

## OECD QSAR Toolbox v.3.4

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

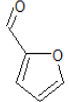
- **Remember the Toolbox has 6 modules which are used in a sequential workflow:**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category Definition
  - Filling Data Gaps
  - Report

# Outlook

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

# Chemical Input

The screenshot shows the QSAR Toolbox software interface. The 'Input' menu is highlighted with a red callout '1'. A search dialog box is open, showing '98011' entered in the 'Search by CAS #' field, with a red callout '2' pointing to the input field and another red callout '3' pointing to the 'Search' button. The dialog box also shows 'Tautomeric sets' as an unchecked option and 'OK' and 'Cancel' buttons. Below the dialog box, a table displays the search results for CAS # 98011-1, including columns for Selected, CAS, Smiles, Depiction, Names, and CAS/2D. The table shows a single entry for CAS # 98011-1 with a 'Yes' selection and a chemical structure depiction.

Selected	CAS	Smiles	Depiction	Names	ID/Name	CAS/2D
1. Yes	98-01-1	O=C=C1=				

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

# Chemical Input

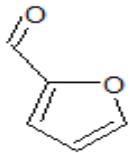
## Target chemical identity

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=			1:: High 1:: Av 2:: Ec 3:: Ec 4:: E	1:: High 1:: U 2:: Ec 3:: Av 4:: R	: High
				1: 2: 3: 4: 5: 6:	5:: Es 6:: G 7:: G 8:: H 9:: M 10:: F 11:: U 12:: U	5:: E 6:: U 7:: H 8:: M 9:: G 10:: E 11:: C 12:: E	

**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 navigation bar includes modules: Input, Profiling (highlighted with a red box), Endpoint, Category Definition, Data Gap Filling, and Report. The main window is divided into several sections:

- Documents:** A list of documents, currently showing 'Document\_1' with CAS: 98-01-1.
- Filter endpoint tree...:** A search filter set to '[target]'.
- Structure:** A chemical structure diagram of 2-furaldehyde.
- Substance Identity:** A table of identifiers and names for the chemical.
 

CAS Number	98-01-1
Chemical IDs	EINECS:2026277
Chemical Name	2-furaldehyde furfural 2-furancarboxaldeh...
Molecular Formula	C5H4O2
Structural Formula	O=CC1=CC=CO1
- Physical Chemical Properties, Environmental Fate and Transport, Ecotoxicological Information, Human Health Hazards:** These sections are currently collapsed.

At the bottom of the main window, there is a status bar showing '1 Document\_1' and '1/0/0'.

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

**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 software interface. The main window is titled "QSAR TOOLBOX" and has a menu bar with "Input", "Profiling", "Endpoint", "Category Definition", "Data Gap Filling", and "Report". The "Profiling" menu is open, showing "Apply", "New", "View", and "Delete" options. The "Profiling methods" section is visible, with "Predefined" and "General Mechanistic" categories. The "Filter endpoint tree..." window is open, showing a tree structure of endpoints. The "Profile" endpoint is highlighted with a red circle and a callout box containing the number "1". The "Structure" endpoint shows the chemical structure of furfural. The "Substance Identity" endpoint shows various identifiers for furfural, including CAS Number (98-01-1), EINECS (2026277), and Chemical Name (2-furaldehyde, furfural, 2-furancarboxalde..., furfural, furan-2-aldehyde, furfural (2-furaldehy...)). The "Molecular Formula" is C5H4O2 and the "Structural Formula" is O=CC1=CC=CO1.

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

# Profiling Profiles of "Furfural"

The screenshot shows the QSAR Toolbox interface with the 'Profiling' menu open. The 'Filter endpoint tree...' window displays a list of endpoints for the target '1 [target]'. The chemical structure of Furfural is shown at the top right. The endpoints list includes various physicochemical and toxicological parameters. A red circle highlights the 'Protein binding by OASIS v1.4' endpoint, which is associated with the value 'High (Class III)'. A callout box with the number '1' points to this endpoint.

Endpoint	Value
Hydrolysis half-life (Kb, pH 7)(Hydrowin)	Not calculated
Hydrolysis half-life (Kb, pH 8)(Hydrowin)	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 v1.4	High (Class III)
Protein binding by OECD	High (Class III)
Protein binding potency	No superfragment
Superfragments	No superfragment
Toxic hazard classification by Cramer (extension)	Moderately reactiv...
Toxic hazard classification by Cramer (original)	Moderately reactiv...
Ultimate biodeg	> 100 days 1 to 10 days

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).

1

# Profiling Profiles of "Furfural"

The screenshot displays the QSAR Toolbox interface. On the left, the 'Profiling methods' panel lists various methods, with 'Protein binding by OASIS v1.4' checked. The 'Filter endpoint tree...' panel shows a list of endpoints, with 'Protein binding by OASIS v1.4' selected. The 'Profiling results' panel shows a hierarchical tree of alerts, with 'Schiff base formation >> Schiff base formation with carbonyl compounds >> Aldehydes' highlighted in red. A callout box with the number '1' points to this highlighted alert. Another callout box with the number '2' points to the 'Details' button at the bottom of the results panel.

The Protein binding by OASIS v.1.4 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

# Outlook

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

# Endpoints

## Extracting endpoint values

The screenshot displays the QSAR Toolbox interface during the 'Endpoint' workflow. On the left, the 'Databases' panel shows four categories checked: Physical Chemical Properties, Environmental Fate and Transport, Ecotoxicological Information, and Human Health Hazards. A callout box with the number '1' points to these checked items. In the top toolbar, the 'Gather' button is highlighted with a callout box containing the number '2'. The main window shows a 'Filter endpoint tree...' with a search filter '[1 target]' and a list of endpoints including 'Acute Oral Toxicit...', 'Aquatic OASIS', 'Bacterial mutageni...', 'Biodegradation NITE', 'Carcinogenic Pote...', 'Carcinogenicity&m...', 'ECHA CHEM', 'ECOTOX', 'Estrogen Receptor...', 'Genotoxicity OASIS', 'GSH Experimental...', 'MUNRO non-canc...', 'Phys-chem EPISU...', 'Rodent Inhalation ...', 'ToxCastDB', 'Canada DSL', 'DSSTOX', 'ECHA PR', 'EINECS', 'HPVC OECD', 'METI Japan', 'NICNAS', 'REACH ECB', and 'TSCA'. A chemical structure is shown in the 'Structure' field.

**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 shows a tree view of endpoints under the heading "Filter endpoint tree...". A dialog box titled "Read data?" is overlaid on the interface. The dialog contains the following options:

- All endpoints
- Choose...
- from Tautomers
- OK
- Cancel

A red callout bubble with the number "1" points to the "OK" button. The background interface shows a list of endpoints including "Bacterial mutagen...", "Biodegradation NITE", "Carcinogenic Pote...", "Carcinogenicity&m...", "ECHA CHEM", "ECOTOX", "Estrogen Receptor...", "Genotoxicity OASIS", "GSH Experimental...", "MUNRO non-canc...", "Phys-chem EPISU...", "Rodent Inhalation ...", "ToxCastDB", "Canada DSL", "DSSTOX", "ECHA PR", "EINECS", "HPVC OECD", "METI Japan", "NICNAS", "REACH ECB", "TSCA", and "US HPV Challenge Program".

**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: 94 data-points, 36 groups, 1 chemicals

Data points...

	Endpoint	CAS	Structure	Value	additional_comme
<input checked="" type="checkbox"/>	LC50	98-01-1		3.2E4 micrograms per liter	
<input checked="" type="checkbox"/>	LC50	98-01-1		3.2E4 micrograms per liter	
<input checked="" type="checkbox"/>	NOAEL	98-01-1		100 mg/kg bw/day (nominal)	
<input checked="" type="checkbox"/>	NOAEL	98-01-1		100 mg/kg bw/day (nominal)	
<input checked="" type="checkbox"/>	gene mutation	98-01-1		negative	
<input checked="" type="checkbox"/>	gene mutation	98-01-1		negative	
<input checked="" type="checkbox"/>	gene mutation	98-01-1		negative	

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 menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Category Definition' button is highlighted with a red box. A blue callout bubble with the number '1' points to this button. The main window displays the chemical structure of 2-furaldehyde and its associated properties. The 'Environmental Fate and Transport' section is circled in red.

Property	Value
CAS Number	98-01-1
Chemical IDs	EINECS:2026277
Chemical Name	2-furaldehyde furfural 2-furancarboxaldehyde furfural furan-2-aldehyde
Molecular Formula	furfural (2-furaldehyde) (2-fur... C5H4O2
Structural Formula	C=C1=CC=CC1
Physical Chemical Properties	(1/10) M: 61.7 °C, 5.1E3 mg/L, 4.1...
Environmental Fate and Transport	(1/10) M: 93.5 %, 93.5 %, 100 %, ...
Ecotoxicological Information	(1/294) M: 20.5 mg/L, 10.5 mg/L, 14...
Human Health Hazards	(1/153) M: <50 mg/kg bw/day (nomi...

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 displays the QSAR Toolbox interface with the following components:

- Top Bar:** Navigation tabs for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report.
- Second Bar:** Action buttons for Data, Import, Export, Delete, and Tautomerize.
- Third Bar:** Sub-action buttons for Gather, Import, IUCLID5, Export, IUCLID5, Database, Inventory, and Database.
- Databases Panel:** Includes 'Select All', 'Unselect All', 'Invert', and 'About' buttons. A tree view shows categories like 'Physical Chemical Properties', 'Environmental Fate and Transport', 'Ecotoxicological Information', and 'Human Health Hazards' (checked).
- Inventories Panel:** Includes 'Select All', 'Unselect All', 'Invert', and 'About' buttons. A list of inventories is shown, including 'Canada DSL', 'COSING', 'DSSTOX', 'ECHA PR', and 'EINECS'.
- Main Workspace:**
  - Filter:** '1 [target]'
  - Structure:** Chemical structure of furfural.
  - Tree View:**
    - Substance Identity
      - CAS Number: 98-01-1
      - Chemical IDs: EINECS:2026277
      - Chemical Name: 2-furaldehyde, furfural, 2-furancarboxaldehyde, fufural, furan-2-aldehyde, furfural (2-furaldehyde) (2-fur...
      - Molecular Formula: C5H4O2
      - Structural Formula: O=CC1=CC=CO1
    - Ecotoxicological Information
      - Aquatic Toxicity
        - Growth
          - 48 h
            - Protozoa
              - Ciliophora
                - Ciliata
                  - Tetrahymena pyriformis (1/1) M: 145 mg/L

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 'Category Definition' process in the QSAR Toolbox. The left sidebar shows the 'Grouping methods' tree, where 'US-EPA New Chemical Categories' is highlighted (1). The 'Define' button in the toolbar is circled in red (2). The 'Target(s) profiles' dialog box is open, showing 'Aldehydes (Acute toxicity)' selected. The 'Strict' checkbox is checked (3), and the 'OK' button is highlighted (4). The main workspace displays the chemical structure of Tetra (SMILES: O=C1=CC=CO1) and its associated data, including EINECS:2026277 and CAS:98-01-1.

**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

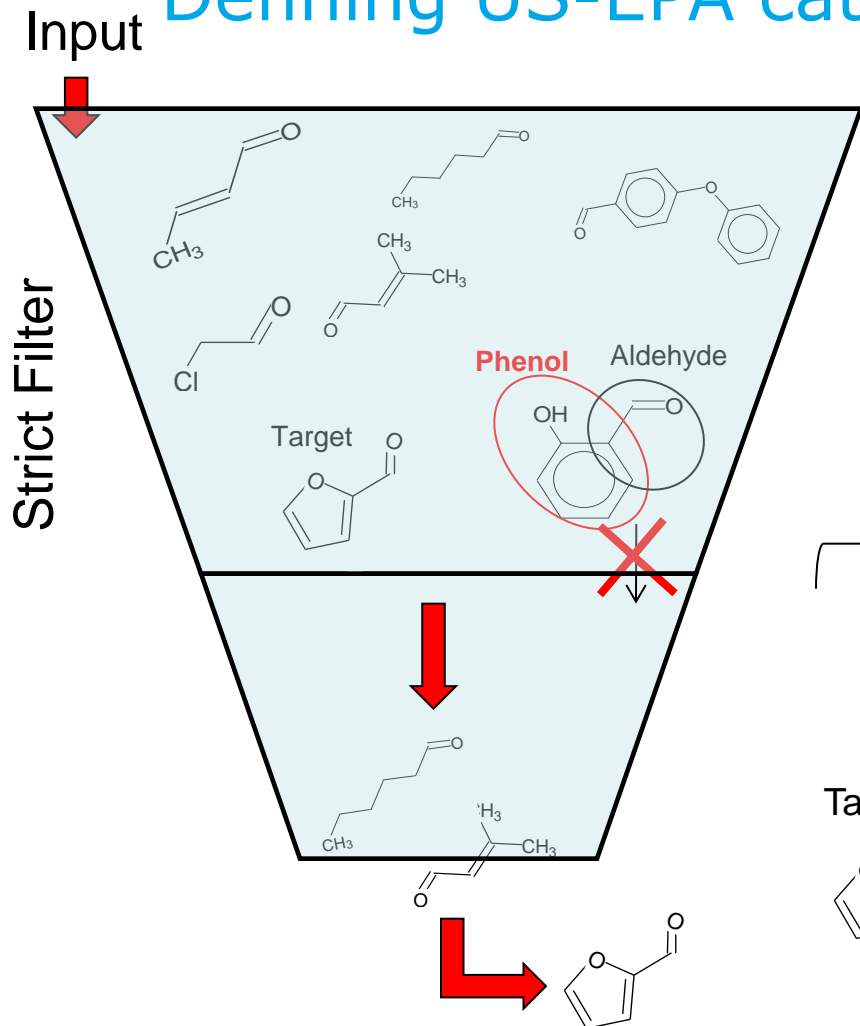
### Defining US-EPA category strict functionality

- 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)



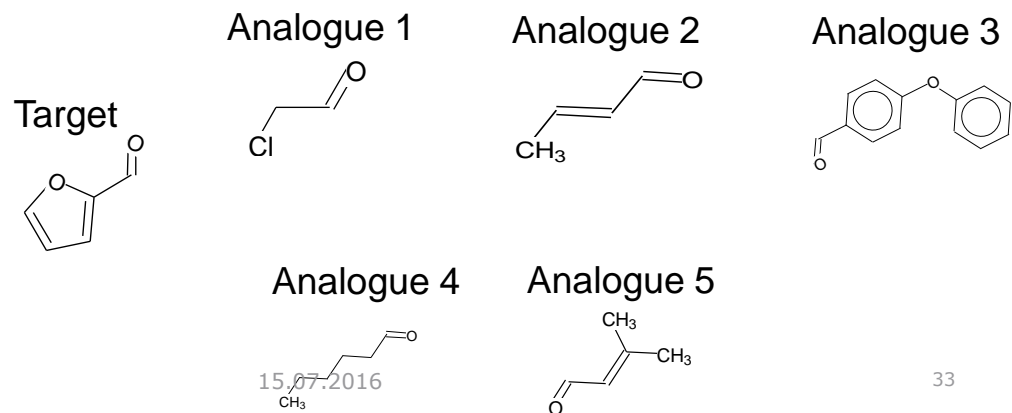
# Category definition

## 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

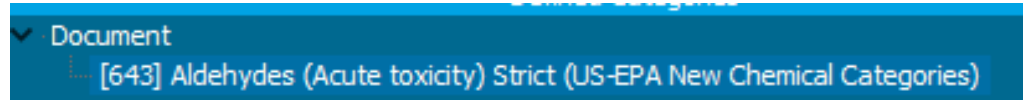
The screenshot displays the QSAR Toolbox software interface during the 'Category Definition' step. The main window shows a tree view of chemical properties for a substance named 'Tetra'. The 'Ecotoxicological Information' section is expanded, showing 'Aquatic Toxicity' with sub-categories like 'Growth', 'GC50', and 'Immobilisation'. A dialog box titled 'Define category name' is open, showing the category name 'toxicity Strict (US-EPA New Chemical Categories)' and an 'OK' button highlighted with a red circle and the number '1'. The background shows a tree view of chemical properties for 'Tetra' and a chemical structure of 2-furaldehyde.

**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”.
- 643 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.

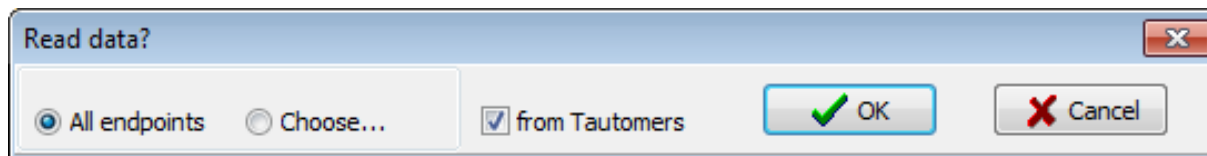


Document  
[643] Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories)

# 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)







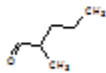
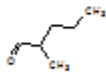
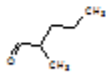

# Category definition

## 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: 3896 data-points, 1199 groups, 973 chemicals

Data points...

|                                     | Endpoint | CAS      | Structure  | Value   | Abnormality |
|-------------------------------------|----------|----------|--|---------|-------------|
| <input checked="" type="checkbox"/> |          | 123-11-5 |    | 0 °C    |             |
| <input checked="" type="checkbox"/> |          | 123-11-5 |    | 0 °C    |             |
| <input checked="" type="checkbox"/> |          | 100-52-7 |   | -26 °C  |             |
| <input checked="" type="checkbox"/> |          | 100-52-7 |   | -26 °C  |             |
| <input checked="" type="checkbox"/> |          | 123-15-9 |  | -100 °C |             |
| <input checked="" type="checkbox"/> |          | 123-15-9 |  | -100 °C |             |
| <input checked="" type="checkbox"/> |          | 123-15-9 |  | -100 °C |             |
| <input checked="" type="checkbox"/> |          | 104-55-2 |  | -7.5 °C |             |

1

2

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 displays the QSAR Toolbox interface with the 'Category Definition' tab active. The main window shows a matrix of chemical analogues. A red box highlights a row of experimental results for *Tetrahymena pyriformis* (72/72) at a concentration of 145 mg/L. The results are as follows:

| Analogue                              | 1 (target)  | 2 | 3           | 4            | 5 | 6            | 7 | 8 |
|---------------------------------------|-------------|---|-------------|--------------|---|--------------|---|---|
| <i>Tetrahymena pyriformis</i> (72/72) | M: 145 mg/L |   | M: 152 mg/L | M: 59.4 mg/L |   | M: 10.9 mg/L |   |   |

# 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*)

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

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

Structure

Tetra

Tetrahymena pyriformis... (72/72) M: 145 mg/L M: 152 mg/L M: 59.4 mg/L M: 10.9 mg/L M: 114 mg/L M: 88.7 mg/L M: 194 mg/L M: 2.18 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

3 Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories) Create prediction by gap filling 0/1 1/10

## 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 Acute aquatic toxicity MOA by OASIS

- After the available data has been retrieved, the user can then further subcategorize the results according to the following endpoint-specific subcategorizations:
  - Acute aquatic toxicity MOA by OASIS
  - Protein binding by OASIS v1.4
  - Aquatic toxicity classification by ECOSAR
- These steps are summarized in the next screen shots.

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorization 1: Acute aquatic toxicity MOA by OASIS

The screenshot displays the QSAR Toolbox interface for subcategorization. The main window shows a list of chemical structures and their predicted IGC50 values. A trend analysis plot is visible, showing a linear relationship between log Kow and IGC50. The model equation is  $IGC50 = +2.65 + 0.395 * \log Kow$ . The observed target value is 145 mg/L, and the predicted target value is 101 mg/L. The 'Return to matrix' panel on the right contains the following options:

- Accept prediction
- Return to matrix
- Select/filter data
  - Subcategorize
  - Mark chemicals by WS
  - Mark chemicals by descriptor value
  - Mark outlier points
  - Filter points by test conditions
  - Mark focused chemical
  - Mark focused points
  - Invert existing marks
  - Remove marked chemicals/points
  - Clear existing marks
- Selection navigation
- Gap filling approach

The 'Grouping methods' list on the left includes:

- Hydrolysis half-life (Kb, pH 8)(Hydrowin)
- Hydrolysis half-life (pH 6.5-7.4)
- Ionization at pH = 1
- Ionization at pH = 4
- Ionization at pH = 7.4
- Ionization at pH = 9
- Protein binding by OASIS v1.4
- Protein binding by OECD
- Protein binding potency
- Superfragments
- Toxic hazard classification by Cr
- Toxic hazard classification by Cr
- Ultimate biodeg
- Endpoint Specific
  - Acute aquatic toxicity classification by Verhaar (Modified)
  - Acute aquatic toxicity MOA by OASIS
  - Aquatic toxicity classification by ECOSAR
  - Bioaccumulation - metabolism alerts
  - Bioaccumulation - metabolism half-lives
  - Biodegradation fragments (BioWIN MITI)
  - Carcinogenicity (genotox and nongenotox) alerts by ISS
  - DART scheme v.1.0
  - DNA alerts for AMES by OASIS v.1.4
  - DNA alerts for CA and MNT by OASIS v.1.1
  - Eye irritation/corrosion Exclusion rules by BfR
  - Eye irritation/corrosion Inclusion rules by BfR
  - in vitro mutagenicity (Ames test) alerts by ISS
  - in vivo mutagenicity (Micronucleus) alerts by ISS
  - Keratinocyte gene expression
  - Oncologic Primary Classification
- Metabolism/Transformations
  - Do not account metabolism
  - Documented
    - Observed Mammalian metabolism
    - Observed Microbial metabolism
    - Observed Rat In vivo metabolism
    - Observed Rat Liver S9 metabolism
  - Simulated
    - Autoxidation simulator
    - Autoxidation simulator (alkaline medium)
    - Dissociation simulation
    - Hydrolysis simulator (acidic)
    - Hydrolysis simulator (basic)
    - Hydrolysis simulator (neutral)
    - in vivo Rat metabolism simulator
    - Microbial metabolism simulator

The 'Selected 4 (67/71)' panel at the bottom shows the 'Remove' button being highlighted.

1. Click Select filter data
2. Select Subcategorize;
3. Select Acute aquatic toxicity MOA by OASIS
4. Click Remove to eliminate dissimilar to the target chemicals

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorization 2: Protein binding by OASIS v1.4

**1**

**2**

**1. Select Protein binding by OASIS v1.4**  
**2. Click Remove to eliminate dissimilar to the target Chemicals.**

| Chemical Structure          | M: 145 mg/L           | M: 152 mg/L                 | M: 59.4 mg/L                | M: 10.9 mg/L          | M: 114 mg/L           | M: 88.7 mg/L          | M: 194 mg/L           | M: 134 mg/L           |
|-----------------------------|-----------------------|-----------------------------|-----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| <chem>C1=CC=C(C=C1)O</chem> | <chem>CCCCCCCC</chem> | <chem>C1=CC=C(C=C1)O</chem> | <chem>C1=CC=C(C=C1)O</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> |

Trend analysis prediction of IGC50, making a linear approximation, based on 67 values from 67 analogue chemicals, Observed target value: 145 mg/L, Predicted target value: 161 mg/L, Model equation:  $IGC50 = +2.37 + 0.488 * \log Kow$

# Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorization 3: Aquatic toxicity classification by ECOSAR

The screenshot displays the 'Subcategorization' window in the QSAR Toolbox. The 'Grouping methods' list on the left includes 'Acute aquatic toxicity classification by ECOSAR (Modified)', which is highlighted with a red callout box labeled '1'. The 'Target' dropdown is set to 'Aldehydes (Mono)'. The 'Data Gap Filling' section shows a table of predicted values for 48 chemicals, with the first value being 145 mg/L. The 'Trend analysis' section shows a scatter plot of observed vs. predicted values with a regression line, and a red callout box labeled '2' pointing to the 'Remove' button in the 'Selected 19 (29/48)' list.

| Chemical   | M: 145 mg/L | M: 152 mg/L | M: 59.4 mg/L | M: 114 mg/L | M: 88.7 mg/L | M: 194 mg/L | M: 193 mg/L | M: 104 mg/L |
|------------|-------------|-------------|--------------|-------------|--------------|-------------|-------------|-------------|
| 1 [target] | 3           | 4           | 9            | 10          | 12           | 27          | 29          |             |

**1. Select Aquatic toxicity classification by ECOSAR.**  
**2. Click Remove to eliminate dissimilar to the target chemicals.**

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

The screenshot displays the QSAR Toolbox interface for Data Gap Filling. The central plot shows a scatter plot of log(I<sub>GC50</sub>) vs log K<sub>ow</sub> with a red regression line. An information dialog box is open over the plot, stating 'The current prediction was accepted'. The right sidebar contains the 'Accept prediction' section with 'Return to matrix' selected. The top navigation bar shows 'Data Gap Filling' as the active tab.

1. **Click** Accept prediction
3. **Click** Return to matrix

2. **Click** OK



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

QSAR Toolbox 3.4.0.14 [Document\_1]

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

Structure

Tetra

1 [target] 3 4 9 10 12

M: 145 mg/L M: 152 mg/L M: 59.4 mg/L M: 114 mg/L M: 88.7 mg/L M: 194 mg/L

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

**Trend analysis prediction of IGC50, making a linear approximation, based on 29 values from 29 analogue chemicals, Observed target value: 145 mg/L, Predicted target value: 262 mg/L**

Model equation:  $IGC50 = +2.13 + 0.528 * \log Kow, \log(1/mol/L)$

Descriptor X: log Kow

**Accept prediction**

**Return to matrix**

- Select/filter data
  - Subcategorize
    - Mark chemicals by WS
    - Mark chemicals by descriptor value
    - Mark outlier points
  - Filter points by test conditions
    - Mark focused chemical
    - Mark focused points
- Selection navigation
  - Gap filling approach
  - Descriptors/data
  - Model/(Q)SAR
  - Calculation options
  - Visual options
  - Information
  - Miscellaneous

643 Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories) Create prediction by gap filling 0/100 1/10

## 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 during the 'Data Gap Filling' process. The 'Adequacy' tab is active, showing a scatter plot of predicted versus observed IGC50 values. A callout box with the number '1' highlights the 'Adequacy' tab label. The plot includes a regression line and model statistics:  $R^2 = 0.798$ ,  $R^2_{adj} = 0.791$ , and  $s = 0.303$ . The x-axis is labeled 'IGC50 (obs.), log(1 mol/L)' and the y-axis is 'IGC50 (pred.), log(1 mol/L)'. The interface also shows a list of chemical structures and their predicted values, and a sidebar with various data manipulation options.

**1. Click Adequacy**

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

## Evaluation of the model cumulative frequency

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 main workspace displays a list of chemical structures with their predicted values (M) in mg/L. A callout box with the number '1' points to the 'Cumul. freq.' tab in the bottom navigation bar. Below the list is a cumulative frequency plot titled '95% of Residuals = < 0.510, log(1/mol/L)'. The plot shows a step-like cumulative frequency curve. On the right side, there are panels for 'Accept prediction', 'Return to matrix', and 'Select/filter data'.

**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 displays the QSAR Toolbox interface. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Data Gap Filling' section is active, showing a list of chemicals with their predicted values. A callout box with the number '1' points to the 'Statistics' column header in the table below.

**Chemical List:**

| Chemical                       | M. 152 mg/L | M. 59.4 mg/L | M. 114 mg/L | M. 88.7 mg/L | M. 194 mg/L | M. 193 mg/L |
|--------------------------------|-------------|--------------|-------------|--------------|-------------|-------------|
| Tetrahymena pyriformis (30/32) |             |              |             |              |             |             |

**Model Statistics Table:**

| Descriptors   | Prediction | Adequacy | Cumul. freq. | Statistics    | Residuals |
|---|------------|----------|--------------|---------------|-----------|
| Statistical characteristics                         |            |          |              | TA model      |           |
| Number of data points, (N)                          |            |          |              | 29            |           |
| Coefficient of determination, (R2)                  |            |          |              | 0.798         |           |
| Adjusted coefficient of determination, (R2adj)      |            |          |              | 0.791         |           |
| Coefficient of determination - leave one out, (Q2)  |            |          |              | 0.773         |           |
| Coefficient of correlation for external set, (r2)   |            |          |              | -             |           |
| Sum of squared residuals, (SSR)                     |            |          |              | 2.48          |           |
| Standard deviation of residuals, (sH)               |            |          |              | -             |           |
| Sample standard deviation of residuals, (s)         |            |          |              | 0.303         |           |
| Fisher function, (F)                                |            |          |              | 107           |           |
| Fisher threshold for statistical significance, (Fa) |            |          |              | 5.99          |           |
| b0  |            |          |              | Intercept     |           |
| - model descriptor                                  |            |          |              | 2.13          |           |
| - coeff. value                                      |            |          |              | ± 0.25        |           |
| - coeff. range                                      |            |          |              | Yes           |           |
| - significance                                      |            |          |              | 0.248 (vs b1) |           |
| - max. covariation                                  |            |          |              |               |           |
| b1  |            |          |              | log Kow       |           |
| - model descriptor                                  |            |          |              | 0.528         |           |
| - coeff. value                                      |            |          |              | ± 0.105       |           |
| - coeff. range                                      |            |          |              | Yes           |           |
| - significance                                      |            |          |              | 0.248 (vs b0) |           |
| - max. covariation                                  |            |          |              |               |           |

**Right Panel: Accept prediction / Return to matrix**

- Select/filter data
  - Subcategorize
  - Mark chemicals by WS
  - Mark chemicals by descriptor value
  - Mark outlier points
  - Filter points by test conditions
  - Mark focused chemical
  - Mark focused points
- Selection navigation
  - Gap filling approach
  - Descriptors/data
  - Model/(Q)SAR
  - Calculation options
  - Visual options
  - Information
  - Miscellaneous

**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 Ciliatea Tetrahymena pyriformis

tetra

Structure

Tetrahymena pyriformis (30/32)

M: 145 mg/L M: 152 mg/L M: 59.4 mg/L M: 114 mg/L M: 88.7 mg/L M: 194 mg/L M: 193 mg/L

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

Distribution of residuals for IGC50 vs descriptors in use

Descriptor X: log Kow

Accept prediction

Return to matrix

- Select/filter data
  - Subcategorize
  - Mark chemicals by WS
  - Mark chemicals by descriptor value
  - Mark outlier points
  - Filter points by test conditions
  - Mark focused chemical
  - Mark focused points
- Selection navigation
  - Gap filling approach
  - Descriptors/data
  - Model/(Q)SAR
  - Calculation options
  - Visual options
  - Information
  - Miscellaneous

**1. Click Residuals**

## 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 interface during the 'Data Gap Filling' process. The main window displays a chemical structure of *Tetrahymena pyriformis* and a table of predicted values. A scatter plot shows the relationship between  $\log(Kow)$  (Descriptor X) and  $IGC50 (obs.) \cdot \log(I) (mg/L)$ . An 'Edit model' dialog box is open, allowing the user to save the model. The dialog box contains the following fields and options:

- Model name:** IGC50 Tetrahymena Furfural
- Model version:** (empty)
- QMRf file:** C:\Users\Nina\Documents\QSAR Toolbox\Ver 3.4\UserDir\IGC50 Tetrahymena...
- generate XML QMRf file
- 1.1. Model identifier:** IGC50 Tetrahymena Furfural
- 1.2. Data gap filling approach:** Trend analysis
- 1.3. Other related models:** (empty)
- 1.4. Software coding the model:** QSAR Toolbox 3.4.0.14

The right sidebar shows the 'Accept prediction' panel with the following options:

- Select/filter data
- Selection navigation
- Gap filling approach
- Descriptors/data
- Model/(Q)SAR
- Save model
- Save domain as category
- Save JRC XML QMRf
- Calculation options
- Visual options
- Information
- Miscellaneous

**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 the Data Gap Filling process. 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 'Read-across', 'Trend analysis', and '(Q)SAR models' options. The 'Target Endpoint' is set to 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis'. The 'Relevant (Q)SAR models' panel lists '<< CREATE A NEW QSAR >>' and 'IGC50 Tetrahymena Furfural'. The main area shows a tree view of 'tetra' with 'Ecotoxicological Information' expanded to 'Aquatic Toxicity' and 'Growth'. A data matrix table shows predicted values for various endpoints. Callout boxes 1, 2, and 3 highlight key features: 1 points to a prediction value in the matrix, 2 points to the '(Q)SAR models' radio button, and 3 points to the 'Relevant (Q)SAR models' list.

| Structure                    | 1 [target] | 2                                     | 3           | 4            | 5 | 6            | 7 |
|------------------------------|------------|---------------------------------------|-------------|--------------|---|--------------|---|
| Structure                    |            |                                       |             |              |   |              |   |
| Substance Identity           |            |                                       |             |              |   |              |   |
| Ecotoxicological Information |            |                                       |             |              |   |              |   |
| Aquatic Toxicity             |            |                                       |             |              |   |              |   |
| Avoidance                    | (1/2)      |                                       |             |              |   |              |   |
| Growth                       |            |                                       |             |              |   |              |   |
| EC50                         | (3/6)      |                                       |             |              |   |              |   |
| IGC50                        |            |                                       |             |              |   |              |   |
| 48 h                         |            |                                       |             |              |   |              |   |
| Protozoa                     |            |                                       |             |              |   |              |   |
| Ciliophora                   |            |                                       |             |              |   |              |   |
| Ciliata                      |            |                                       |             |              |   |              |   |
| Tetrahymena pyriformis       | (72/73)    | M: 145 mg/L<br>T: 262(59-1.16E3) mg/L | M: 152 mg/L | M: 59.4 mg/L |   | M: 10.9 mg/L |   |
| Growth Inhibition            | (4/7)      |                                       |             |              |   |              |   |
| Immobilisation               |            |                                       |             |              |   |              |   |
| Population                   | (19/39)    |                                       |             |              |   |              |   |
| Human Health Hazards         | (1/1)      |                                       |             |              |   |              |   |

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 during the 'Data Gap Filling' process. The main window displays a tree structure for 'tetra' (Tetrahymina pyriformis) with various endpoints like 'Aquatic Toxicity', 'Growth', and 'EC50'. A 'Training set of: IGC50 Tetrahymina Furfural' window is open, showing a grid of 15 chemical entries. Each entry includes a CAS number and an IGC50 value. The grid is as follows:

| Entry | CAS#      | IGC50     |
|-------|-----------|-----------|
| 1     | 66-25-1   | 152 mg/L  |
| 2     | 66-77-3   | 59.4 mg/L |
| 3     | 111-71-7  | 114 mg/L  |
| 4     | 123-05-7  | 88.7 mg/L |
| 5     | 123-72-8  | 194 mg/L  |
| 6     | 96-17-3   | 193 mg/L  |
| 7     | 110-62-3  | 104 mg/L  |
| 8     | 123-15-9  | 296 mg/L  |
| 9     | 590-86-3  | 188 mg/L  |
| 10    | 613-45-6  | 191 mg/L  |
| 11    | 874-42-0  | 16 mg/L   |
| 12    | 4460-86-0 | 247 mg/L  |
| 13    | 6361-21-3 | 54.8 mg/L |
| 14    | 112-44-7  | 3.9 mg/L  |
| 15    | 124-13-0  | 44.5 mg/L |

Callout 1: CREATE A NEW QSAR (IGC50 Tetrahymina Furfural)  
 Callout 2: Display Training Set  
 Callout 3: Training set of: IGC50 Tetrahymina Furfural

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

**1** Right Click on the derived QSAR model;

**2** Select Display Domain (see next screen shot)

| Model                          | M: 145 mg/L | T: 262(69-1.16E3) mg/L | M: 152 mg/L | M: 59.4 mg/L | M: 10.9 mg/L |
|--------------------------------|-------------|------------------------|-------------|--------------|--------------|
| Tetrahymena pyriformis (72/73) |             |                        |             |              |              |

# Data Gap Filling

## Visualisation of model domain

Domain Boundaries Browser

In Domain

Target

Metabolism

Simulator  
Do not apply metabolism

Process

Parent

Metabolites  Use parent if none

All

Ignore inorganic metabolites

Match

Any

All

Accumulatively

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 with **GREEN** tick; 3. otherwise is labelled with **RED** cross.

# Data Gap Filling

## Visualisation of the training set of the model

The screenshot shows the 'Domain Boundaries Browser' application. On the left, a 'Target' chemical structure is displayed. The main window is titled 'In Domain' and contains a list of chemicals with columns for CAS, Name, and SMILES. A callout '1' points to the 'Training set' tab. A callout '2' points to the list of chemicals. A callout '3' points to the 'Data points' window, which displays a table of data points.

| # | endpoint      | Value                      | al value                   | Strain | Organ | Effect    | Source |
|---|---------------|----------------------------|----------------------------|--------|-------|-----------|--------|
| 1 | LC50          | 17.8 mg/L                  | 0.000178 mol/L             |        |       | Mortality |        |
| 2 | LC50          | 9.79 mg/L                  | 9.77E-5 mol/L              |        |       | Mortality |        |
| 3 | IGC50         | 152 mg/L                   | 0.00151 mol/L              |        |       | Growth    |        |
| 4 | BOD           | 50 %                       | 50 % (Biodegradability)    |        |       |           |        |
| 5 | Gene mutation | Negative (Gene mutation I) | Negative (Gene mutation I) | TA 98  |       |           |        |

**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 displays the QSAR Toolbox interface. The top navigation bar shows the 'Data Gap Filling' step is active. The left sidebar contains a 'Data Gap Filling Method' section with 'Read-across' selected. Below it, the 'Target Endpoint' is set to 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliatea Tetrahymena pyriformis'. A context menu is open over the 'Display Domain' option. The main workspace shows a tree view of the model structure and a data matrix table. The table has columns for chemical IDs (59-65) and rows for various toxicity endpoints. A blue box highlights the cell for chemical #62 in the 'Growth IGC50 48 h Protozoa Ciliophora Ciliatea Tetrahymena pyriformis' row. Callout 1 points to this cell, callout 2 points to the right-click action, and callout 3 points to the 'Display Domain' menu option.

| Structure                    | 59 | 60          | 61 | 62 | 63           | 64         | 65          |
|------------------------------|----|-------------|----|----|--------------|------------|-------------|
| Substance Identity           |    |             |    |    |              |            |             |
| Ecotoxicological Information |    |             |    |    |              |            |             |
| Aquatic Toxicity             |    |             |    |    |              |            |             |
| Avoidance                    |    |             |    |    |              |            |             |
| Growth                       |    |             |    |    |              |            |             |
| EC50                         |    |             |    |    |              |            |             |
| IGC50                        |    |             |    |    |              |            |             |
| 48 h                         |    |             |    |    |              |            |             |
| Protozoa                     |    |             |    |    |              |            |             |
| Ciliophora                   |    |             |    |    |              |            |             |
| Ciliatea                     |    |             |    |    |              |            |             |
| Tetrahymena pyriformis       |    | M: 3.9 mg/L |    |    | M: 44.5 mg/L | M: 22 mg/L | M: 103 mg/L |
| Growth Inhibition            |    |             |    |    |              |            |             |
| Immobilisation               |    |             |    |    |              |            |             |
| Population                   |    |             |    |    |              |            | M: 569 mg/L |
| Human Health Hazards         |    |             |    |    |              |            |             |

**1. Highlight** the cell of one of the analogues (e.g., chemical # 62 in the data matrix; **2. Right 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.
- The chemical is an aldehyde as required by Acute aquatic toxicity MOA by OASIS group.
- It can react with protein by Schiff-base formation and should not belong to any of the eliminated mechanistic domains according to Protein binding by OASIS v.1.4:
  - Michael addition
  - AN2
  - Schiff base formation << Aldehydes
- The chemical is an aldehyde as required by Aquatic toxicity classification by ECOSAR
- Another requirement is Log Kow to be  $\geq 0.3156$  and  $\leq 4.75$ .

# 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 belonging to Michael addition mechanism by Protein binding by OASIS v1.4, which have been eliminated from the domain (see boundary 5 and 6)

# Data Gap Filling

## Edit QMRF data

The screenshot displays the QSAR Toolbox interface with the 'Data Gap Filling' method selected. A context menu is open over the 'IGC50 Tetrahymena Furfural' model, with 'Edit QMRF data' selected. An 'Edit model' dialog box is open, showing fields for model name, version, and QMRF file, along with sections for species and endpoints.

**1** Right click above the model; **2** Select Edit QMRF data.

The 'Edit model' dialog box contains the following information:

- Model name:** IGC50 Tetrahymena Furfural (editable field)
- Model version:** (editable field)
- QMRF file:** C:\Users\Ksenia\Documents\QSAR Toolbox\Ver 3.3\userDir\IGC50 Tetrahymena Furfural.xml (Browse ...)
- generate XML QMRF file
- 3.1. Species (one per line):** Tetrahymena pyriformis (editable field)
- 3.2. Endpoints (one per line):** IGC50 (editable field)
- Endpoint classification:** (not selected)

1. Right click above the model; 2. Select Edit QMRF data. 3. Fill in/edit fields of QMRF template

# Data Gap Filling

## Perform prediction for chemicals in domain

The screenshot displays the QSAR Toolbox interface during a Data Gap Filling operation. The left sidebar shows the 'Data Gap Filling Method' set to 'Read-across' and the 'Target Endpoint' as 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis'. The 'Relevant (Q)SAR models' list includes 'IGC50 Tetrahymena Furfural'. A context menu is open over this model, with the 'Predict Chemicals in Domain' option selected. The central tree view shows the classification hierarchy, with 'Tetrahymena p...' (78/92) highlighted. The right-hand table shows predicted values for various endpoints:

| Endpoint                     | 62                             | 63                    | 64                    | 65                    | 66                    | 67                    | 68                    | 69                    |
|------------------------------|--------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Structure                    | <chem>CC1=CC=C(C=C1)C=C</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> | <chem>CCCCCCCC</chem> |
| Substance Identity           |                                |                       |                       |                       |                       |                       |                       |                       |
| Ecotoxicological Information |                                |                       |                       |                       |                       |                       |                       |                       |
| Aquatic Toxicity             |                                |                       |                       |                       |                       |                       |                       |                       |
| Avoidance                    |                                |                       |                       |                       |                       |                       |                       |                       |
| Growth                       |                                |                       |                       |                       |                       |                       |                       |                       |
| EC50                         |                                |                       |                       |                       |                       |                       |                       |                       |
| IGC50                        |                                |                       |                       |                       |                       |                       |                       |                       |
| 48 h                         |                                |                       |                       |                       |                       |                       |                       |                       |
| Protozoa                     |                                |                       |                       |                       |                       |                       |                       |                       |
| Ciliophora                   |                                |                       |                       |                       |                       |                       |                       |                       |
| Ciliata                      |                                |                       |                       |                       |                       |                       |                       |                       |
| Tetrahymena p...             | M: 44.5 mg/L                   | M: 22 mg/L            | M: 103 mg/L           |                       |                       |                       |                       |                       |
| Growth Inhibition            |                                |                       |                       |                       |                       |                       |                       |                       |
| Immobilisation               |                                |                       |                       |                       |                       |                       |                       |                       |
| Population                   |                                |                       |                       | M: 569 mg/L           |                       |                       |                       |                       |
| Human Health Hazards         |                                |                       |                       |                       |                       |                       |                       |                       |

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

# Data Gap Filling

## Perform prediction for chemicals in domain

The screenshot displays the QSAR Toolbox interface during a Data Gap Filling operation. The top navigation bar shows the 'Data Gap Filling' step is active. The left sidebar contains options for 'Read-across', 'Trend analysis', and '(Q)SAR models'. The main workspace shows a tree view of chemical categories, with 'Tetrahymina ... (78/106)' selected. A table of results is visible, showing predicted values for various chemicals. A red box highlights the status bar at the bottom, which displays '643 Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories)'. A callout box with the number '2' points to an 'OK' button in a prediction message that says 'Predicted 7 out of 643 chemicals'.

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. The top navigation bar includes buttons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The left sidebar contains sections for Data Gap Filling Method (Read-across, Trend analysis, QSAR models), Target Endpoint (Ecotoxicological Information, Aquatic Toxicity Growth, IGC50 48 h Protozoa, Ciliophora Ciliatea), and relevant QSAR models. The main area displays a tree structure of endpoints and a table of results for various chemical structures. A red box highlights the 'Export' option in a context menu that appears over the 'Tetrahymena pyriformis' endpoint row. A blue callout '1' points to the row, and another blue callout '2' points to the 'Export' menu item.

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. Browse to save 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.



## 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.