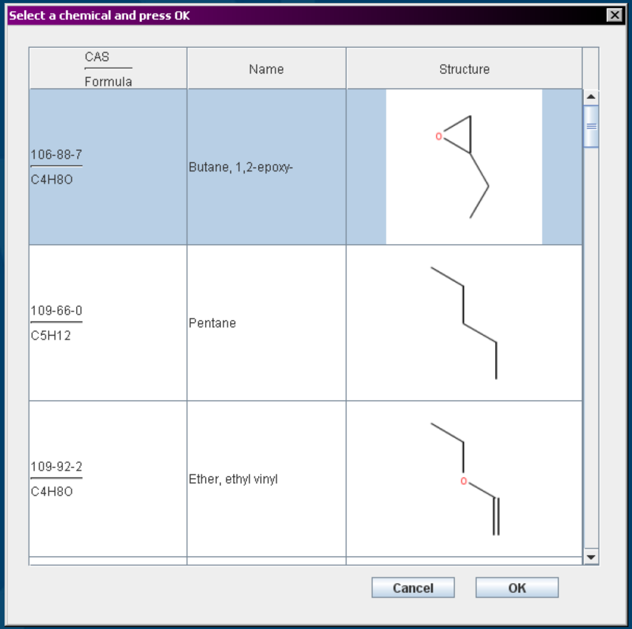
Search structure database

○ CAS # (e.g. 71-43-2):

● Molecular weight: 72

○ Formula (e.g. C6H6):

○ Currently drawn structure

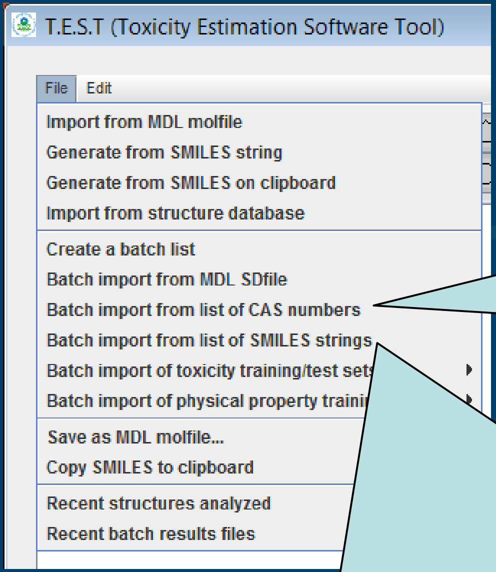


Select a chemical and press OK

CAS Formula	Name	Structure
106-88-7 C4H8O	Butane, 1,2-epoxy-	
109-66-0 C5H12	Pentane	
109-92-2 C4H8O	Ether, ethyl vinyl	

EPA
United States
Environmental Protection
Agency

Batch Importing

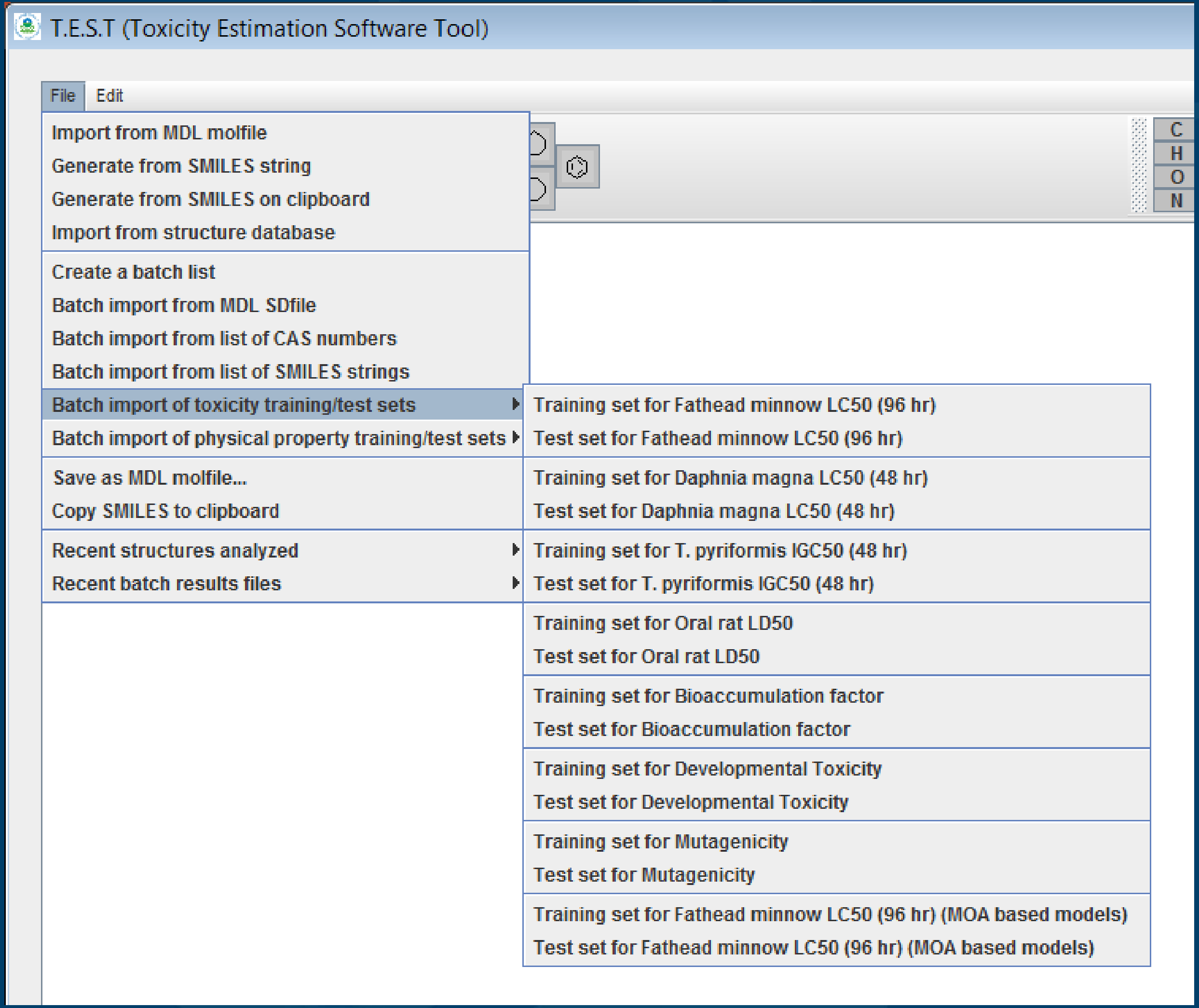

T.E.S.T (Toxicity Estimation Software Tool)
File
Edit
Import from MDL molfile
Generate from SMILES string
Generate from SMILES on clipboard
Import from structure database
Create a batch list
Batch import from MDL SDfile
Batch import from list of CAS numbers
Batch import from list of SMILES strings
Batch import of toxicity training/test sets
Batch import of physical property traini
Save as MDL molfile...
Copy SMILES to clipboard
Recent structures analyzed
Recent batch results files

```
53-96-3
62-44-2
80-08-0
87-62-7
90-94-8
91-59-8
...
```

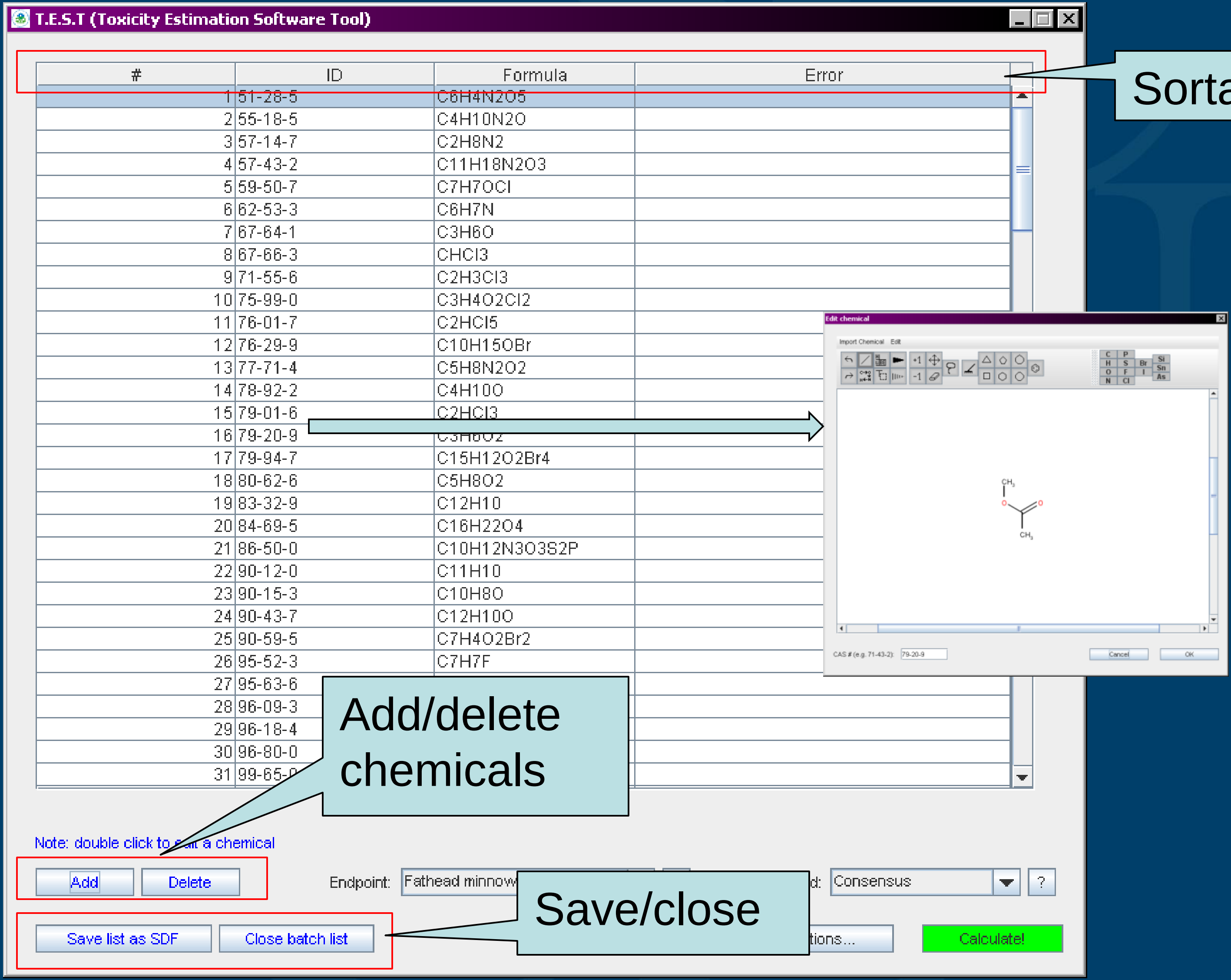
```
CC(=O)NC1=CC=C2C(CC3=C2C=CC=C3)=C1   53-96-3
CCOC1=CC=C(NC(C)=O)C=C1  62-44-2
NC1=CC=C(C=C1)S(=O)(=O)C1=CC=C(N)C=C1        80-08-0
CC1=CC=CC(C)=C1N   87-62-7
CN(C)C1=CC=C(C=C1)C(=O)C1=CC=C(C=C1)N(C)C  90-94-8
NC1=CC2=C(C=CC=C2)C=C1    91-59-8
NC1=C(Cl)C=C(C=C1)C1=CC(Cl)=C(N)C=C1         91-94-1
...
```

Batch importing continued

You can import training and test sets used for each endpoint


T.E.S.T (Toxicity Estimation Software Tool)
File
Edit
Import from MDL molfile
Generate from SMILES string
Generate from SMILES on clipboard
Import from structure database
Create a batch list
Batch import from MDL SDfile
Batch import from list of CAS numbers
Batch import from list of SMILES strings
Batch import of toxicity training/test sets
Batch import of physical property training/test sets
Save as MDL molfile...
Copy SMILES to clipboard
Recent structures analyzed
Recent batch results files
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 Bioaccumulation factor
Test set for Bioaccumulation 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)
C
H
O
N

Batch mode



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

#	ID	Formula	Error
1	51-28-5	C6H4N2O5	
2	55-18-5	C4H10N2O	
3	57-14-7	C2H8N2	
4	57-43-2	C11H18N2O3	
5	59-50-7	C7H7OCl	
6	62-53-3	C6H7N	
7	67-64-1	C3H6O	
8	67-66-3	CHCl3	
9	71-55-6	C2H3Cl3	
10	75-99-0	C3H4O2Cl2	
11	76-01-7	C2HCl5	
12	76-29-9	C10H15OBr	
13	77-71-4	C5H8N2O2	
14	78-92-2	C4H10O	
15	79-01-6	C2HCl3	
16	79-20-9	C3H6O2	
17	79-94-7	C15H12O2Br4	
18	80-62-6	C5H8O2	
19	83-32-9	C12H10	
20	84-69-5	C16H22O4	
21	86-50-0	C10H12N3O3S2P	
22	90-12-0	C11H10	
23	90-15-3	C10H8O	
24	90-43-7	C12H10O	
25	90-59-5	C7H4O2Br2	
26	95-52-3	C7H7F	
27	95-63-6		
28	96-09-3		
29	96-18-4		
30	96-80-0		
31	99-65-0		

Note: double click to edit a chemical

Endpoint: Fathead minnow

d: Consensus

CAS # (e.g. 71-43-2): 79-20-9

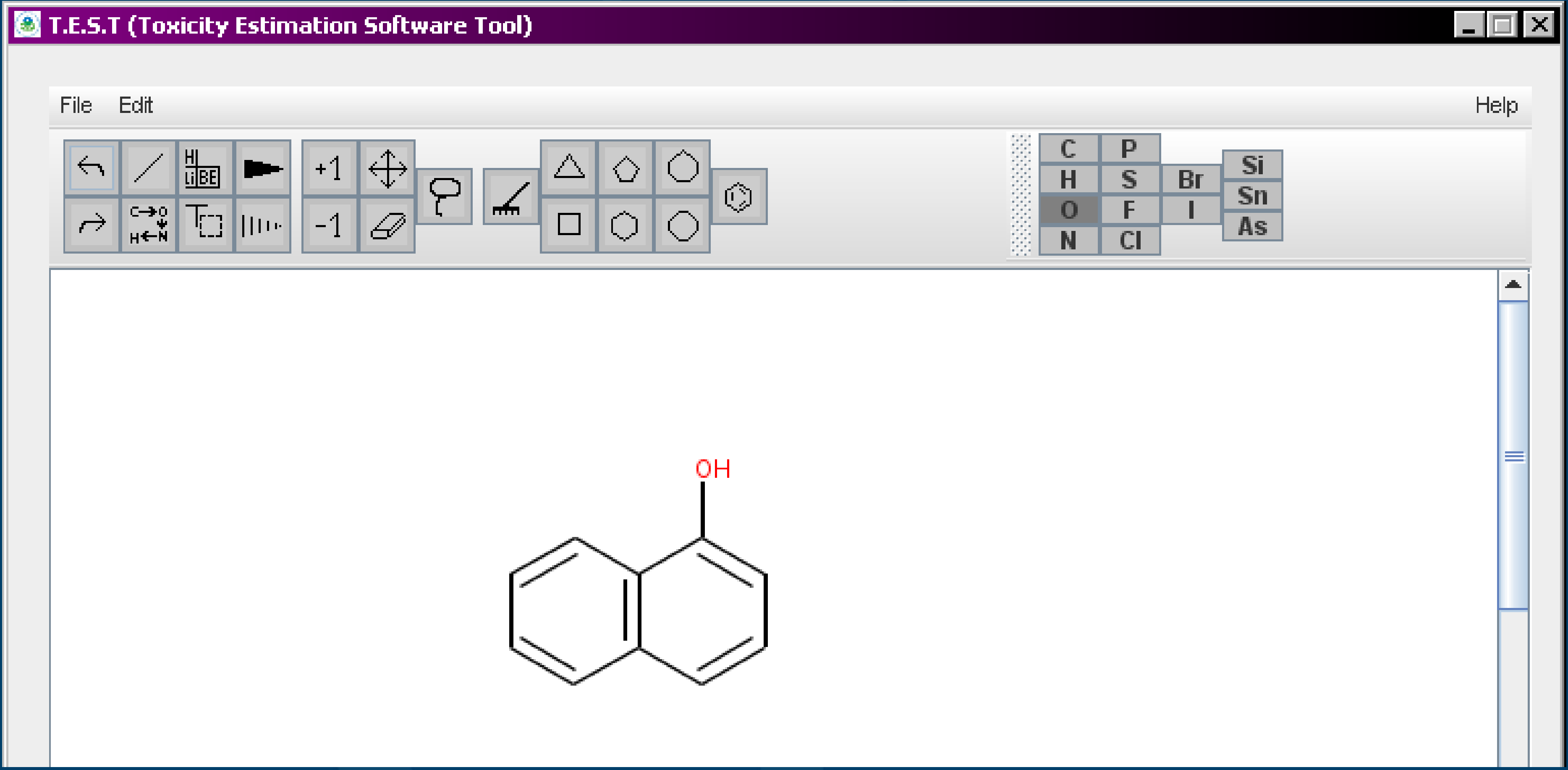
Sortable

Add/delete chemicals

Save/close

Drawing structures

Structures can also be drawn using graphical user interface:

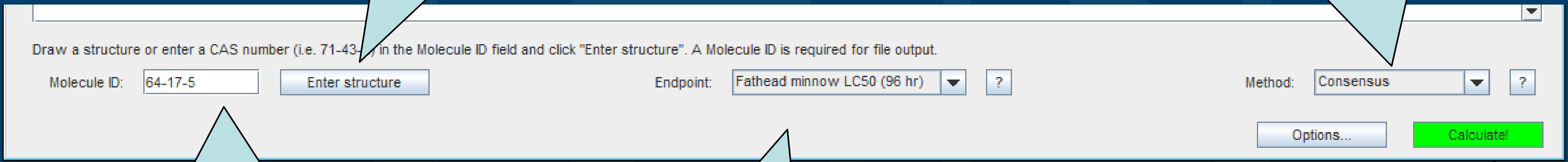


EPA United States Environmental Protection Agency

Bottom of interface

Load structure from molecule ID (CAS only)

Selects QSAR method



Draw a structure or enter a CAS number (i.e. 71-43-) in the Molecule ID field and click "Enter structure". A Molecule ID is required for file output.

Molecule ID: 64-17-5

Endpoint: Fathead minnow LC50 (96 hr)

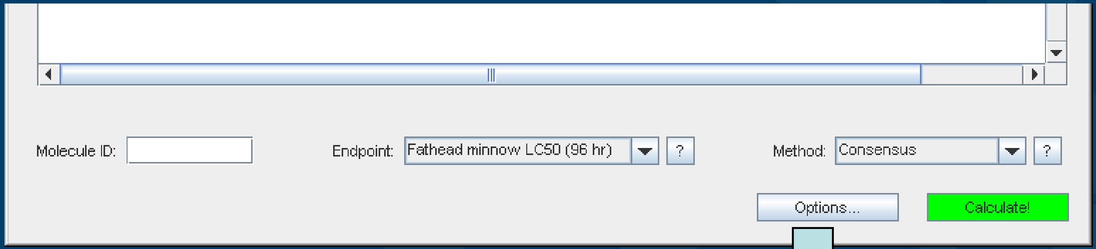
Method: Consensus

Enter molecule ID

Selects endpoint

Runs the QSAR calculation

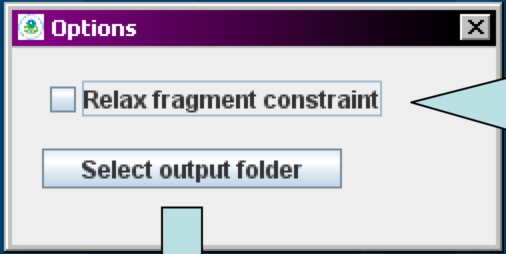
Options button



Molecule ID:

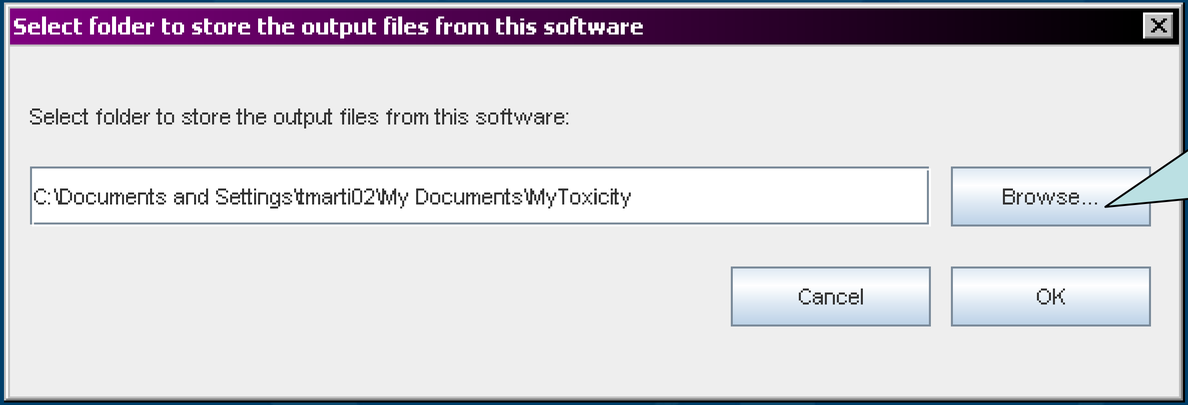
Endpoint: Fathead minnow LC50 (96 hr)

Method: Consensus



□ Relax fragment constraint

Checking this box will remove the fragment constraint from determination of applicability domain



Select folder to store the output files from this software

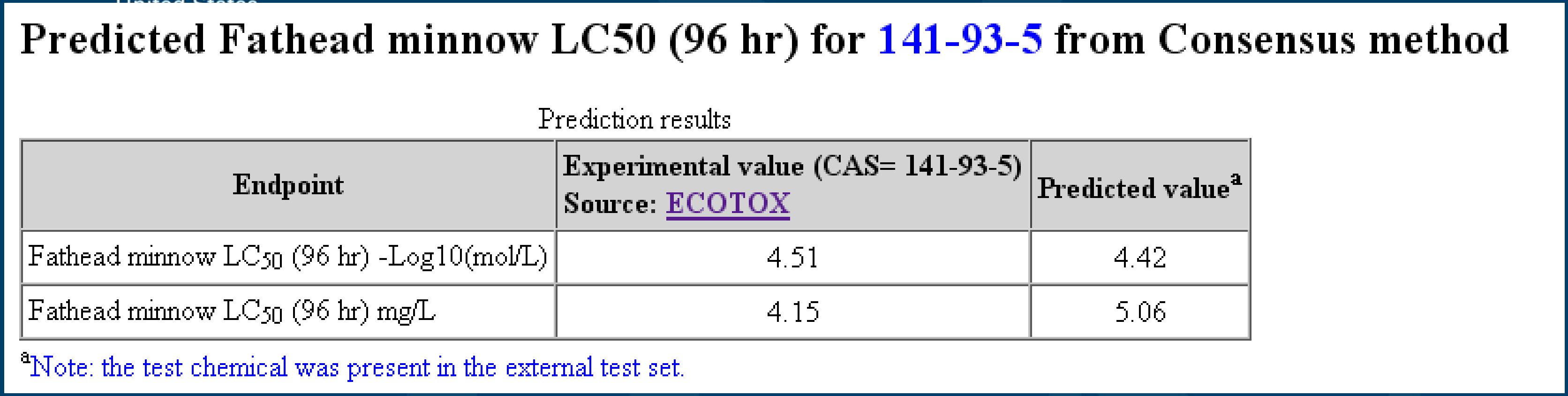
Select folder to store the output files from this software:

C:\Documents and Settings\tmarti02\My Documents\MyToxicity

Sets main folder where all results web pages will be stored

Examples

Well predicted chemical

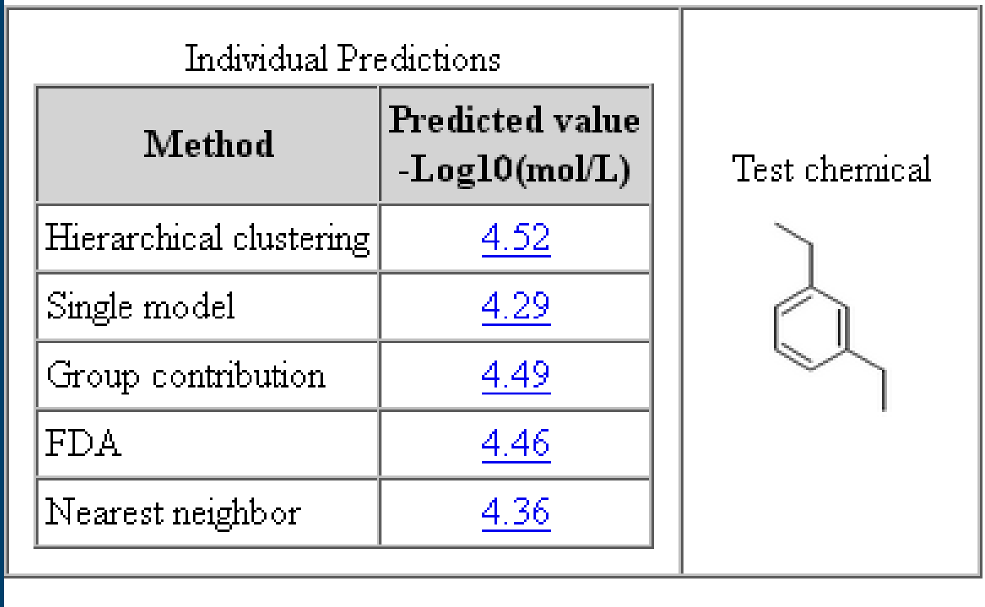


Predicted Fathead minnow LC50 (96 hr) for 141-93-5 from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 141-93-5) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	4.51	4.42
Fathead minnow LC_{50} (96 hr) mg/L	4.15	5.06

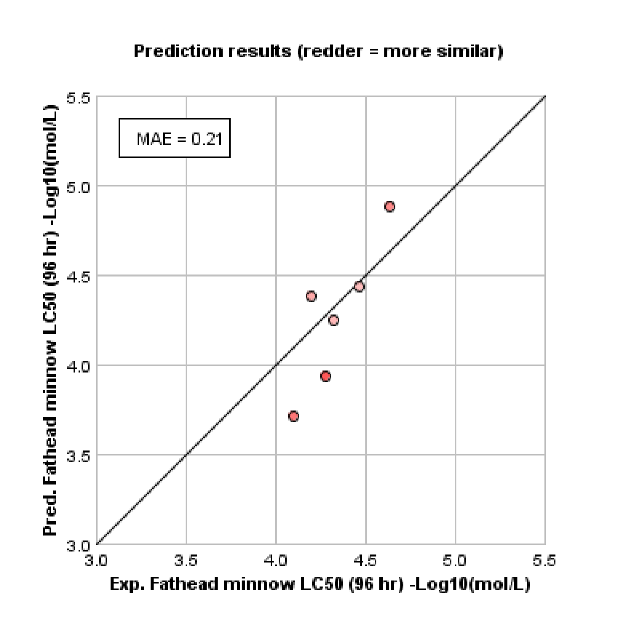
[a]Note: the test chemical was present in the external test set.



Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	4.52
Single model	4.29
Group contribution	4.49
FDA	4.46
Nearest neighbor	4.36

Test chemical



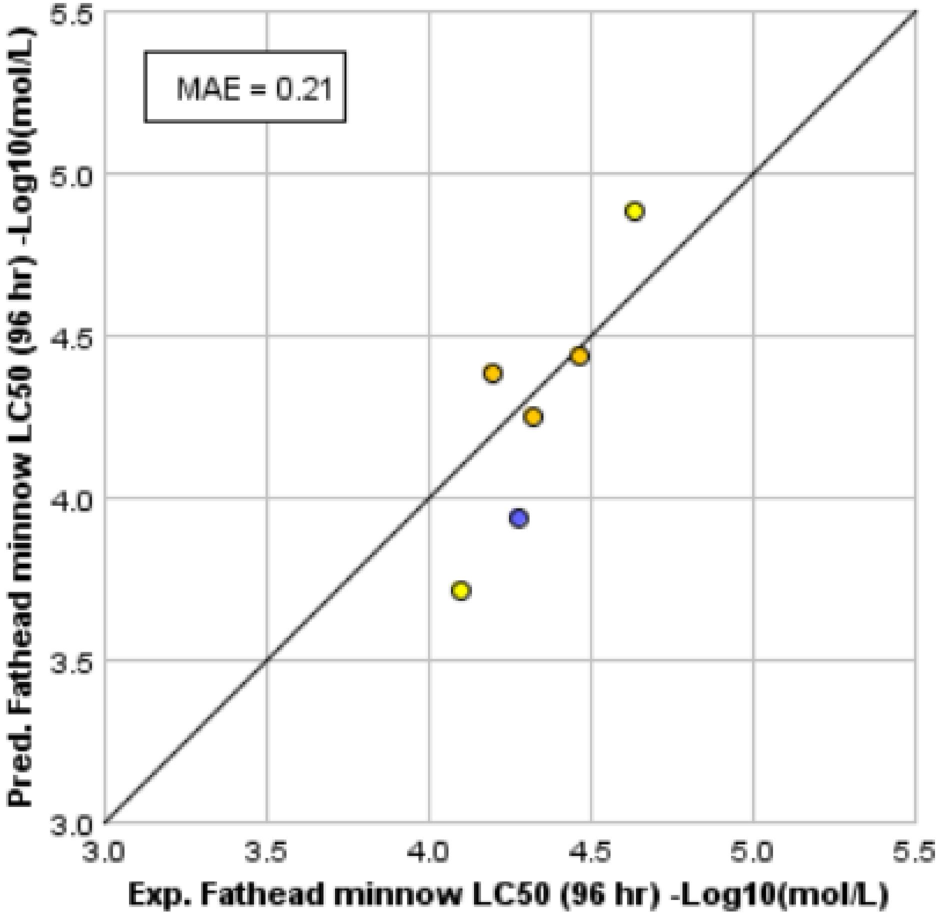
Prediction results (redder = more similar)

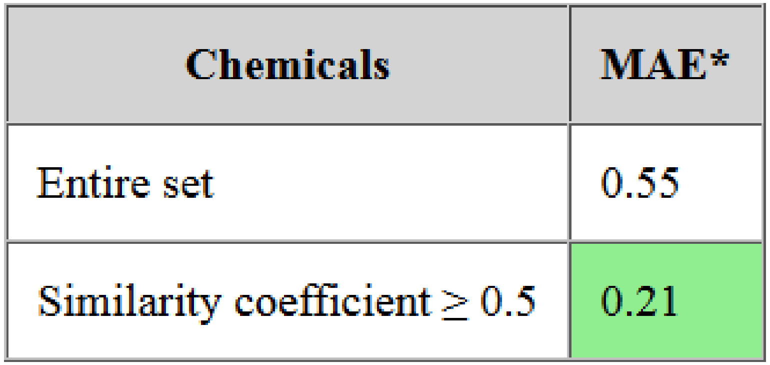
- Predictions are consistent

- Similar test set chemicals are predicted well

Well predicted chemical, cont.

Prediction results (colors defined in table below)

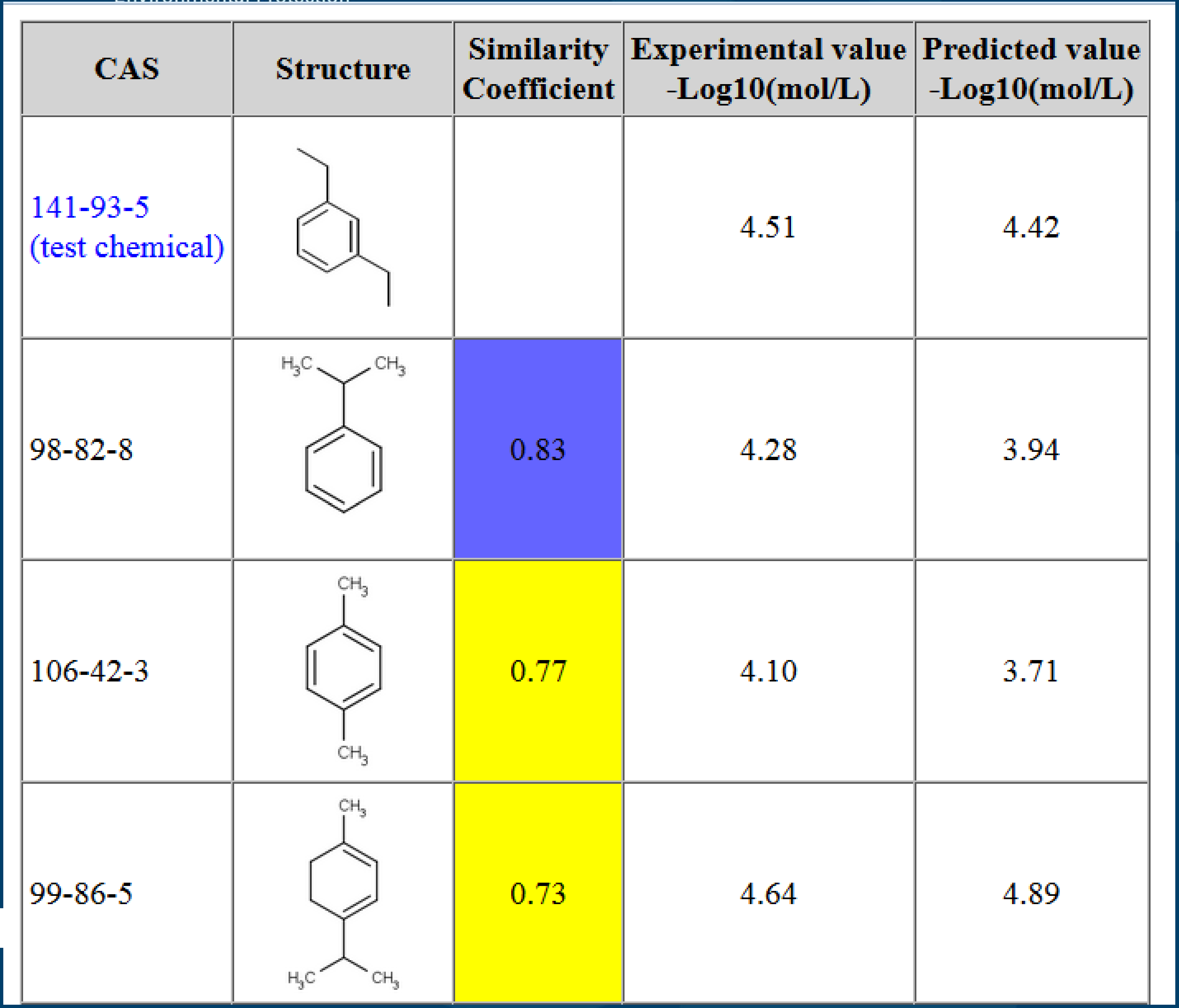




Chemicals	MAE*
Entire set	0.55
Similarity coefficient ≥ 0.5	0.21

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

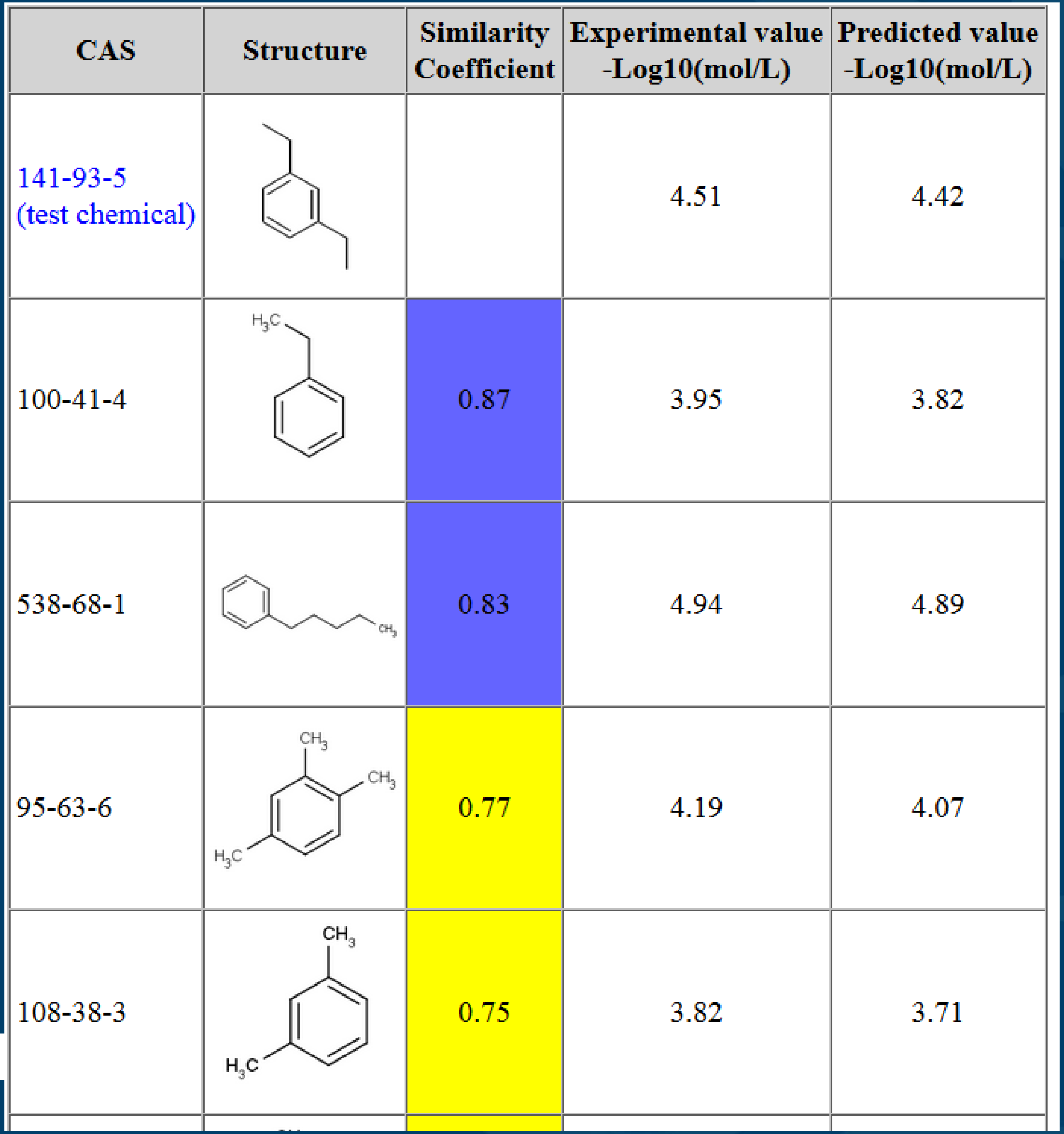
Well predicted chemical, cont.



CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
141-93-5 (test chemical)			4.51	4.42
98-82-8		0.83	4.28	3.94
106-42-3		0.77	4.10	3.71
99-86-5		0.73	4.64	4.89

- Similar chemicals in the test set

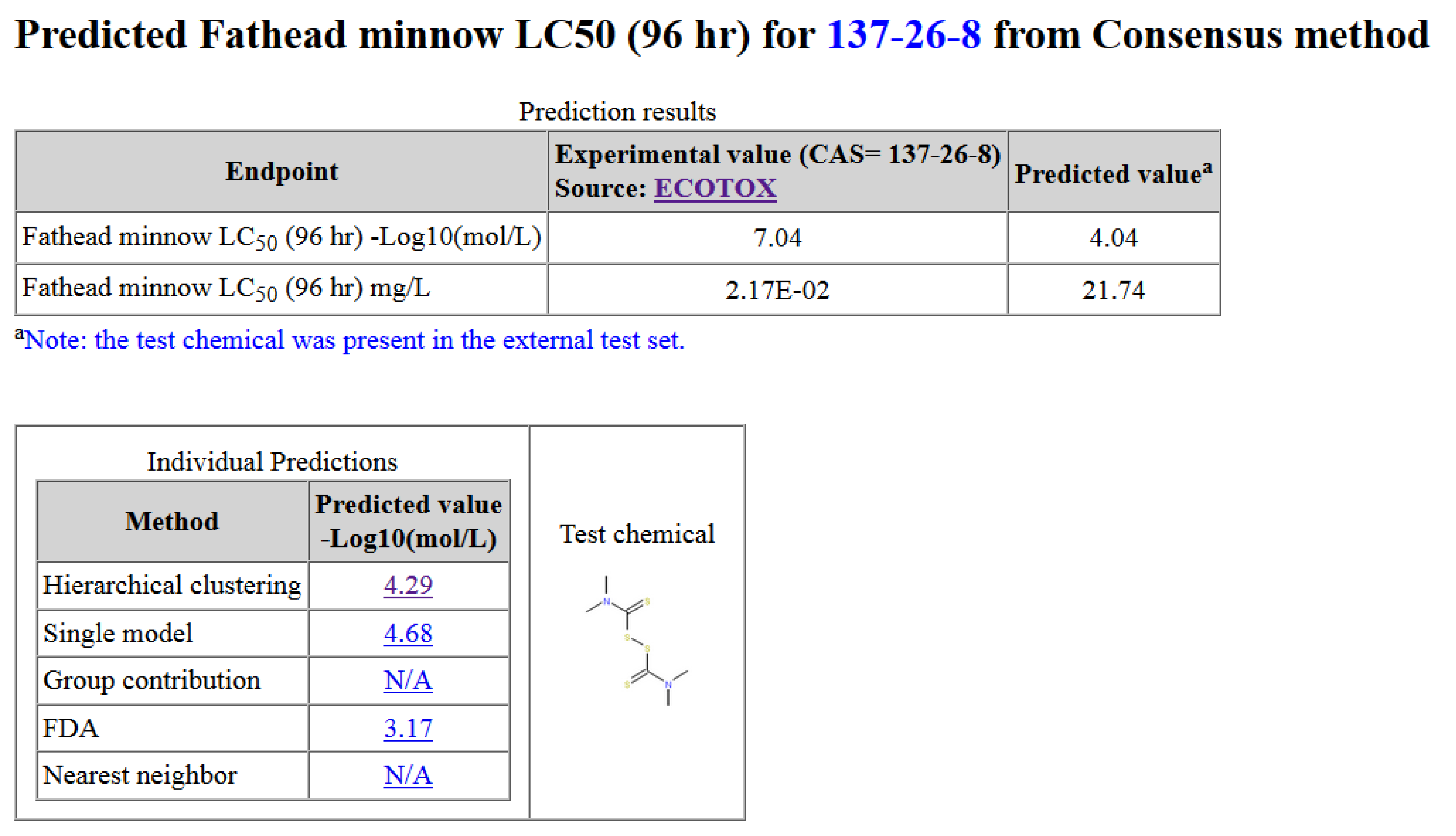
Well predicted chemical, cont.


H$_3$C
CH$_3$
CH$_3$
CH$_3$
H$_3$C
CH$_3$
H$_3$C

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
141-93-5 (test chemical)			4.51	4.42
100-41-4		0.87	3.95	3.82
538-68-1		0.83	4.94	4.89
95-63-6		0.77	4.19	4.07
108-38-3		0.75	3.82	3.71

- Similar chemicals are present in the training set

Ex. poorly predicted chemical



Predicted Fathead minnow LC50 (96 hr) for 137-26-8 from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 137-26-8) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	7.04	4.04
Fathead minnow LC_{50} (96 hr) mg/L	2.17E-02	21.74

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

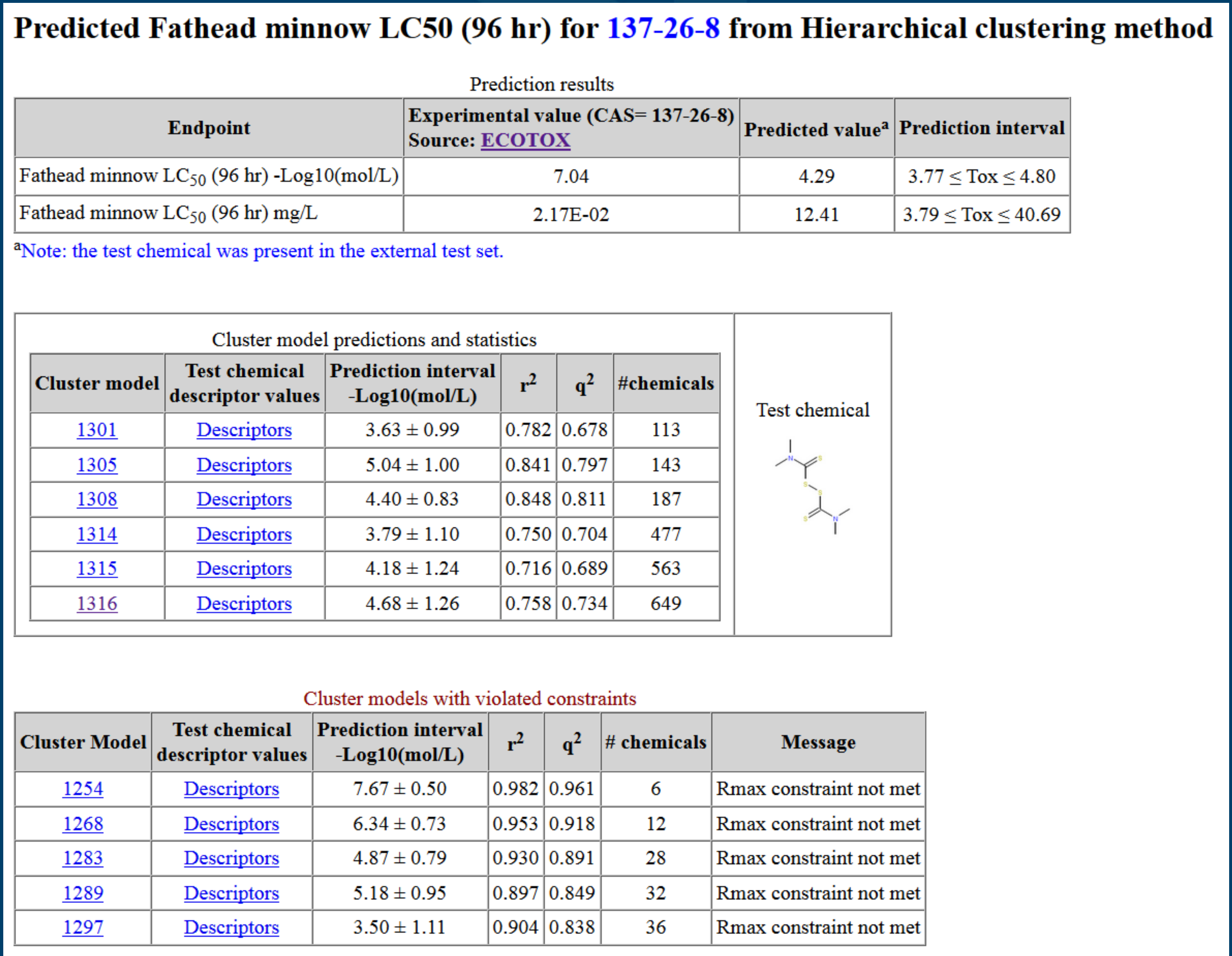
Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	4.29
Single model	4.68
Group contribution	N/A
FDA	3.17
Nearest neighbor	N/A

Test chemical

- ➢Predictions are not consistent or some methods are outside their applicability domain

Ex. poorly predicted chemical, cont.



Predicted Fathead minnow LC50 (96 hr) for 137-26-8 from Hierarchical clustering method

Prediction results

Endpoint	Experimental value (CAS= 137-26-8) Source: ECOTOX	Predicted value[a]	Prediction interval
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	7.04	4.29	3.77 ≤ Tox ≤ 4.80
Fathead minnow LC_{50} (96 hr) mg/L	2.17E-02	12.41	3.79 ≤ Tox ≤ 40.69

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

Cluster model predictions and statistics

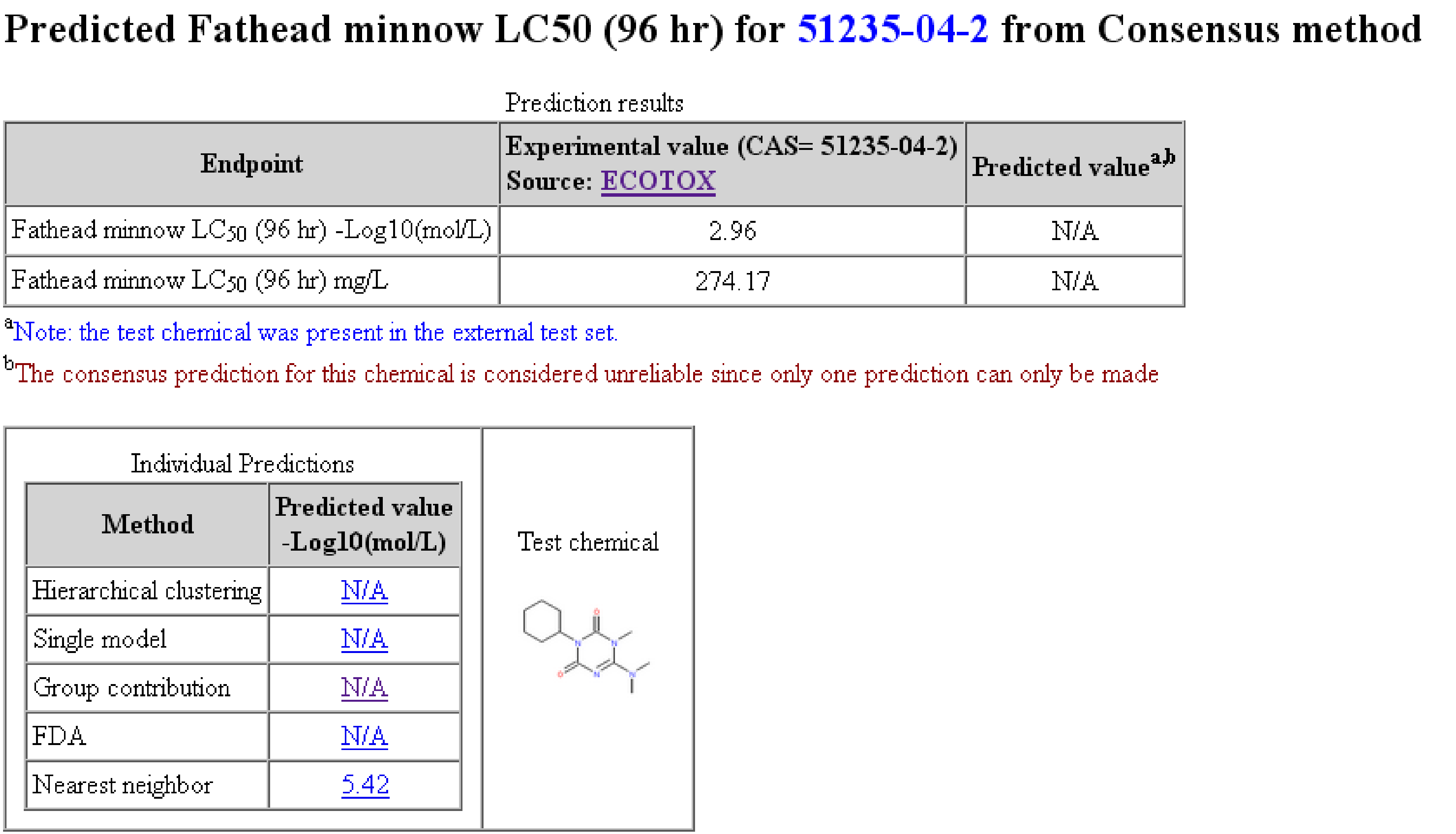
Cluster model	Test chemical descriptor values	Prediction interval -Log10(mol/L)	r^2	q^2	#chemicals
1301	Descriptors	3.63 ± 0.99	0.782	0.678	113
1305	Descriptors	5.04 ± 1.00	0.841	0.797	143
1308	Descriptors	4.40 ± 0.83	0.848	0.811	187
1314	Descriptors	3.79 ± 1.10	0.750	0.704	477
1315	Descriptors	4.18 ± 1.24	0.716	0.689	563
1316	Descriptors	4.68 ± 1.26	0.758	0.734	649

Test chemical

Cluster models with violated constraints

Cluster Model	Test chemical descriptor values	Prediction interval -Log10(mol/L)	r^2	q^2	# chemicals	Message
1254	Descriptors	7.67 ± 0.50	0.982	0.961	6	Rmax constraint not met
1268	Descriptors	6.34 ± 0.73	0.953	0.918	12	Rmax constraint not met
1283	Descriptors	4.87 ± 0.79	0.930	0.891	28	Rmax constraint not met
1289	Descriptors	5.18 ± 0.95	0.897	0.849	32	Rmax constraint not met
1297	Descriptors	3.50 ± 1.11	0.904	0.838	36	Rmax constraint not met

Chemical which can’t be predicted



Predicted Fathead minnow LC50 (96 hr) for 51235-04-2 from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 51235-04-2) Source: ECOTOX	Predicted value[a,b]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	2.96	N/A
Fathead minnow LC_{50} (96 hr) mg/L	274.17	N/A

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

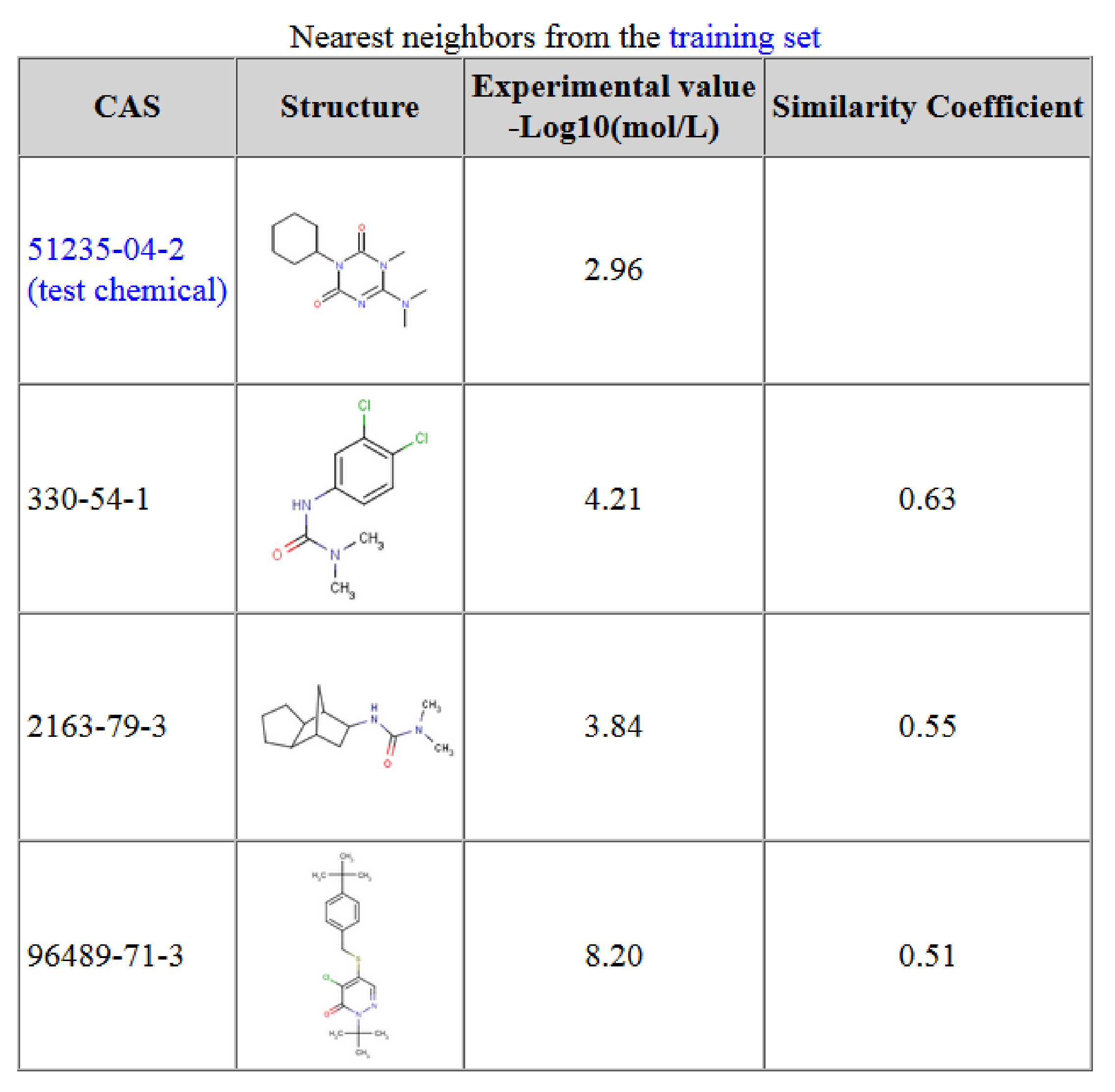
[b] The consensus prediction for this chemical is considered unreliable since only one prediction can only be made

Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	N/A
Single model	N/A
Group contribution	N/A
FDA	N/A
Nearest neighbor	5.42

Test chemical

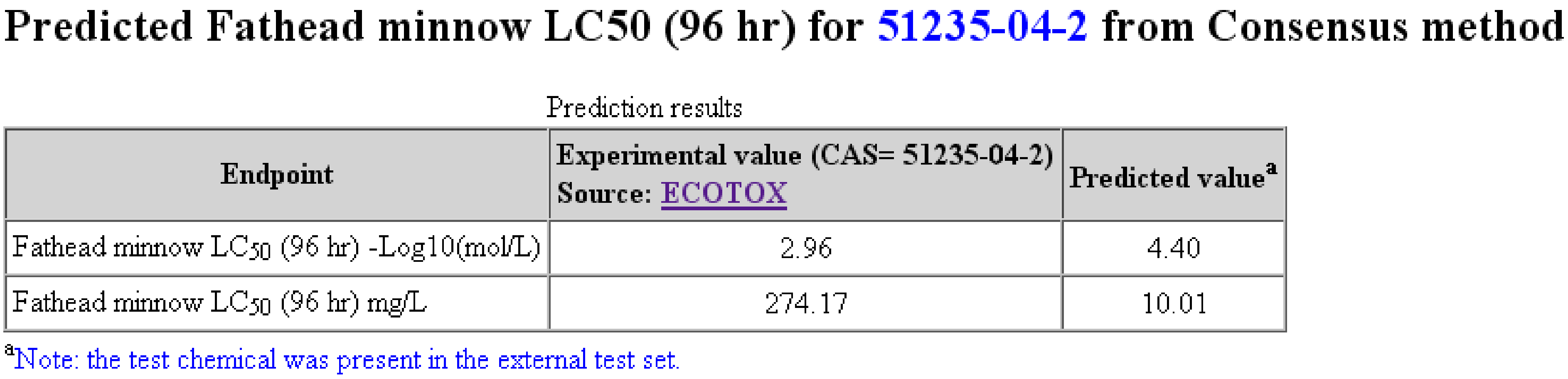
Nearest neighbor prediction



Nearest neighbors from the training set

CAS	Structure	Experimental value -Log10(mol/L)	Similarity Coefficient
51235-04-2 (test chemical)		2.96	
330-54-1		4.21	0.63
2163-79-3		3.84	0.55
96489-71-3		8.20	0.51

After relaxing fragment constraint

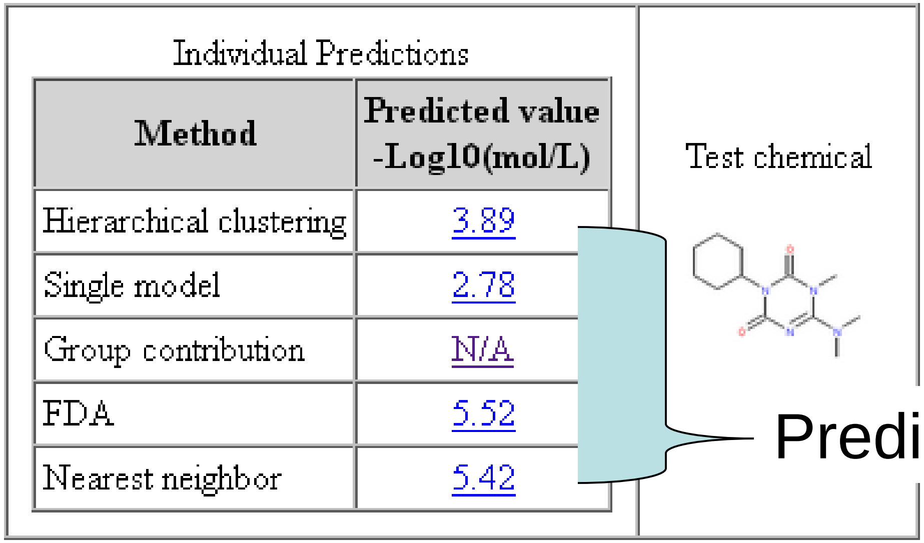


Predicted Fathead minnow LC50 (96 hr) for 51235-04-2 from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 51235-04-2) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	2.96	4.40
Fathead minnow LC_{50} (96 hr) mg/L	274.17	10.01

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



Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	3.89
Single model	2.78
Group contribution	N/A
FDA	5.52
Nearest neighbor	5.42

Test chemical

Predictions are inconsistent!

EPA
United States
Environmental Protection
Agency

Questions???

Email: martin.todd@epa.gov

The views expressed in this presentation are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency