

OECD QSAR Toolbox v.4.1

Step-by-step example for building QSAR model

Outlook

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

Background

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

Outlook

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

Objectives

- **This presentation demonstrates building a QSAR model for predicting acute toxicity of aldehydes to *Tetrahymena pyriformis*. The presentation addresses specifically:**
 - predicting acute toxicity for a target chemical;
 - building a 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 98-01-1;
 - 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**

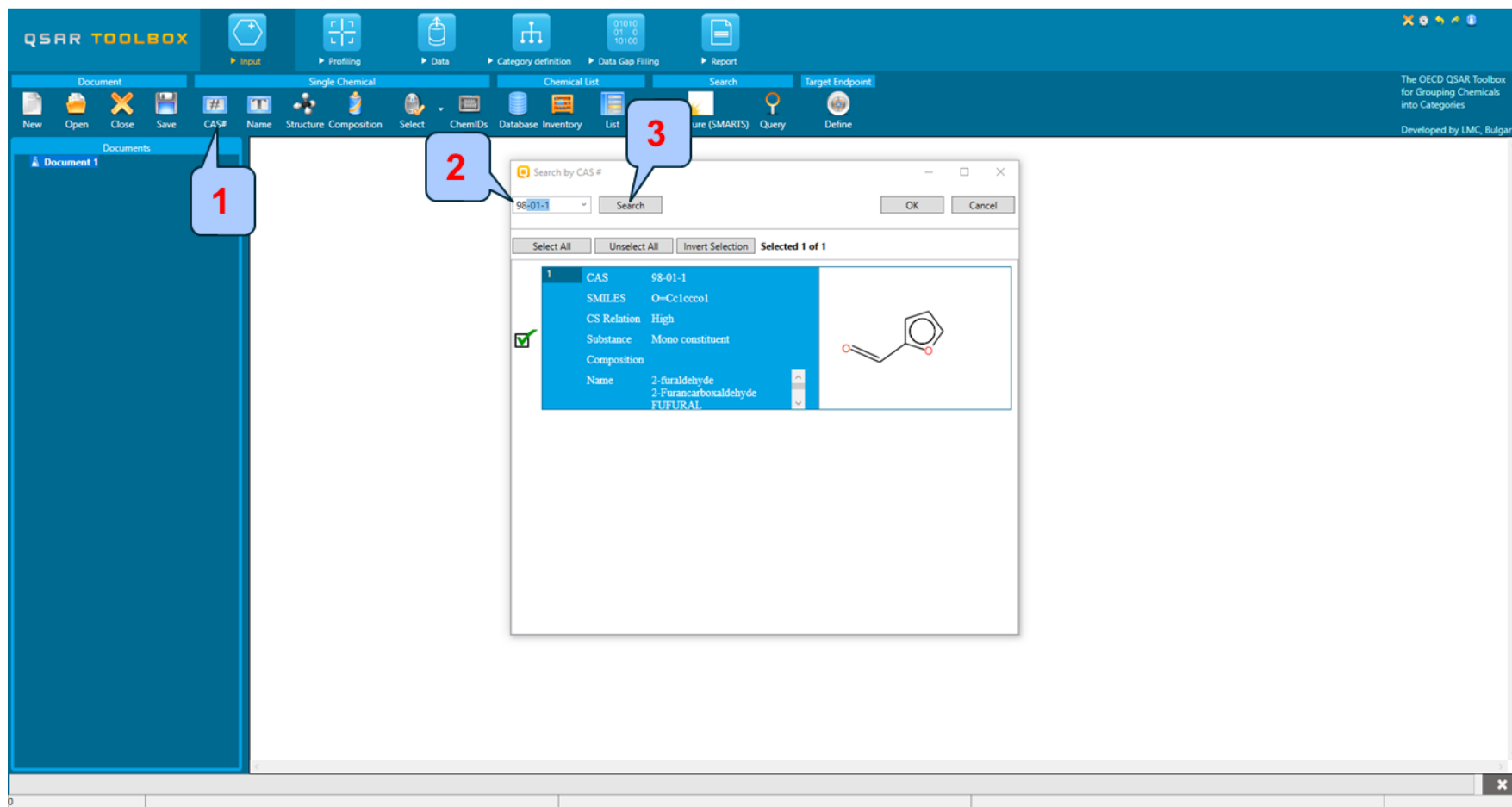
Workflow of the exercise

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

Outlook

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

Input



The screenshot shows the QSAR Toolbox software interface. The main window has a menu bar and a toolbar. A search dialog box is open, showing the search results for CAS# 98-01-1. The dialog box has a search bar with the text "Search by CAS #" and a "Search" button. Below the search bar, there are buttons for "Select All", "Unselect All", and "Invert Selection". The search results are displayed in a table with the following columns: CAS, SMILES, CS Relation, Substance, Composition, and Name. The results show that CAS# 98-01-1 corresponds to 2-furaldehyde, with the SMILES string O=Cc1ccco1 and the chemical structure of 2-furaldehyde.

CAS	SMILES	CS Relation	Substance	Composition	Name
98-01-1	<chem>O=Cc1ccco1</chem>	High	Mono constituent		2-furaldehyde 2-Furanaldehyde FUFURAL

1. **Click** on CAS# 2. **Enter** CAS# 98-01-1; 3. **Click** Search;

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. Note it is unselected by default.

The screenshot shows the QSAR Toolbox software interface. The 'Search by CAS #' dialog box is open, displaying search results for CAS 98011. The results table shows the following information:

1	CAS	98-01-1
	SMILES	O=Cc1ccoc1
	CS Relation	High
	Substance	Mono constituent
	Composition	
	Name	2-furaldehyde 2-Furanicarboxaldehyde FUFURAL

A chemical structure of 2-furaldehyde is displayed next to the results. Callout 1 points to the checkbox next to the result, and callout 2 points to the OK button.



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 to be retained for the subsequent workflow.

1. **Mark** desired chemical (in case there is only one chemical it is marked by default);
2. **Click** OK to add chemical in data matrix;

Input

Target chemical identity

- Target chemical is displayed on the data matrix.
- To see chemical identification click on the box next to "Structure info" (see next screen shot).

Chemical Input

Target chemical identity

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes 'Document', 'Single Chemical', 'Chemical List', 'Search', and 'Target Endpoint'. The 'Target Endpoint' menu is currently active, showing options like 'New', 'Open', 'Close', 'Save', 'CAS#', 'Name', 'Structure', 'Composition', 'Select', 'ChemIDs', 'Database', 'Inventory', 'List', 'Substructure (SMARTS)', 'Query', and 'Define'. The 'Documents' panel on the left shows 'Document 1' with CAS number 98011. The 'Filter endpoint tree...' panel in the center lists various chemical properties and parameters. The 'Structure' panel on the right shows the chemical structure of 2-furaldehyde, with a red box highlighting the '98-01-1' CAS number and the 'High' hazard level.

Documents

- Document 1
 - # CAS: 98011

Filter endpoint tree...

- Structure
- Structure info
 - CAS Number
 - CAS Smiles relation
 - Chemical name(s)
 - Composition
 - Molecular Formula
 - Predefined substance type
 - Structural Formula
- Parameters
 - Physical Chemical Properties
 - Environmental Fate and Transport
 - Ecotoxicological Information
 - Human Health Hazards

1 [target]

Chemical structure: 2-furaldehyde

98-01-1
High
2-furaldehyde
C5H4O2
Mono constituent
O=Cc1ccco1

Outlook

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

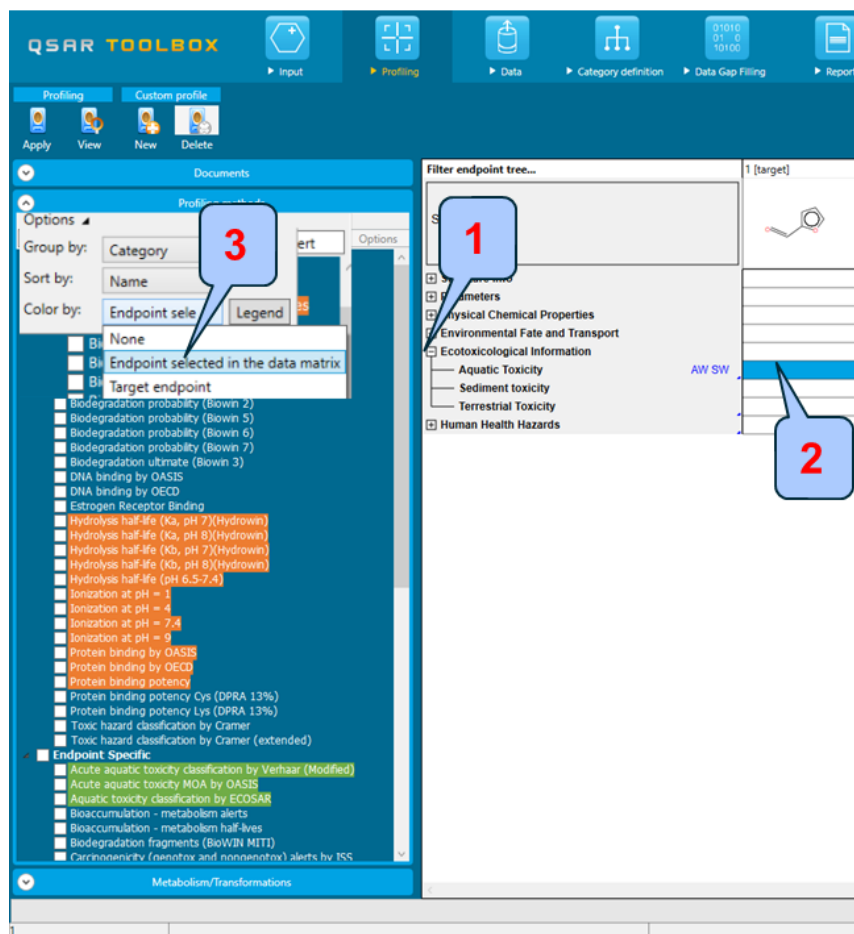
Profiling

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.
- To help the user to choose suitable profiling methods, a new feature has been developed – see next slide

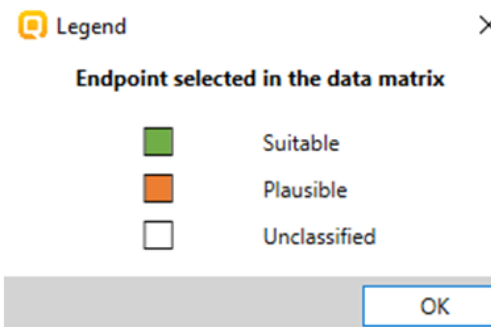
Profiling

Profiling the target chemical



1. **Click** on the box to open the nodes of the tree;
2. **Mark the** box in the data matrix related to the target endpoint of chemical;
3. From Profiling / Profiling methods / Options / **Color by** (pop-up menu) select **Endpoint selected in the data matrix**

According to the legend profilers related to the selected endpoint are highlighted in **Green** or **Orange**



Profiling

Profiling the target chemical

For this example, select all profilers.

The screenshot shows the QSAR Toolbox Profiling interface. The 'Profiling methods' panel on the left has the 'Predefined' section expanded, with the 'Select All' button checked. A red circle highlights the 'Select All' button, and a blue callout box with the number '1' points to it. The 'Metabolism/Transformations' panel below it also shows 'Select All' checked. The main panel displays the chemical structure of 2-furaldehyde and its associated data.

1. Check Select All; 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).
- Green-white rectangles in some result boxes indicate there is more than one profiling result and the field needs to be expanded.

Profiling Profiles of "Furfural"

1. Right click to see why the target is Protein binder by "Protein binding by OASIS";

2. Left click on the "Explain";

3. From the list of the profiling results Click on the structural alert "Aldehydes";

4. Click Details;

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 as aldehyde. This step is critical for next grouping of analogues.

1. **Right click** to see why the target is Protein binder by "Protein binding by OASIS";

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

2. **Left click** on the "Explain";
3. From the list of the profiling results **Click** on the structural alert "Aldehydes";
4. **Click** Details;

1. Structural boundary of the category;
2. Definition of the used common fragments;
3. Mechanistic justification of the category (Literature tab)

Outlook

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

Data

Extracting endpoint values

A new functionality for specifying bases containing data with desired endpoint is available (similar to this one for suitable profiling methods). In our case we will use all databases.

1 Go to Data

2 Select all databases

3 Click Gather

The screenshot shows the QSAR Toolbox interface with the 'Data' tab selected. The 'Filter endpoint tree...' panel is open, displaying a list of endpoints and their corresponding values. The endpoints are grouped into categories such as 'Physical Chemical Properties', 'Environmental Fate and Transport', and 'Toxicity'. The values are listed in a table format, with some entries showing 'No alert found' or 'Not possible to classify according to these rules'. The interface includes a top toolbar with icons for Input, Profiling, Data, Category definition, Data Gap Filling, and Report. The left sidebar contains sections for Documents, Databases, and Inventories, each with a 'Gather' button. The 'Databases' section is currently selected, showing a list of databases with checkboxes for selection.

Data

Process of collecting data

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

The screenshot displays the QSAR Toolbox software interface. The 'Data' tab is selected, showing a list of endpoints on the left and a list of chemical classes on the right. A 'Read data?' dialog box is open, asking the user to choose between 'All endpoints' (selected) and 'Choose...'. A blue callout box with the number '1' points to the 'OK' button in the dialog box. Below the dialog box, a message box states '570 points added across 1 chemicals.' and has an 'OK' button. A blue callout box with the number '1' points to the 'OK' button in the message box.

1. Click OK to read all available data

A window with "Read data?" appears. Now the user could choose (via radio button) to collect "All" or "Endpoint specific" data. In our case collect "All".

Data

Read data for analogues

In this example, an insert window appears stating that there were found 570 data points available for the target chemical appears. Click OK.

The screenshot shows the QSAR TOOLBOX software interface. The top navigation bar includes icons for Input, Profiling, Data, Category definition, Data Gap Filling, and Report. Below this, the 'Data' tab is selected, showing options for Gather, Import, IUCLID6, and IUCLID6. The main window is divided into several sections:

- Documents:** A section for managing documents.
- Databases:** A section for selecting databases, with a list of properties and checkboxes. The list includes:
 - Physical Chemical Properties: Chemical Reactivity COLIPA, ECHA CHEM, Experimental pKa, GSH Experimental RC50, n syn, Phys-chem EPISUITE.
 - Environmental Fate and Transport: Bioaccumulation Canada, Bioaccumulation fish CEFIC LRI, Bioconcentration NITE, Biodegradation in soil OASIS, Biodegradation NITE, Biota-Sediment Accumulation Factor L, ECHA CHEM, ECOTOX, Hydrolysis rate constant OASIS, KM database Environment Canada, n syn, Phys-chem EPISUITE.
 - Ecotoxicological Information: Aquatic ECETOC, Aquatic Japan MoE, Aquatic OASIS, ECHA CHEM, ECOTOX, n syn.
 - Human Health Hazards.
- Filter endpoint tree...:** A section for filtering endpoints, showing a tree structure with categories like Structure, Parameters, Physical Chemical Properties, Environmental Fate and Transport, Ecotoxicological Information, Human Health Hazards, and Profile.
- 1 [target]:** A section showing the chemical structure of the target compound.
- Data Matrix:** A table displaying data points for the target chemical. The table has two columns: 'M' (Molecular Weight) and 'M' (Molecular Weight). The data points are:

M	M
(1/10)	M: 0.41
(1/12)	M: 0
AW SW (1/161)	M: <0.426 mg/L
(1/215)	M: <100 ug/org
(1/172)	M: >0.54+1.63 mg/L

Data are displayed into the data matrix

Outlook

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

Category definition

Target endpoint

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

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

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

Category definition

Navigate to the target endpoint

The image shows two screenshots of the QSAR Toolbox software interface, illustrating the process of navigating to a target endpoint.

Screenshot 1 (Left): The 'Categorize' menu is open, and the 'Define' option is selected. A search filter 'tetra' is entered in the 'Documents' section. The 'Structure' tab is active, showing the chemical structure of Tetrahydrocannabinol (THC). The 'Ecotoxicological Information' section is expanded, showing various endpoints.

Screenshot 2 (Right): The 'Documents' section shows the 'tetra' node selected. The 'Ecotoxicological Information' section is expanded, showing a hierarchy of endpoints. The 'M: 145 mg/L' cell is highlighted, indicating the target endpoint.

1. Use Filter (in green) to type "Tetra" in the empty field and **click** Enter; 2. This will open nodes of data matrix to the target endpoint; 3. **Highlight** the cell that will be filled in (in this case we will reproduce the observed data);

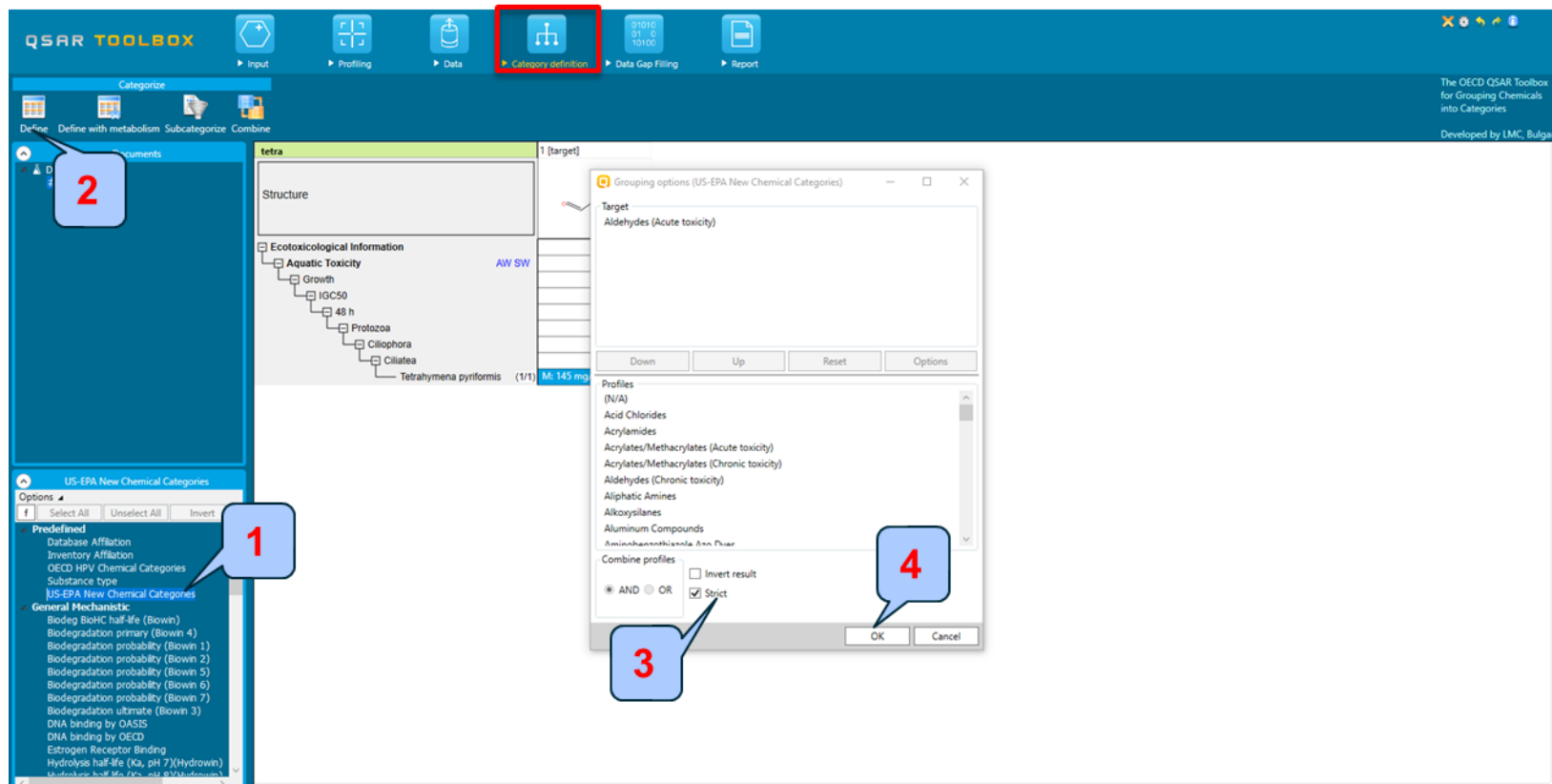
Category definition

Defining US-EPA category

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

Category definition

Defining US-EPA category



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

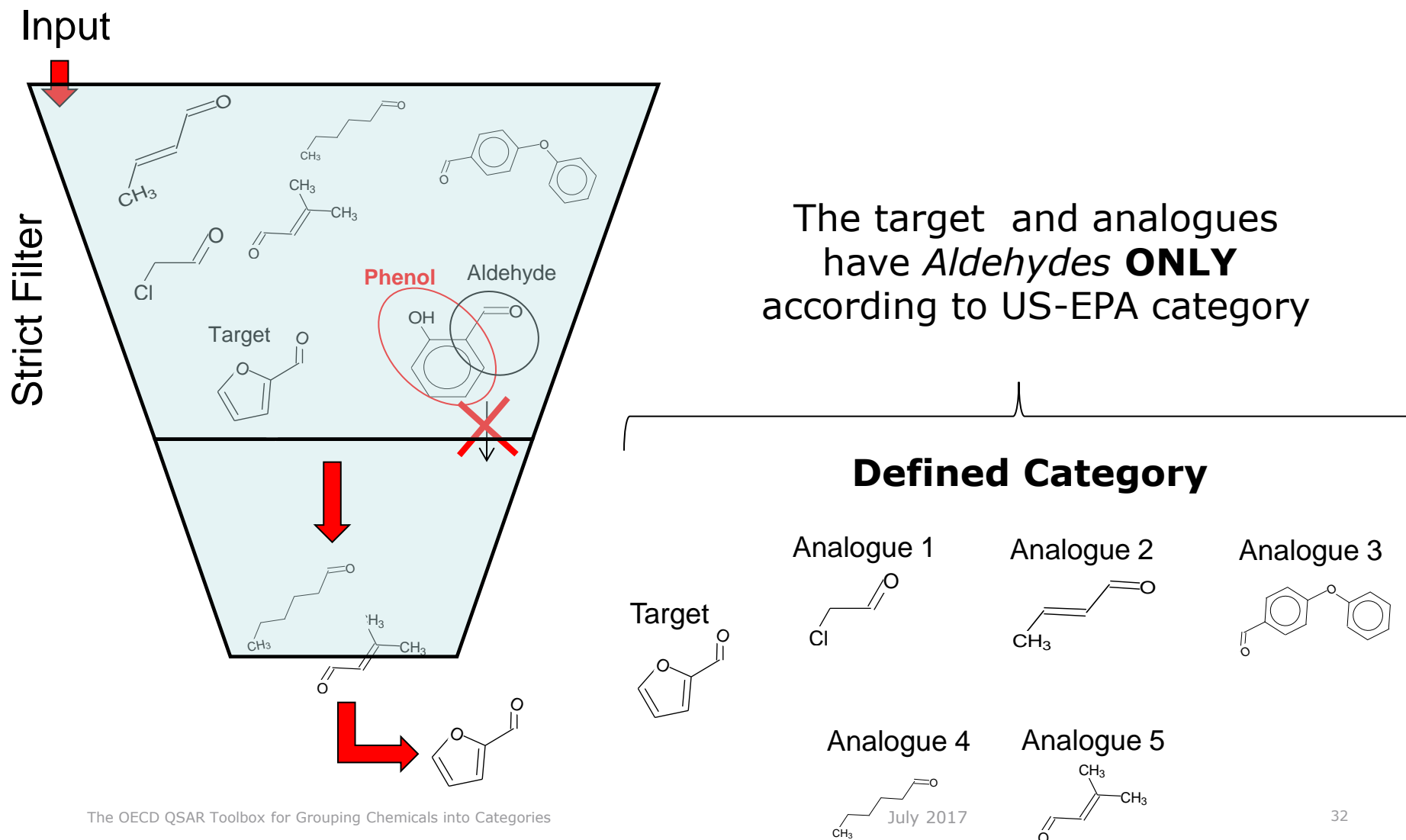
Category definition

Defining US-EPA category strict functionality

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

Category definition

Defining US-EPA category strict functionality



Category definition

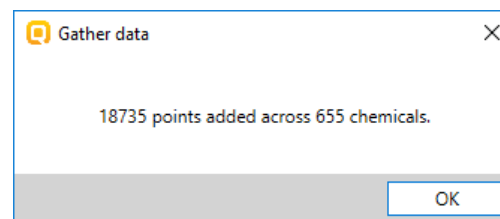
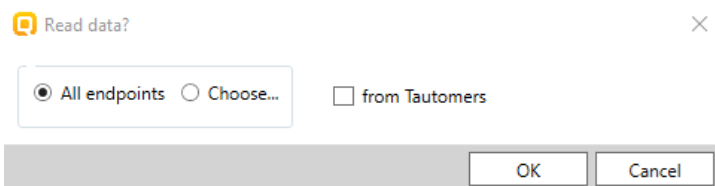
Analogues

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

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

Summary information for Analogues

After a message for number of data collected the experimental results for the target and analogues are inserted into the matrix.

The screenshot displays the QSAR TOOLBOX software interface. The top menu bar includes options like 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling', and 'Report'. The 'Category definition' tab is active, showing a 'Filter endpoint tree' on the left and a data matrix on the right. The matrix has columns for different chemical structures (1-10) and rows for various endpoints. A red box highlights the row for 'Tetrahymena pyriformis (72/72)', which shows experimental results for endpoints 1, 6, and 10.

Endpoint	1 [target]	2	3	4	5	6	7	8	9	10
Structure										
EC50 (10/18)										
IGC50										
48 h										
Protozoa										
Ciliophora										
Tetrahymena pyriformis (72/72)	M: 145 mg/L					M: 31.7 mg/L				M: 112 mg/L
LOEL										
Undefined Endpoint (3/12)										
Growth Curve (1/1)										
Growth Inhibition (23/65)										
Growth Rate (52/173)										
Growth Rate and Biomass (2/6)										
Growth Rate, Biomass and Yield (1/4)										
Growth Rate, Cell Density, Biomass (1/1)										
Growth Rate, Cell Number & Biomass (1/1)										
Growth Rate, Cell Yield (1/2)										
Histology (3/11)										
Immobilization (12/13)										
Immobilization (1/1)										
Immunological (2/4)										
Inhibition of Photosynthesis (1/3)										
Injury (1/3)										
Larval Development (1/1)										
Length (1/4)										
Loss of Schooling Behaviour, Hyperactivity and (1/1)										
Morphology (2/3)										
Mortality and Behaviour (1/2)										
Mortality of Embryos/larvae (1/1)										
Mortality, Activity (1/1)										
Mortality, Target Organ Pathologies (1/2)										
No Data (3/4)										
Observations (1/1)						M: 6.62 mg/L				

Outlook

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

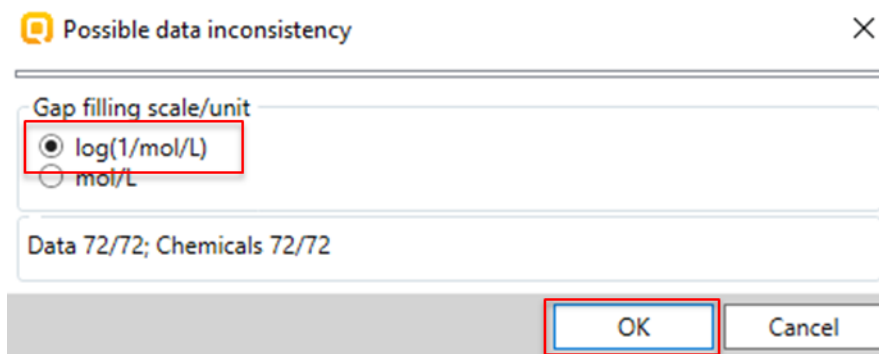
Data Gap Filling (IGC 50 48h of *T. pyriformis*) Apply Trend analysis

The screenshot displays the QSAR TOOLBOX interface. The top menu bar includes 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling', and 'Report'. The 'Data Gap Filling' menu is highlighted. Below the menu, the 'Workflow' section shows 'Gap Filling' and 'Standardized Automated'. The main area is divided into three panes. The left pane, 'Documents', shows 'New Chemical Categories' and 'Data Gap Filling Settings'. The center pane, 'Filter endpoint tree...', shows a hierarchical tree of endpoints, with 'Tetrahymena pyriformis' selected. The right pane shows a data table with chemical structures and their corresponding data values. Callout boxes 1, 2, and 3 highlight specific steps: 1 points to the 'Data Gap Filling' menu item, 2 points to the 'Tetrahymena pyriformis' entry in the filter tree, and 3 points to the 'Trend analysis' option in the 'Documents' pane.

1. **Go to** Data Gap Filling 2. **Highlight** the Data gap corresponding to IGC50, *Tetrahymena pyriformis* under the target chemical; 3. **Select** Trend analysis;

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

- A message for possible data inconsistency appears
- It is recommended the $\log(1/\text{mol/L})$ scale to be chosen



- The resulting plot can be seen on next screen shot

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

QSAR TOOLBOX

Input Profiling Data Category definition Data Gap Filling Report

Gap Filling Workflow

Trend analysis Read across (QSAR) Standardized Automated

Documents

ent 1
: 98011
JS-EPA New Chemical Categories
Y Enter GF (IA) with 72 chemicals, 72 data points

Filter endpoint tree...

Structure

Behaviour (4/4)
Biochemistry (1/5)
Biomass (8/21)
Cell Yield (1/4)
Enzyme(s) (1/2)
Growth (2/3)
EC50
IGC50
48 h
Protozoa
Ciliophora
Ciliata
Tetrahymena pyriformis (72/72)
LOEL (1/6)
Undefined Endpoint (1/2)
Growth inhibition (10/73)

1 [target] 6 10 11 14 18 21 29 30 44

M: 145 mg/L M: 31.7 mg/L M: 112 mg/L M: 3.9 mg/L M: 14 mg/L M: 7.96 mg/L M: 8.2 mg/L M: 937 mg/L M: 191 mg/L M: 167 mg/L

Descriptors

Prediction

Adequacy

Cumulative frequency

Residuals

Statistics

Trend analysis prediction for IGC50, based on 71 values
Observed: 145 mg/L; Predicted: 101 mg/L
Model equation: $IGC50 = 2.65 (\pm 0.301) + 0.395 (\pm 0.135) \cdot \log Kow, \log(1/mol/L)$

log Kow

IGC50 (log(1/mol/L))

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

Accept prediction

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 previous screen shot).
- The **RED** dot represents the predicted value for target chemical.
- The **BLUE** dots represent the experimental results available for the analogues.
- The **LIGHT BLUE** 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

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

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

Subcategorization 1: Acute aquatic toxicity MOA by OASIS

The screenshot displays the QSAR Toolbox Subcategorization interface. The 'Options' panel on the left lists various endpoints, with 'Acute aquatic toxicity MOA by OASIS' highlighted. The 'Adjust options' panel on the right shows a list of chemical classes, with 'Aldehydes' selected. The 'Trend analysis prediction for IGC50, based on 71 values' plot at the bottom shows a scatter plot of observed vs. predicted IGC50 values. The 'Select / filter data' panel on the right contains buttons for 'Subcategorize', 'Mark chemicals by WS', 'Mark chemicals by descriptor value', 'Mark outliers', 'Filter points by test conditions', 'Mark focused chemical', 'Mark focused points', and 'Remove marked data'. Numbered callouts 1 through 4 indicate the sequence of actions: 1. Click 'Select / filter data'; 2. Click 'Subcategorize'; 3. Select 'Acute aquatic toxicity MOA by OASIS'; 4. Click 'Remove selected'.

1. **Click** Select / filter data; 2. **Select** Subcategorize; 3. **Select** "Acute aquatic toxicity MOA by OASIS" (note there is the same suggestion of appropriate for subcategorization profiles and metabolic simulators); 4. **Click** "Remove selected" to eliminate dissimilar to the target chemicals

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

The screenshot displays the QSAR Toolbox Subcategorization interface. On the left, the 'Options' panel shows 'Protein binding by OASIS' selected under the 'Endpoint Specific' category. A red circle with the number '1' highlights this selection. In the center, the 'Adjust options' panel shows 'Differ from target by' set to 'At least one category'. A red circle with the number '2' highlights the 'Remove selected' button. The main window displays a table of chemicals with their predicted IGC50 values. Below the table, a scatter plot shows the trend analysis prediction for IGC50, with a red line indicating the model equation: $IGC50 = 2.37 (\pm 0.242) + 0.488 (\pm 0.107) \cdot \log Kow, \log(1/mol/L)$. The plot shows observed values (blue dots) and predicted values (red dots) for 67 chemicals. A red circle with the number '2' highlights the 'Remove selected' button in the 'Adjust options' panel.

1. **Select** "Protein binding by OASIS";
2. **Click** "Remove selected" to eliminate dissimilar to the target chemicals.

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

Subcategorization 3: Aquatic toxicity classification by ECOSAR

The screenshot displays the QSAR Toolbox Subcategorization interface. On the left, the 'Endpoint Specific' list includes 'Aquatic toxicity classification by ECOSAR', which is highlighted. A red circle with the number '1' points to this selection. Below this, the 'Options' panel shows various simulation methods. A red circle with the number '2' points to the 'Remove selected' button in the 'Selected 1 (27/28)' list. The main window shows a table of chemical structures and their predicted IGC50 values. A scatter plot at the bottom shows the relationship between log Kow and IGC50 (log1/mg/L), with a red regression line. The plot includes data points for the selected chemicals and a legend for the regression line.

1. Select "Aquatic toxicity classification by ECOSAR";

2. Click "Remove selected" to eliminate dissimilar to the target chemicals;

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

The screenshot displays the OECD QSAR Toolbox interface for Data Gap Filling. The 'Subcategorization' window on the left shows various options for defining chemical categories. The 'Filter endpoint tree' in the center lists endpoints such as Growth, Growth Inhibition, and Growth Rate, with associated data values. The 'Trend analysis prediction' plot at the bottom shows a scatter plot of log Kow vs. IGC50 (log1/moL/L) with a red trend line. A confirmation dialog asks 'Are you sure you want to accept this prediction?' with 'Yes' and 'No' buttons. A green checkmark and 'Accept prediction' button are also visible.

1. Close "Subcategorization" window

2. Click "Accept prediction"

3. Click "Yes" ("No" allows to continue with the subcategorization)

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

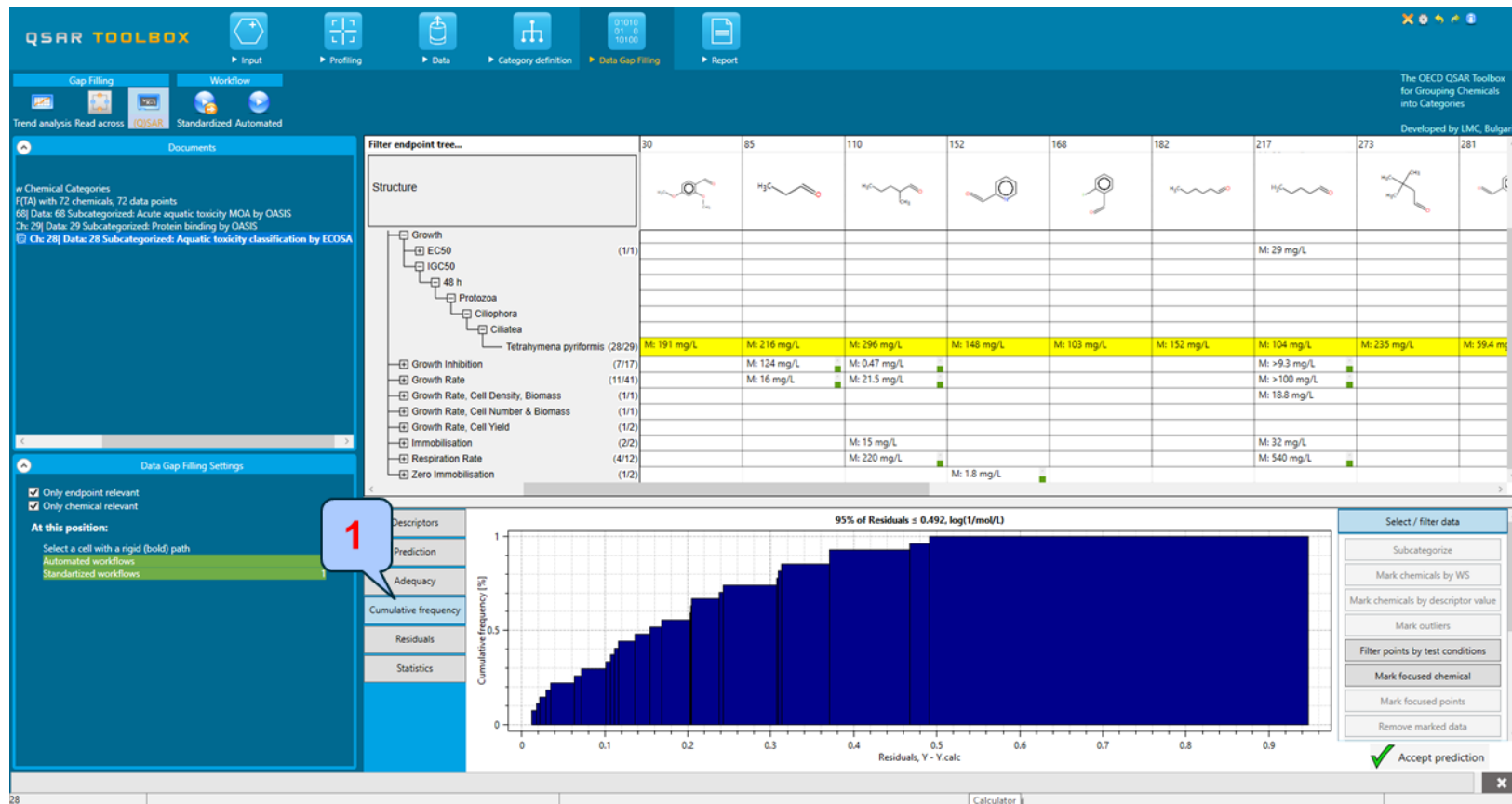
Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model - Adequacy

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes options like Input, Profiling, Data, Category definition, Data Gap Filling, and Report. The left sidebar shows a document tree with a highlighted path: Chemical Categories > (TIA) with 72 chemicals, 72 data points > (68) Data: 68 Subcategorized: Acute aquatic toxicity MOA by OASIS > Ch: 29] Data: 29 Subcategorized: Protein binding by OASIS > Ch: 28] Data: 28 Subcategorized: Aquatic toxicity classification by ECOSA. A red circle with the number '1' points to this path. The central workspace shows a 'Filter endpoint tree...' on the left and a data table on the right. The data table has columns for chemical structures and their corresponding IGC50 values. A red circle with the number '2' points to the 'Adequacy' button in the bottom right sidebar. The bottom section features a scatter plot titled 'Adequacy of prediction' showing the relationship between observed and predicted IGC50 values, with a regression line and statistical data (R2 = 0.802, R2adj = 0.795, s = 0.303). The bottom right sidebar contains buttons for 'Select / filter data', 'Subcategorize', 'Mark chemicals by WS', 'Mark chemicals by descriptor value', 'Mark outliers', 'Filter points by test conditions', 'Mark focused chemical', 'Mark focused points', 'Remove marked data', and 'Accept prediction'.

1. Position on the last level of document tree 2. Click "Adequacy";

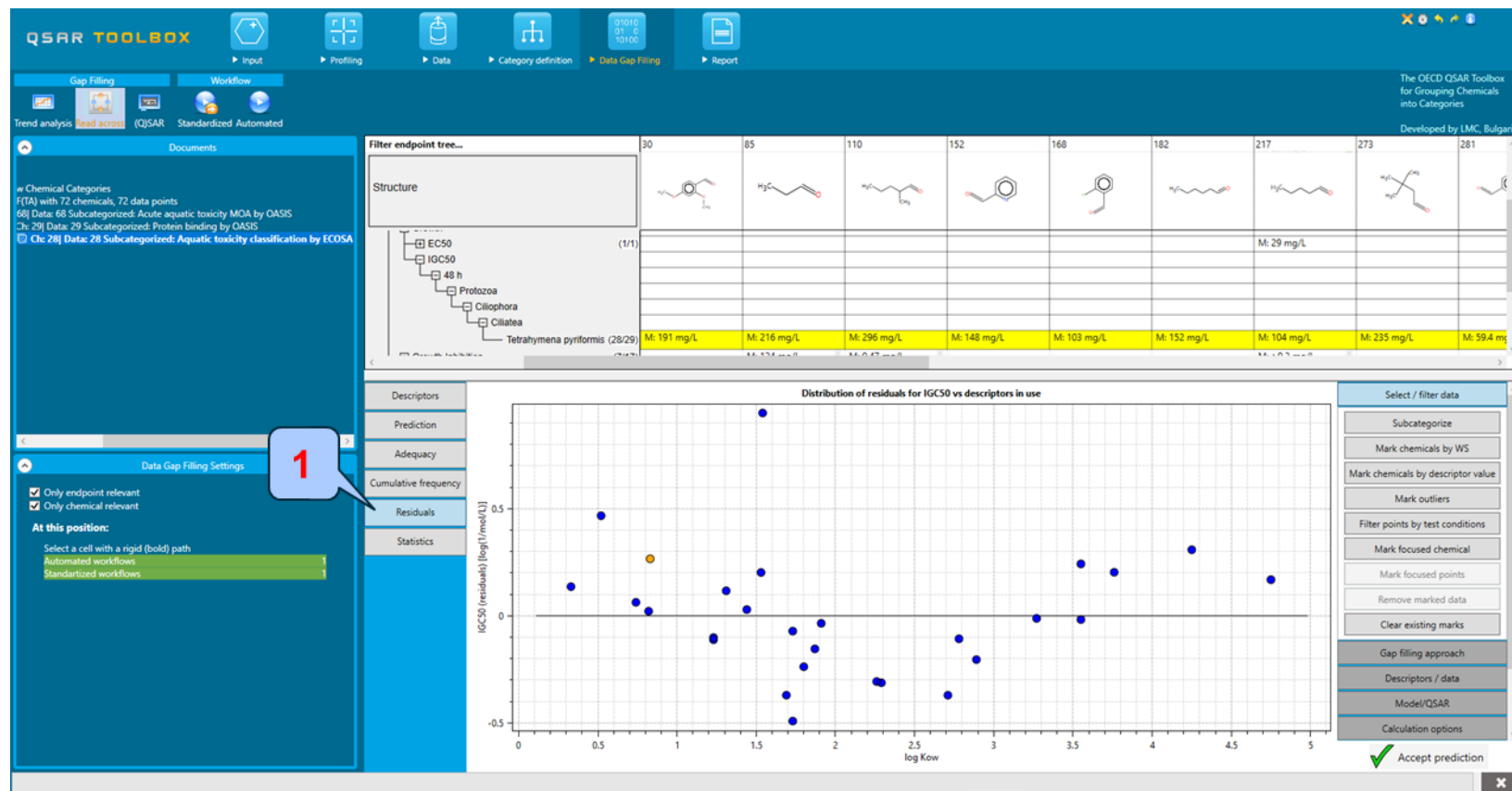
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

Evaluation of the model - Cumulative frequency



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

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



1. Click "Residuals";

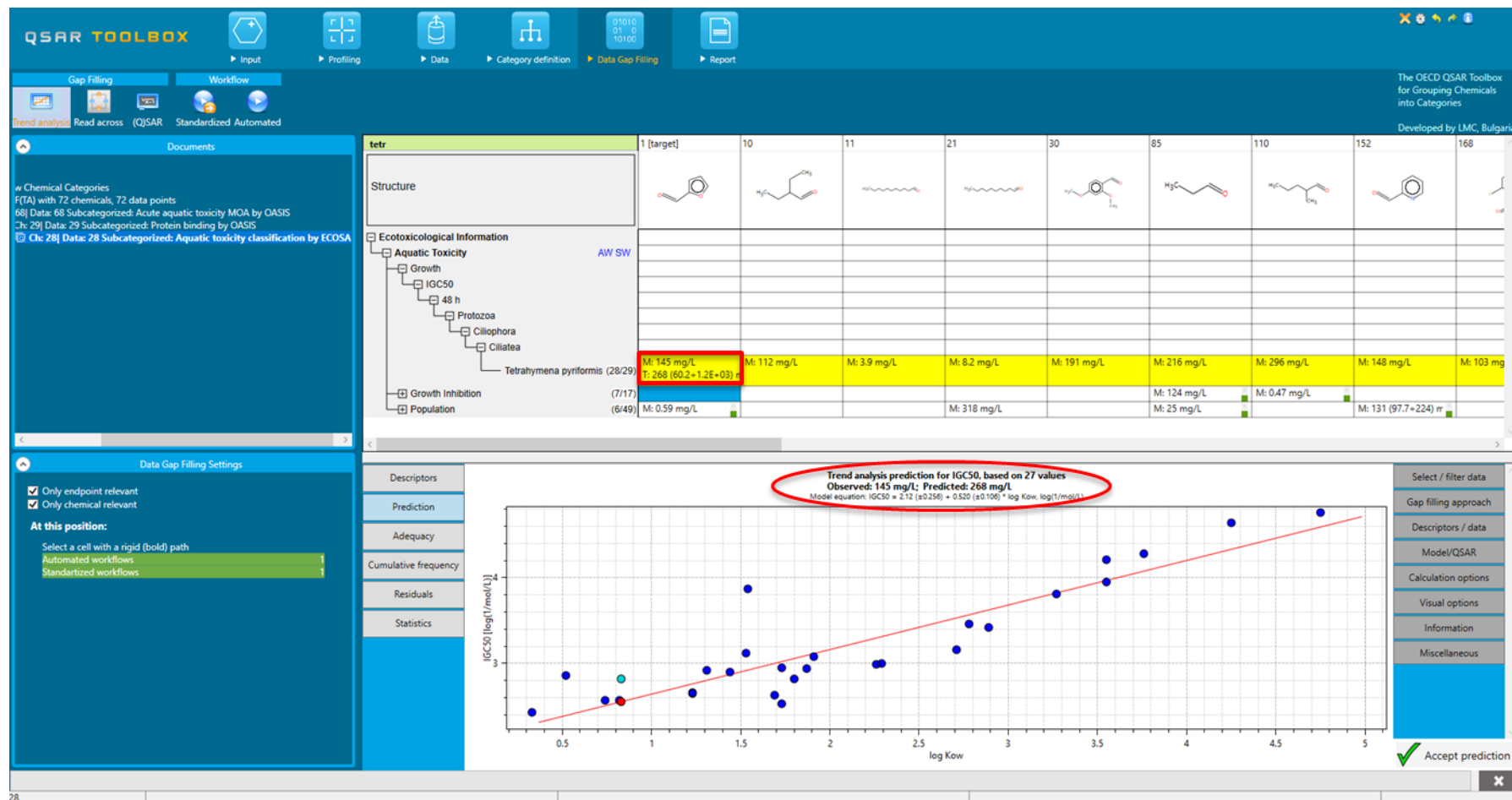
Data Gap Filling (IGC 50 48h of *T. pyriformis*) Evaluation of the model - Statistics

The screenshot shows the QSAR Toolbox interface during the Data Gap Filling workflow. The 'Statistics' tab is selected in the left sidebar, indicated by a red circle with the number 1. The main window displays a table of statistical results for the model, including Descriptors, Prediction, Adequacy, Cumulative frequency, Residuals, and Statistics. The 'Statistics' section shows the model descriptor, coefficient value, range, significance, and max covariation for both b0 and b1. The right sidebar contains a 'Select / filter data' panel with various options like 'Subcategorize', 'Mark chemicals by WS', and 'Filter points by test conditions'.

Descriptors	Statistical characteristics	TA model
Prediction	Number of data points, (N)	27
	Coefficient of determination, (R2)	0.802
	Adjusted coefficient of determination, (R2adj)	0.795
Adequacy	Coefficient of determination - leave one out, (Q2)	N/A
	Sum of squared residuals, (SSR)	2.29
Cumulative frequency	Standard deviation of residuals, (sN)	0.291
	Sample standard deviation of residuals, (s)	0.303
Residuals	Fisher function, (F)	102
Statistics	Fisher threshold for statistical significance, (Fa)	6.06 (95.0%)
	b0	
	- model descriptor	Intercept
	- coeff. value	2.12
	- coeff. range	±0.256
	- significance	No
	- max covariation	0.249 vs b1
	b1	
	- model descriptor	log Kow
	- coeff. value	0.520
	- coeff. range	±0.106
	- significance	No
	- max covariation	0.249 vs b0

1. Click "Statistics";

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



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

- To save the new regression model follow these steps:
 - Go to the last row on the Document tree
 - Click on "Model/QSAR"
 - Select Save model
 - Enter the model name and fill editable fields if necessary
 - Click on OK

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

The screenshot displays the QSAR Toolbox interface. The 'Customize model content' wizard is open, showing the 'General information' page. The title field is populated with 'IGC50_Tetrahyena_Furfural'. A scatter plot at the bottom shows log Kow values on the x-axis (0.5 to 5.0) and log Kow values on the y-axis (0.5 to 5.0). A red regression line is shown. A message box at the bottom left states 'The model was saved successfully!' with an 'OK' button. A sidebar on the right contains a list of options: 'Model/QSAR', 'Save model', 'Save domain as category', 'Calculate Q2', 'Calculation options', 'Visual options', and 'Information'. A green checkmark and the text 'Accept prediction' are at the bottom right.

Numbered callouts indicate the following steps:

- Click "Model/QSAR"
- Select "Save model"
- Type Name of the model and fill fields in the Wizard if necessary (Use Next/Back buttons to navigate within it)
- Click "Save model"
- Click OK on the message

1. **Click** "Model/QSAR"; 2. **Select** "Save model"; 3. **Type** Name of the model and fill fields in the Wizard if necessary (Use Next/Back buttons to navigate within it); 4. **Click** "Save model"; 5. **Click** OK on the message;

Outlook

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

Data Gap Filling

How to see the derived QSAR?

3

1

2

4

QSAR name	#	Predicted	Class	Domain
Mortality Aldehydes (Mono) (1.0)	1	62.1 mg/L	branchiopoda (branchiopods)	No domain available
ECOSAR: DAPHNID ChV Aldehydes (Mono) (1.0)	2	7.61 mg/L	Branchiopoda (branchiopods)	No domain available
ECOSAR: Fish (SW) 96 h LC50 Mortality Aldehydes (Mono) (1.0)	3	32.0 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) ChV Aldehydes (Mono) (1.0)	4	3.20 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	5	22.5 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	6	5.40 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: GREEN ALGAE 96 h EC50 Aldehydes (Mono) (1.0)	7	76.2 mg/L		No domain available
ECOSAR: GREEN ALGAE ChV Aldehydes (Mono) (1.0)	8	21.3 mg/L		No domain available
IGC50_Tetrahymena_Furfural (1.0)	9	268 mg/L	Ciliata	In domain
M1 - LC50 - Pimephales promelas (fathead minnow) (1.0)	10	71.6 mg/L		No domain available
M2 - LC50 - Pimephales promelas (fathead minnow) (1.0)	11	372 mg/L		No domain available
M3 - LC50 - Pimephales promelas (fathead minnow) (1.0)	12	865 mg/L		No domain available
M4 - LC50 - Pimephales promelas (fathead minnow) (1.0)	13	167 mg/L		No domain available
Photoinduced toxicity of PAHs (1.0)	14	Not Phototoxic		Out of domain

1. Select a non-Gap filling list from the document tree; 2. Note the accepted prediction is inserted into data matrix 3. **Click** "(Q)SAR"; 4. 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:
- Visualize the domain of the model:
- Visualize whether a chemical is in the applicability domain of the model:
- Enter in Data Gap filling
- Perform predictions for:
 - Selected chemical
 - All chemicals (in the matrix)
 - Chemicals in domain:

Data Gap Filling

Visualisation of the training set

Details for 15 (QSAR models)

QSAR name	#	Predicted	Class	Domain
ECOSAR: DAPHNID 48 h LC50 Mortality Aldehydes (Mono) (1.0)	1	65.1 mg/L	Branchiopoda (branchiopods)	No domain available
ECOSAR: DAPHNID ChV Aldehydes (Mono) (1.0)	2	7.61 mg/L	Branchiopoda (branchiopods)	No domain available
ECOSAR: Fish (SW) 96 h LC50 Mortality Aldehydes (Mono) (1.0)	3	32.0 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) ChV Aldehydes (Mono) (1.0)	4	3.20 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	5	22.5 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	6	5.40 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	7	76.2 mg/L		No domain available
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	8	21.3 mg/L		No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	9	268 mg/L	Ciliata	In domain
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	10	71.6 mg/L		No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	11	372 mg/L		No domain available
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	12	865 mg/L		No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	13	167 mg/L		No domain available

File

CAS#	Predicted	Chemical Structure
97-96-1	112 mg/L	<chem>CCCCC=O</chem>
112-44-7	mg/L	<chem>CCCCC=O</chem>
112-31-2	8.20 mg/L	<chem>CCCCC=O</chem>
613-45-6	191 mg/L	<chem>CCCCC=O</chem>
123-38-6	216 mg/L	<chem>CCCCC=O</chem>
123-15-9	296 mg/L	<chem>CCCCC=O</chem>
1121-60-4	148 mg/L	<chem>CCCCC=O</chem>
446-52-6	103 mg/L	<chem>CCCCC=O</chem>
66-25-1	152 mg/L	<chem>CCCCC=O</chem>
110-62-3	104 mg/L	<chem>CCCCC=O</chem>
2987-16-8	235 mg/L	<chem>CCCCC=O</chem>
66-77-3	59.4 mg/L	<chem>CCCCC=O</chem>
590-86-3	188 mg/L	<chem>CCCCC=O</chem>
4460-86-0	247 mg/L	<chem>CCCCC=O</chem>
78-84-2	194 mg/L	<chem>CCCCC=O</chem>
529-20-4	123 mg/L	<chem>CCCCC=O</chem>
124-13-0	44.5 mg/L	<chem>CCCCC=O</chem>
111-71-7	114 mg/L	<chem>CCCCC=O</chem>
123-72-8	194 mg/L	<chem>CCCCC=O</chem>
112-54-9	3.20 mg/L	<chem>CCCCC=O</chem>
123-05-7	88.7 mg/L	<chem>CCCCC=O</chem>
21661-97-2	17.3 mg/L	<chem>CCCCC=O</chem>
122-78-1	16.2 mg/L	<chem>CCCCC=O</chem>
124-19-6	22.0 mg/L	<chem>CCCCC=O</chem>
65405-70-1	9.51 mg/L	<chem>CCCCC=O</chem>

Save to smi

1. **Right Click** on the derived QSAR model; 2. **Select** Show training Set; 3. Note the experimental data is displayed under CAS# of each chemical; 4. The training set can be saved as *.smi file.

Data Gap Filling

Visualisation of model domain

Details for 15 (QSAR) models

QSAR name	#	Predicted	Class	Domain
ECOSAR: DAPHNID 48 h LC50	1	65.1 mg/L	Branchiopoda (branchiopod)	No domain available
Mortality Aldehydes (Mono) (1.0)	2	7.61 mg/L	Branchiopoda (branchiopod)	No domain available
ECOSAR: DAPHNID CHV Aldehydes (Mono) (1.0)	3	32.0 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) 96 h LC50	4	3.20 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
Mortality Aldehydes (Mono) (1.0)	5	22.5 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) CHV Aldehydes (Mono) (1.0)	6	5.40 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) 96 h LC50	7	76.2 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	8	21.3 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	9	268 mg/L	Ciliates	In domain
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	10	71.6 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	11	372 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	12	865 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	13	167 mg/L		No domain available
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	14			
ECOSAR: GREEN AUGAE CHV Aldehydes (Mono) (1.0)	15			

Category tree

```

graph TD
    1((1)) -- NOT --> 2((2))
    1 -- NOT --> 3((3))
    2 -- AND --> 4((4))
    3 -- AND --> 4
    4 -- AND --> 5((5))
    
```

Query details

Reference Query: Metabolism

Profiling schemes:

- Custom
- Empiric
- Endpoint Specific
- General Mechanistic
- Predefined
 - Database Affiliation
 - Inventory Affiliation
 - OECD HPV Chemical Categories
 - Substance type
 - US-EPA New Chemical Categories
 - Toxicological

Selected categories: Aldehydes (Acute toxicity)

Available categories:

- (N/A)
- Acid Chlorides
- Acrylamides
- Acrylates/Methacrylates (Acute toxicity)
- Acrylates/Methacrylates (Chronic toxicity)
- Aldehydes (Chronic toxicity)
- Aliphatic Amines
- Aliphatic Amides

Multiple categories: ☒ Strict ☐ OR-ed ☐ AND-ed

1. Right click on the derived QSAR model; **2. Select "Display Domain"**; **3.** Note the boundaries of the domain are combined logically; **4.** If the chemical answers the query of the domain then the current query is a labelled with **GREEN** tick; **5.** Otherwise is labelled with **RED** cross.

Data Gap Filling

Visualisation whether a chemical is in the domain of the model

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling', and 'Report'. The left sidebar shows the 'Filter endpoint tree...' and 'Data Gap Filling Settings'. The central data matrix shows chemical data, with chemical # 94 highlighted. The 'Details for 21 (Q)SAR models' window is open, showing a table of models and their domains. The 'Display Domain' option is highlighted in the context menu.

QSAR name	#	Predicted	Class	Domain
Imidazoles (1.0)	8	21.1 mg/L	fishes, spiny rayed fishes	No domain available
ECOSAR: FISH CHV Aldehydes (Mono) (1.0)	9	26.3 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH CHV Imidazoles (1.0)	10	0.396 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: GREEN ALGAE 96 h EC50 Aldehydes (Mono) (1.0)	11	341 mg/L		No domain available
ECOSAR: GREEN ALGAE 96 h EC50 Imidazoles (1.0)	12	0.622 mg/L		No domain available
ECOSAR: GREEN ALGAE CHV Aldehydes (Mono) (1.0)	13	82.0 mg/L		No domain available
ECOSAR: GREEN ALGAE CHV Imidazoles (1.0)	14	0.715 mg/L		No domain available
ECOSAR: MYSID (SW) 96 h LC50 Mortality Imidazoles (1.0)	15	15.3 mg/L	Malacostraca	No domain available
ECOSAR: Tetrahymena Furfural (1.0)	16	990 mg/L	Ciliata	Out of domain
M1 (fr)		93.5 mg/L		No domain available
M2 (fr)		2.17E3 mg/L		No domain available
M3 (fr)		5.5E3 mg/L		No domain available
M4 (fr)		850 mg/L		No domain available
Ph		Not Phototoxic		Out of domain

1. **Highlight** the cell of one of the analogues (e.g., chemical # 94 in the data matrix;
2. Click on "(Q)SAR";
3. **Right click** above the model; 3. **Left click** on Display domain (see next screen shot).

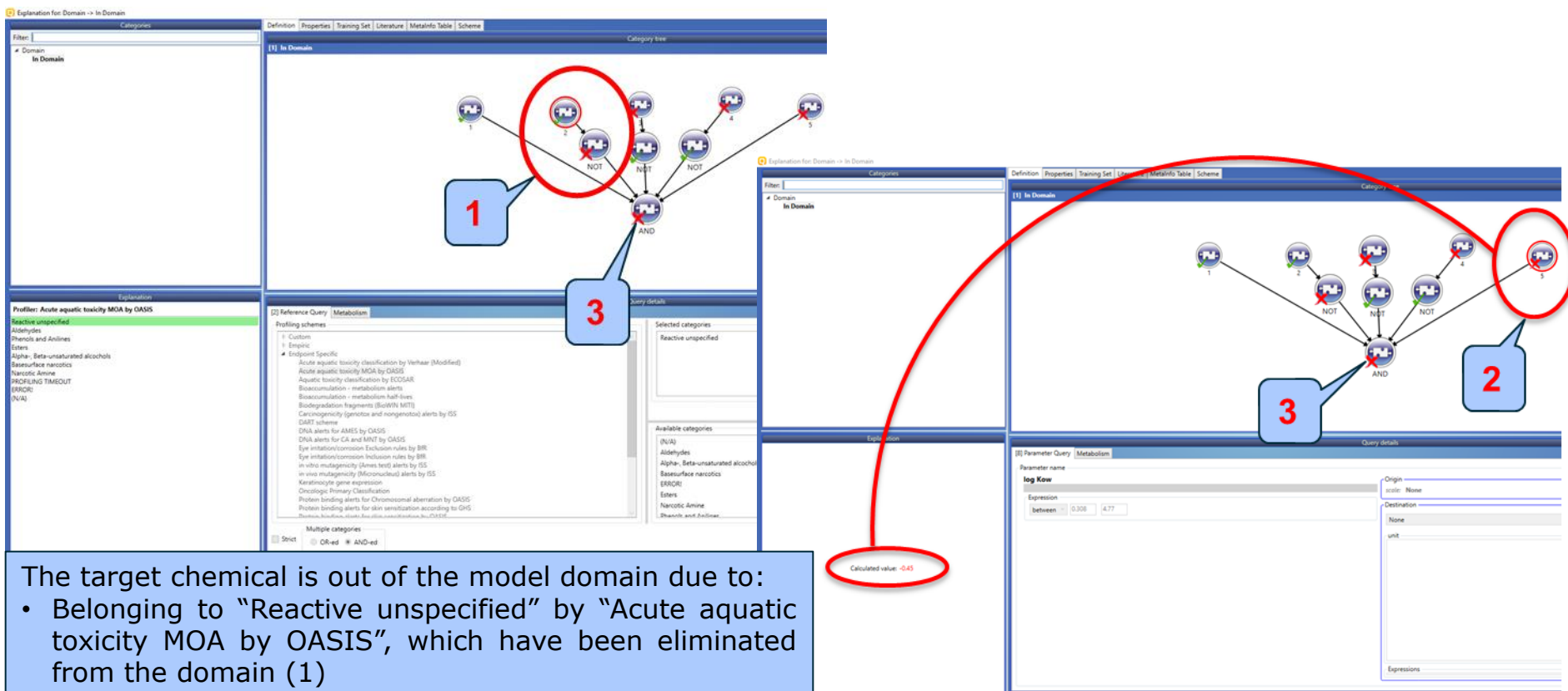
Data Gap Filling

Visualisation whether a chemical is in the domain of the model

- The chemical is an aldehyde as required by US-EPA categorization group (boundary 1 on next screen shot).
- The chemical is an aldehyde as required by Acute aquatic toxicity MOA by OASIS group (boundary 2).
- It can react with protein by Schiff-base formation and should not belong to any of the eliminated mechanistic domains according to Protein binding by OASIS (boundary 3):
 - Michael addition (α,β -Aldehydes, Conjugated systems with electron withdrawing groups)
 - S_NAr (Activated aryl and heteroaryl compounds)
 - Schiff base formation (Bis aldehydes, Di-substituted α,β -unsaturated aldehydes and Aromatic carbonyl compounds)
- The chemical is an aldehyde as required by Aquatic toxicity classification by ECOSAR (boundary 4).
- Another requirement is Log Kow to be ≥ 0.308 and ≤ 4.77 (boundary 5):

Data Gap Filling

Visualisation whether a chemical is in the domain of the model



The target chemical is out of the model domain due to:

- Belonging to "Reactive unspecified" by "Acute aquatic toxicity MOA by OASIS", which have been eliminated from the domain (1)
- Parameter log Kow different from the model boundaries (2)

The definitive designation for belonging or not to the domain is the collectible boundary (3) which is red crossed in case of "Out of domain" (green checked in case of "In domain")

Data Gap Filling

Enter Gap filling

5

QSAR TOOLBOX

Input Profiling Data Category definition Data Gap Filling Report

Gap Filling Workflow

Trend analysis Read across (QSAR) Standardized Automated

Documents

1: 98011

JS-EPA New Chemical Categories

Enter Gf(TA) with 72 chemicals, 72 data points

Ch: 68 Data: 68 Subcategorized: Acute aquatic

Ch: 29 Data: 29 Subcategorized: Protein bi

Ch: 28 Data: 28 Subcategorized: Aquatic

Enter Gf(ExternalQSAR) with 72 chemicals, 99

Filter endpoint tree...

Structure

Tetrahymena pyriformis

Ciliata

Ciliophora

Protozoa

EC50

IC50

Growth

48 h

Tetrahymena thermophila

Toxodina sp.

Turbellaria

Unonema parduzci

Unonema parduzci

Vibrio fischeri

Xenopus laevis

1 target

6

14

18

ECOSAR: FISH ChV Aldehydes (Mono) (1.0)

ECOSAR: FISH ChV Imidazoles (1.0)

ECOSAR: GREEN ALGAE 96 h EC50 Aldehydes (Mono) (1.0)

ECOSAR: GREEN ALGAE 96 h EC50 Imidazoles (1.0)

ECOSAR: GREEN ALGAE ChV Aldehydes (Mono) (1.0)

ECOSAR: GREEN ALGAE Imidazoles (1.0)

ECOSAR: MYSID (SW) Mortality Imidazoles (1.0)

IGC50_Tetrahymena_Furfural (1.0)

M1 - LC50 - Pimephales promelas (fathead minnow) (1.0)

M2 - LC50 - Pimephales promelas (fathead minnow) (1.0)

M3 - LC50 - Pimephales promelas (fathead minnow) (1.0)

M4 - LC50 - Pimephales promelas (fathead minnow) (1.0)

Photoinduced toxicity of PAHs (1.0)

Observed: 145 mg/L Predicted: 21

Model equation: $IGC50 = -2.12 + 0.520 \cdot \log Kow$

Active descriptor X log Kow

Descriptors / data

Model/QSAR

Calculation options

Visual options

Information

Miscellaneous

Accept prediction

1

2

3

4

Run Cancel

Go to target chemical and call (Q)SAR;
 1. **Mark** the model; 2. **Click** Run; 3. **Select** Enter Gap filling; 4. **Click** OK; 5. You are in Gap filling and can operate;

Data Gap Filling

Perform prediction for chemicals in domain (for selected chemical and all chemicals - analogically)

The screenshot displays the QSAR Toolbox software interface. The top menu bar includes options like Input, Profiling, Data, Category definition, Data Gap Filling, and Report. The left sidebar shows 'Documents' and 'Data Gap Filling Settings'. The central area shows 'Details for 14 (Q)SAR models' with a table of models. A 'Select QSAR method' dialog box is open, showing options for gap filling or prediction. Red callout boxes numbered 1 through 4 highlight the steps: 1. Marking a model in the table, 2. Clicking the 'Run' button, 3. Selecting 'Predict chemicals in domain' in the dialog, and 4. Clicking 'OK' in the dialog.

QSAR name	#	Predicted	Class	Domain
ECOSAR: DAPHNID 48 h LC50 Mortality Aldehydes (Mono) (1.0)	1	497 mg/L	Branchiopoda (branchiopods)	No domain available
ECOSAR: DAPHNID ChV Aldehydes (Mono) (1.0)	2	62.9 mg/L	Branchiopoda (branchiopods)	No domain available
ECOSAR: Fish (SW) 96 h LC50 Mortality Aldehydes (Mono) (1.0)	3	301 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: Fish (SW) ChV Aldehydes (Mono) (1.0)	4	32.3 mg/L	Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH 96 h LC50 Mortality Aldehydes (Mono) (1.0)	5		Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: FISH ChV Aldehydes (Mono) (1.0)	6		Actinopterygii (ray-finned fishes, spiny rayed fishes)	No domain available
ECOSAR: GREEN ALGAE 96 h EC50 Aldehydes (Mono) (1.0)				No domain available
ECOSAR: GREEN ALGAE ChV Aldehydes (Mono) (1.0)				No domain available
ECOSAR: Tetrahymena Furfural (1.0)	9		Ciliata	In domain
M1 - LC50 - Pimephales promelas (fathead minnow) (1.0)	10			No domain available
M2 - LC50 - Pimephales promelas (fathead minnow) (1.0)	11	265		No domain available
M3 - LC50 - Pimephales promelas (fathead minnow) (1.0)	12	58		No domain available
M4 - LC50 - Pimephales promelas (fathead minnow) (1.0)	13	130 mg/L		No domain available

1. **Mark** the model; 2. **Click** Run; 3. **Select** Predict Chemicals in domain; 4. **Click** OK;

Data Gap Filling

Perform prediction for chemicals in domain

[illegible]

The process of applying the model is indicated by status bar on the bottom of the window. The predictions are placed on the matrix. Note there are different signs for the origin of the data: M for experimental data, T for result of Trend analysis, Q for originated from QSAR data.

Outlook

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

Export QSAR results

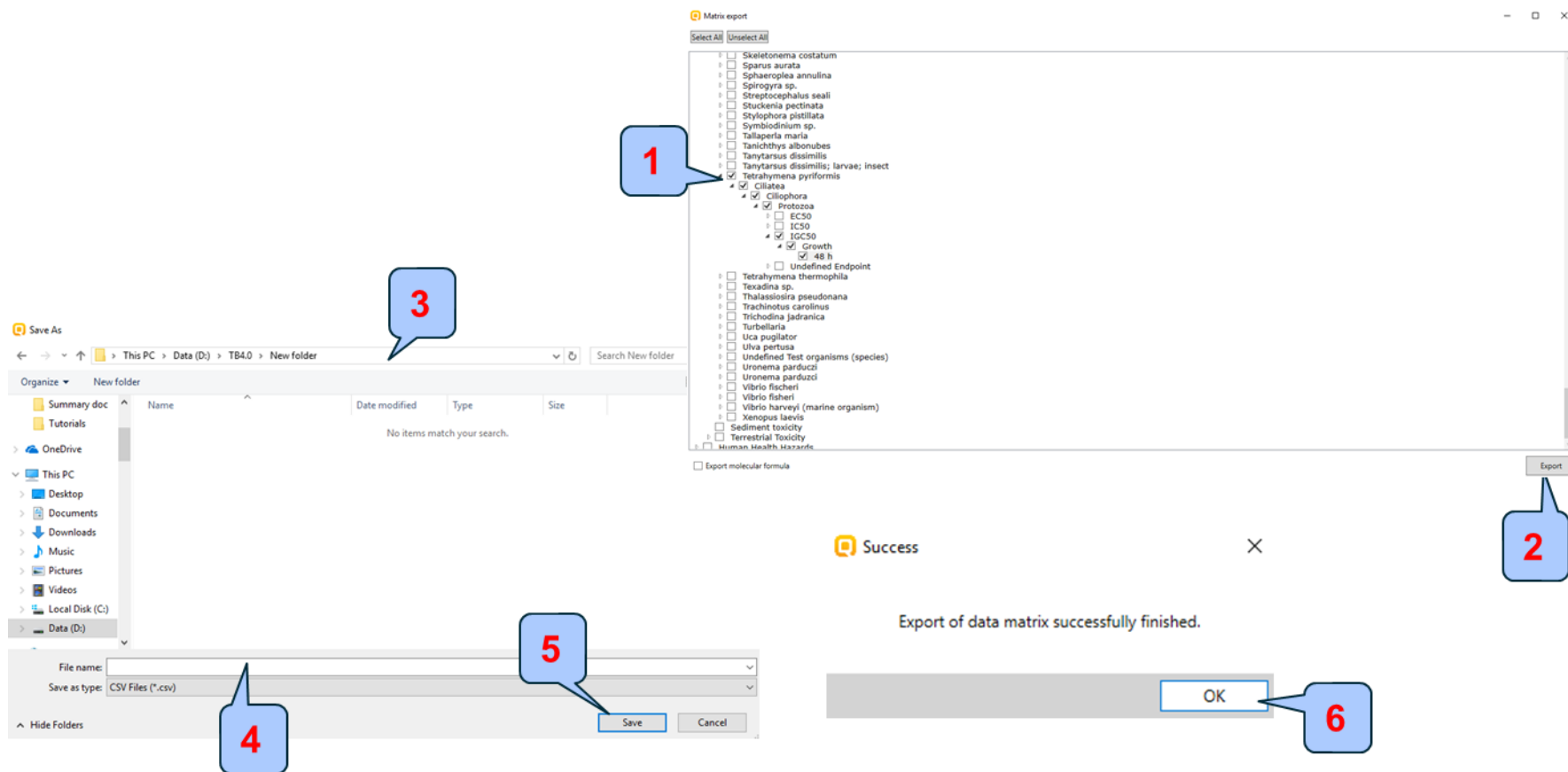
- The predictions for the chemicals in the matrix can be exported into a file.
- In the Endpoint tree **right click** on Tetrahymena pyriformis (for the endpoint IGC50 48h for Tetrahymena pyriformis) and **select** Export Data matrix from the context menu (see next three screen shots).

Export QSAR results

The screenshot shows the QSAR Toolbox software interface. The top toolbar contains icons for Input, Profiling, Data, Category definition, Data Gap Filling, and Report. Below the toolbar are tabs for Gap Filling and Workflow. The main window displays a hierarchical tree of chemical categories on the left, a central table of chemical structures and their corresponding QSAR results, and a right sidebar with various settings. A red box highlights the '48 h' endpoint in the tree, and a blue callout '1' points to it. A blue callout '2' points to the 'Export Data matrix' option in the context menu that appears when right-clicking on the highlighted endpoint.

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

Export QSAR results



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

Export QSAR results

The resulting file in *.csv format can be opened via Microsoft Excel and further analysed.

FileHomeInsertPage LayoutFormulasDataReviewView

CutCopyFormat PainterClipboard

Font

AlignmentMerge & Center

General

Number

Conditional Formattingas Table

Styles

NormalBadGoodNeutralCalculation

Check CellExplanatory...InputLinked CellNote

InsertDeleteFormat

Cells

AutoSumFillClear

Sort & Find & Filter & Select

Editing

A1	fx CAS Number																												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	CAS Num	Structural	Endpoint	Year	Title	Author	Effect	Comment	Endpoint	Reference Test	organ Database	Assigned	Kingdom	Phylum	Class	Subclass	Order	Suborder	Family	Genus	Duration	Duration	Duration	Duration	Duration	Duration	Data Mean	Data Unit	
2	98-01-1	O=Cc1cccc Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h					Time		145.4256	mol/L
3	70201-42-	Brc1cc(C=O)cc(Br)n1																											
4	119-67-5	OC(=O)c1cccc1C=O																											
5	63282-01-	CC1CCCC(C=O)=C1																											
6	5703-26-4	O=Cc1ccc(Cc=O)cc1																											
7	122-02-3	C(C)(Cl)co Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		31.68359	mol/L	
8	494-08-6	OCc1cc(C=O)ccc1OC1OC(CO)(CO)(C)O C1O																											
9	5614-52-8	CN1C(=O)N(CC=O)C(=O)c2c1nnc2C																											
10	432-25-7	CC1CCCC(C)(C)C=C1=O																											
11	97-96-1	CCC(CO)C Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		112.3761	mol/L	
12	112-44-7	CCCCCCCC Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		3.900997	mol/L	
13	5362-56-1	CC(C)C=CC=O																											
14	98506-67-	CC(O)C(C)C=O																											
15	123-73-9	CC=CC=O Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		13.98432	mol/L	
16	63034-44-	FC(Cl)(Cl)C=O																											
17	1849-55-4	Oc1cccncc1C=O																											
18	Invalid CA	O=CCC(=O)c1cccc(cc1)C1CCCCC1																											
19	2548-87-0	CCCCCC=O Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		7.962124	mol/L	
20	51575-61-	CC(C=O)=CC1OC(C)(C)C1CO1																											
21	624-67-9	O=CC#C																											
22	112-31-2	CCCCCCCC Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		8.200578	mol/L	
23	41468-25-	Cc1ncdCOP(O)(O)=O)c(C=O)c1O																											
24	488-11-9	OC(=O)C(Br)=C(Br)C=O																											
25	533-49-3	OCC(O)C(O)C=O																											
26	93943-58-	CC(CCl1C(O)CC=C1C=O)C1CC(C)(C)CC=1																											
27	Invalid CA	Oc1ccc2ccccc2c1C=O																											
28	13289-18-	CC1OC1OC2CCCC3(C=O)C4CCS(C)(C)CC5(O)C4CCC3(O)C2C2C=CC(=O)OC=2)C(O)C1OC1OC(CO)(CO)C(O)C1O																											
29	17422-74-	O=CC1=COC2cccc2C1=O																											
30	68282-53-	Cc1[nH]cn Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		937.2531	mol/L	
31	613-45-6	Oc1ccc(C Ecotoxico	1997	Tetratox:	Schultz, T.	Growth	Impairme	IGC50	Toxicolog	Tetrahym Aquatic O	FALSE	Protozoa	Ciliophora	Ciliata	Rhabdoph	Hymenost	Tetrahym	Tetrahym	Tetrahym	Tetrahym	48 h				Time		190.788	mol/L	
32	930-60-9	CC(Cl)C(=O)CC=O																											
33	1210-05-5	O=Cc1cccc1-c1cccc1C=O																											
34	58402-14-	OC(=O)C(C(=O)O)(Cc1cccc1)c1cccc1																											
35	128946-65	CCCCC(O)C=CC=O																											

Outlook

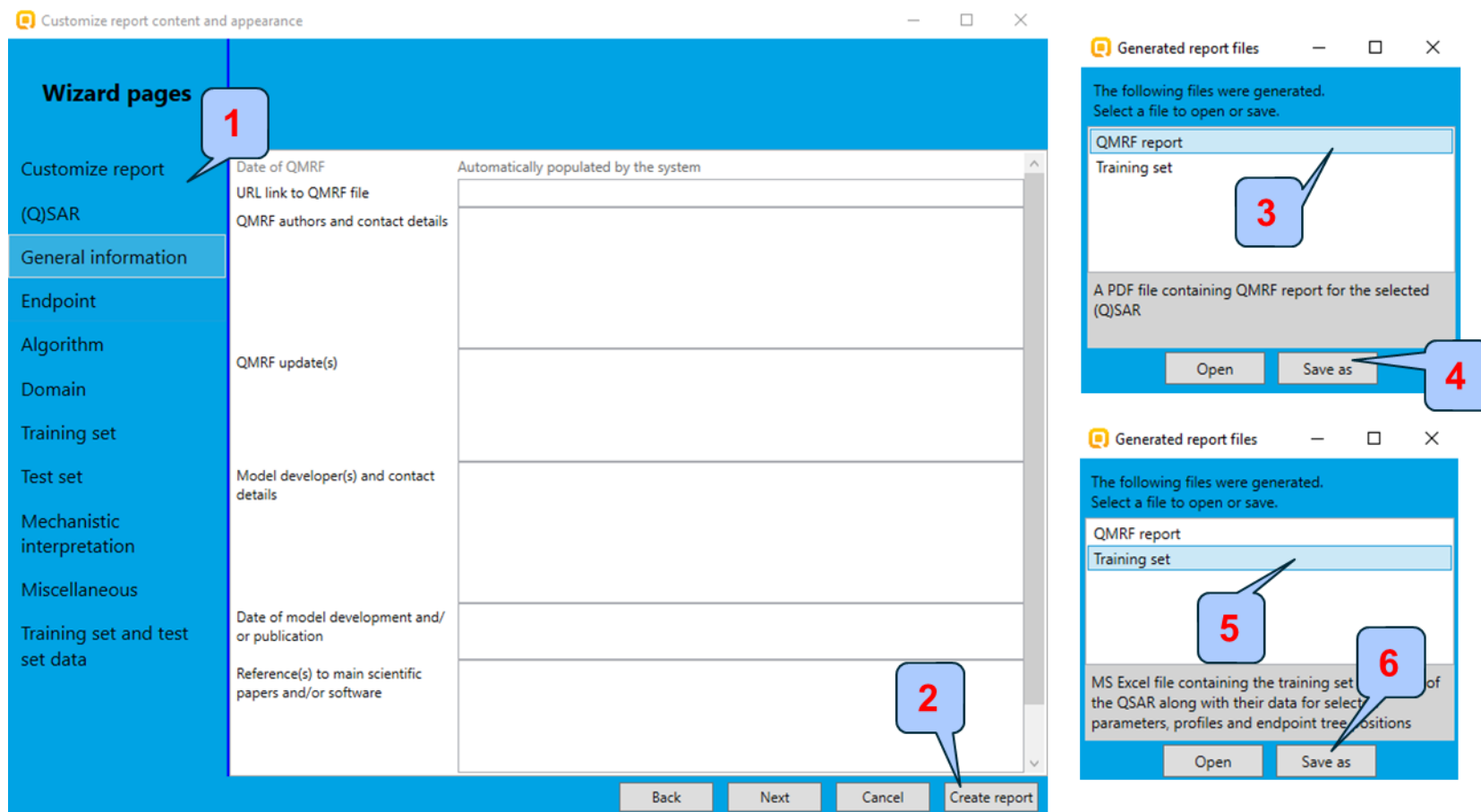
- Background
- Objectives
- The exercise
- **Workflow of the exercise**
 - Input
 - Profiling
 - Data
 - Category definition
 - Data gap filling
 - QSAR model
 - Export QSAR prediction
 - **Report**

Report

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling', and 'Report'. The 'Reports' section on the left has 'Prediction Data Matrix', 'Category', and 'QMRF'. The 'Documents' panel on the left shows a tree structure of chemical categories. The main window displays a table of chemical structures and their associated data. A dialog box titled '(Q)SARs' is open, showing a list of models with '[1] IGC50_Tetrahymena_Furfural (1.0)' selected. Red callout boxes with numbers 1 through 4 indicate the steps: 1. Click 'Report' in the top menu; 2. Select 'QMRF' in the Reports section; 3. Mark the user-defined QSAR model in the (Q)SARs dialog; 4. Click 'OK' in the dialog.

1. **Go** to "Report"; 2. **Select** QMRF; 3. **Mark** the user-defined QSAR model; 4. **Click** OK;

Report



1. **Navigate** through the Wizard to customize the report; 2. **Select** Create report; 3. **Choose** QMRf report to create a PDF format of the report; 4. **Click** Save as; 5. Choose Training set in order to create a MS Excel file (training set of the QSAR along with their data); 6. **Click** Save as;

Report

IGC50_Tetrahymena_Furfural

1 / 4

IGC50_Tetrahymena_Furfural

QMRF Report

A (Q)SAR model

1. (Q)SAR identifier

- 1.1. (Q)SAR identifier (title):
IGC50_Tetrahymena_Furfural (v.1.0)
- 1.2. Other related models:
Not available
- 1.3. Software coding the model:
QSAR Toolbox 4.1

2. General information

- 2.1. Date of QMRF:
Not available
- 2.2. QMRF author(s) and contact details:
Not available
- 2.3. QMRF update(s):
Not available
- 2.4. Model developer(s) and contact details:
Not available
- 2.5. Date of model development and/or publication:
Not available
- 2.6. Reference(s) to main scientific papers and/or software package:
Not available
- 2.7. Availability of information about the model:
Not available
- 2.8. Availability of another QMRF for exactly the same model:
Not available

3. Defining the endpoint (OECD Principle 1)

- 3.1. Species:
Tetrahymena pyriformis

Training set

Training set.xlsx - Microsoft Excel

Training set #1														Training set #2				Training set #3				Training set #4			
1	Substance identity																								
2	Structure		Invalid structure																						
3	CAS number		97-96-1				112-44-7				112-31-2				613-45-6										
4	Chemical name		Ethylbutanal				Undecanal				Decanal				2,4-DIMETHOXYBENZALDEHYD										
5	Other identifier																								
6	SMILES		CCC(CC)C=O				CCCCCCCCC=O				CCCCCCCCC=O				COC1ccc(C=O)c(OC)c1										
7	Parameters		unit																						
8	Profilers																								
9	Training set data and user gathered data																								
10	Training set data																								
11	environment		endpoint		value	unit	species, duration, test type, type of method, assay, strain, test guideline, year, reference	value	unit	species, duration, test type, type of method, assay, strain, test guideline, year, reference	value	unit	species, duration, test type, type of method, assay, strain, test guideline, year, reference	value	unit										
12	Aquatic Toxicity		IGC50	112	mg/L	Tetrahymena pyriformis	3.9	mg/L	Tetrahymena pyriformis	8.2	mg/L	Tetrahymena pyriformis	191	mg/L	Tetrahymena pyriformis										

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.