

QSAR TOOLBOX

The OECD QSAR Toolbox
for Grouping Chemicals
into Categories

OECD QSAR Toolbox v.3.0

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 are 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**

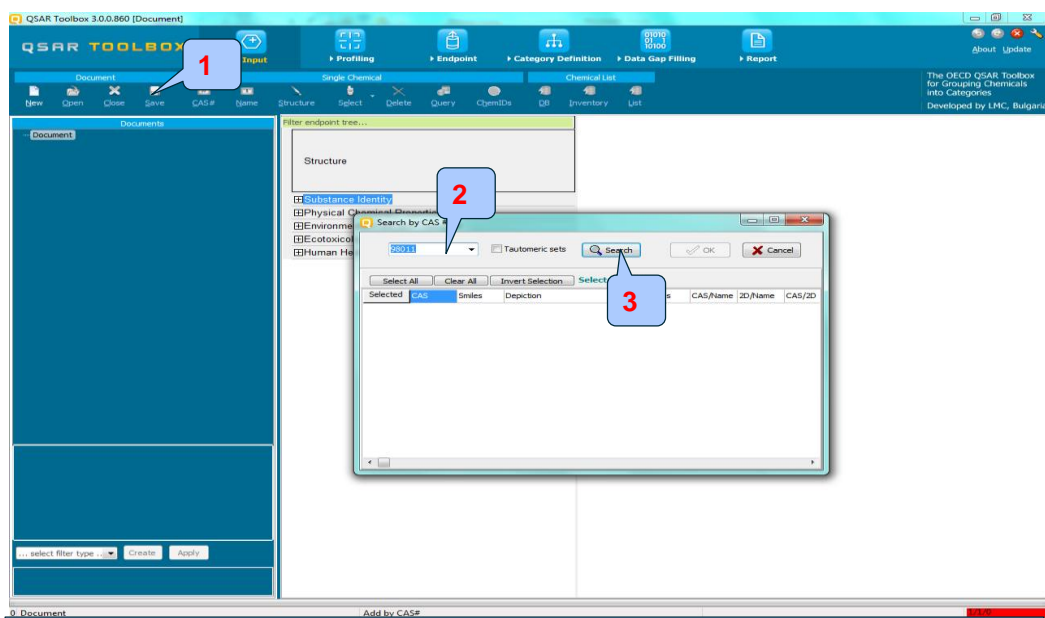
Workflow of the exercise

- **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

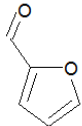


1. Click on CAS # 2. Enter 98011; 3. Click Search

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.

Selected	CAS	Smiles	Depiction	Names	CAS/Name	2D/Name	CAS/2D
1: Yes	98-01-1	C1(C=O)=C		1: 2-furalde 2: 2-furanc 3: fufural 4: furan-2- 5: furfural	8: METI 9: REAC 10: US I 2: High Que 1: Bacte 2: Carci 3: Carci 4: DSST 5: Geno 6: MUNI 7: Rep I 8: Repe 9: Rode 3: High Que	8: US HI 9: ECHA 10: GSH 2: High Que 1: Bacte 1: Bacte 2: Carci 3: Carci 4: MUNI 5: Rode 6: Geno 7: DSST 8: Repe 9: Rep I 3: High Que	5: BI 6: C 7: C 8: C 9: D 10: I 11: I 12: I 13: I 14: I 15: I 16: I 17: I 18: I

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 displays the QSAR Toolbox 3.0.0.860 interface. The 'Input' module is selected in the top navigation bar. The main window shows a chemical structure of furan-2-carbaldehyde with its SMILES string O=C1C=CC=CO1 and a list of identifiers including CAS Number 98-01-1, EC Number 202-627-7, and Chemical Name furan-2-aldehyde.

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

QSAR Toolbox 3.0.0.860 [Document]

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Profiling Profiling Schemes

Apply [1] [2] Delete

Select All Unselect All Invert About

Hydrolysis half-life (kb, pH 7)(Hydrowin)
 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
 Protein binding by OASIS v 1.1
 Protein binding by OECD
 Protein binding potency
 Subfragments
 Toxic hazard classification by Cramer (original)
 Toxic hazard classification by Cramer (with extended)
 Uptake biodeg

Endpoint Specific

Acute aquatic toxicity classification by Verhaar
 Acute aquatic toxicity MCA by OASIS
 Acute toxicity classification by ECOSAR
 Bioaccumulation – metabolism alerts
 Bioaccumulation – metabolism half-lives
 Biodegradation fragments (BioWIN MITI)
 Carcinogenicity (genotox and nongenotox) alerts by
 DNA alerts for Ames, MN and CA 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

Filter endpoint tree... 1 [target]

Structure

Substance Identity
 CAS Number
 Chemical IDs
 Chemical Name
 Structural Formula
 Physical Chemical Properties
 Environmental Fate and Transport
 Ecotoxicological Information
 Human Health Hazards

98-01-1
 EC Number:202-627-7
 Eines Number:202...
 2-furaldehyde
 2-furancarboxaldehyde
 furfural
 furan-2-aldehyde
 furfural
 C1(C=O)=CC=CO1

Metabolism/Transformations

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 3.0.0.860 interface. The 'Profiling methods' list on the left includes various endpoints such as 'Hydrolysis half-life', 'Protein binding by OASIS', and 'Acute aquatic toxicity classification by Verhaar'. The 'Filter endpoint tree...' window on the right shows a tree structure with nodes like 'Structure', 'Substance Identity', 'Chemical IDs', 'Chemical Name', 'Structural Formula', 'Physical Chemical Properties', 'Environmental Fate and Transport', 'Ecotoxicological Information', 'Human Health Hazards', and 'Profile'. A red circle highlights the 'Profile' node, and a callout box with the number '1' points to it.

1. Double click on the box  to open the nodes of the tree (see insert)

Filter endpoint tree...	1 [target]
Structure	
— 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
— Keratinocyte gene expression	Moderate gene exp...
— Lipinski Rule Oasis	Moderate gene exp...
— New profiler	Bioavailable
— OECD HPV Chemical Categories	(N/A)
— Oncologic Primary Classification	Not categorized
— Organic functional groups	Aldehyde Type Co...
— Organic functional groups (US EPA)	Alpha,beta unsatur...
— Organic functional groups(nested)	Furane
— Organic functional groups, Norbert Haider (checkmol)	Aromatic Oxygen
— Protein binding alerts for skin sensitization by OASIS...	Carbonyl, olefinic at...
— Protein binding by OASIS	Miscellaneous sulfi...
— Protein binding by OASIS v1.1	Olefinic carbon [=C...
— Protein binding by OASIS-pilot v.1	Oxygen, two olefini...
— Protein binding by OECD	Alpha,beta unsatur...
	Furane
	Overlapping groups
	Aromatic compound
	Heterocyclic comp...
	Schiff base formatio...
	Schiff base formatio...
	Schiff base formatio...
	Schiff base formatio...
	Schiff base formatio...
	Schiff base formatio...
	Schiff base formatio...
	No alert found

The OECD QSAR Toolbox for Grouping Chemicals into Categories
05.10.2012 19

QSAR TOOLBOX

Profiling Profiles of "Furfural"

The screenshot shows the QSAR Toolbox 3.0.0.860 interface. The 'Profiling' tab is selected, and the 'Profiling methods' list on the left includes various predefined and general mechanisms. The 'Filter endpoint tree...' on the right shows a hierarchical view of endpoints, with 'Structure' selected. A chemical structure of furfural is shown next to it. A right-click context menu is open over the 'Structure' node, with the 'Explain' option highlighted by a red circle and a callout box containing the number '1'.

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"

QSAR Toolbox 3.0.0.860 [Document]

QSAR TOOLBOX

Profiling

Profiling methods

Predefined

- Database Affiliation
- Inventory Affiliation
- OECD HPV Chemical Categories
- Substance Type
- US-EPA New Chemical Categories

General Mechanistic

- Biodeg BioHC half-life (Bowen)
- Biodeg primary (Bowen 4)
- Biodeg probability (Bowen 1)
- Biodeg probability (Bowen 2)
- Biodeg probability (Bowen 5)
- Biodeg probability (Bowen 6)
- Biodeg probability (Bowen 7)
- Biodeg ultimate (Bowen 3)
- DNA binding by OASIS v.1.1
- DNA binding by OECD
- DRA Cysteine peptide depletion
- DRA Lysine peptide depletion
- Estrogen Receptor Binding
- Hydrolysis half-life (90, pH 7)(Hydrowin)
- Hydrolysis half-life (90, pH 8)(Hydrowin)
- Hydrolysis half-life (90, pH 7)(Hydrowin)
- Hydrolysis half-life (90, pH 8)(Hydrowin)
- Hydrolysis half-life (pH 6.3-7-9)
- Ionization at pH = 1
- Ionization at pH = 4

Metabolism/Transformations

Documented

- Observed Mammalian metabolism
- Observed Microbial metabolism
- Observed Rat Liver metabolism
- Observed Rat Liver S9 metabolism

Filter endpoint tree...

Structure

Organic functional groups, Norbert Haider (checkmol)

Protein binding alerts for skin sensitization by OASIS v1.1

Protein binding by OASIS v1.1

Protein binding by OASIS-pit

Protein binding by OECD

Protein binding potency

Protein binding potency base

Protein binding potency base

Protein binding potency base

Protein binding potency base

Repeated dose (HESS)

Skin irritation/corrosion Exclusion

Skin irritation/corrosion Inclusion

Substance Type

Profiling results

- Chemical profile
 - Protein binding by OASIS v1.1
 - Schiff base formation
 - Schiff base formation >> Schiff base formation with carbonyl compounds
 - Schiff base formation >>> Schiff base formation with carbonyl compounds >>> Aldehydes

The Protein binding by OASIS v.1.1 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.1.1 of target chemical

Protein binding by OASIS v1.1

Aldehydes

Target

Boundaries Training set

Boundary Options Metabolism

Fragment: CC(=O)H Edit

Profile Description

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

This category includes chemicals that potentially can cause skin sensitization effect as a result of protein conjugation via Schiff base formation with aldehydes.

The possible structural alert acting by this mechanism is illustrated below:

$$\text{R}-\text{C}(=\text{O})\text{H} + \text{Pr}-\text{NH}_2 \rightarrow \text{R}-\text{C}(\text{H})=\text{N}-\text{Pr}$$

R = H, alkyl

Outlook

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

Endpoints

Extracting endpoint values

1

2

QSAR Toolbox 3.0.0.860 [Document]

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Data Import Export Delete Tautomerize

Gather Import IJCLID5 Export IJCLID5 Database Inventory Database

Databases

Select All Unselect All Invert About

- Physical Chemical Properties
- Environmental Fate and Transport
- Ecotoxicological Information
- Human Health Hazards

Filter endpoint tree...

Structure

Organic functional groups, Norbert Haider (checkmol) Aromatic compound
Heterocyclic comp...
Schiff base formation

Protein binding alerts for skin sensitization by OASIS... Schiff base formatio...
Schiff base formatio...
Schiff base formatio...

Protein binding by OASIS Schiff base formatio...
Schiff base formatio...

Protein binding by OASIS v1.1 Schiff base formatio...
Schiff base formatio...
Schiff base formatio...

Protein binding by OASIS-pilot v.1 Schiff base formatio...
Schiff base formatio...

Protein binding by OECD No alert found

Protein binding potency Moderately reactive...
Moderately reactive...

Protein binding potency based on Cys(1:10) depleti... Minimal reactivity >...
Minimal reactivity >...

Protein binding potency based on Cys(1:10) depleti... Non reactive
Non reactive >> No...

Protein binding potency based on Lys(1:50) depletion... Minimal reactivity >...
Minimal reactivity >...

Protein binding potency based on Lys(1:50) depletion... Non reactive
Non reactive >> No...

1 [target]

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 shows the QSAR Toolbox 3.0.0.860 interface. The 'Endpoint' menu is open, displaying a list of endpoints such as 'Organic functional groups, Norbert Haider (checkmol)', 'Protein binding alerts for skin sensitization by OASIS', and 'Protein binding by OASIS'. A 'Read data?' dialog box is overlaid on the interface, with the 'All endpoints' radio button selected and the 'from Tautomers' checkbox checked. A red callout box with the number '1' points to the 'OK' button in the dialog box. A blue callout box at the bottom right contains the text '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: 232 data-points, 54 groups, 1 chemicals

	Endpoint	CAS	Structure	Value	Administration
<input checked="" type="checkbox"/>	LC50	98-01-1	<chem>O=C1C=CC=C1</chem>	3,2E4 micrograms per liter	
<input checked="" type="checkbox"/>	LC50	98-01-1	<chem>O=C1C=CC=C1</chem>	3,2E4 micrograms per liter	
<input checked="" type="checkbox"/>	Summary carcinogenicity	98-01-1	<chem>O=C1C=CC=C1</chem>	Positive	
<input checked="" type="checkbox"/>	Summary carcinogenicity	98-01-1	<chem>O=C1C=CC=C1</chem>	Positive	
<input checked="" type="checkbox"/>	LC50	98-01-1	<chem>O=C1C=CC=C1</chem>	>5E4 micrograms per liter	
<input checked="" type="checkbox"/>	LC50	98-01-1	<chem>O=C1C=CC=C1</chem>	>5E4 micrograms per liter	
<input checked="" type="checkbox"/>	LC50	98-01-1	<chem>O=C1C=CC=C1</chem>	3,7E4 micrograms per liter	

1. Click Select one

2. Click OK

1. Click Select one 2. Click OK

QSAR TOOLBOX

Endpoints

Inserting data for target in data matrix

QSAR Toolbox 3.0.0.860 [Document]

QSAR TOOLBOX

Input Profiling Endpoint **Category Definition** Data Gap Filling Report

Data Import Export Delete Database Inventory Data

Databases: Select All Unselect All Invert About

- Physical Chemical Properties
- Environmental Fate and Transport
- Ecotoxicological Information
- Human Health Hazards

filter endpoint tree...

Structure

Substance Identity

- CAS Number
- Chemical IDs
- Chemical Name
- Structural Formula

Physical Chemical Properties

Environmental Fate and Transport

Ecotoxicological Information

Human Health Hazards

Profile

1 [target]

O=Cc1ccoc1

98-01-1

EC Number:202-627-7

Einecs Number:202...

2-furaldehyde

2-furan-carboxaldehyde

fural

furan-2-aldehyde

furfural

O=Cc1ccoc1

(1/10)M: 162 °C, 0.41, -38

(1/9)M: 93.5 %, 0.382 Pa...

(1/204)M: 20.5 mg/L, 10.5

(1/520)M: Negative, Negativ...

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#Cili
atea#

Tetrahymena pyriformis

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

Category definition

Navigate to the target endpoint

Q^SAR Toolbox 2.3.0.1130 [Document_1]

Q^SAR TOOLBOX

Profiling Endpoint Category Definition Data Gap Filling Report

Define Subcategorize Combine Clustering Delete Delete All

Grouping methods

Predefined

- Database Affiliation
- Inventory Affiliation
- OECD HPV Chemical Categories
- Substance Type
- US-EPA New Chemical Categories

General Mechanistic

- DNA binding by OASIS
- DNA binding by OECD
- Estrogen Receptor Binding
- Protein binding by OASIS
- Protein binding by OECD
- Protein Binding Potency
- Superfragments
- Toxic hazard classification by Cramer (original)
- Toxic hazard classification by Cramer (with e.g.)

Endpoint Specific

- Acute aquatic toxicity classification by Verhaar
- Acute aquatic toxicity MOA by OASIS
- Aquatic toxicity classification by ECoSAR
- Bioaccumulation – metabolism alert by
- Bioaccumulation – metabolism half-lives
- Biodegradation fragments (BioWIN MITI)
- Carcinogenicity (genotox and nongenotox) alert by
- Eye irritation/corrosion Exclusion rules by BR
- Eye irritator/corrosion Inclusion rules by BR
- in vivo mutagenicity (Ames test) alerts by IS
- in vivo mutagenicity (Micronucleus) alerts by
- OncoLogic Primary Classification
- Skin irritator/corrosion Exclusion rules by BR
- Skin irritator/corrosion Inclusion rules by BR

Empiric

- Chemical elements

tetra

Structure

Substance Identity

- CAS Number 98-01-1
- Chemical Name furfural, 2-furancarboxaldehyde, 2-furaldehyde, furan-2-aldehyde
- Structural Formula furan-2-aldehyde, C1(C=O)=CC=CO1

Ecotoxicological Information

- Aquatic Toxicity
- Growth
- IGC50
- 48 h
- Protozoa
- Ciliophora
- Ciliata
- Tetrahymena pyriformis (1/1) M. 145 mg/L
- Immobilisation
- Profile

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 displays the QSAR Toolbox 3.0.0.860 interface. The 'Category Definition' tab is active. In the 'Predefined' list on the left, 'US-EPA New Chemical Categories' is highlighted with a blue box labeled '1'. The 'Define' button is circled in red and labeled '2'. The 'US-EPA New Chemical Categories' dialog box is open, showing a list of categories. The 'Strict' radio button is selected and circled in red, labeled '3'. The 'OK' button is highlighted with a blue box labeled '4'. The main window shows a chemical structure and its properties, including CAS Number 99-01-1 and EC Number 202-627-7.

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

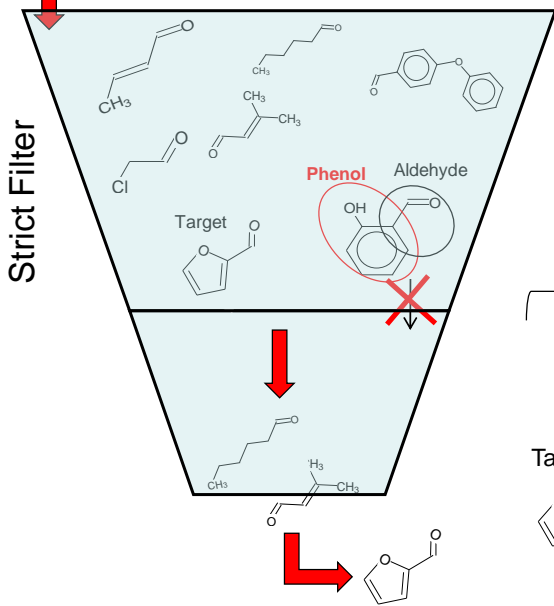
- The **Strict** functionality means that the software will group analogues having **ONLY** the categories of the target and will exclude the analogues having any other categories according to the profiler used in the grouping method.
- For example, if the profiling for the target results in *Aldehydes(Acute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

QSAR TOOLBOX

Category definition

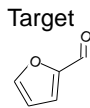
Defining US-EPA category strict functionality

Input



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

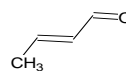
Defined Category



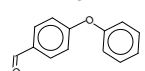
Analogue 1



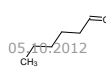
Analogue 2



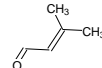
Analogue 3



Analogue 4



Analogue 5



Category definition

Defining US-EPA category

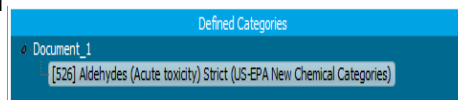
The screenshot displays the QSAR Toolbox software interface. The main window is titled 'QSAR Toolbox 3.0.0.860 [Document]'. The 'Category Definition' tab is active, showing a hierarchical tree of chemical properties on the left and a table of chemical identifiers on the right. A dialog box titled 'Define category name' is open, with 'US-EPA New Chemical Categories' selected as the category name. A red '1' in a blue box points to the 'OK' button. The background shows a hierarchical tree of chemical properties and a table of chemical identifiers.

1. Click OK to confirm the name of the category

Category definition

Analogues

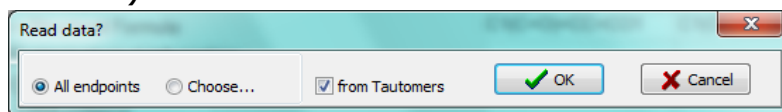
- The Toolbox now identifies all chemicals corresponding to *Aldehydes (Acute toxicity)* by US-EPA listed in the databases selected under “Endpoints”.
- 526 analogues including the target chemical are identified; they form a mechanistic category “**Aldehydes (Acute toxicity)**”, which will be used for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



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: 2969 data-points, 732 groups, 481 chemicals

	Endpoint	CAS	Structure	Value	Abnormality
<input checked="" type="checkbox"/>		50-00-0	<chem>H2C=O</chem>	155 K	
<input checked="" type="checkbox"/>		50-00-0	<chem>H2C=O</chem>	155 K	
<input checked="" type="checkbox"/>		50-00-0	<chem>H2C=O</chem>	181 K	
<input checked="" type="checkbox"/>		50-00-0	<chem>H2C=O</chem>	181 K	
<input checked="" type="checkbox"/>	Melting point	542-78-9	<chem>C=CC=O</chem>	72 deg C	
<input checked="" type="checkbox"/>	Melting point	542-78-9	<chem>C=CC=O</chem>	72 deg C	
<input checked="" type="checkbox"/>		124-13-0	<chem>CCCCC</chem>	246 K	
<input checked="" type="checkbox"/>		124-13-0	<chem>CCCCC</chem>	246 K	

1. Click Select one; 2. Click OK

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

Category definition

Summary information for Analogues

The experimental results for the analogues are inserted into the matrix

The screenshot displays the QSAR Toolbox 3.0.0.860 interface. The main workspace shows a comparison matrix for chemical analogues. The matrix has columns for different target substances and rows for various chemical and biological properties. A red box highlights the entry for *Tetrahymena pyriformis* in the matrix.

Property	1 [target]	2	3	4	5
Structure					
Substance Identity					
CAS Number	99-01-1	63-42-3	66-25-1	66-77-3	66-99
EC Number 202-627-7		NA			
Eines Number 202...			Einecs Number 200...	Einecs Number 200...	Einecs
Chemical Name	2-furaldehyde 2-furanicarboxaldehyde furalan furan-2-aldehyde furfural	d-glucose, 4-o-7-d-... lactose	hexaldehyde hexanal	1-naphthaldehyde 1-naphthalene carb...	2-nap... 2-napit... naphit... naphit...
Structural Formula	C1(C=O)=CC=CO1	C1(OC(C(O)C(O)C=O)C=O)C(=O)CCCC	C(=O)CCCCC	c12c(cc(C=O)ccc1)c...	c12c(c...
Ecotoxicological Information					
Aquatic Toxicity					
Avoidance	(1/2)				
Growth					
BEC50	(3/6)				
GC50					
48 h					
Protozoa					
Ciliophora					
Gillataea					
Tetrahymena pyriformis (1/1/1)	M: 145 mg/L		M: 152 mg/L	M: 59.4 mg/L	
Growth Inhibition	(4/6)				
Immobilisation					
Profile					

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 software interface. The 'Data Gap Filling' method is selected in the left sidebar. The main table displays data for several target chemicals. The following table represents the data shown in the screenshot:

Structure	1 Target	2	3	4	5
Structure					
Substance Identity	98-01-1	63-42-3	66-25-1	66-77-3	66-99
CAS Number	EC Number 202.627.7	NA	Einecs Number 200...	Einecs Number 200...	Einecs
Chemical IDs	2-furaldehyde	d-glucose, 4-o-7-d-...	hexaldehyde	1-naphthaldehyde	2-napht...
Chemical Name	2-furanicarboxaldehyde	lactose	hexanal	1-naphthalene carb...	2-napht...
Structural Formula	furalen-2-aldehyde	furalen-2-aldehyde	furalen-2-aldehyde	1-naphthalene-1-carb...	1-napht...
Ecotoxicological Information	C1(C=O)=CC=CO1	C1(OC(CO)C)C(O)C=	C(=O)CCCCC	c12c(c(C=O)ccc1)c...	c12c(c...
Aquatic Toxicity	(1/2)				
EC50	(3/6)				
IGC50					
48 h					
Protozoa					
Ciliophora					
Ciliata					
Tetrahymena pyriformis	(7/7) M: 145 mg/L		M: 152 mg/L	M: 59.4 mg/L	
Growth Inhibition	(4/6)				
Immobilisation					
Profile					

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 3.0.0.860 [Document]

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Filing Apply

Data Gap Filling Method

- Read-across
- Trend analysis
- QSAR models

Target Endpoint

Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Proboscis Cladocera Cladocera *Tetrahymena pyriformis*

Structure

Tetrahymena pyriformis (71/71) M: 145 mg/L M: 162 mg/L M: 69.4 mg/L M: 10.9 mg/L

Descriptors Prediction Adequacy Cumul. freq. Statistics Residuals

Trend analysis prediction of Ecotoxicological Information#Aquatic Toxicity, making a linear approximation, based on 70 values from 70 analogues. Observed target value: 145 mg/L, Predicted target value: 101 mg/L.

Model equation: Ecotoxicological Information#Aquatic Toxicity = +2,64 + 0,405 * log Kow

Accept prediction

Return to matrix

- Select/filter data
- Selection navigation
- Gap filling approach
- Descriptors/data
- Model/QSAR
- Calculation options
- Visual options
- Information
- Miscellaneous

The OECD QSAR Toolbox for Grouping Chemicals into Categories

05.10.2012

42

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.1.1

- **To improve the data by subcategorizing, the Protein binding by OASIS v.1.1 profiler is used:**
- **Click** on Select filter data then **click** Subcategor.
- **Select** Protein binding by OASIS v.1.1 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).

QSAR TOOLBOX

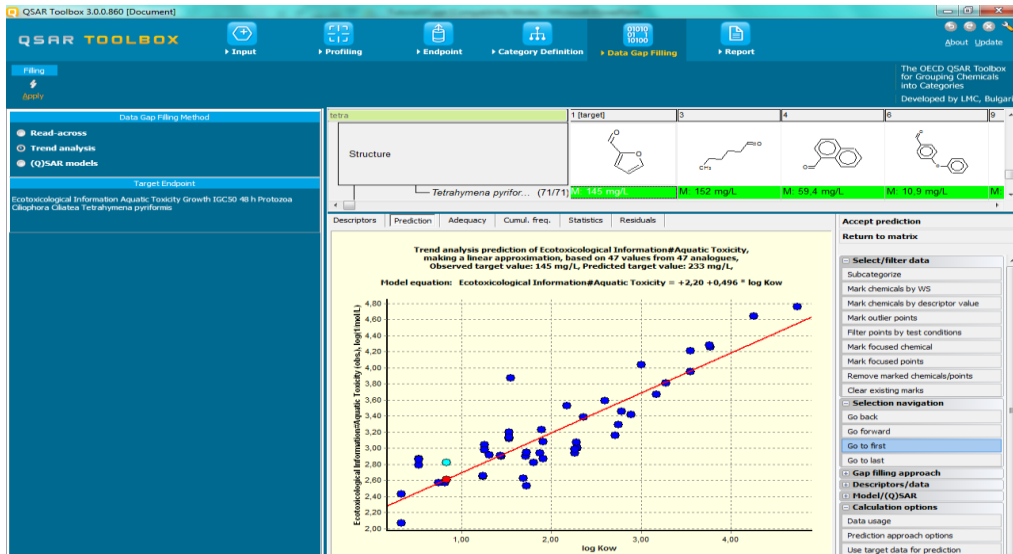
Data Gap Filling (IGC 50 48h of *T. pyriformis*) Subcategorisation by Protein binding by OASIS v.1.1

The screenshot displays the QSAR Toolbox software interface. On the left, a 'Grouping methods' list is visible, with 'Protein binding by OASIS' highlighted. The main window shows a 'Trend analysis' plot with the equation: $\text{Ecotoxicological Information\#Aquatic Toxicity} = +2,64 + 0,405 * \log \text{Kow}$. The plot shows a positive correlation between log Kow and aquatic toxicity. On the right, the 'Subcategorization' panel is open, showing a list of chemicals and their predicted values. A 'Remove' button is highlighted at the bottom of the panel.

1. Click Select filter data
2. Select Subcategorize;
3. Select Protein binding by OASIS v.1.1
4. Click Remove to eliminate dissimilar to the target chemicals

The OECD QSAR Toolbox for Grouping Chemicals into Categories

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



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

An accurate trend analysis of data set

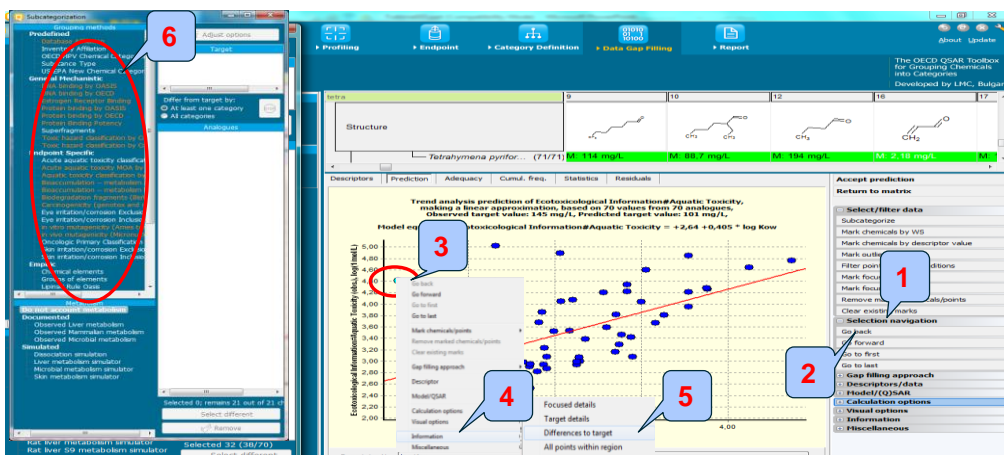
- The chemicals which differ from the target according to Protein binding by OASIS are:
 - Michael addition <<alpha, beta – unsaturated carbonyl compounds << alpha, beta-unsaturated aldehydes (22);
 - SNAr <<Nucleophilic aromatic substitution on activated halogens <<Activated(1).
- Another way for refining 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.1.1 from the Grouping methods list
- **Click** on Remove (see next three screen shots).

QSAR TOOLBOX

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

QSAR TOOLBOX

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

The screenshot displays the QSAR Toolbox 3.0.0.860 interface. The 'Subcategorization' panel on the left lists various methods, with 'Protein binding by OASIS v.1.1' highlighted. The central plot shows a scatter of data points with a red regression line. The 'Data gap filling' table at the bottom lists chemicals with their predicted values and target values. A callout box with a blue background contains the following text:

1. Select Protein binding by OASIS v.1.1; 2. Click Remove to eliminate chemicals dissimilar (those noted in green) to the target.

The OECD QSAR Toolbox for Grouping Chemicals into Categories

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

The screenshot displays the QSAR Toolbox 3.0.0.860 interface. The main window is titled 'Data Gap Filling Method' and shows the following components:

- Menu Bar:** Input, Profiling, Endpoint, Category Definition, Data Gap Filling (active), Report.
- Left Sidebar:**
 - Data Gap Filling Method: Read-across, Trend analysis, (Q)SAR models.
 - Target Endpoint: Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa *Cladophora glauca* *Tetrahymena pyriformis*.
- Main Workspace:**
 - Structure: Tetrahymena pyriformis... (7/1/71)
 - Table of predicted values:

Target	1	2	3	4	5
M	145 mg/L	152 mg/L	59.4 mg/L	10.9 mg/L	M
 - Descriptors: Prediction, Adequacy, Cumul. freq., Statistics, Residuals.
 - Graph: Trend analysis prediction of Ecotoxicological Information@Aquatic Toxicity. Model equation: $\text{Ecotoxicological Information@Aquatic Toxicity} = +2.20 + 0.496 \cdot \log \text{Kow}$. The graph plots Ecotoxicological Information@Aquatic Toxicity (log₁₀ mg/L) against log Kow, showing a positive linear correlation with a red regression line.
- 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, Remove marked chemicals/points, Clear existing marks.
 - Selection navigation: Go back, Go forward, Go to first, Go to last.
 - Gap filling approach: Descriptors/data, Model/(Q)SAR, Calculation options.
 - Data usage: Prediction approach options, Use target data for prediction.

The OECD QSAR Toolbox for Grouping Chemicals into Categories

05.10.2012

52

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

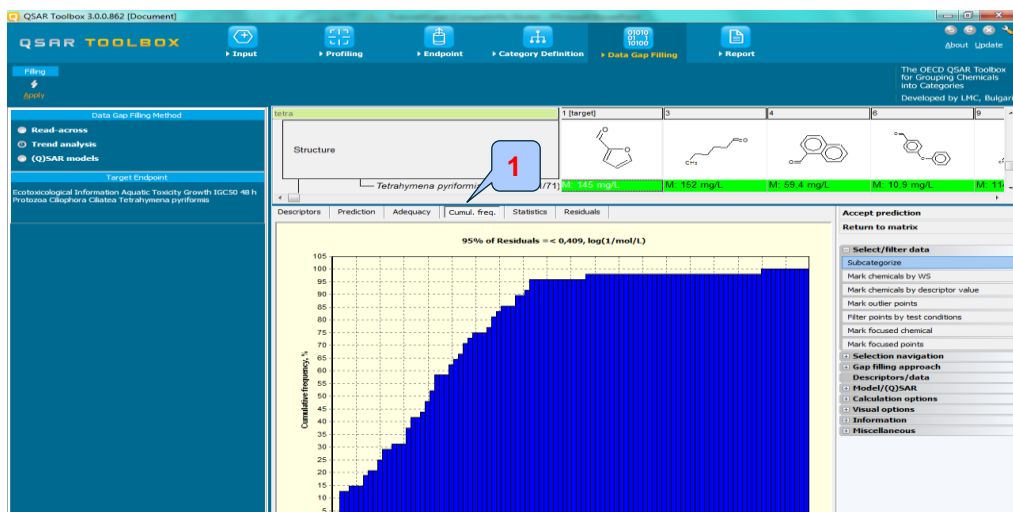
Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model

The screenshot displays the QSAR Toolbox 3.0.0.862 interface. The main window shows the 'Data Gap Filling' method selected. The central workspace displays the chemical structure of Tetrahymena pyriformis and a table of data points for various concentrations (M, 145 mg/L, M, 152 mg/L, M, 18.4 mg/L, M, 10.8 mg/L, M, 11 mg/L). The table includes columns for Descriptors, Prediction, Adequacy, Cumul. freq., Statistics, and Residuals.

Descriptors	Prediction	Adequacy	Cumul. freq.	Statistics	Residuals
Statistical characteristics					
Number of data points, (N)					
47					
Coefficient of determination, (R ²)					
0,781					
Adjusted coefficient of determination, (R ² adj)					
0,777					
Coefficient of determination - leave one out, (Q ²)					
0,763					
Coefficient of correlation for external set, (Q ² ext)					
-					
Sum of squared residuals, (SSR)					
3,49					
Standard deviation of residuals, (sR)					
-					
Sample standard deviation of residuals, (s)					
0,278					
Fisher function, (F)					
161					
Fisher threshold for statistical significance, (F _a)					
5,67					
b0					
- model descriptor					
Intercept					
- coeff. value					
2,20					
- coeff. range					
# 0,18					
- significance					
Yes					
- max. covariation					
0,200 (vs b1)					
b1					
- model descriptor					
log Kow					
- coeff. value					
0,496					
- coeff. range					
# 0,077					
- significance					
Yes					
- max. covariation					
0,200 (vs b0)					

The right-hand panel shows the 'Accept prediction' section with options to 'Return to matrix', 'Select/filter data', and 'Selection navigation'. The 'Selection navigation' section includes options for 'Gap filling approach', 'Descriptors/data', 'Model/QSAR', 'Calculation options', 'Visual options', 'Information', and 'Miscellaneous'.

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



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

QSAR TOOLBOX

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' workflow selected. The 'Statistics' tab is active, displaying a table of model characteristics and descriptors. A red box with the number '1' highlights the 'Statistics' tab in the top navigation bar. A callout box with the number '1' points to the 'Statistics' column header in the table.

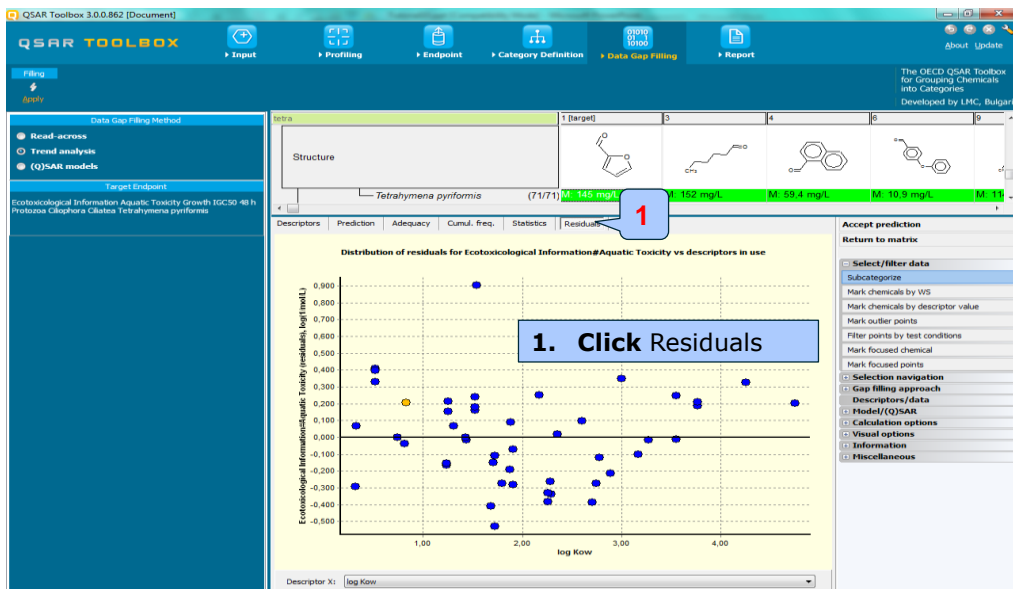
Descriptors	Prediction	Adequacy	Cumul. freq.	Statistics	Residuals
Statistical characteristics					
TA model					
Number of data points, (N)					
47					
Coefficient of determination, (R ²)					
0,781					
Adjusted coefficient of determination, (R ² adj)					
0,777					
Coefficient of determination - leave one out, (Q ²)					
0,763					
Coefficient of correlation for external set, (Q ²)					
-					
Sum of squared residuals, (SSR)					
3,49					
Standard deviation of residuals, (sR)					
0,278					
Sample standard deviation of residuals, (s)					
161					
Fisher function, (F)					
5,67					
Fisher threshold for statistical significance, (F _a)					
5,67					
b0					
- model descriptor					
Intercept					
- coeff. value					
2,20					
- coeff. range					
# 0,18					
- significance					
Yes					
- max. covariation					
0,200 (vs b1)					
b1					
- model descriptor					
log Kow					
- coeff. value					
0,496					
- coeff. range					
# 0,077					
- significance					
Yes					
- max. covariation					
0,200 (vs b0)					

The OECD QSAR Toolbox for Grouping Chemicals into Categories

05.10.2012

56

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



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, a target endpoint 'Tetrahymena pyriformis', and a matrix of predicted values. A 'Save model' dialog box is open, allowing the user to name the model and save it as a QMRF file. A 'Return to matrix' dialog box is also visible, with a 'Select/filter data' section. Red callout boxes numbered 1 through 6 highlight specific actions: 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.

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 3.0.0.1059 alpha interface. The 'Data Gap Filling Method' panel on the left shows three options: 'Read-across', 'Trend analysis', and '(Q)SAR model'. The '(Q)SAR model' option is selected and highlighted with a red circle and a callout '3'. The 'Target Endpoint' panel shows 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h' and 'Protozoa ciliophora ciliates Tetrahymena pyriformis'. The 'Relevant (Q)SAR models' panel shows '<< CREATE A NEW QSAR >>' and 'IGC50 Tetrahymena', which is also highlighted with a red circle and a callout '3'. The 'Filter endpoint tree...' panel shows a tree structure with 'Tetrahymena pyriformis' selected, with a callout '1' pointing to the 'M: 145 mg/L' and 'I: 233(64.4,845) mg/L' values. The 'Data matrix' table shows the following data:

Structure	1 [Target]	2	3	4
EC50 (10/26)				
IGC50				
48 h				
Protozoa				
Ciliophora				
Ciliata				
Tetrahymena pyriformis (7/172)	M: 145 mg/L I: 233(64.4,845) mg/L		M: 152 mg/L	M: 59.4 mg/L
LOEC (2/7)				
MATC (2/6)				
NOEC (8/18)				
Undefined Endpoint (3/12)				
Growth Curve (1/1)				
Growth Inhibition (13/58)				
Growth Rate (33/103)				
Growth Rate - Average Specific Growth Rate (1/2)				

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

QSAR Toolbox 3.0.0.1058 alpha [Document_1]

Training set of: IGC50 Tetrachymene

1 CAS# 66-25-1 Ecotoxicological Informatic	2 CAS# 66-77-3 Ecotoxicological Informatic	3 CAS# 67-36-7 Ecotoxicological Informatic	4 CAS# 111-71-7 Ecotoxicological Informatic	5 CAS# 123-05-7 Ecotoxicological Informatic
6 CAS# 123-72-8 Ecotoxicological Informatic	7 CAS# 100-52-7 Ecotoxicological Informatic	8 CAS# 96-17-3 Ecotoxicological Informatic	9 CAS# 104-88-1 Ecotoxicological Informatic	10 CAS# 110-62-3 Ecotoxicological Informatic
11 CAS# 123-15-9 Ecotoxicological Informatic	12 CAS# 500-22-1 Ecotoxicological Informatic	13 CAS# 555-16-8 Ecotoxicological Informatic	14 CAS# 590-86-3 Ecotoxicological Informatic	15 CAS# 613-45-6 Ecotoxicological Informatic

Save to smi Search OK

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

QSAR TOOLBOX

Data Gap Filling Visualisation of model domain

The screenshot shows the QSAR Toolbox 3.0.0.1058 alpha interface. The 'Data Gap Filling' method is selected. The 'Domain Boundaries Browser' is open, showing a tree structure of models. A right-click context menu is visible over the 'IGC50' model, with 'Display Domain' selected. The 'In Domain' section shows a network of chemical structures. The 'Metabotom' section is also visible.

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

Data Gap Filling

Visualisation of model domain

The screenshot displays the 'Domain Boundaries Browser' interface. The main area shows a network of domain boundaries represented by nodes and logical connectors (AND, NOT). A specific node is highlighted with a red circle and labeled '1'. Other nodes are labeled '2' and '3'. On the left, a chemical structure is shown in a green box, indicating it is within the domain. Below the diagram, there are settings for 'Metabolism' and 'Simulator'.

1. Note the boundaries of the domain are combined logically; 2. If the chemical answers the query of the domain then the current query is 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' window. On the left, the 'Target' section displays a chemical structure and a green highlighted box. The 'Boundaries' section shows a logic tree with nodes labeled 'AND', 'OR', and 'NOT'. A callout '1' points to the 'Training set' button. Below the logic tree, a list of chemicals is shown with columns for CAS, Name, and SMILES. A callout '2' points to this list. A 'Data points' table is open, showing columns for Endpoint, Value, Original value, Effect, Source, Creation Date, and Exposure F. Type. A callout '3' points to the 'Display data' button above the table.

Endpoint	Value	Original value	Effect	Source	Creation Date	Exposure F. Type
LC50	0,000178 mol/L	0,000178 mol/L	Mortality			
LC50	9,778-8 mol/L	9,778-8 mol/L	Mortality			
NOEC	0,00151 mol/L	0,00151 mol/L	Growth			
EC10	95 %	95 %				
EC50	125 °C	125 °C	121 day C	SARC		
EC10	21,8 Pa in 30 min			SARC		
EC10	4,41	4,41		SARC		
EC10	1,79	1,79		SARC		

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 3.0.0.105B alpha interface. The top navigation bar includes 'Input', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The left sidebar shows 'Data Gap Filling Method' with options for 'Read-across', 'Trend-analysis', and '(U)SAR models'. The central pane shows a 'Model endpoint tree...' with a tree structure including 'EC50', 'EC10', 'EC10 h', 'Protozoa', 'Ciliophora', and 'Ciliata', with 'Tetrahymena pyriformis' selected. The right pane shows a data matrix with columns for 'Target' and chemical structures. A callout '1' points to a cell in the matrix. A callout '2' points to the 'CREATE A New Model' button in the left sidebar. A callout '3' points to the 'Display Domain' option in the model menu.

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 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.1.1:
 - Michael addition
 - No alert found
 - SNAr
- Another requirement is Log Kow to be ≥ 0.3187 and ≤ 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

The screenshot displays the QSAR Toolbox interface. On the left, a chemical structure of a polyphenol is shown in a red circle. Below it, a green image of a leaf is labeled 'Target'. In the center, a flowchart titled 'Out of Domain' shows a path from the chemical structure to a red circle labeled 'Aldehydes (Acute toxicity) [FOUND]'. Below the flowchart, a list of 'Available categories' is shown, with 'Aldehydes (Acute toxicity)' highlighted in red. A blue callout box with the number '1' points to the 'Aldehydes (Acute toxicity)' category in the list.

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

QSAR TOOLBOX

Data Gap Filling Edit QMRF data

The screenshot shows the QSAR Toolbox interface during a 'Data Gap Filling' session. The 'Data Gap Filling Method' sidebar is active, and a context menu is open over the 'Tetraachlymene' model. The main window displays a hierarchical tree of models and a table of predicted values for four targets.

Structure	1 (target)	2	3	4
[-]BEC60 (10/26)				
[-]BEC50				
[-]BEC48 h				
[-]Protozoa				
[-]Ciliophora				
[-]Ciliata				
[-]Tetraachlymene pyriformis (7/172)	M: 145 mg/L T: 233(64,4,845) mg/L		M: 152 mg/L	M: 59,4 mg/L
[-]BLOEC (2/7)				
[-]BMATC (2/6)				
[-]BNOEC (8/18)				
[-]Undefined Endpoint (3/12)				
[-]EGrowth Curve (1/1)				
[-]EGrowth Inhibition (13/39)				
[-]EGrowth Rate (33/103)				
[-]EGrowth Rate - Average Specific Growth Rate (1/2)				
[-]EGrowth Rate 6.4 (1/1)				
[-]EGrowth Rate and Biomass (2/6)				
[-]EGrowth Rate, Biomass and Yield (1/4)				
[-]EGrowth Rate, Cell Density, Biomass (1/1)				
[-]EGrowth Rate, Cell Yield (1/2)				
[-]Growth Rate, Yield, Biomass (1/3)				
[-]Immunotoxicology (3/11)				
[-]Immunotoxicology (8/9)				
[-]Immunotoxicology (2/4)				
[-]Inhibition of Reproduction (1/3)				
[-]Injury (1/3)				
[-]Intoxication (12/71)	M: 36 mg/L, 25 mg/L...			

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

Data Gap Filling Perform prediction

The screenshot shows the QSAR Toolbox software interface. The 'Data Gap Filling' tab is selected. An 'Information' dialog box is open, displaying 'Predicted 283 out of 524 structures.' and an 'OK' button. A red box highlights the status bar at the bottom of the window, which shows '524 Aldehydes (Acute toxicity) Strict (US-EPA New Chemical Categories) QSAR predict chemicals in domain'. A red circle highlights the 'OK' button in the dialog box. A blue callout box with the number '1' points to the status bar, and another blue callout box with the number '2' points to the 'OK' button.

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 with the following elements:

- 1:** A red circle highlights the selected nodes in the tree view: Growth, IGCS0, IGCS0 48 h, Protozoa, Ciliophora, and Tetrachymena pyriformis.
- 2:** A callout points to the folder icon in the 'Export to...' dialog box.
- 3:** A callout points to the file name 'IGCS0 Tetrachymena...' in the 'Save As' dialog box.
- 4:** A callout points to the 'Save' button in the 'Save As' dialog box.
- 5:** A callout points to the 'Start' button in the 'Export to...' dialog box.
- 6:** A callout points to the 'OK' button in the 'Export completed successfully' dialog box.

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.

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.