

## OECD QSAR Toolbox v.3.2

Step-by-step example of how to build a user-defined QSAR

# Outlook

- **Background**
- Objectives
- The exercise
- Workflow of the exercise

## Background

- This is a step-by-step presentation designed to take you through the workflow of the Toolbox for building a QSAR model for predicting aquatic toxicity.
- By now you have some experience in using the Toolbox so there will be multiple key strokes between screen shots.

# Outlook

- Background
- **Objectives**
- The exercise
- Workflow of the exercise

## Objectives

- **This presentation demonstrates building a QSAR model for predicting acute toxicity to *Tetrahymena pyriformis* of aldehydes. The presentation addresses specifically:**
  - predicting acute toxicity for a target chemical;
  - building QSAR model based on the prediction;
  - applying the model to other aldehydes;
  - exporting the predictions to a file.

# Outlook

- Background
- Objectives
- **The exercise**
- Workflow of the exercise

## The Exercise

- **This exercise includes the following steps:**
  - select a target chemical – Furfural, CAS 98011;
  - extract available experimental results;
  - search for analogues;
  - estimate the 48h-IGC50 for *Tetrahymena pyriformis* by using trend analysis;
  - improve the data set by either:
    - subcategorizing by “Protein binding” mechanisms, or
    - assessing the difference between outliers and the target chemical
  - evaluate and save the model;
  - use the model to display its training set, visualize its applicability domain and perform predictions.

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**



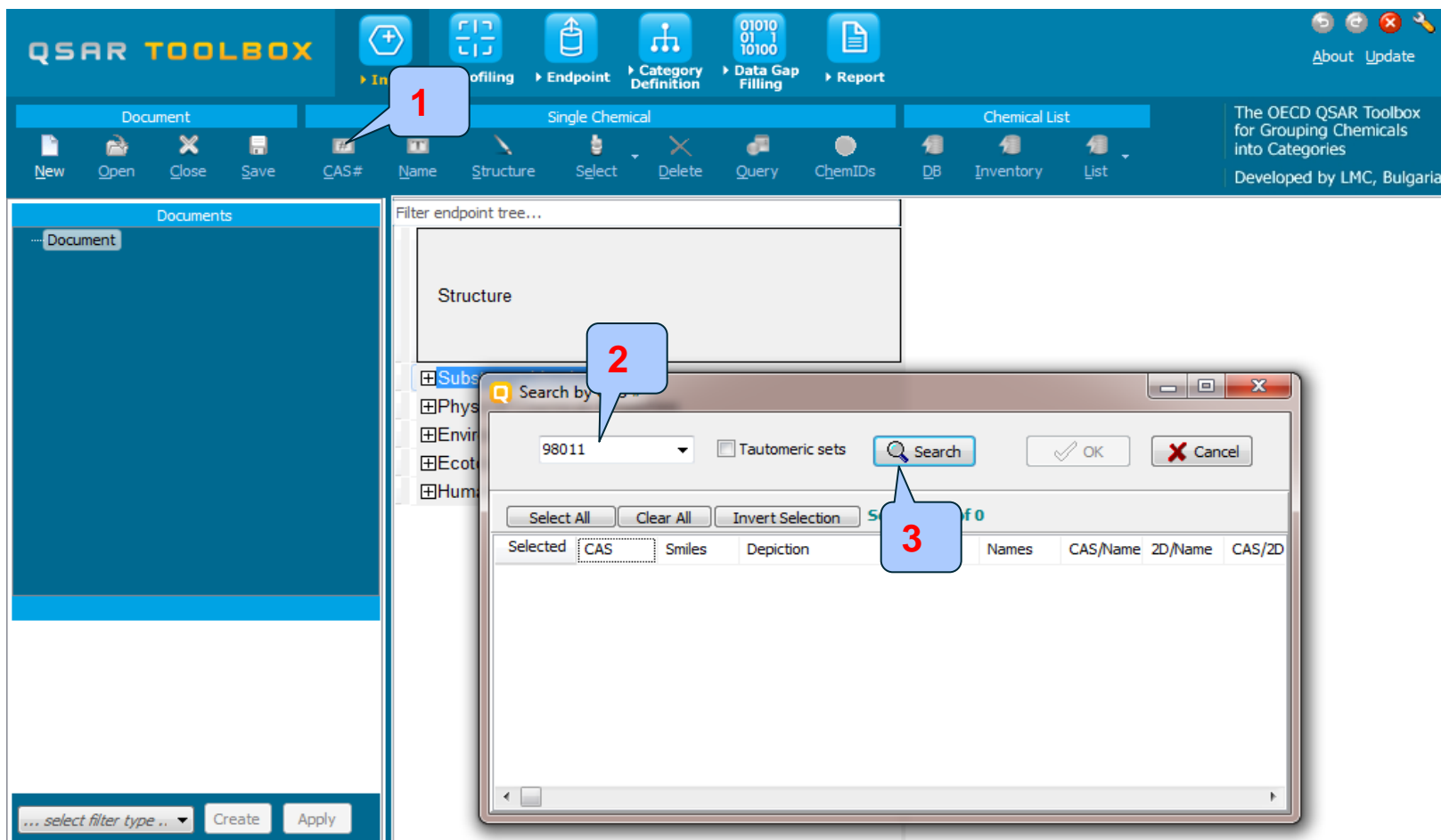
## Workflow of the exercise

- **Remember the Toolbox has 6 modules which are used in a sequential workflow:**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category Definition
  - Filling Data Gaps
  - Report

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - **Chemical Input**

# Chemical Input



**1. Click on CAS # 2. Enter 98011; 3. Click Search**

# Chemical Input

## Target chemical identity

The Toolbox now searches the Toolbox databases and inventories for the presence of the chemical with structure related to the current CAS number. It is displayed as a 2D image.

Search by CAS #

98011  Tautomeric sets

Select All Clear All Invert Selection Selected 1 of 1

Selected	CAS	Smiles	Depiction	Names	CAS/Name	2D/Name	CAS/2D
1. Yes	98-01-1	O=CC1=COC1		1: 2-fural 2: furfural 3: 2-fural 4: fufural 5: furan-	8: ME 9: RE 10: U 2: High C 1: Ba 2: Ca 3: Ca 4: DS 5: Ge 6: ML	8: US 9: GS 10: E 2: High C 1: Ba 2: Ca 3: ML 4: R 5: Ge 6: DS	

**1. Click OK to add chemical in data matrix**



In case a structure has several CAS numbers or a structure could be related to more than one substance (e.g. in the case of compounds), more than one chemical identity could be retrieved. In this case the user can decide which substance is to be retained for the subsequent workflow.

# Chemical Input

## Target chemical identity

- You have now your target chemical with its structure.
- **Click** on the box next to “Substance Identity”; this displays the chemical identification information. (see next screen shot)

# Chemical Input

## Target chemical identity

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Profiling' menu item is highlighted with a red box. Below the menu bar, there are tabs for 'Document', 'Single Chemical', and 'Chemical List'. The 'Single Chemical' tab is active, showing a 'Filter endpoint tree...' on the left and a list of identifiers on the right. The chemical structure of 2-furaldehyde is displayed in the center. The identifiers list includes: 98-01-1, EINECS Number: 200-001-2, 2-furaldehyde, furfural, 2-furancarboxaldehyde, furfural, furan-2-aldehyde, and O=CC1=CC=CO1. The 'Substance Identity' section is expanded, showing fields for CAS Number, Chemical IDs, Chemical Name, and Structural Formula. The 'Physical Chemical Properties', 'Environmental Fate and Transport', 'Ecotoxicological Information', and 'Human Health Hazards' sections are collapsed.

The workflow on the first module is now complete; click "Profiling" to move to the next module.

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - **Profiling**

# Profiling

## Profiling the target chemical

- **Select** the “Profiling methods” related to the target endpoint
- This selects (a **green** check mark appears) or deselects(**green** check disappears) profilers.
- For this example, select all profilers (see next screen shot)



# Profiling

## Profiling the target chemical

The screenshot displays the QSAR Toolbox software interface during the Profiling step. The top navigation bar includes buttons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The Profiling Schemes panel on the left shows a list of predefined profiling methods under the 'General Mechanistic' category, with the 'Select All' button circled in red and labeled '1'. The 'Apply' button is labeled '2'. The main workspace shows a 'Filter endpoint tree...' with a 'Structure' tab displaying the chemical structure of furfural. Below the structure, a list of identifiers is shown, including CAS Number (98-01-1), EINECS Number, and Chemical Name (2-furaldehyde, furfural, 2-furancarboxaldehyde, fufural, furan-2-aldehyde). The SMILES string O=CC1=CC=CO1 is also displayed.

**1. Check Select All profilers 2. Click Apply**


# Profiling

## Profiling the target chemical

- The actual profiling will take several seconds depending on the number and type of selected profilers.
- The results of profiling automatically appeared as a dropdown box under the target chemical. (see next screen shot)

# Profiling Profiles of "Furfural"

The screenshot shows the QSAR Toolbox interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Profiling' menu is open, showing options like 'Apply', 'New', 'View', and 'Delete'. On the left, a list of 'Profiling methods' is shown, with 'General Mechanistic' expanded. The main window displays a 'Filter endpoint tree...' for the target '1 [target]' (Furfural). The tree includes nodes for 'Structure', 'Substance Identity', 'Physical Chemical Properties', 'Environmental Fate and Transport', 'Ecotoxicological Information', 'Human Health Hazards', and 'Profile'. The 'Profile' node is circled in red, and a callout box with the number '1' points to it.

**1. Double click on the box  to open the nodes of the tree**

# Profiling Profiles of "Furfural"

The screenshot shows the QSAR Toolbox interface. On the left, the 'Profiling methods' panel is expanded to 'General Mechanistic'. The 'Filter endpoint tree...' window is active, showing a list of endpoints. A right-click context menu is open over the 'Protein binding by OASIS v1.2' endpoint, with the 'Explain' option highlighted. A blue callout box with the number '1' points to the 'Explain' option.

Endpoint	Value
Structure	<chem>O=Cc1ccoc1</chem>
Hydrolysis half-life (Ka, pH 7)(...)	Not calculated
Hydrolysis half-life (Ka, pH 8)(...)	Not calculated
Hydrolysis half-life (Kb, pH 7)(...)	Not calculated
Hydrolysis half-life (Kb, pH 8)(...)	Not calculated
Hydrolysis half-life (pH 6.5-7.4)	Not calculated
Ionization at pH = 1	Basic [0,10] No pKa value
Ionization at pH = 4	Basic [0,10] No pKa value
Ionization at pH = 7.4	Basic [0,10] No pKa value
Ionization at pH = 9	Basic [0,10] No pKa value
Protein binding by OASIS v1.2	Schiff base formation Schiff base formation >> Schiff Schiff base formation >> Schiff
Protein binding by OECD	No alert found
Protein binding potency	Moderately re... Moderately re...
Superfragments	No superfragm...
Toxic hazard classification by ...	High (Class III)
Toxic hazard classification by ...	High (Class III)

In this case there is structural evidence that the target could interact to DNA and proteins, it has also mode of action and it is aldehyde. This step is critical for next grouping of analogues.

**1. Right click** to see why the target is Protein binder (see next screen shot).

# Profiling Profiles of "Furfural"

The screenshot shows the QSAR Toolbox interface. The main window displays a filter endpoint tree with '1 [target]' and a chemical structure of furfural. Below the structure, there are three entries for 'Ionization at pH = 4', '7.4', and '9', each with associated pKa values. A 'Profiling results' dialog box is open, showing a hierarchical tree of alerts. The path 'Protein binding by OASIS v1.2 >> Schiff base formation >> Schiff base formation with carbonyl compounds >> Aldehydes' is highlighted. Callout boxes '1' and '2' point to the 'Aldehydes' node and the 'Details' button, respectively.

The Protein binding by OASIS v.1.2 profiler has hierarchical structure consisting of three levels: Structural alert, Mechanistic alert and Mechanistic domain

1. From the list of the profiling results **Click** on the structural alert Aldehydes
2. **Click** Details

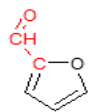
# Profiling

## Protein binding by OASIS v.1.2 of target chemical

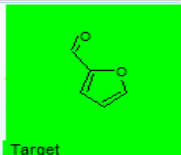
Protein binding by OASIS v.1.2

### Aldehydes

Target



Target







Boundaries: Training set Options

Boundary Options: Metabolism

Fragment

```
<m:c1(C{H}=O)c{H}cccc{H}1;m><m:c1(C{H}=O)c(O{H})c{H}c(C)cc1O{H};m><m:c1(C{H}=O)c{H}c{~1}{@}c{~1}{@}c{~1}{@}c{H}1*~@
```

Common Fragments

Definition	1	2
1 He1		
2 RP-V <sub>1</sub>		

Profile Description

**Mechanistic Domain:** Schiff base formation  
**Mechanistic Alert:** Schiff base formation with carbonyl compounds  
**Structural Alert:** Aldehydes

Profile Comments

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - **Endpoints**

# Endpoints

## Extracting endpoint values

The screenshot shows the QSAR Toolbox interface. The top navigation bar includes buttons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The 'Endpoint' button is highlighted with a callout box containing the number '2'. Below this, the 'Data' menu is open, showing options like Gather, Import, IUCLID5, Export, IUCLID5, Database, Inventory, and Database. The 'Gather' button is highlighted with a callout box containing the number '1'. The main window displays a list of endpoints for a target chemical structure (1 [target]). The endpoints listed include: Acute aquatic toxicity MOA by..., Aquatic toxicity classification..., Bioaccumulation – metabolism..., Bioaccumulation – metabolism..., Biodegradation fragments (Bio..., Carcinogenicity (genotox and n..., DNA alerts for AMES, MN and..., Eye irritation/corrosion Exclusi..., Eye irritation/corrosion Inclusio..., in vitro mutagenicity (Ames tes..., in vivo mutagenicity (Micronucl..., Keratinocyte gene expression, Oncologic Primary Classification, and Protein. The endpoint 'No alert found' is highlighted in blue. A bottom callout box contains the instructions: '1. Select all databases 2. Click Gather'.



# Endpoints

## Process of collecting data

Toxicity information on the target chemical is electronically collected from the selected datasets.

A window with "Read data?" appears. Now the user could choose to collect "all" or "endpoint specific" data.

The screenshot displays the QSAR Toolbox software interface. The main window is titled "Endpoint" and shows a workflow with steps: Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The "Endpoint" step is active, showing a list of endpoints for a target chemical. A "Read data?" dialog box is overlaid on the interface, with a blue callout box containing the number "1" pointing to the "OK" button. The dialog box has radio buttons for "All endpoints" (selected) and "Choose...", and a checked checkbox for "from Tautomers". The "OK" button is highlighted with a green checkmark.

The background interface shows a list of endpoints on the right side, including:

- Acute aquatic toxicity MOA by... Aldehydes
- Aquatic toxicity classification ... Aldehydes (Mon...
- Biodegradation fragments (Bio... Aldehyde [-CHO]
- Aromatic-H
- Carcinogenicity (genotox and n... Simple aldehyde (Genotox)
- DNA alerts for AMES, MN and... Structural alert for genotoxic carcin...
- Eye irritation/corrosion Exclusi... No alert found
- Eye irritation/corrosion Inclusio... (Undefined)Group All Lipid Solubilit...
- in vitro mutagenicity (Ames tes... Inclusion rules not met
- in vivo mutagenicity (Micronucl... Simple aldehyde
- Keratinocyte gene expression H-acceptor-path3-H-acceptor
- Oncologic Primary Classification Simple aldehyde
- Protein binding alerts for skin s... Moderate gene expression
- Schiff base formation Moderate gene expression >> alph...
- Schiff base formation >> Schiff bas... Aldehyde Type Compounds
- Schiff base formation >> Schiff bas...

**1. Click OK to read all available data**







# Endpoints

## Read data for analogues

Due to the overlap between the Toolbox databases same data for intersecting chemicals is found simultaneously in more than one database. The data redundancy is identified and the user has the opportunity to select either a single data value or all data values.

Repeated values for: 208 data-points, 48 groups, 1 chemicals

Data points...

	CAS	Structure	Value	additional_co...	Administration peri...	Any other informa
<input checked="" type="checkbox"/>	98-01-1		Positive			
<input checked="" type="checkbox"/>	98-01-1		Positive			
<input checked="" type="checkbox"/>	98-01-1		45 mg/kg/day		91	
<input checked="" type="checkbox"/>	98-01-1		45 mg/kg/day		91	
<input checked="" type="checkbox"/>	98-01-1		90 mg/kg/day		91	
<input checked="" type="checkbox"/>	98-01-1		90 mg/kg/day		91	

1. Click Select one 2. Click OK

Buttons: Select one, Invert, Check All, Uncheck All, OK, Cancel

# Endpoints

## Inserting data for target in data matrix

The screenshot shows the QSAR Toolbox interface. The top toolbar has several icons, with 'Category Definition' highlighted by a red box. A callout bubble with the number '1' points to this icon. The main window is divided into several panes. On the left, there are 'Databases' and 'Inventories' sections. The central pane shows a 'Filter endpoint tree...' with a search for '1 [target]'. Below this, there is a 'Structure' section with a chemical structure of 2-furaldehyde. The right pane shows a table of chemical properties for 2-furaldehyde. A red circle highlights the 'Physical Chemical Properties' row in the table.

Property	Value
Substance Identity	
— CAS Number	98-01-1
— Chemical IDs	Einecs Number:2026277
— Chemical Name	2-furaldehyde furfural 2-furancarboxaldehyde furfural furan-2-aldehyde
— Structural Formula	O=CC1=CC=CO1
Physical Chemical Properties (1/10)	M: 162 °C, 0.41, -38.1 °C, 2.21 m...
Environmental Fate and Trans... (1/10)	M: 93.5 %, 0.382 Pa-m3/mole, 3.5...
Ecotoxicological Information (1/218)	M: 20.5 mg/L, 10.5 mg/L, 145 mg/...
Human Health Hazards (1/518)	M: Negative, Negative, Negative, N...
Profile	

Now the data is inserted into data matrix; 1. **Click** Category Definition

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - Endpoints
  - **Category definition**

# Category definition

## Target endpoint

- In this exercise we will build a QSAR model to estimate the following endpoint :

Ecotoxicological Information#Aquatic

Toxicity#Growth#IGC50#48h#Protozoa#Ciliophora#Ciliat  
ea#Tetrahymena pyriformis

- The initial search for analogues is based on structural similarity, of US EPA categorization

# Category definition

## Navigate to the target endpoint

The screenshot shows the QSAR Toolbox interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The toolbar below has 'Data', 'Import', 'Delete', and 'Tautomerize' sections. The left sidebar contains 'Databases' and 'Inventories' sections. The central tree view shows a search filter 'Tetra' applied. The tree structure is as follows:

- Tetra
  - Substance Identity
    - CAS Number
    - Chemical IDs
    - Chemical Name
    - Structural Formula
  - Ecotoxicological Information
    - Aquatic Toxicity
      - Growth
        - IGC50
          - 48 h
            - Protozoa
              - Ciliophora
                - Ciliatea
                  - Tetrahymena pyriformis... (1/1) M: 145 mg/L

The right-hand data table shows the following information for the selected endpoint:

| Structure | Substance Identity               | Chemical Name  | Structural Formula |
|-----------|----------------------------------|--|--------------------|
|           | 98-01-1<br>Einecs Number:2026277 | 2-furaldehyde<br>fufural<br>2-furancarboxaldehyde<br>fufural<br>furan-2-aldehyde | O=CC=CC=CO1        |

1. **Type** "Tetra" in the empty filter field; 2. **Open** the nodes to target endpoint; 3. **Highlight** the cell that will be filled in (in this case we will reproduce the observed data).

# Category definition

## Defining US-EPA category

- The initial search for analogues is based on structural similarity, of US EPA categorization
- **Select** US-EPA category
- **Click** Define (see next screen shot)

# Category definition

## Defining US-EPA category

The screenshot illustrates the steps for defining a category in the QSAR Toolbox. The 'Define' button in the toolbar is circled in red (2). In the 'Predefined methods' list, 'US-EPA New Chemical Categories' is highlighted (1). The 'US-EPA New Chemical Categories' dialog box is shown, with 'Aldehydes (Acute toxicity)' selected as the target profile (3). The 'Strict' checkbox is checked (4).

**1. Highlight** "US-EPA New Chemical Categories"; **2. Click** Define; **3. Select** Strict (see next screen shot); **4. Click** OK to confirm the category **Aldehydes (Acute toxicity)** Defined from US-EPA category.



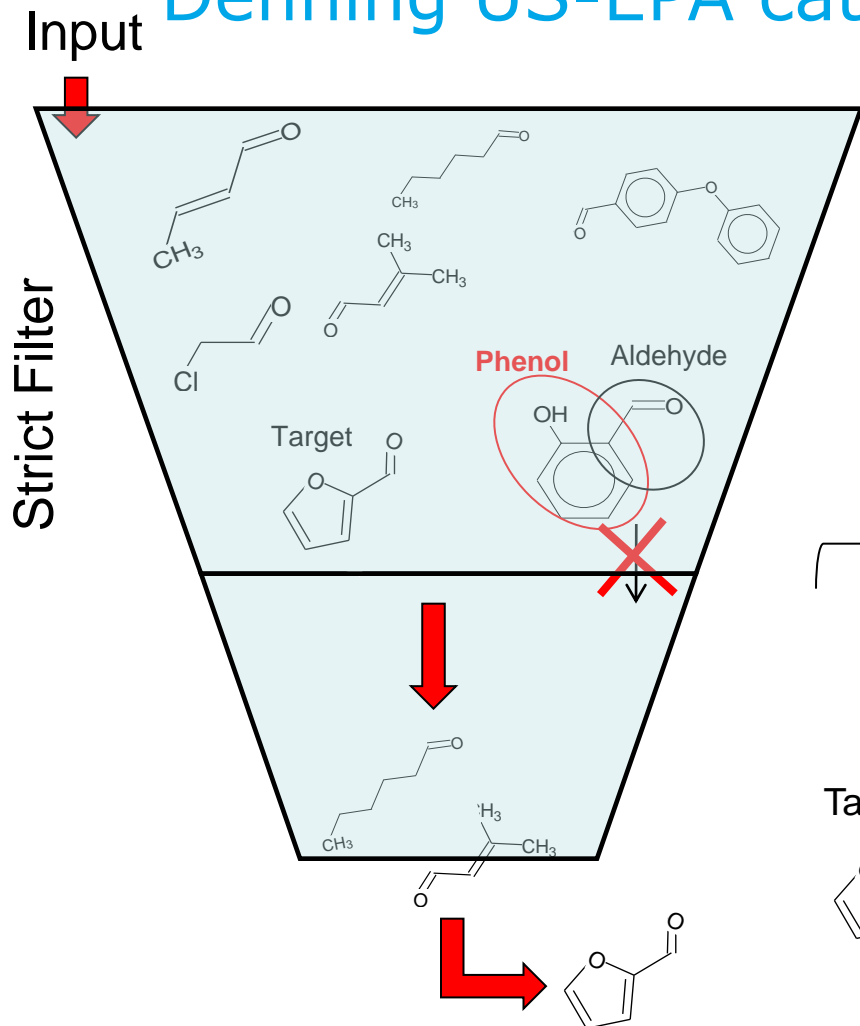
## Category definition

### Defining US-EPA category strict functionality

- The **Strict** functionality means that the software will group analogues having **ONLY** the categories of the target and will exclude the analogues having any other categories according to the profiler used in the grouping method.
- For example, if the profiling for the target results in *Aldehydes(Acute toxicity)* **ONLY** according to US-EPA category, the group of analogues will include *Aldehydes(Acute toxicity)* **ONLY**. (See next screen shot)

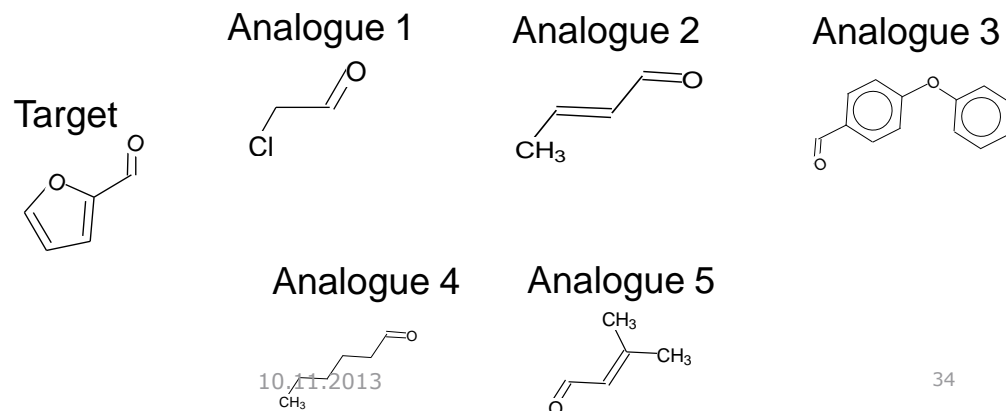
# Category definition

## Defining US-EPA category strict functionality



The target among with analogues have *Aldehydes* **ONLY** according to US-EPA category

### Defined Category



# Category definition

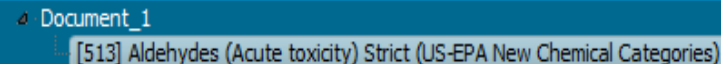
## Defining US-EPA category

The screenshot displays the QSAR Toolbox software interface. The main window shows a tree view of chemical properties for a selected chemical, identified as '2-furaldehyde' (CAS Number: 98-01-1, EINECS Number: 2026277). The tree view includes sections for Substance Identity, Environmental Fate and Transport, and Ecotoxicological Information. A dialog box titled 'Define category name' is overlaid on the screen, with the text 'toxicity) Strict (US-EPA New Chemical Categories)' entered in the 'Category name' field. A red circle with the number '1' points to the 'OK' button in the dialog box. A blue box at the bottom of the image contains the text: '1. Click OK to confirm the name of the category'.

## Category definition

### Analogues

- The Toolbox now identifies all chemicals corresponding to *Aldehydes(Acute toxicity)* by US-EPA listed in the databases selected under “Endpoints”.
- 513 analogues including the target chemical are identified; they form a mechanistic category “**Aldehydes (Acute toxicity)**”, which will be used for gap filling.
- The name of the analogues and name of the category appear in the “Defined Categories” window.

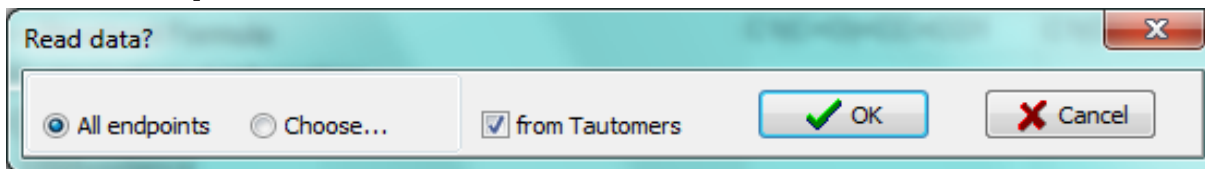


Document\_1  
[513] Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories)

## Category definition

### Reading data for Analogues

- The Toolbox will now retrieve those chemicals that have the same structural alert as the target
- The Toolbox automatically request the user to select the endpoint that should be retrieved
- The user can either select the specific endpoint or by default choose to retrieve data on all endpoints (see bellow)



# Category definition

## Reading data for Analogues

Due to the overlap between the Toolbox databases same data for intersecting chemicals is found simultaneously in more than one database. The data redundancy is identified and the user has the opportunity to select either a single data value or all data values.

Repeated values for: 2348 data-points, 565 groups, 423 chemicals

Data points...

|                                     | Endpoint                | CAS      | Structure                           | Value                          | Abnormality | Action |
|-------------------------------------|-------------------------|----------|-------------------------------------|--------------------------------|-------------|--------|
| <input checked="" type="checkbox"/> | LC50                    | 107-02-8 | <chem>CC(F)=O</chem>                | 27(24;30) micrograms per liter |             | 1      |
| <input checked="" type="checkbox"/> | LC50                    | 107-02-8 | <chem>CC(F)=O</chem>                | 27(24;30) micrograms per liter |             |        |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 50-00-0  | <chem>H2C=O</chem>                  | Positive                       |             | 2      |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 50-00-0  | <chem>H2C=O</chem>                  | Positive                       |             |        |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 75-07-0  | <chem>CC=O</chem>                   | Positive                       |             |        |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 75-07-0  | <chem>CC=O</chem>                   | Positive                       |             |        |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 78-84-2  | <chem>CC1=CC=C(C=C1)C(F)(F)F</chem> | Negative                       |             |        |
| <input checked="" type="checkbox"/> | Summary carcinogenicity | 78-84-2  | <chem>CC1=CC=C(C=C1)C(F)(F)F</chem> | Negative                       |             |        |

Buttons: Select one, Invert, Check All, Uncheck All, OK, Cancel

**1. Click Select one; 2. Click OK**

# Category definition

## Summary information for Analogues

The experimental results for the analogues are inserted into the matrix

The screenshot shows the QSAR Toolbox software interface. The main window displays a matrix of experimental results for analogues. The interface includes a menu bar with 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. Below the menu is a toolbar with 'Define', 'Subcategorize', 'Combine', 'Clustering', 'Delete', and 'Delete All'. The main window displays a tree view of 'tetra' with various subcategories like 'Substance Identity', 'Ecotoxicological Information', and 'Human Health Hazards'. A red box highlights a row in the matrix for 'Tetrahymena pyriformis' with values: (72/73), M: 937 mg/L, M: 296 mg/L, M: 247 mg/L, M: 235 mg/L, M: 216.

| Structure                      | 2           | 3           | 4           | 5           | 6      |
|--------------------------------|-------------|-------------|-------------|-------------|--------|
| Structure                      |             |             |             |             |        |
| Substance Identity             |             |             |             |             |        |
| Ecotoxicological Information   |             |             |             |             |        |
| Aquatic Toxicity               |             |             |             |             |        |
| Avoidance (1/2)                |             |             |             |             |        |
| Growth                         |             |             |             |             |        |
| EC50 (3/6)                     |             |             |             |             |        |
| IGC50                          |             |             |             |             |        |
| 48 h                           |             |             |             |             |        |
| Protozoa                       |             |             |             |             |        |
| Ciliophora                     |             |             |             |             |        |
| Amoebozoa                      |             |             |             |             |        |
| Tetrahymena pyriformis (72/73) | M: 937 mg/L | M: 296 mg/L | M: 247 mg/L | M: 235 mg/L | M: 216 |
| Growth Inhibition (2/4)        |             |             |             |             |        |
| Immobilisation                 |             |             |             |             |        |
| Population (19/42)             |             |             |             |             |        |
| Human Health Hazards (1/1)     |             |             |             |             |        |

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category definition
  - **Data gap filling**



# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Apply Trend analysis

The screenshot shows the QSAR Toolbox interface. The top navigation bar includes icons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The left sidebar shows 'Data Gap Filling Method' with options: Read-across, Trend analysis, and Q models. The main area displays a tree view of 'tetra' with 'Tetrahymena pyriformis' selected under 'IGC50' and '48 h'. A table shows chemical structures and their corresponding data points. A callout box '1' points to the 'Tetrahymena pyriformis' entry in the tree, and another callout box '2' points to the 'Trend analysis' option in the sidebar. A third callout box '3' points to the 'Apply' button in the top left.

**1. Highlight** the Data gap corresponding to *Tetrahymena pyriformis* IGC50 under the target chemical; **2. Select** Trend analysis; **3. Click** Apply

# Data Gap Filling (IGC 50 48h of *T. pyriformis*)

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

The OECD QSAR Toolbox for Grouping Chemicals into Categories  
Developed by LMC, Bulgaria

**Data Gap Filling Method**

- Read-across
- Trend analysis
- (Q)SAR models

**Target Endpoint**

Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis

tetra [target] 4 10 13

Structure

Tetrahymena pyriformis (T2/T2) M: 145 mg/L M: 104 mg/L M: 2.18 mg/L M: 134 mg/L

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

**Trend analysis prediction of IGC50, making a linear approximation, based on 71 values from 71 analogue chemicals, Observed target value: 145 mg/L, Predicted target value: 101 mg/L,**

**Model equation:  $IGC50 = +2.65 + 0.395 * \log Kow$**

Descriptor X: log Kow

**Accept prediction**

Return to matrix

- Select/filter data
- Selection navigation
- Gap filling approach
- Descriptors/data
- Model/(Q)SAR
- Calculation options
- Visual options
- Information
- Miscellaneous

## Data Gap Filling (IGC 50 48h of *T. pyriformis*) Interpreting dots on the graph

- The resulting plot outlines the experimental results of all analogues (Y axis) according to a descriptor (X axis) with LogKow being the default descriptor (see next screen shot)
- The **RED** dot represents the predicted value for target chemical.
- The **BLUE** dots represent the experimental results available for the analogues
- The **GREEN** dots (see the following screen shots) represent analogues belonging to different subcategories

## Data Gap Filling (IGC 50 48h of *T. pyriformis*) An accurate analysis of data set

- In this example, the mechanistic properties of the analogues are consistent.
- Subcategorization can be performed based on protein binding mechanisms. This is the second stage of analogue search - requiring the same interaction mechanism.
- Acute effects are associated with covalent interaction of chemicals within cell proteins, i.e. with protein binding.
- Chemicals with a different protein binding mechanism/reactions compared to the target chemical will be removed.

## Data Gap Filling (IGC 50 48h of *T. pyriformis*)

### Subcategorisation by Protein binding by OASIS v.1.2

- **To improve the data by subcategorizing, the Protein binding by OASIS v.1.2 profiler is used:**
- **Click** on Select filter data then **click** Subcategorize
- **Select** Protein binding by OASIS v.1.2 from the Grouping methods list.
- All chemicals which have a potential protein binding mechanism different from the target chemical are **GREEN** coloured.
- **Click** on Remove (see next two screen shots).

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorisation by Protein binding by OASIS v.1.2

The screenshot displays the 'Subcategorization' window of the QSAR Toolbox. On the left, a list of grouping methods is shown, with 'Protein binding by OASIS v.1.2' selected (indicated by callout 3). The 'Target' section shows 'Schiff base formation' selected. The 'Differ from target by' section has 'All categories' selected. The 'Analogues' list shows several chemical structures with their corresponding IGC50 values: formis (72/73) M: 145 mg/L, M: 296 mg/L, M: 247 mg/L, M: 235 mg/L, and M: ... (indicated by callout 4). Below the list is a scatter plot showing the relationship between predicted and target values, with a regression line and data points (indicated by callout 2). On the right, the 'Accept prediction' and 'Return to matrix' sections are visible, with 'Select/filter data' selected (indicated by callout 1). A callout box at the bottom provides instructions: '1. Click Select filter data 2. Select Subcategorize; 3. Select Protein binding by OASIS v.1.2 4. Click Remove to eliminate dissimilar to the target chemicals'.

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Results after subcategorisation

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

The OECD QSAR Toolbox for Grouping Chemicals into Categories  
Developed by LMC, Bulgaria

**Data Gap Filling Method**

- Read-across
- Trend analysis
- (Q)SAR models

**Target Endpoint**

Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliatea Tetrahymena pyriformis

Structure

1 [target] 4 14 16

Tetrahymena pyriformis (72/72) M: 145 mg/L M: 104 mg/L M: 194 mg/L M: 152 mg/L

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

**Trend analysis prediction of IGC50, making a linear approximation, based on 30 values from 30 analogue chemicals, Observed target value: 145 mg/L, Predicted target value: 275 mg/L, Model equation:  $IGC50 = +2.10 + 0.539 * \log Kow$**

Descriptor X: log Kow

**Accept prediction**

**Return to matrix**

- Select/filter data
  - Subcategorize
  - Mark chemicals by WS
  - Mark chemicals by descriptor v...
  - Mark outlier points
  - Filter points by test conditions
  - Mark focused chemical
  - Mark focused points
  - Remove marked chemicals/points
  - Clear existing marks
- Selection navigation
  - Gap filling approach
  - Descriptors/data
  - Model/(Q)SAR

# Data Gap Filling

## (IGC 50 48h of *T. pyriformis*)

### An accurate trend analysis of data set

- The chemicals which differ from the target according to Protein binding by OASIS v1.2 are:
  - Michael addition << alpha, beta – unsaturated carbonyl compounds << alpha, beta-unsaturated aldehydes (20);
  - Michael addition << Michael addition on conjugated systems with electron withdrawing group << alpha, beta-Carbonyl compounds with polarized (2);
  - No alert found (17);
  - SNAr << Nucleophilic aromatic substitution on activated halogens << Activated aryl and heteroaryl compounds (1).
- Another way for refining the data set is to ask what makes the obvious outliers different from the target.
- **Click** on Selection navigation then, click Back (see next screen shot).



## Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorisation by using “Difference to target” functionality

- **Right-Click** on any of the outlying analogues colored in **BLUE**.
- **Select** Differences to target from the context menu. The profilers by which the analogues differ to the target are colored in **ORANGE**.
- **Select** Protein binding by OASIS v.1.2 from the Grouping methods list
- **Click** on Remove (see next three screen shots).

# Data Gap Filling (IGC 50 48h of *T. pyriformis*)

## Subcategorisation by using "Difference to target" functionality

**1. Click** Selection navigation; **2. Click** Go back; **3. Right click** above one of the outliers on the graph; **4. Select** Information from the context menu; **5. From** the newly appeared menu **Select** Difference to target **6. The** profilers coloured in orange are those by which the analogues differ to the target; **Go to the next screen shot**

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorisation by using "Difference to target" functionality

**1** Select Protein binding by OASIS v.1.2; 2. Click Remove to eliminate chemicals dissimilar (those noted in green) to the target.

| Category                     | 72          | 73           | 74         | 75    |
|------------------------------|-------------|--------------|------------|-------|
| <i>T. pyriformis</i> (72/72) | M: 3.2 mg/L | M: 16.2 mg/L | M: 14 mg/L | M: 22 |

Trend analysis prediction of IGC50, approximation, based on 71 values from 71 analogue chemicals, target value: 145 mg/L, Predicted target value: 101 mg/L, model equation:  $IGC50 = +2.65 + 0.395 * \log Kow$

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorisation by using "Difference to target" functionality

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

The OECD QSAR Toolbox for Grouping Chemicals into Categories  
Developed by LMC, Bulgaria

Filling Apply

Data Gap Filling Method

- Read-across
- Trend analysis**
- (Q)SAR models

Target Endpoint

Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis

Structure

|              |              |             |
|--------------|--------------|-------------|
| 100          | 102          | 103         |
|              |              |             |
| M: 17.3 mg/L | M: 9.51 mg/L | M: 937 mg/L |

Tetrahymena pyriformis (72/72)

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

Trend analysis prediction of IGC50, making a linear approximation, based on 30 values from 30 analogue chemicals, Observed target value: 145 mg/L, Predicted target value: 275 mg/L, Model equation:  $IGC50 = +2.10 + 0.539 * \log Kow$

Descriptor X: log Kow

Accept prediction  
Return to matrix

- Select/filter data
- Selection navigation
  - Go back
  - Go forward
  - Go to first
  - Go to last
- Gap filling approach
  - Descriptors/data
  - Model/(Q)SAR
  - Calculation options
  - Visual options
  - Information
  - Miscellaneous

## Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model

- To assess the model accuracy use:
  - Adequacy (predictions after leave-one-out)
  - Statistics
  - Cumulative frequency
  - Residuals
- See next four screen shots

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model

The screenshot displays the QSAR Toolbox interface. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The left sidebar shows the 'Data Gap Filling Method' with options for 'Read-across', 'Trend analysis', and '(Q)SAR models'. The main workspace shows a table of chemical structures and their predicted IGC50 values. A callout box with the number '1' points to the 'Adequacy' tab in the bottom navigation bar. The 'Adequacy of prediction' plot shows a scatter plot of predicted vs. observed IGC50 values with a regression line. The model statistics are:  $R^2 = 0.815$ ,  $R^2_{adj} = 0.809$ ,  $s = 0.300$ .

| Chemical Name       | IGC50 (pred.) | IGC50 (obs.) |
|---------------------|---------------|--------------|
| Tetrahydrofuran     | 17.3 mg/L     | 9.51 mg/L    |
| Tetrahydrothiophene | 9.51 mg/L     | 937 mg/L     |

**1. Click Adequacy**

# Data Gap Filling (IGC 50 48h of *T. pyriformis*)

## Evaluation of the model cumulative frequency

The screenshot displays the QSAR Toolbox interface. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Data Gap Filling' method is selected. The interface shows a table of chemical structures and their predicted values for *Tetrahymena pyriformis* (72/72). A callout box with the number '1' points to the 'Cumul. freq.' column in the table. Below the table, a cumulative frequency plot is shown, with the text '95% of Residuals =< 0.499, log(1/mol/L)'. The plot shows a cumulative frequency of 100% for the predicted values. The right sidebar contains options for 'Accept prediction', 'Return to matrix', and various data selection and calculation options.

| Structure                      | 100                   | 102                   | 103                         |
|--------------------------------|-----------------------|-----------------------|-----------------------------|
| <chem>CCCCCCCC</chem>          | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>C1=NC=C(C=C1)C</chem> |
| Tetrahymena pyriformis (72/72) | M: 17.3 mg/L          | M: 9.51 mg/L          | M: 937 mg/L                 |

**1. Click** Cumul.freq.; The residuals abs (obs-predicted) for 95% of analogues are comparable with the variation of experimental data.

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model statistics

The screenshot shows the QSAR Toolbox interface with the 'Data Gap Filling' method selected. The target endpoint is 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis'. The model used is '(Q)SAR models'. The chemical structure of Tetrahymena pyriformis is shown, along with its predicted toxicity values: 17.5 mg/L (highlighted in green), 9.51 mg/L (highlighted in green), and 937 mg/L (highlighted in blue). A blue callout box with the number '1' points to the 'Statistics' column in the table below.

| Descriptors   | Prediction | Adequacy | Cumul. freq. | Statistics    | Residuals |
|---|------------|----------|--------------|---------------|-----------|
| <b>Statistical characteristics</b>                  |            |          |              |               |           |
| TA model  |            |          |              |               |           |
| Number of data points, (N)                          |            |          |              | 30            |           |
| Coefficient of determination, (R2)                  |            |          |              | 0.815         |           |
| Adjusted coefficient of determination, (R2adj)      |            |          |              | 0.809         |           |
| Coefficient of determination - leave one out, (Q2)  |            |          |              | 0.794         |           |
| Coefficient of correlation for external set, (r2)   |            |          |              | -             |           |
| Sum of squared residuals, (SSR)                     |            |          |              | 2.52          |           |
| Standard deviation of residuals, (sN)               |            |          |              | -             |           |
| Sample standard deviation of residuals, (s)         |            |          |              | 0.300         |           |
| Fisher function, (F)                                |            |          |              | 124           |           |
| Fisher threshold for statistical significance, (Fa) |            |          |              | 5.97          |           |
| <b>b0</b>   |            |          |              |               |           |
| - model descriptor                                  |            |          |              | Intercept     |           |
| - coeff. value                                      |            |          |              | 2.10          |           |
| - coeff. range                                      |            |          |              | ± 0.24        |           |
| - significance                                      |            |          |              | Yes           |           |
| - max. covariation                                  |            |          |              | 0.234 (vs b1) |           |

**1. Click Statistics**



# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model statistics

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

The OECD QSAR Toolbox for Grouping Chemicals into Categories  
Developed by LMC, Bulgaria

Filling  
Apply

Data Gap Filling Method

- Read-across
- Trend analysis
- (Q)SAR models

Target Endpoint

Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis

tetra

|                                |              |              |             |
|--------------------------------|--------------|--------------|-------------|
| Structure                      | 100          | 102          | 103         |
|                                |              |              |             |
| Tetrahymena pyriformis (T2/T2) | M: 17.3 mg/L | M: 9.51 mg/L | M: 937 mg/L |

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

1. Click Residuals

Distribution of residuals for IGC50 vs descriptors in use

IGC50 (residuals), log(l/mol/L)

log Kow

Descriptor X: log Kow

Accept prediction  
Return to matrix

- Select/filter data
- Selection navigation
- Gap filling approach
- Descriptors/data
- Model/(Q)SAR
- Calculation options
- Visual options
- Information
- Miscellaneous

## Data Gap Filling (IGC 50 48h of *T. pyriformis*) Save the derived QSAR model

- To save the new regression model follow these steps:
  - **Click** on Model (Q)SAR
  - **Select** Save model
  - **Enter** the model name and fill editable fields if necessary
  - **Click** on OK and
  - **Accept** the value
  - **Click** on Return to the matrix (see next screen shot)

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Save the derived QSAR model

The screenshot shows the QSAR Toolbox 3.0.0.862 interface. The 'Data Gap Filling' tab is active. The main window displays a table with columns for chemical structures and their predicted values. The table includes the following data:

| Structure                        | 1 [target]  | 3           | 4            | 6            | 9     |
|----------------------------------|-------------|-------------|--------------|--------------|-------|
| <chem>C1=CC=C(C=C1)C(=O)O</chem> | M: 145 mg/L | M: 152 mg/L | M: 59.4 mg/L | M: 10.9 mg/L | M: 11 |

The 'Edit model - IGC50' dialog box is open, showing the following fields:

- Model name: IGC50 (editable field)
- Model version: (editable field)
- QMRf file: C:\Users\Ioanna\Documents\QSAR Toolbox\Ver 3.0\UserDir\IGC50.xml (Browse ...)
- generate XML QMRf file:
- 1.1. Model identifier: IGC50
- 1.2. Data gap filling approach: Trend analysis
- 1.3. Other related model: (editable field)

The right sidebar shows the 'Accept prediction' and 'Return to matrix' options. The 'Return to matrix' section includes options for 'Select/filter data' and 'Save model'.

**1. Click** Model (Q)SAR; **2. Select** Save model; **3. Type** Name of the model and fill fields if necessary; **4. Click** Save; **5. Click** Accept prediction; **6. Select** Return to the matrix

# Outlook

- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category definition
  - Data gap filling
    - **QSAR model**

# Data Gap Filling

## How to see the derived QSAR?

The screenshot displays the QSAR Toolbox interface during a Data Gap Filling operation. The 'Data Gap Filling Method' section on the left has '(Q)SAR models' selected. Under 'Relevant (Q)SAR model', 'IGC50 Tetrahymena' is highlighted with a red circle and callout 3. The main tree view shows 'Tetrahymena pyriformis' selected, with callout 2 pointing to the '(Q)SAR models' list. The data matrix table shows predicted values for 'M: 145 mg/L' and 'T: 275(63.3;1.19E3...', with callout 1 pointing to this prediction.

1. Note the accepted prediction is inserted into data matrix; 2. **Click** (Q)SAR models; 3. The derived QSAR is listed in the panel with Relevant (Q)SAR models.

# Data Gap Filling

## How to see the derived QSAR?

As seen in the next five screen shots the derived model can be used to:

- **Visualize training set of the model:**
  - **Right-click** on the QSAR model IGC50 48h *Tetrahymena pyriformis*; **Select** Display Training Set from the context menu;
- **Visualize the domain of the model:**
  - **Right-click** on the QSAR model IGC50 48h *Tetrahymena pyriformis*; **Select** Display Domain from the context menu;
- **Visualize whether a chemical is in the applicability domain of the model:**
  - In the data matrix **highlight** the empty cell of one of the analogues (e.g. chemical no 2 in the matrix) for the endpoint 48h IGC50 *Tetrahymena pyriformis*; **Right-click** on the QSAR model IGC50 48h *Tetrahymena pyriformis*; **Select** Display domain;
- **Edit QMRF data** – the user could change the data already saved in the QMRF form
- **Perform predictions for:**
  - All chemicals in the matrix.
  - Current chemical
  - Chemicals in domain:
    - **Right-click** on the QSAR model IGC50 48h *Tetrahymena pyriformis*; **Select** the desired option

# Data Gap Filling

## Visualisation of the training set

The screenshot shows the QSAR Toolbox interface. On the left, a menu is open with 'Display Training Set' selected, indicated by a blue callout box with the number '2'. Below the menu, the 'IGC50 Tetrahymena' model is highlighted with a red circle and a blue callout box with the number '1'. On the right, a window titled 'Training set of: IGC50 Tetrahymena' displays a grid of 15 chemical structures, each with its CAS number and IGC50 value.

| Chemical ID | CAS #     | IGC50 (mg/L) |
|-------------|-----------|--------------|
| 1           | 110-62-3  | 104          |
| 2           | 123-72-8  | 194          |
| 3           | 66-25-1   | 152          |
| 4           | 96-17-3   | 193          |
| 5           | 111-71-7  | 114          |
| 6           | 123-05-7  | 88.7         |
| 7           | 123-15-9  | 296          |
| 8           | 123-38-6  | 216          |
| 9           | 446-52-6  | 103          |
| 10          | 529-20-4  | 123          |
| 11          | 552-89-6  | 115          |
| 12          | 590-86-3  | 188          |
| 13          | 613-45-6  | 191          |
| 14          | 874-42-0  | 16           |
| 15          | 4460-86-0 | 247          |

1. **Right Click** on the derived QSAR model; 2. **Select** Display Training Set; 3. Note the experimental data is displayed under CAS # of each chemical

# Data Gap Filling

## Visualisation of model domain

The screenshot shows the QSAR Toolbox interface. On the left, a menu is open with 'Display Domain' circled in red and labeled with a blue callout box containing the number '2'. Below this, in the 'Relevant (Q)SAR models' section, 'IGC50 Tetrahymena' is circled in red and labeled with a blue callout box containing the number '1'. The main window displays a 'Domain Boundaries Browser' with a metabolic pathway diagram showing various metabolites (e.g., IGC50, 48 h, Protozoa, Ciliophora, Ciliata, Tetrahymena pyriformis) and their relationships. A table at the bottom right shows model statistics for 'Growth Inhibition' (2/4) and 'Immobilisation' (19/39). A blue box at the bottom of the screenshot contains the following instructions:

- 1. Right Click** on the derived QSAR model;
- 2. Select Display Domain** (see next screen shot)



# Data Gap Filling

## Visualisation of model domain

Target

In Domain

Boundaries Training set Options

2 1 3 3 4 5

1 9 6 7 8 10

AND AND AND NOT AND

Metabolism

Simulator  
Do not apply metabolism

Process

Parent  
 Metabolites  Use parent if none  
 All

Match

Any  
 All  
 Accumulatively

Ignore inorganic metabolites

1. Note the boundaries of the domain are combined logically; 2. If the chemical answer the query of the domain then the current query is a labelled **GREEN**; 3. otherwise is labelled **RED**.

# Data Gap Filling

## Visualisation of the training set of the model

The screenshot shows the 'Domain Boundaries Browser' software. On the left, the 'Target' section displays a chemical structure of furfural. The main area shows a hierarchical tree of nodes representing the training set, with nodes 1 through 10. Node 1 is highlighted with a red callout '1'. Node 2 is highlighted with a red callout '2'. Node 3 is highlighted with a red callout '3'. Below the tree is a list of chemicals with their CAS numbers and SMILES. The 'Data points' table is also visible, showing experimental data for various chemicals.

| #  | endpoint      | Value          | Original value | Strain | Organ | Effect     | DATA QUALITY |
|----|---------------|----------------|----------------|--------|-------|------------|--------------|
| 1  | LC50          | 0.00158 mol/L  | 0.00158 mol/L  |        |       | Mortality  |              |
| 2  | LC50          | 0.000204 mol/L | 0.000204 mol/L |        |       | Mortality  |              |
| 3  | LC50          | 0.000191 mol/L | 0.000191 mol/L |        |       | Mortality  |              |
| 4  | IGC50         | 0.00269 mol/L  | 0.00269 mol/L  |        |       | Growth     |              |
| 5  | pT            | 1.51E-5 mol/L  | 1.51E-5 mol/L  |        |       | Physiology |              |
| 6  | pT            | 0.00123 mol/L  | 0.00123 mol/L  |        |       | Physiology |              |
| 7  | EC50          | 25.8 mg/L      | 25.8 mg/L      |        |       | Mortality  | 2,2          |
| 8  | EC50          | 25.8 mg/L      | 25.8 mg/L      |        |       | Mortality  | 1            |
| 9  | Gene mutation | Negative (Gene | Negative (Gene | TA 100 |       |            |              |
| 10 | Gene mutation | Negative (Gene | Negative (Gene | TA 102 |       |            |              |

- 1. Click** Training set to see training set of the model; **2.** The training set is presented as a list of chemicals; Click above the chemical from the list and **3. Select** Display data to see all available data.

# Data Gap Filling

Visualisation whether a chemical is in the domain of the model

The screenshot shows the QSAR Toolbox interface. On the left, the 'Data Gap Filling Method' is set to 'Read-across'. Under 'Relevant (Q)SAR models', 'ECOSAR (US EPA)' and 'IGC50 Tetrahymena' are listed, with 'IGC50 Tetrahymena' circled in red and labeled '2'. A context menu is open over the 'IGC50 Tetrahymena' model, with 'Display Domain' selected and labeled '3'. The main window displays a data matrix with columns for target endpoints and rows for chemical analogues. The second column contains the chemical structure of formaldehyde (H<sub>2</sub>C=O) and is highlighted in blue, labeled '1'. Below the matrix, a hierarchical tree shows the model structure, including 'Tetrahymena pyriformis' and its associated endpoints like 'Growth Inhibition' and 'Immobilisation'.

**1. Highlight** the cell of one of the analogues (e.g., chemical # 2 in the data matrix; **2. Click** above the model; **3. Select** Display domain (see next screen shot).

# Data Gap Filling

Visualisation whether a chemical is in the domain of the model

The screenshot shows the QSAR Toolbox interface. On the left, the 'Data Gap Filling Method' is set to '(Q)SAR models'. The 'Target Endpoint' is 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis'. The 'Relevant (Q)SAR models' list includes 'EC50', 'IGC50', '48 h', 'Protozoa', 'Ciliophora', 'Ciliata', and 'Tetrahymena pyriformis'. The 'Display Domain' menu option is highlighted with callout 3. The data matrix table shows chemical structures and their corresponding model results. Callout 1 points to a highlighted cell in the matrix, and callout 2 points to the 'IGC50 Tetrahymena' model in the tree.

| Chemical                      | 3 | 4 | 5 | 6 |
|-------------------------------|---|---|---|---|
| <chem>O=C(O)O</chem>          |   |   |   |   |
| <chem>CCCCC=O</chem>          |   |   |   |   |
| <chem>CCCCC(=Cc1ccc...</chem> |   |   |   |   |
| <chem>COc1ccc...</chem>       |   |   |   |   |

1. **Highlight** the cell of one of the analogues (e.g., chemical # 5 in the data matrix; 2. **Click** above the model; 3. **Select** Display domain (see next screen shot).

## Data Gap Filling

### Visualisation whether a chemical is in the domain of the model

- The chemical is an aldehyde as required by US-EPA categorization group.
- It can react with protein by Schiff-base formation and does not belong to any of the eliminated mechanistic domains according to Protein binding by OASIS v.1.2:
  - Michael addition
  - No alert found
  - SNAr
- Another requirement is Log Kow to be  $\geq 0.3187$  and  $\leq 4.75$ .
- The second requirement is violated because the chemical is not protein binder and therefore it is outside of the applicability domain of the model (see next screen shot).

# Data Gap Filling

Visualisation whether a chemical is in the domain of the model

1. The target chemical is out of model domain due to be non protein binder

# Data Gap Filling

## Edit QMRF data

The screenshot shows the QSAR Toolbox interface. On the left, a tree view displays the model hierarchy for 'IGC50 Tetrahymena', which is circled in red. A context menu is open over the 'IGC50' node, with 'Predict' selected and its sub-menu 'Predict Chemicals in Domain' highlighted. Two callouts, labeled '1' and '2', point to the 'Predict' and 'Predict Chemicals in Domain' options respectively. The main window displays a table with chemical structures and their corresponding SMILES strings.

| Chemical Structure            | SMILES           |
|-------------------------------|------------------|
| <chem>O=C(O)O</chem>          | C=OOO            |
| <chem>CCCC=O</chem>           | CCCC=O           |
| <chem>CCCCC(=Cc1ccc...</chem> | CCCCC(=Cc1ccc... |
| <chem>COc1ccc...</chem>       | COc1ccc...       |

1. Right click above the model; 2. Select Predict Chemicals in Domain.

# Data Gap Filling

## Perform prediction

The screenshot displays the QSAR Toolbox software interface during a prediction process. The main window shows a tree view of 'Ecotoxicological Information' with 'Tetrahymena pyriformis' selected. An 'Information' dialog box is open, displaying 'Predicted 56 out of 130 chemicals' and an 'OK' button. A status bar at the bottom shows '112/130 IGC50 Tetrahymena: predicting chemical(s)' and '17/70'. A red box highlights the status bar area, and a blue callout box with the number '1' points to it. Another blue callout box with the number '2' points to the 'OK' button in the dialog box.

1. The process of applying the model is indicated by status bar on the bottom of the window; the message with number of predicted chemicals appears; 2. **Click** OK.



# Outlook




- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category definition
  - Data gap filling
    - QSAR model
  - **Export QSAR prediction**

## Export QSAR results

- The predictions for the chemicals in the matrix can be exported into text file.
- In the data tree **right-click** on Tetrahymena pyriformis (for the endpoint IGC50 48h for Tetrahymena pyriformis) and **select** Export from the context menu (see next three screen shots).

# Export QSAR results

The screenshot shows the QSAR Toolbox interface. On the left, a tree view under 'tetra' shows the endpoint 'Tetrahymena pyriformis' selected, highlighted with a red box and a callout '1'. A context menu is open over this selection, with 'Export' highlighted, and a blue callout with a '2' points to it. The main window displays chemical structures and SMILES strings for the selected endpoint.

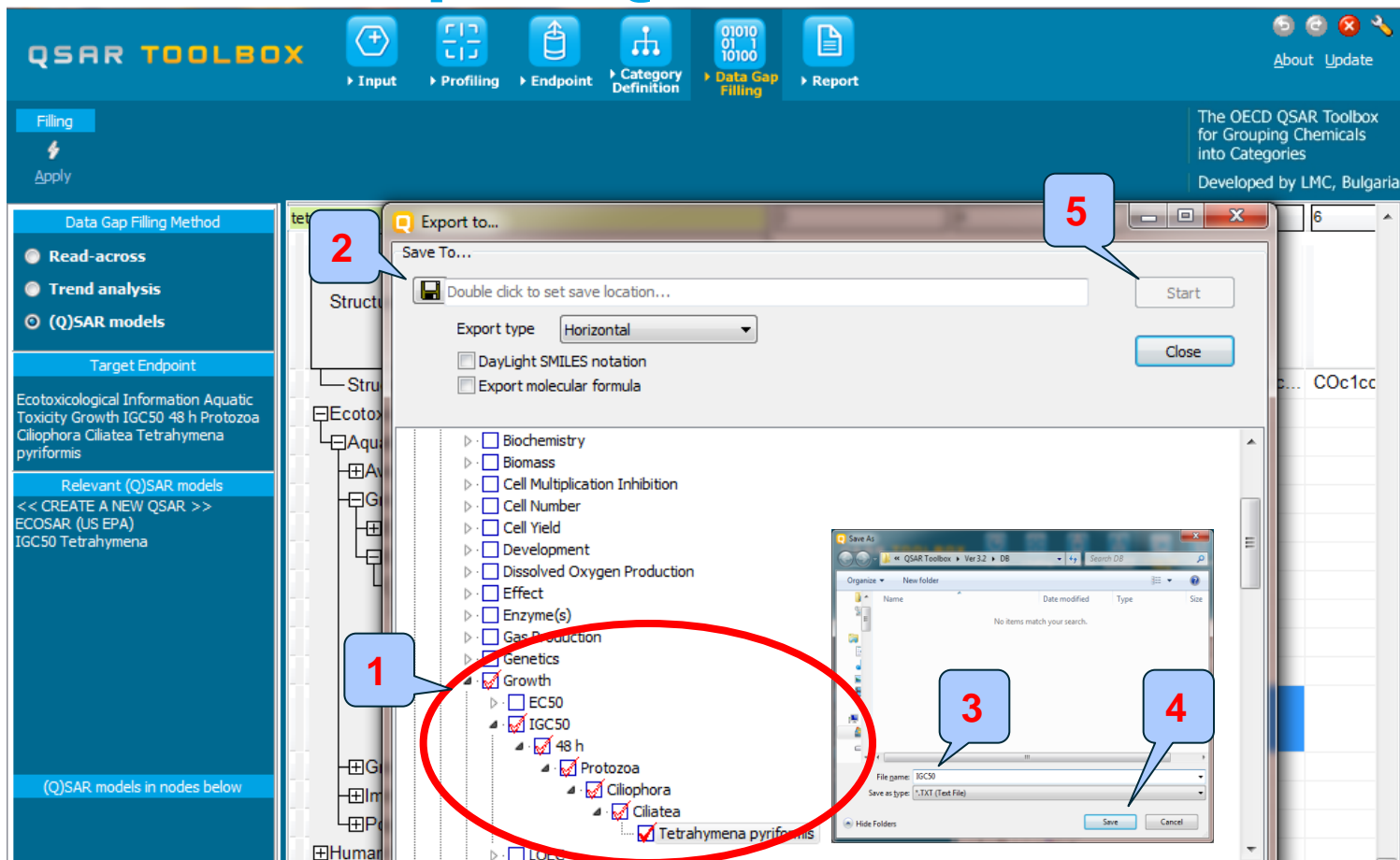
| 1 [target]  | 2        | 3   | 4   |
|---|----------|---|---|
|  | $H_2C=O$ |  |  |
| O=CC1=CC=CO1  | C=O      | C=OOO   | CCCC  |


Endpoint details for *Tetrahymena pyriformis* (264/317):

- M: 145 mg/L
- T: 275(63.3;1.19E3...
- Q: 156(35;694) mg...
- M: 104
- Q: 136(

1. **Right click** on the row of endpoint tree associated with predictions from the QSAR model; 2. **Select** Export (see next screen shot).

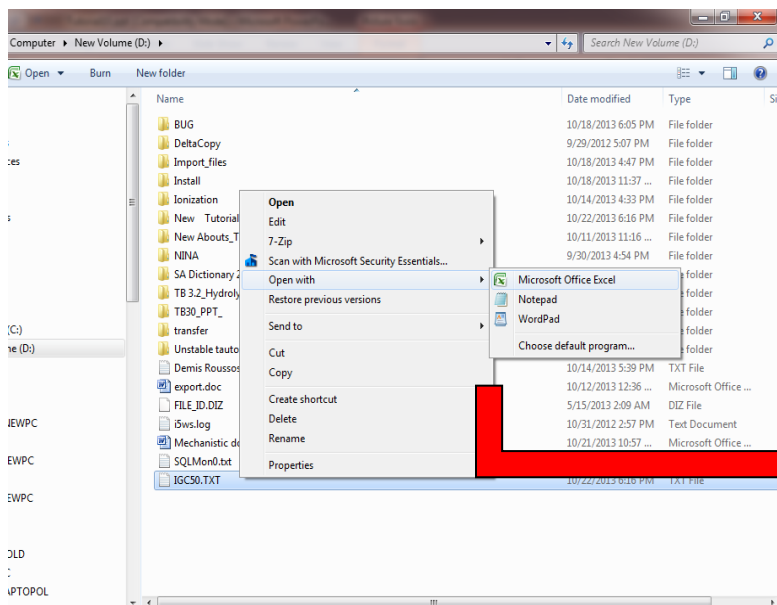
# Export QSAR results



1. The nodes from the tree associated with QSAR predictions which will be exported are labelled with **RED** check marks; 2. **Click**  to browse the folder on your PC; 3. Give name of the file; 4. **Click** Save; 5. **Click** Start; 6. **Click** OK when the file is exported.

# Export QSAR results

The resulting text file can be loaded into a spreadsheet and further analysed.



| CAS       | NAME      | SMILES              | Data       | Unit | Duration | Endpoint | Endpoint   | Endpoint   | Database  | QA (CAS-2) | Assigned    | Author      | Comment  | Data from |
|-----------|-----------|---------------------|------------|------|----------|----------|------------|------------|-----------|------------|-------------|-------------|----------|-----------|
| 98-01-1   | 2-furalde | O=CC1=CC            | 145        | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 98-01-1   | 2-furalde | O=CC1=CC275(63.3;1  | mg/L       | 48 h | IGCS0    |          | Ecotoxicol | Aquatic O  | High Qual | NO         | Schultz, T. | Impairme    | No       |           |
| 98-01-1   | 2-furalde | O=CC1=CC275(63.3;1  | mg/L       | 48 h | IGCS0    |          | Ecotoxicol | Aquatic O  | High Qual | NO         | Schultz, T. | Impairme    | No       |           |
| 98-01-1   | 2-furalde | O=CC1=CC275(63.3;1  | mg/L       | 48 h | IGCS0    |          | Ecotoxicol | Aquatic O  | High Qual | NO         | Schultz, T. | Impairme    | No       |           |
| 98-01-1   | 2-furalde | O=CC1=CC275(63.3;1  | mg/L       | 48 h | IGCS0    |          | Ecotoxicol | Aquatic O  | High Qual | NO         | Schultz, T. | Impairme    | No       |           |
| 50-00-0   | formalde  | C=O                 | 156(35;69  | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 50-00-0   | formalde  | C=O                 | 156(35;69  | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| *5-09-6   | peroxy    | C=OO                |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 110-62-3  | valeralde | CCCCC=O             | 104        | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 110-62-3  | valeralde | CCCCC=O             | 136(31.8;5 | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 110-62-3  | valeralde | CCCCC=O             | 136(31.8;5 | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 122-40-7  | alpha-am  | CCCCC(=C1cccc1)C=O  |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 123-11-5  | 4-methox  | CO1cccc(C=O)ccc1    |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 298-12-4  | glyoxylic | OC(=O)C=O           |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 3268-49-3 | 3-(methyl | CSCCC=O             | 502(113;2  | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 3268-49-3 | 3-(methyl | CSCCC=O             | 502(113;2  | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 4170-30-3 | crotonald | CC=CC=O             |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 107-02-8  | acrolein  | (;C=CC=O            | 2.18       | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 75-07-0   | ethanal;  | CC=O                |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 6287-38-3 | benzalde  | O=C1cccc(Cl)c(Cl)c1 |            |      |          |          |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 100-52-7  | benzalde  | O=C1cccc            | 134        | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |
| 123-72-8  | butyralde | CCCC=O              | 194        | mg/L | 48 h     | IGCS0    |            | Ecotoxicol | Aquatic O | High Qual  | NO          | Schultz, T. | Impairme | No        |

## Congratulations

- You have used the Toolbox to build a user-defined QSAR model.
- You now know another useful tool in the Toolbox.
- Continue to practice with this and other tools. Soon you will be comfortable dealing with many situations where the Toolbox is useful.