

OECD QSAR Toolbox v.3.4

Step-by-step example of how to predict acute aquatic toxicity to *Tetrahymena pyriformis* by trend analysis using category pruning capabilities

Outlook

- **Background**
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- Workflow
- Save the prediction result

Background

- This is a step-by-step presentation designed to take you through the workflow of the Toolbox in a data-gap filling exercise using trend analysis based on a category formed with data pruning.
- If you are a novice user of the Toolbox you will want to review the “Getting Started” document available at [[http://www.oecd.org/chemicalsafety/risk-assessment/TB3%20 GettingStarted_rev2.pdf](http://www.oecd.org/chemicalsafety/risk-assessment/TB3%20GettingStarted_rev2.pdf)] as well as go through tutorials 1 - 3.

Outlook

- Background
- **Objectives**
- Specific Aims
- Trend analysis
- The exercise
- Workflow
- Save the prediction result

Objectives

- **This presentation demonstrates a number of functionalities of the Toolbox :**
 - Entering a target chemical by SMILES notation and Profiling
 - Identifying analogues for a target chemical by molecular similarity
 - Retrieving experimental results available for those analogues, and for multiple endpoints
 - Filling data gaps by trend analysis

Outlook

- Background
- Objectives
- **Specific Aims**
- Trend analysis
- The exercise
- Workflow
- Save the prediction result

Specific Aims

- To reacquaint the user with the workflow and the six modules of the Toolbox.
- To reacquaint the user with the basic functionalities within each module.
- To introduce the user to new functionalities of selected modules.
- To explain to the rationale behind each step of the exercise.

Outlook

- Background
- Objectives
- Specific Aims
- **Trend analysis**
- The exercise
- Workflow
- Save the prediction result

Trend analysis

- For a given (eco)toxicological endpoint, the members of a category are often related by a trend (e.g., increasing, decreasing or constant) in potency.
- The trend could be related to molecular mass, carbon chain length, or to some other property.
- A demonstration of consistent trends in the toxicity of a group of chemicals is one of the desirable attributes of a chemical category and one of the indicators a common mechanism of action for all chemicals in the category.
- When some chemicals in a category have measured values and a consistent trend is observed, missing values can be estimated by simple scaling from the measured values to fill in the data gaps.

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- **The exercise**
- Workflow
- Save the prediction result

The Exercise

- In this exercise we will predict the toxicity towards the *Tetrahymena pyriformis*, of the substance, 4-Methylhexanal [CAS# 41065-97-8], which will be the “target” chemical.
- This prediction will be accomplished by collecting experimental results for a set of chemicals considered to be in the same category as the target molecule.
- The category will initially be defined based on a protein binding mechanism.
- The initial category will be pruned via subcategorisation.
- Trend-analysis will be used for data gap filling.

Side-Bar On *Tetrahymena*

Tetrahymena pyriformis is a free-living ciliated protozoan that is found in fresh-water ecosystems throughout the world.

It is a well-studied genus with understood nutrient requirements, growth characteristics, and cell biology and genomics.

The ability to culture it axenically in simple media, coupled with its rapid doubling time of makes it a good test organism.

Strain GL-C is amiconuclated and thereby genetically extremely stable.

Side-Bar On TETRATOX

Population growth impairment testing with *Tetrahymena pyriformis* is conducted in a static assay uses population density quantified spectrophotometrically as its endpoint.

The endpoint 50% inhibitory growth concentration from this assay is one of the largest aquatic toxicity databases in the Toolbox.

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
- Save the prediction result

Workflow

- **By now you should be familiar with the six modules which are used in a sequential workflow within the Toolbox. These are:**
 - Chemical Input
 - Profiling
 - Endpoints
 - Category Definition
 - Filling Data Gaps
 - Report

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - **Chemical Input**

Chemical Input Overview

- This module provides the user with several ways of entering the target chemical into the Toolbox.
- This is important because all subsequent functions are based on chemical structure.
- The goal of this module is to make sure the molecular structure assigned to the target chemical is the correct one.

Chemical Input

Ways of Entering a Chemical

User Alternatives for Chemical ID:

A. Single target chemical

- Chemical Name
- Chemical Abstract Services (CAS) number (#)
- SMILES (simplified molecular information line entry system) notation/InChi
- Drawing chemical structure
- Select from User List/Inventory/Databases
- Chemical IDs such as EC number, Einecs number

B. Group of chemicals

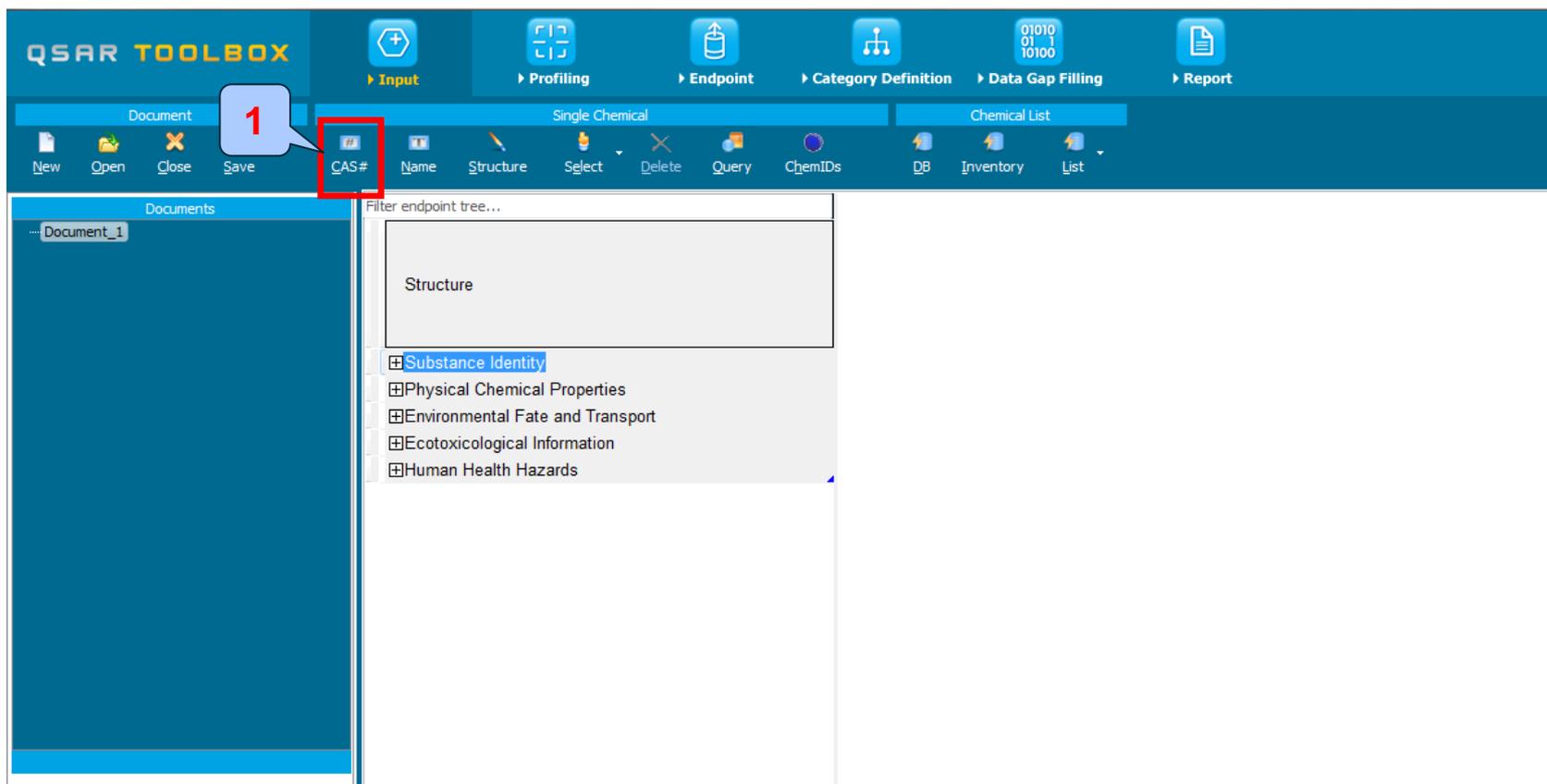
- User List/Inventory
- Specialized Databases

Chemical Input

- Turn on the computer and open the Toolbox. This will take some time as the databases will be populated.
- Note that the Toolbox opens to the first Module “Input” (see next screen shot).
- Since we have described this modules in the earlier tutorials, we will go through it rapidly.

Chemical Input Screen

Input target chemical by CAS#



1. Click on CAS# and a dropdown box will appear (see next screen shot)

Chemical Input Screen

Enter CAS# 41065-97-8 of 4-Methylhexanal

Search by CAS #

41065-97-8 Tautomeric sets

Select All Clear All **1** Invert Selection Selected **2** 1 **3**

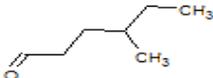
Selected	CAS	Smiles	Depiction	Names	CAS/Name	2D/Name	CAS/2D
1. Yes	41065-97	CCC(C)C		1: 1:: Low (1:: Low (: Low (1: 1:: M 1: 1:: M		

1. **Type** the CAS# In the field; 2. **Click** Search button; 3. **Press** OK

Chemical Input

Target chemical identity

The Toolbox searches the databases to find out if the CAS# you entered is linked to a molecular structure stored in the Toolbox. It is displayed as a 2D image.

Selected	CAS	Smiles	Depiction	Names	CAS/Name	2D/Name	CAS/2D
1. Yes	41065-97-8	CCC(C)C		1: 1:: Low (1:: Low (: Low (1:: M 1:: M			



In case a structure has several CAS numbers or a structure could be related to more than one substance 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

- **Double left click** "Substance Identity"; this displays the chemical identification information (see next screen shot).
- Note that existing in the Toolbox names of target chemical are in different colours.
- The workflow on the first module is now completed, and the user can proceed to the next module (see next screen shot).

Chemical Input

Target chemical identity

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes options for Document, Single Chemical, and Chemical List. The main workspace is divided into several panels:

- Documents Panel:** Shows a document titled "CAS: 41065-97-8".
- Filter endpoint tree...:** A search box contains "1 [target]". Below it, a chemical structure is shown.
- Substance Identity Panel:** A tree view lists various properties:
 - CAS Number: **41065-97-8** (highlighted with a red circle)
 - Chemical IDs: NA
 - Chemical Name: **hexanal, 4-methyl-** (highlighted with a red circle)
 - Molecular Formula: C7H14O
 - Structural Formula: CCC(C)CCC=O
- Physical Chemical Properties, Environmental Fate and Transport, Ecotoxicological Information, and Human Health Hazards:** These sections are currently collapsed.

At the bottom of the interface, the status bar shows "1 Document" and "1/0/0".

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - Chemical Input
 - **Profiling**

Profiling

Overview

- As in the other tutorials, “Profiling” refers to the electronic process of retrieving relevant information on the target compound stored in the Toolbox database, other than fate and toxicity data.
- Available information includes likely mechanism(s) of action, as well as observed or simulated metabolites.

Profiling Side-Bar to Profiling

Summary information of the different profilers is provided in the “**About**”.

1. Select the name of the profiler and perform a **right click** on it;

2. Select **About**;

3. After reading the material **click** on **Close**.

Details	
Version	1.4
Adopted	QSAR Toolbox 2.0, October 2010
Number of categories	146
Number of help files	128

Profiling

Profiling the target chemical

- **Click** in the box next to the Profiling methods related to the target endpoint
- This selects (a **green** check mark appears) or deselects (**green** check mark disappears) profilers
- For this example **check** all the mechanistic methods.

Profiling

Profiling the target chemical

The screenshot displays the QSAR Toolbox software interface during the Profiling step. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Profiling' section is active, showing a list of 'Profiling methods'. The 'General Mechanistic' category is expanded, and all 30 items are checked. A red circle highlights the 'Apply' button in the top-left corner of the Profiling methods panel, with a callout bubble containing the number '2'. Another callout bubble with the number '1' points to the 'General Mechanistic' section. The right-hand side of the interface shows the 'Structure' panel for the target chemical, displaying its chemical structure and a table of properties:

Property	Value
CAS Number	41065-97-8
Chemical IDs	NA
Chemical Name	hexanal, 4-methyl-
Molecular Formula	C7H14O
Structural Formula	CCC(C)CCC=O

At the bottom of the interface, a blue box contains the following instructions:

1. Check all the General Mechanistic profilers; **2. Click** Apply

Profiling

Profiling the target chemical

- The actual profiling will take several seconds depending on the number and type of profilers selected.
- The results of profiling automatically appear as a dropdown box under the target chemical (**see next slide**)
- Please note the specific protein-binding profiler – Protein binding by OASIS v1.4 - background information can be retrieved by double click on the box with Protein binding by OASIS v1.4 result (**see next slide**)
- This result will be used to search for suitable analogues in the next steps of the exercise.
- Move to next module “Endpoint”.

Profiling

Profiling the target chemical

The screenshot shows the QSAR Toolbox software interface. The main window displays a list of profiling methods, including 'Protein binding by OASIS v1.4'. A callout box with the number '1' points to this entry. A 'Details' button is visible next to the entry. A second callout box with the number '2' points to the 'Details' button. The 'Profiling results' window is open, showing a chemical profile for 'Protein binding by OASIS v1.4'. A red box highlights the 'Aldehydes' category under 'Schiiff base formation >> Schiff base formation with carbonyl compounds >> Aldehydes'.

1. **Double click** on Protein binding by OASIS profiling result
2. **Click Details** to see the explanation of **Aldehydes** category

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - Chemical Input
 - Profiling
 - **Endpoints**

Endpoints

Overview

- As you may remember, “Endpoints” refer to the electronic process of retrieving the measured data for fate and toxicity that are stored in the Toolbox database.
- Data gathering can be executed in a global fashion (i.e., collecting all data of all endpoints) or on a more narrowly defined basis (e.g., collecting data for a single or limited number of endpoints).

Endpoints

Case study

- We will gather data for acute aquatic toxicity endpoints
- For this purpose, data are collected from all Ecotox
- **Click** on "Select All" from Ecotoxicological part
- **Click** on "Gather data" (see next screen shot).

Endpoints

Gather data

The screenshot shows the QSAR Toolbox software interface. At the top, there is a menu bar with icons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. Below this is a toolbar with buttons for Data, Import, Export, Delete, and Tautomerize. The 'Gather' button is circled in red and labeled with a '2'. In the 'Databases' panel on the left, the 'Ecotoxicological Information' section is expanded, and all its sub-items (Aquatic ECETOC, Aquatic Japan MoE, Aquatic OASIS, ECHA CHEM, and ECOTOX) are checked, labeled with a '1'. The main window displays a 'Filter endpoint tree...' dialog with a search box containing '1 [target]'. The 'Structure' field shows a chemical structure of 1,1,1-trifluorobutane. Below the structure, a list of endpoint categories is shown with checkboxes: Substance Identity, Physical Chemical Properties, Environmental Fate and Transport, Ecotoxicological Information, and Human Health Hazards.

1. **Select** all databases from Ecotoxicological Information section.
2. **Click** Gather

Endpoints

Case study

- Aquatic Toxicity information on the target chemical is electronically collected from the selected dataset(s).
- In this example, an insert window appears stating there was “no data found” for the target chemical .
- Close the insert window by **right clicking** on “OK”.

Endpoints Recap

- The workflow in this exercise is the same as you have observed with the preceding tutorials.
- You have entered the target chemical being sure of the correct structure.
- You have profiled the target chemical.
- You have found that no experimental data is currently available for this structure.
- In other words, you have identified a data gap, which you would like to fill.
- Proceed by **right clicking** on “Category definition” which move you to the next module.

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - Chemical Input
 - Profiling
 - Endpoint
 - **Category definition**

Category Definition Overview

- As before, this module provides the user with several means of grouping chemicals into a category that includes the target molecule.
- Remember this is the critical step in the workflow.
- Several options are available in the Toolbox to assist the user in refining the category definition. We will use several of these options in this tutorial.

Category Definition

Grouping methods

- You may want to review the information about grouping chemicals which is in (Chapter 4) on the following link.
<http://www.oecd.org/dataoecd/58/56/46210452.pdf>
- Re-examine slide 31-32; you see that the target chemical (4-Methyl hexanal) could react with proteins via Schiff base formation and thus it has a potential for exhibiting aquatic toxicity in excess of baseline potency.
- The reaction by which a target chemical binds with proteins is relevant to deriving a group of chemicals that may mechanistically act the same way to elicit aquatic toxicity.

Category Definition

Protein binding by OASIS v1.4 grouping method

- Protein binding is one of the best grouping methods in the Toolbox. It is based on conventional organic chemical mechanisms and reactions, and as such is qualitative in character.
- This grouping method is particularly relevant for respiratory and skin sensitization and acute aquatic toxicity, but also relevant to chromosomal aberration and acute inhalation toxicity.

Category Definition

Side bar to Protein binding by OASIS v1.4 grouping method

- Protein binding by OASIS v1.4 includes 617 chemical categories.
- Each category is represented by defined 2-dimensional structural alerts that are associated with chemicals that act as electrophiles and covalently react with various moieties, in particular thiol (-SH) and amino (-NH₂) groups in proteins.
- Therefore, there is a sound mechanistic basis for using this grouping method.

Category Definition

Defining Protein binding by OASIS v1.4 category

The screenshot displays the QSAR Toolbox software interface. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Category Definition' tab is active, showing a 'Filter endpoint tree...' with '1 [target]' and a chemical structure. A list of 'Grouping methods' is visible on the left, with 'Protein binding by OASIS v1.4' highlighted. A dialog box titled 'Protein binding by OASIS v1.4' is open, showing a list of target profiles, with 'Schiff base formation' and its sub-profiles circled. The dialog also includes options for 'Combine profiles logically' (AND/OR), 'Invert result', and 'Strict', along with 'OK' and 'Cancel' buttons.

1. Highlight the "Protein binding by OASIS v1.4"; 2. Click Define; 3. Click OK to confirm the category for target chemical.

Category Definition

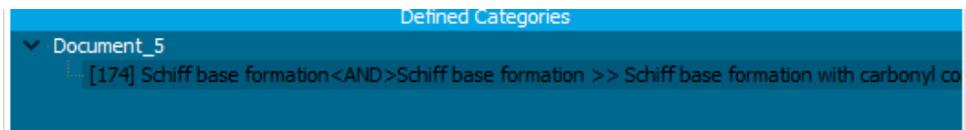
Defining Protein binding by OASIS v1.4 category

The screenshot displays the QSAR Toolbox software interface. The top toolbar includes buttons for 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Category Definition' button is highlighted. Below the toolbar, the 'Define' button is also highlighted. The main window shows a 'Filter endpoint tree...' with a search for '1 [target]' and a chemical structure. A list of 'Grouping methods' is visible on the left, including 'Protein binding by OASIS v1.4'. A dialog box titled 'Define category name' is open, showing a search result for 'Aldehydes (Protein binding by OASIS v1.4)' with 174 chemicals. A blue callout box with the number '1' points to the 'OK' button in the dialog.

1. Click OK to confirm the name of the category.

Category Definition Analogues

- The data is automatically collected.
- Based on the defined category (Schiff base formation<AND>Schiff base formation >> Schiff base formation with carbonyl compounds<AND>Schiff base formation >> Schiff base formation with carbonyl compounds >> Aldehydes) 174 analogues* (plus the target chemical) have been identified.
- These 174 compounds along with the target chemical form a category, which can be used for data filling (see next screen shot).

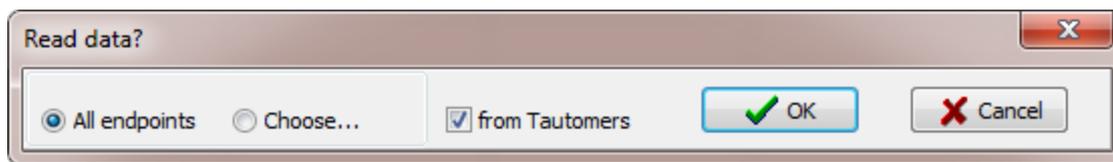


- The name of the category appears in the "Defined Categories" window, indicating the number of substances belonging to the category.

Category Definition

Read data for Analogues

- 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 below).



Category Definition

Read data for Analogues

Due to the overlap between the Toolbox databases same data for intersecting chemicals could be found simultaneously in more than one databases. 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: 2569 data-points, 911 groups, 681 chemicals

	Endpoint	CAS	Structure	Value	Abnormality
<input checked="" type="checkbox"/>		90-02-8	<chem>O=C1C=CC=C1</chem>	-7 °C	
<input checked="" type="checkbox"/>		90-02-8	<chem>O=C1C=CC=C1</chem>	-7 °C	
<input checked="" type="checkbox"/>		123-15-9	<chem>CCOC(=O)CC</chem>	-100 °C	
<input checked="" type="checkbox"/>		123-15-9	<chem>CCOC(=O)CC</chem>	-100 °C	
<input checked="" type="checkbox"/>		123-15-9	<chem>CCOC(=O)CC</chem>	-100 °C	
<input checked="" type="checkbox"/>		104-55-2	<chem>C=CC1=CC=CC=C1</chem>	-7.5 °C	
<input checked="" type="checkbox"/>		104-55-2	<chem>C=CC1=CC=CC=C1</chem>	-7.5 °C	
<input checked="" type="checkbox"/>		89-98-5	<chem>ClC1=CC=CC=C1</chem>	10;12 °C	

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

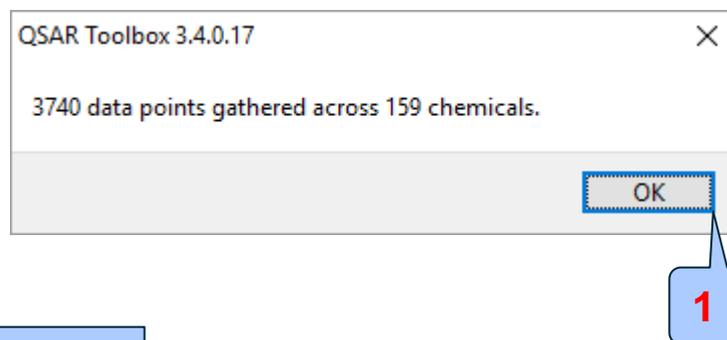
Annotations: 1 (pointing to 'Select one'), 2 (pointing to 'OK')

1. **Click** Select one and then 2. **Click** OK

Category Definition

Read data for Analogues

The system automatically gives indication for the number of gather experimental data points



1. **Click** OK

Category Definition

Summary information for Analogues

- The experimental results for the analogues are inserted into the data matrix.

The screenshot shows the QSAR Toolbox software interface. The main window displays the 'Category Definition' workflow. The 'Filter endpoint tree...' window is open, showing a list of chemical structures and a data matrix. A red circle highlights the value '(159/3740)' in the matrix, and a red box highlights the entire data matrix area. A blue callout box at the bottom explains that this represents chemical statistics.

Chemical statistics presenting the number of chemicals and the available experimental data.

Category Definition

Side bar of experimental data

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. Below the menu are icons for 'Define', 'Define with metabolism', 'Subcategorize', 'Combine', 'Clustering', 'Delete', and 'Delete All'. The left sidebar lists various grouping methods and endpoint-specific categories. The main area shows a 'Structure' view and a table of data points.

#	Endpoint	Value	Original value	Effect	Reliability
1	LC50	12 mg/L	0.000105 mol/L	Mortality	
2	LC50	8.86 mg/L	7.76E-5 mol/L	Mortality	
3	IGC50	114 mg/L	0.001 mol/L	Growth	
4	pT	16.5 mg/L	0.000145 mol/L	Physiology	
5	LC50	12 mg/L	12 mg/L	mortality	2 (reliable with
6	FC50	4.99 mmol	4.99 mmol	mbhilitv	1 (reliable without

1. Double-click on the cell with measured data to see detailed information.

Category Definition

Recap

- You have identified a mechanistic category by Protein binding by OASIS v1.4 for the target chemical (4-Methylhexanal).
- The available experimental results for these 174 analogues are collected from the previously selected databases.
- The user can then proceed to the module “Data gap filling”, but before that the user should navigate through the endpoint tree and find the gap that will be filled in.

Category Definition

Navigation through the endpoint tree

- You can proceed to the module “Data gap filling”, but before that we will navigate through the endpoint tree and find the gap that will be filled.
- As before we navigate through the data tree by closing or opening the nodes of the tree.
- In this example, the IGC50-48h Growth for *Tetrahymena pyriformis* is the target endpoint (see next screen shot).

Category Definition

Navigation through the endpoint tree

QSAR Toolbox 3.4.0.17 [Document_5]

Workflow: Input → Profiling → Endpoint → **Category Definition** → Data Gap Filling → Report

Sub-workflow: Define → Define with metabolism → Subcategorize → Combine → Clustering → Delete → Delete All

Grouping methods:

- DNA binding by OASIS v.1.4
- DNA binding by OECD
- DPRA Cysteine peptide depletion
- DPRA Lysine peptide depletion
- Estrogen Receptor Binding
- Hydrolysis half-life (Ka, pH 7)(Hydrowin)
- Hydrolysis half-life (Kb, pH 8)(Hydrowin)
- 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 v1.4
- Protein binding by OECD
- Protein binding potency
- Superfragments
- Toxic hazard classification by Cramer
- Toxic hazard classification by Cramer 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

Defined Categories: Document_5

Structure	1 [target]	2	3	4	5	6	7
Cell Proliferation Inhibition	(1/1)						
Cell Yield	(1/4)						
Cell(s)	(1/8)						
Development	(13/40)						
Dissolved Oxygen Production	(1/1)						
Effect	(1/1)						
Enzyme(s)	(4/18)						
Feeding Behavior	(1/1)						
Gas Production	(1/2)						
Genetics	(3/10)						
Growth	(9/21)						
EC50	(9/21)						
I50							M: 26 mg
IGC50							
48 h							
Protozoa							
Ciliophora							
Ciliatea							
Tetrahymena pyriformis	(68/68)	M: 152 mg/L	M: 59.4 mg/L	M: 114 mg/L	M: 88.7 mg/L	M: 194 mg/L	
LOEC	(4/8)						
LOEL	(1/6)						
MATC	(3/4)						

1. Double click on the specific nodes to reach the leaf of the target endpoint – Ecotoxicological information; Aquatic toxicity; Growth; IGC50; 48h; Protozoa; Ciliophora; Ciliatea; Tetrahymena pyriformis.

Category Definition

Navigation through the endpoint tree with the filter box

- Remember you can also navigate through the endpoint tree by typing the endpoint of choice in the filter box and then double clicking through the endpoint tree to IGC50-48h Growth for *Tetrahymena pyriformis* (see next screen shot).

Category Definition

Navigation through the endpoint tree - use filter

The screenshot shows the QSAR Toolbox interface with the 'Category Definition' workflow selected. A callout '1' points to the 'Filter' input box where 'igc' is entered. Another callout '2' points to the 'Immobilisation' endpoint in the tree, which is highlighted in blue. The main window displays a table of chemical structures and their corresponding immobilisation values.

Structure	1 [target]	2	3	4	5	6	7

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - Chemical Input
 - Profiling
 - Endpoint
 - Category definition
 - **Data Gap Filling**

Data Gap Filling Overview

- **“Data Gap Filling” module give access to three different data gap filling tools:**
 - Read-across
 - Trend analysis
 - (Q)SAR models
- **Depending on the situation, the most relevant data gap mechanism should be chosen, taking into account the following considerations:**
 - Read-across is the appropriate data-gap filling method for “qualitative” endpoints like skin sensitization or mutagenicity for which a limited number of results are possible (e.g. positive, negative, equivocal). Furthermore read-across is recommended for “quantitative endpoints” (e.g., 96h-LC50 for fish) if only a low number of analogues with experimental results are identified.
 - Trend analysis is the appropriate data-gap filling method for “quantitative endpoints” (e.g., 96h-LC50 for fish) if a high number of analogues with experimental results are identified.
 - (Q)SAR models can be used to fill a data gap if no adequate analogues are found for a target chemical.
- **In our case trend analysis is applied**

Data Gap Filling

Data Gap window

1. Click on Data gap filling.

The screenshot shows the QSAR Toolbox interface. The top navigation bar includes buttons for Input, Profiling, Endpoint, Category Definition, Data Gap Filling (highlighted with a red callout '1'), and Report. The left sidebar shows the 'Data Gap Filling Method' and 'Target Endpoint' sections. The main workspace displays a tree view of 'Substance Identity' and 'Ecotoxicological Information' with 'IGC50' selected. Below the tree is a table with columns for '1 (target)', '2', '3', '4', '5', '6', and '7'. The table contains chemical structures and data for various substances, including molecular weights (M: 152 mg/L, M: 59.4 mg/L, M: 1).

Data Gap Filling

Apply Trend analysis

The screenshot shows the QSAR Toolbox interface. The top navigation bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The left sidebar has 'Data Gap Filling Methods' with 'Read-ones', 'Trend analysis', and '(Q)SAR models'. The 'Target Endpoint' section lists 'Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliata Tetrahymena pyriformis'. The main workspace shows a tree structure for 'IGC50' with sub-nodes for 'Growth', '48 h', 'Protozoa', 'Ciliophora', 'Ciliata', and 'Tetrahymena pyriformis'. A table below the tree shows data for 'M: 152 mg/L' and 'M: 59.4 mg/L'. A callout '1' points to the 'IGC50' endpoint box, callout '2' points to 'Trend analysis', and callout '3' points to the 'Apply' button.

1. Highlight the data endpoint box corresponding to *Tetrahymena pyriformis*/IGC50/48h under the target chemical; **2. Select** Trend analysis; **3. Click** Apply.

Data Gap Filling

Interpreting Trend analysis

- The resulting plot outlines the experimental IGC50 results of all analogues (Y axis) according to a descriptor (X axis) with Log Kow being the default descriptor (see next screen shot).
- The **RED** dot represents the estimated result for the target chemical.
- The **BLUE** dots represent the experimental results available for the other analogues.
- The **GREEN** dots (which we will see later) represent analogues belonging to different subcategories.
- Before accepting the estimated result for the target chemical, the trend analysis should be further refined by subcategorisation.

Data Gap Filling

Results of Trend analysis

QSAR TOOLBOX

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

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

Data Gap Filling Method

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

Target Endpoint

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

IGC50 [1 (target)] [3] [4] [7] [8] [9] [14]

Structure

Tetrahymena pyriformis (68/68)

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 68 values from 68 analogue chemicals, Observed target value: N/A, Predicted target value: 26.2 mg/L, Model equation: IGC50 = +2.64 +0.450 * 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

756 Schiff base formation<AND>Schiff base formation >> Schiff base formation with carbonyl compo Create prediction by oao filling 0/1 1/10

Data Gap Filling

Side Bar of Subcategorization

- Subcategorisation refers to the process of applying additional profilers to a previously defined category.
- The goal here is to identify chemicals which have differing profiling results, eliminating these chemicals from the category, and eventually getting a more homogenous chemical category.
- In this example, two subcategorisations are applied to prune the analogues.

Data Gap Filling Subcategorization

- Due to polyfunctionality of molecules, there are analogues which may undergo protein binding reactions different from those of the target compound.
- In addition, there are analogues which contain organic functional group that differ from those of the target compound.
- These analogues can be identified and later removed via subcategorisation.

Data Gap Filling

Subcategorization 1 (Protein binding by OECD)

The screenshot displays the QSAR Toolbox interface for subcategorization. The 'Data Gap Filling' tab is active, showing a list of chemical structures and their predicted values. A scatter plot titled 'Trend analysis prediction of IGC50, making a linear approximation, based on 68 values from 68 analogue chemicals, Observed target value: N/A, Predicted target value: 26.2 mg/L, Model equation: $IGC50 = +2.64 + 0.450 * \log Kow$ ' is shown. The plot has 'log Kow' on the x-axis (ranging from 0.50 to 4.50) and 'IGC50 (mg/L)' on the y-axis (ranging from 2.00 to 5.00). A red trend line is visible. The 'Select/filter data' panel on the right has 'Subcategorize' selected. The 'Grouping methods' sidebar on the left has 'Protein binding by OECD' selected under the 'General Mechanistic' category.

1. Open Select/filter data; 2. Select Subcategorize; 3. From Grouping methods select Protein binding by OECD. The green dots which represent analogues belonging to different subcategories.

Data Gap Filling

Information on an Analogue

- Additional information may be observed for any analogue in the trend analysis.
- Analogues which are different from the target chemical may be removed from the category.
- In this example, we will examine the upper most data point (see the next series of screen shot).

Data Gap Filling

Subcategorization 1 (Protein binding by OECD)

The screenshot displays the 'Subcategorization' window of the OECD QSAR Toolbox. On the left, a list of grouping methods is shown, with 'Protein binding by OECD' highlighted (callout 2). The central window shows the 'Data Gap Filling' process, including a target selection menu (callout 1) and a table of chemical analogues. The table lists several chemical structures with their corresponding predicted values (M) in mg/L. Below the table is a scatter plot titled 'Trend analysis prediction of IGC50, making a linear approximation, based on 68 values from 68 analogue chemicals, Observed target value: N/A, Predicted target value: 26.2 mg/L, Model equation: IGC50 = +2.64 +0.450 * log Kow'. On the right, a 'Selection navigation' panel is visible. At the bottom of the analogue list, a 'Remove' button is highlighted (callout 3).

1. Subcategorization 2. **Select** Protein binding by OECD 3. **Remove** analogues which have different protein binding mechanism than the target chemical.

Data Gap Filling

Subcategorization 1 (Protein binding by OECD)

The screenshot displays the QSAR Toolbox software interface during a 'Data Gap Filling' process. The main window is titled 'Subcategorization' and shows a workflow with steps: Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. The 'Data Gap Filling' step is active, showing a list of chemical structures and their predicted values for a target property (IGC50). The predicted values are: 152 mg/L, 59.4 mg/L, 114 mg/L, 88.7 mg/L, and 194 mg/L. Below the list, a scatter plot shows the trend analysis prediction of IGC50 based on 42 analogue chemicals. The model equation is $IGC50 = +2.13 + 0.588 * \log Kow$. The plot shows a positive correlation between log Kow and IGC50. The interface also includes a sidebar with various grouping methods and a right-hand panel with options for 'Accept prediction' and 'Return to matrix'.

756 Schiff base formation<AND>Schiff base formation >> Schiff base formation with carbonyl compo Create prediction by gap filling

1/1/0

Data Gap Filling

Selecting the Data Type

- The Toolbox allows the user to decide which type of data to be used in cases where more than one result are available for any analogue.
- This can be all values, average values, minimum or maximum results (see next screen shot).
- It should be noted that averaging results is only useful for quantitative endpoints, which is the case in this example.

Data Gap Filling Selecting the Data Type

The screenshot displays the 'Data Gap Filling' module of the OECD QSAR Toolbox. The interface is divided into several sections:

- Left Sidebar:** Lists various grouping methods such as 'Database Affiliation', 'General Mechanistic', and 'Metabolism/Transformations'.
- Top Panel:** Shows the workflow steps: Profiling, Endpoint, Category Definition, Data Gap Filling, and Report.
- Central Workspace:** Displays chemical structures and a table of predicted values for 'Tetrahymena pyriformis'. The table includes columns for predicted values (e.g., 152 mg/L, 59.4 mg/L, 114 mg/L, 88.7 mg/L, 194 mg/L).
- Bottom Center:** A trend analysis plot titled 'Trend analysis prediction of IGC50, making a linear approximation, based on 42 values from 42 analogue chemicals'. The plot shows a scatter of data points with a red regression line. The model equation is $IGC50 = +2.13 + 0.588 * \log Kow$.
- Right Sidebar:** Contains prediction options, including 'Accept prediction', 'Return to matrix', and 'Calculation options'. Two callout boxes are present: '1' points to the 'Calculation options' section, and '2' points to the 'Data usage' section.

1. Open Calculation options; 2. Select Data usage

Data Gap Filling

Selecting the Data Points

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

Tetrahymena pyriformis (42/42)

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 42 values from 42 analogue chemicals, Observed target value: N/A, Predicted target value: 41.7 mg/L,

Model equation: $IGC50 = +2.13 + 0.588 * \log Kow$

Accept prediction

Return to matrix

- Select/filter data
- Selection navigation
- Gap filling approach
- Descriptors/data
- Model/(Q)SAR
- Calculation options
 - Data usage
 - Prediction approach options
 - Set level of significance
- Visual options
- Information

Set usage of chemical:

- All
- Minimal
- Maximal
- Average**
- Median
- Lower median
- Higher median

OK

1. Select type of data to use (in this case Average is selected (the default option)); **2. Click OK**

1/1/0

Data Gap Filling

Subcategorization 2 (Organic functional groups (nested))

The screenshot shows the QSAR Toolbox interface during a subcategorization task. The 'Data Gap Filling' tab is active, displaying a list of chemical structures and their predicted values. A scatter plot shows a trend analysis prediction of IGCS50 based on 42 values from 42 analogue chemicals. The model equation is $IGCS50 = +2.13 + 0.588 * \log Kow$. The observed target value is N/A, and the predicted target value is 41.7 mg/L. A callout box with the number '1' points to the 'Accept prediction' section on the right. Another callout box with the number '2' points to the 'Organic Functional groups (nested)' option in the sidebar. A third callout box with the number '3' points to the 'Remove' button at the bottom of the sidebar.

1. **Select** Subcategorization; 2. From Grouping methods **select** Organic functional groups(nested); 3. **Press** Remove.

Data Gap Filling

Subcategorization 2 (Organic functional groups (nested))

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

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

Apply

Data Gap Filling Method

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

Target Endpoint

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

IGC50

Structure

Tetrahymena pyriformis (14/14)

M: 152 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

Trend analysis prediction of IGC50, making a linear approximation, based on 14 values from 14 analogue chemicals, Observed target value: N/A, Predicted target value: 69.6 mg/L, Model equation: $IGC50 = +1.93 + 0.577 * \log Kow$

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
 - Data usage
 - Prediction approach options
 - Set level of significance
- Visual options
- Information
- Miscellaneous

756. Schiff base formation<AND>Schiff base formation >> Schiff base formation with carbonyl compo: Create prediction by gap filling

1/1/0

Data Gap Filling

Accepting the prediction result

The screenshot shows the QSAR Toolbox interface during the Data Gap Filling process. The main workspace displays a list of chemical structures with their predicted IGC50 values. A callout box labeled '1' points to the 'Accept prediction' button in the right-hand panel. Below this, a callout box labeled '2' points to the 'Return to matrix' button. A central plot shows a trend analysis of IGC50 vs log Kow with a linear regression line and the equation: $IGC50 = +1.93 + 0.577 * \log Kow$.

1. Select Accept prediction and 2. Click Return to matrix.

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- **Workflow**
 - Chemical Input
 - Profiling
 - Endpoint
 - Category definition
 - Data Gap Filling
 - **Reporting**

Report Overview

- Remember, the report module generates reports on any of predictions made with the Toolbox.
- The report module contains predefined report templates as well as a template editor with which users can define their own “user defined” templates.
- The report can then be printed or saved in different formats.

Report

Generate Report

The screenshot shows the QSAR Toolbox interface during the 'Data Gap Filling' process. The left sidebar is set to '(Q)SAR models'. The main panel shows a tree view of 'Ecotoxicological Information' with 'Growth IGC50' selected. Below the tree is a table of predicted values. A context menu is open over the cell containing '69.6(20.8)', with the 'Report' option highlighted. The 'Report' option is circled in red, and a blue callout box with the number '2' points to it. Another blue callout box with the number '1' points to the cell containing the predicted value.

1. **Right click** on the cell which contains the predicted value; 2. **Select** Report from the menu.

Report

Generate Report

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Report' menu is active, showing options like 'Create', 'Print', 'Close', 'Save as', 'Register', 'Unregister', 'Update', 'Clone', and 'Design'. The main window displays a report titled 'Prediction of IGC50 for hexanal, 4-methyl-' with a page number of 1 / 30. The report content includes the title 'QSAR Toolbox prediction for single chemical' and a paragraph of text describing the template's basis on OECD and ECHA guidance documents. The left sidebar shows available data to report (Predictions, QSARs, Categories) and available report templates (Standard and Custom). The status bar at the bottom shows the chemical name '756 Schiff base formation<AND>Schiff base formation >> Schiff base formation with carbonyl compo' and the page number 1/0/0.

Outlook

- Background
- Objectives
- Specific Aims
- Trend analysis
- The exercise
- Workflow
- **Save the prediction result**

Saving the prediction result

- This functionality allow storing/restoring the current state of Toolbox documents including loaded chemicals, experimental data, profiles, predictions etc, on the same computer. The functionality is implemented based on saving the sequence of actions that led to the current state of the Toolbox document and later executing these actions in the same sequence in order to get the same result(s).
- Saving/Loading the file with TB prediction is shown on next screenshots

Saving the prediction result

The screenshot illustrates the steps to save a prediction result in the QSAR Toolbox. The interface shows the 'Input' section selected in the menu bar. A 'Save' button is highlighted in the 'Document' menu. A 'Save As' dialog box is open, allowing the user to define the file name as 'Tutorial 4.tbw' and save it as a 'Toolbox work file (*.tbw)'. The 'Save' button in the dialog box is also highlighted.

1. **Go** to Input section
2. **Click** on Save button
3. **Define** name of the file;
4. **Click** Save button

Open saved file

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'New', 'Open', 'Close', 'Save', 'CAS#', 'Name', 'Structure', 'Select', 'Delete', 'Query', 'ChemIDs', 'DB', 'Inventory', and 'List'. A 'Select file' dialog box is open, showing a list of files in a folder named 'New folder'. The files listed are: Tutorial 4.tbw, Tutorial 5.tbw, Tutorial 9.tbw, Tutorial 11.tbw, and Tutorial 24.tbw. The 'Open' button in the dialog box is highlighted.

1. **Create** new document
2. **Click** Open;
3. **Find** and **select** file;
4. **Click** Open

Open saved file

The screenshot displays the QSAR Toolbox software interface. The main window shows a 'Filter endpoint tree...' with a 'Structure' tab selected. A table of chemical data is visible, with columns for target, molecular weight, melting point, and other properties. A blue callout box with the number '1' points to the 'OK' button on the 'Information' dialog box that says 'The file was executed successfully'.

Filter endpoint tree...	1 [target]	2	3	4	5	6	7
Structure							
Substance Identity							
Physical Chemical Properties	(453/2172)	M: 201.202 °C, 1.9...	M: 131 °C, 4.41, 1...	M: 292 °C, 33.5 °C	M: 270 °C, 1.91, 5...	M: 129 °C	M: 2...
Environmental Fate and Transport	(111/457)		M: Not ready, 21.6...				M: R
Ecotoxicological Information	(163/4409)	T: 69.6(20.8;232) ...	M: 17.8 mg/L, 9.79...	M: 59.4 mg/L, >10...			M: 12
Human Health Hazards			M: Negative, Negat...	M: 666 mg/kg			M: C
Profile							
General Mechanistic							

1. The file is opened successfully 1. **Click OK**

Congratulation

- By now you should feel comfortable with the six basic modules of the toolbox and how they form the workflow of the Toolbox.
- In this tutorial you have been introduced to several additional function in the Toolbox.
- You have used different profilers in subcategorizing the initial category of the target chemical.
- Remember proficiency in using the Toolbox comes with practice.