QSAR TOOLBOX

The OECD QSAR Toolbox for Grouping Chemicals into Categories

OECD QSAR Toolbox v.3.2

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

Outlook

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

Background

- This is a step-by-step presentation designed to take you through the workflow of the Toolbox for building a QSAR model for predicting aquatic toxicity.

- By now you are have some experience in using the Toolbox so there will be multiple key strokes between screen shots.

Outlook

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

Objectives

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

Outlook

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

The Exercise

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

Outlook

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

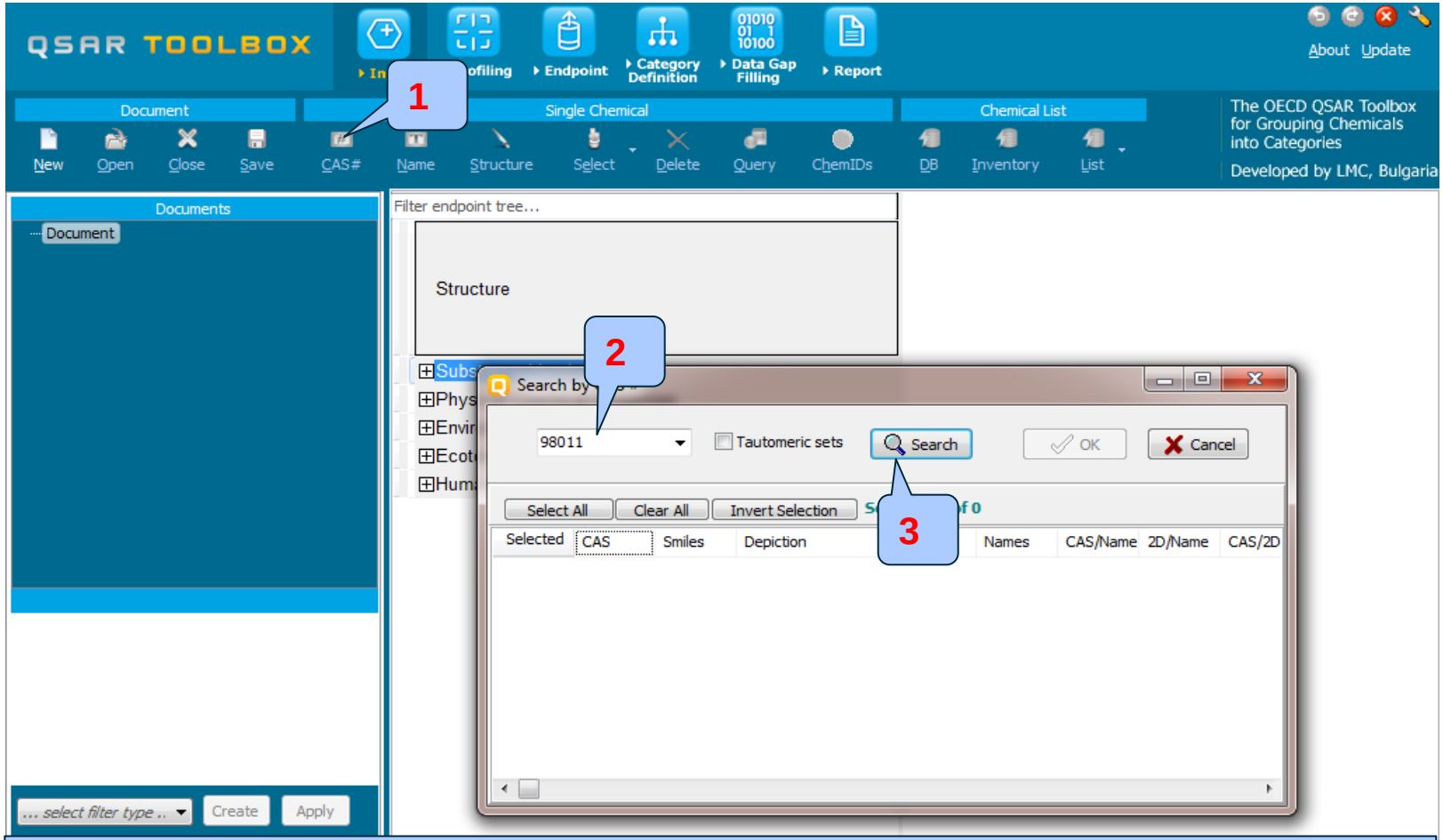
Workflow of the exercise

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

Outlook

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

Chemical Input

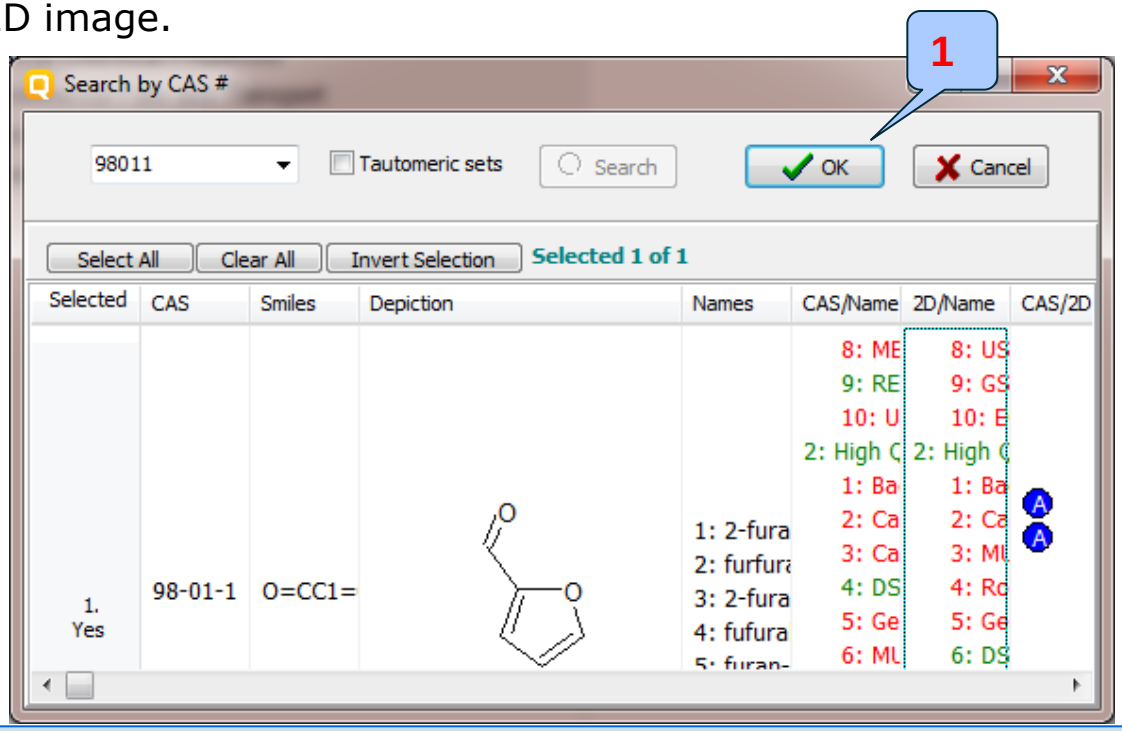


1.**Click** on CAS # 2. **Enter** 98011; 3. **Click** Search

Chemical Input

Target chemical identity

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



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

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

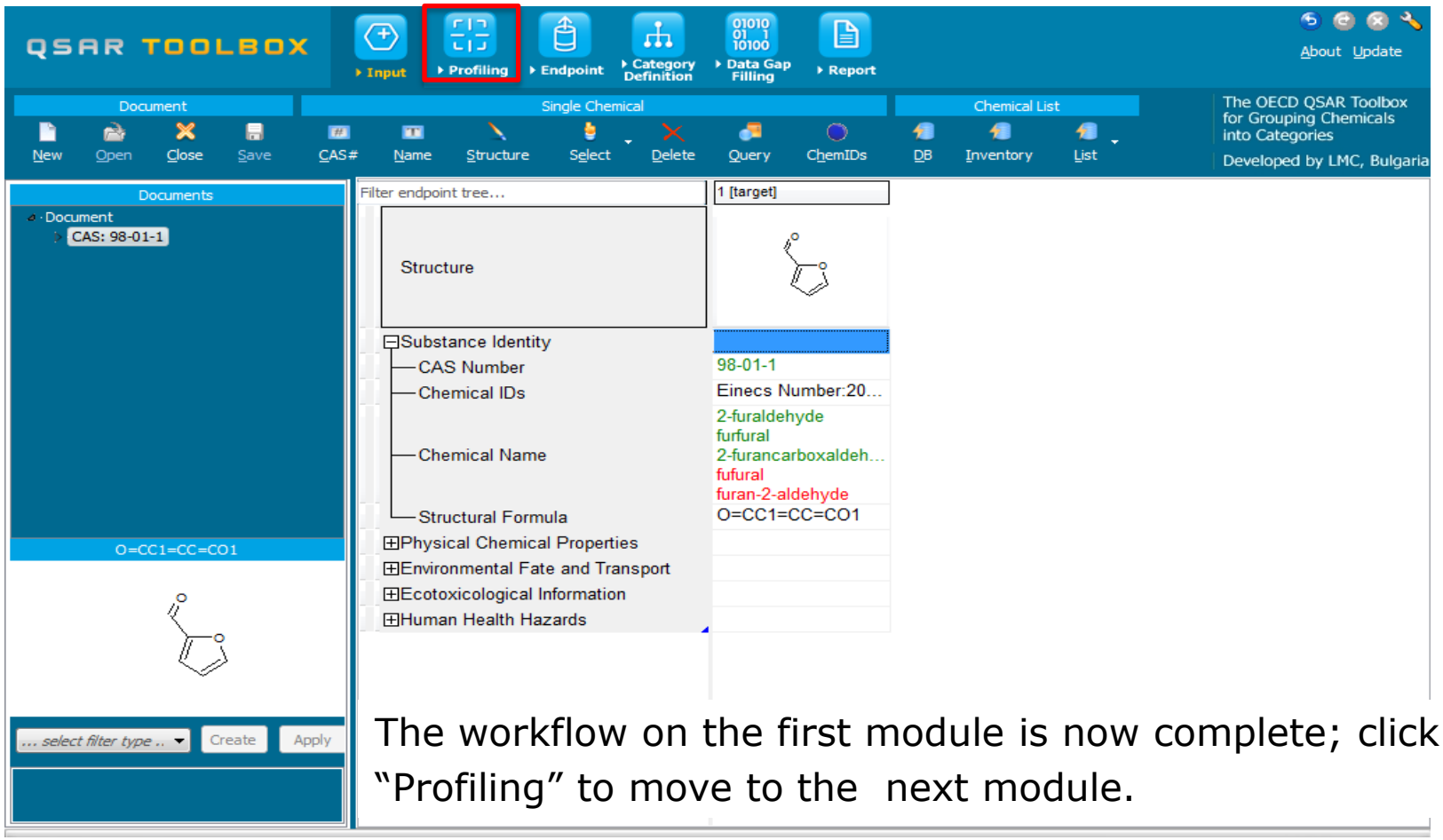
Chemical Input
Target chemical identity

- You have now your target chemical with its structure.

- Click on the box next to “Substance Identity”; this displays the chemical identification information. (see next screen shot)

Chemical Input

Target chemical identity



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

Outlook

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

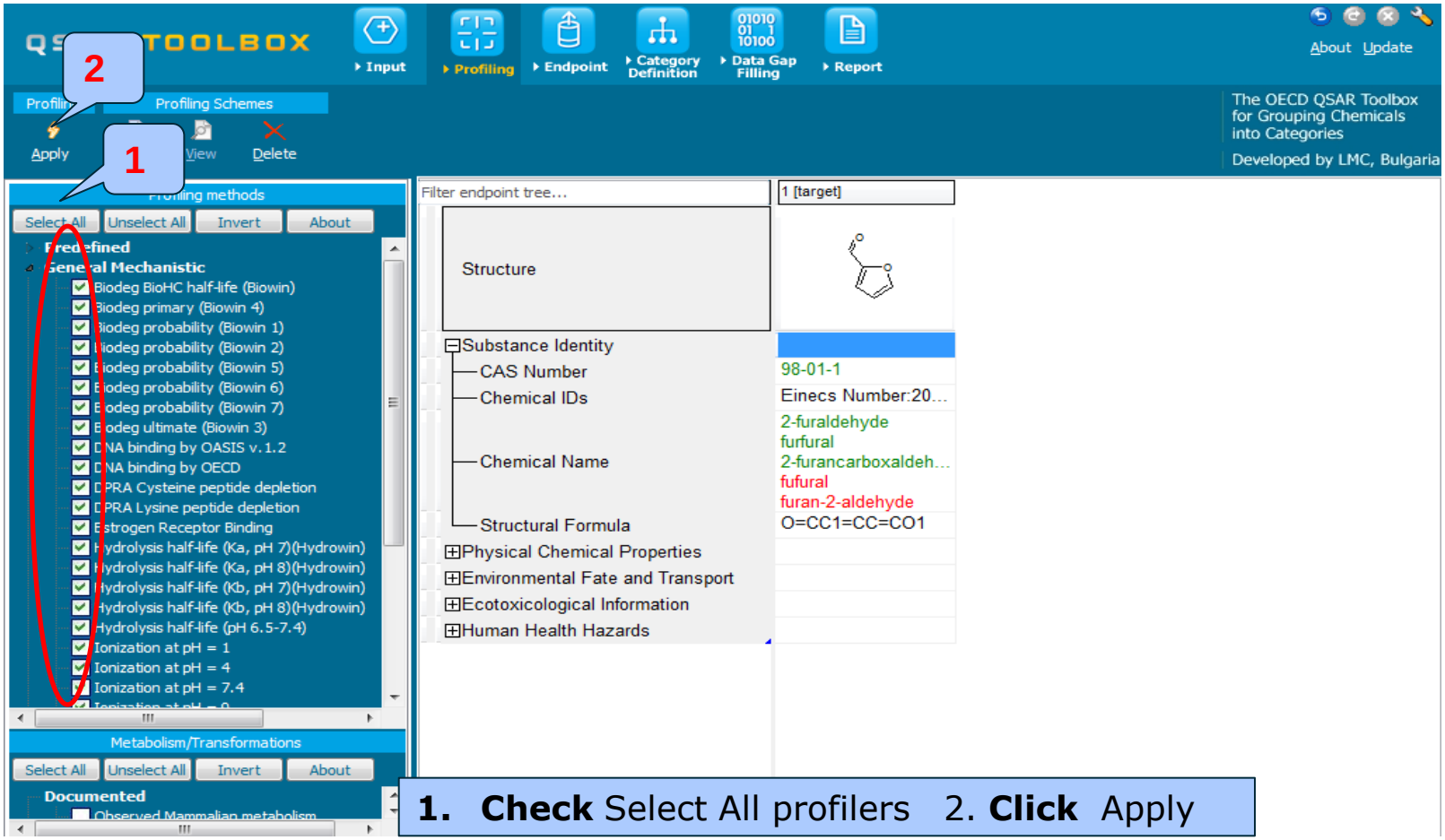
Profiling

Profiling the target chemical

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

Profiling

Profiling the target chemical



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

Profiling

Profiling the target chemical

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

Profiling

Profiles of "Furfural"

QSAR TOOLBOX

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

About Update

Profiling | Profiling Schemes

Apply | New | View | Delete

The OECD QSAR Toolbox for Grouping Chemicals into Categories

Developed by LMC, Bulgaria

Profiling methods

Select All | Unselect All | Invert | About

- Predefined
- General Mechanistic
  - Biodeg BioHC half-life (Biowin)
  - Biodeg primary (Biowin 4)
  - Biodeg probability (Biowin 1)
  - Biodeg probability (Biowin 2)
  - Biodeg probability (Biowin 5)
  - Biodeg probability (Biowin 6)
  - Biodeg probability (Biowin 7)
  - Biodeg ultimate (Biowin 3)
  - DNA binding by OASIS v.1.2
  - 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 (Ka, 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
  - Protein binding by OECD
  - Protein binding potency
  - Superfragments
  - Toxic hazard classificatio

Filter endpoint tree...

1 [target]

Structure

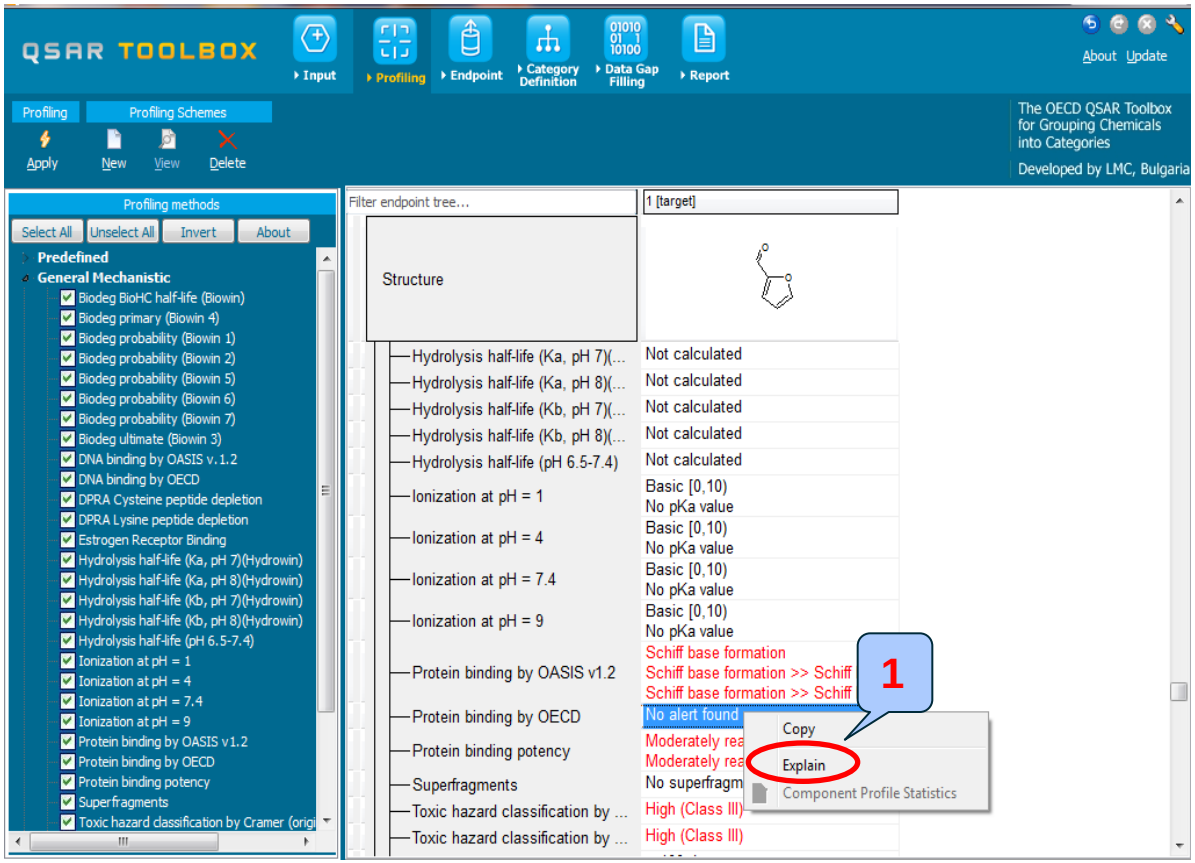
- Substance Identity
  - CAS Number: 98-01-1
  - Chemical IDs: Einecs Number:20...
  - Chemical Name: 2-furaldehyde; furfural; 2-furancarboxaldeh...; fufural; furan-2-aldehyde
  - Structural Formula: O=CC1=CC=CO1
- Physical Chemical Properties
- Environmental Fate and Transport
- Ecotoxicological Information
- Human Health Hazards
- Profile

1

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

Profiling

Profiles of "Furfural"

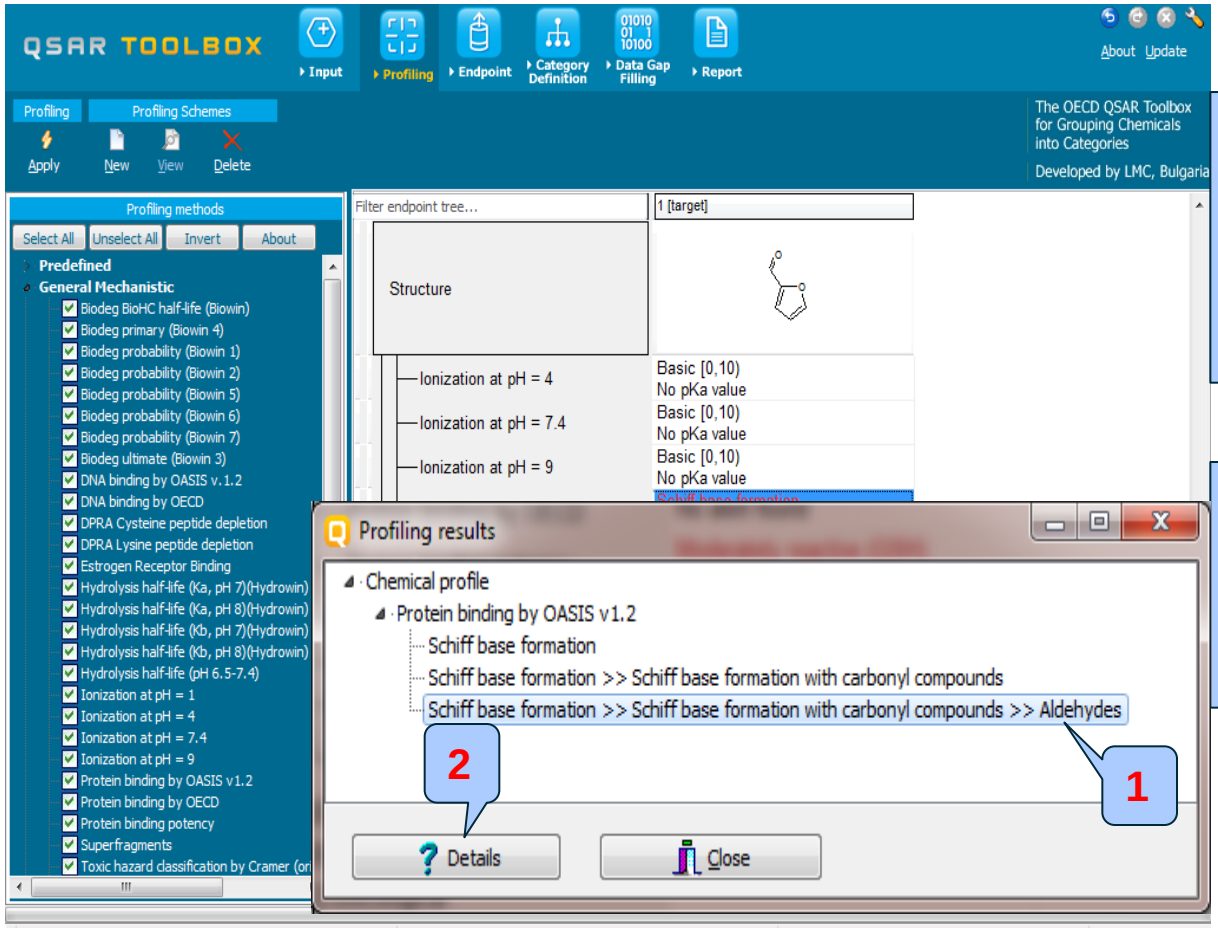


In this case there is structural evidence that the target could interact to DNA and proteins, it has also mode of action and it is aldehyde. This step is critical for next grouping of analogues.

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

Profiling

Profiles of "Furfural"



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

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

Profiling

Protein binding by OASIS v.1.2 of target chemical

Protein binding by OASIS v1.2

Aldehydes

Target

Boundaries | Training set | Options

Boundary Options | Metabolism

Fragment

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

Common Fragments

	Definition	1	2
1	Hetl	N	C
2	RP V1	C X1	H X1

Profile Description

Mechanistic Domain: Schiff base formation

Mechanistic Alert: Schiff base formation with carbonyl compounds

Structural Alert: Aldehydes

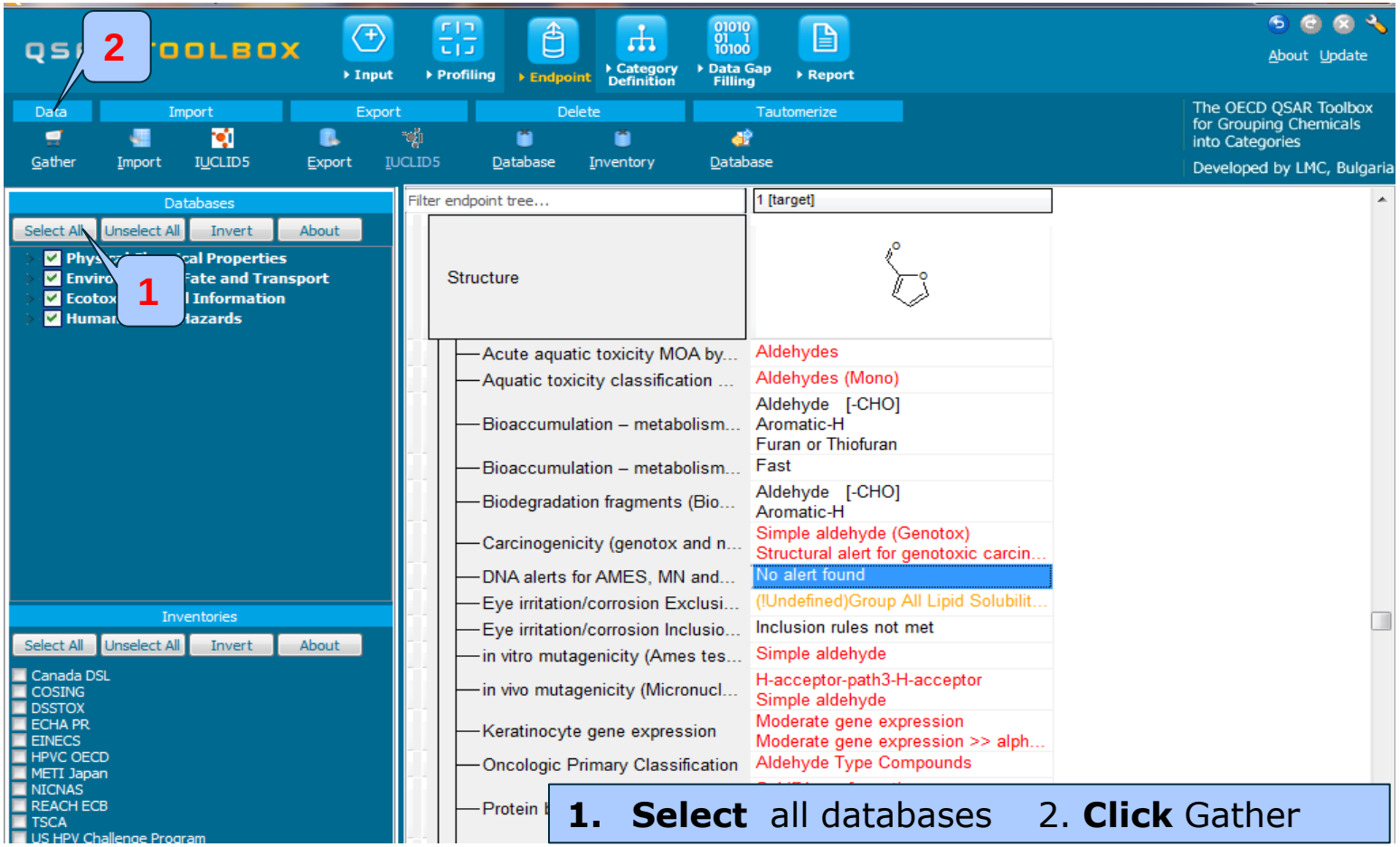
Profile Comments

Outlook

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

Endpoints

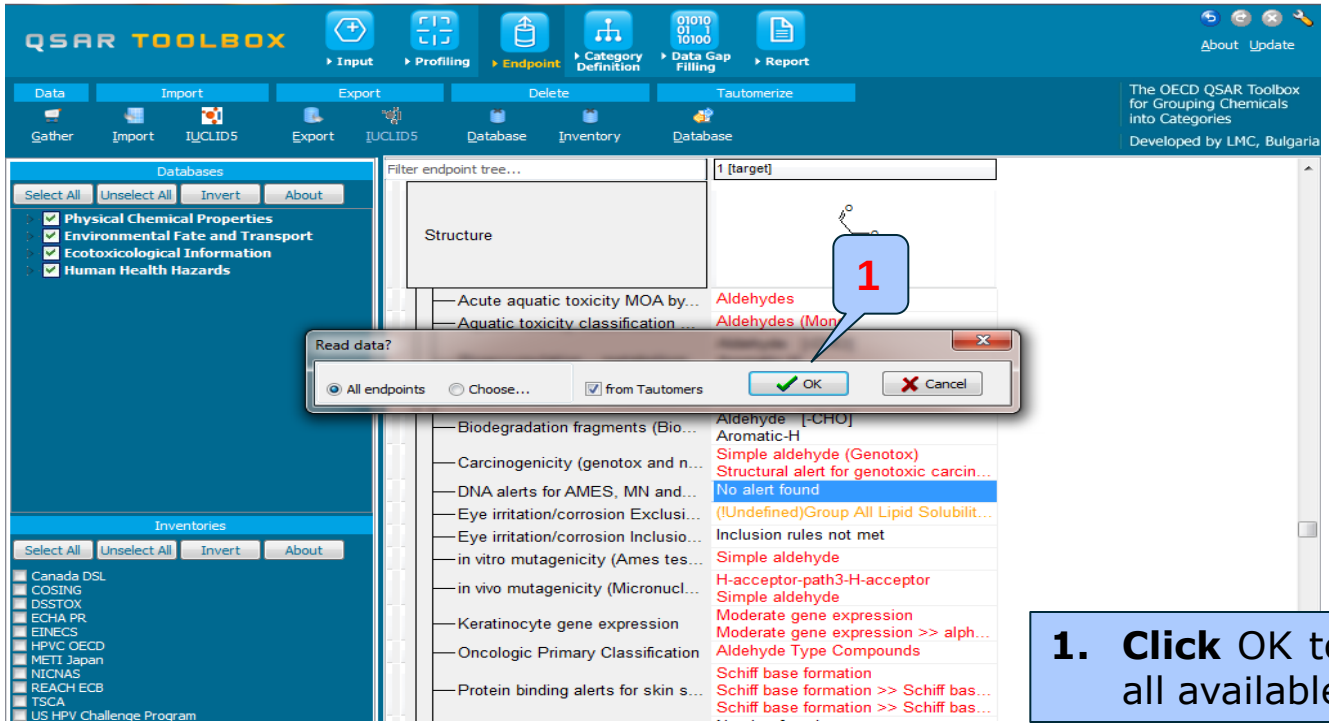
Extracting endpoint values



Endpoints
Process of collecting data

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

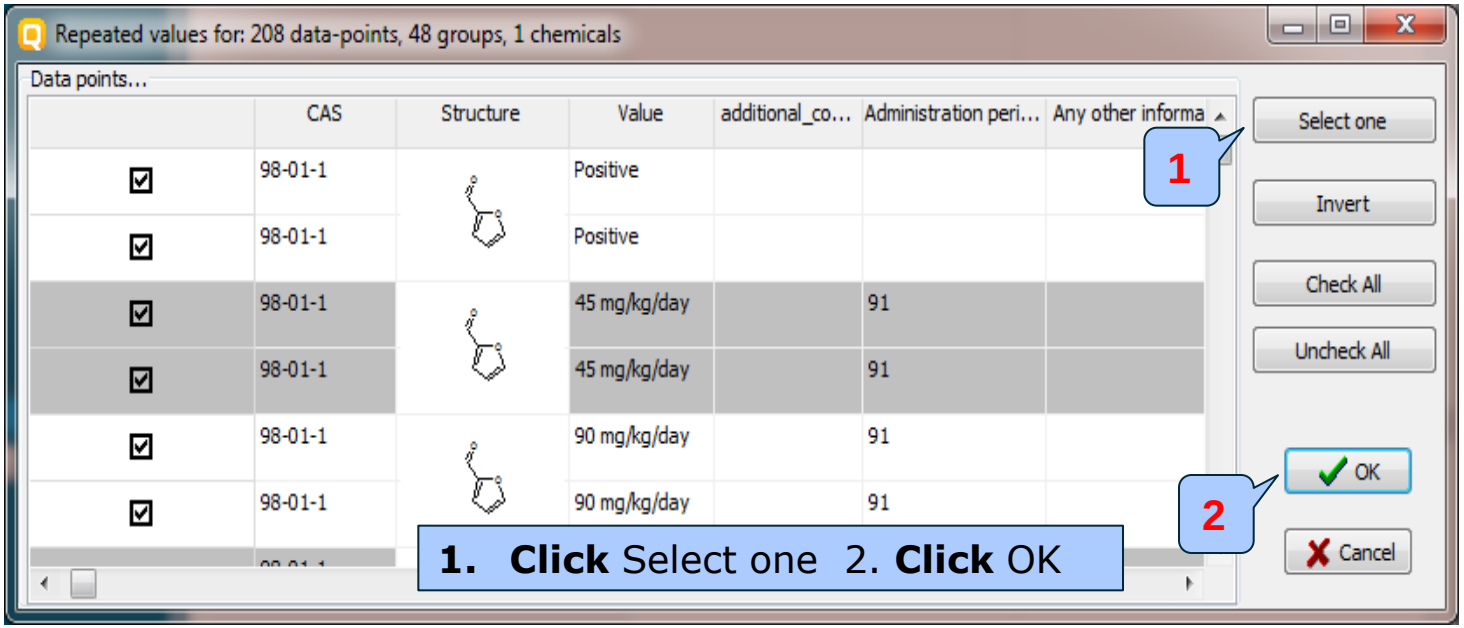
A window with "Read data?" appears. Now the user could choose to collect "all" or "endpoint specific" data.



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

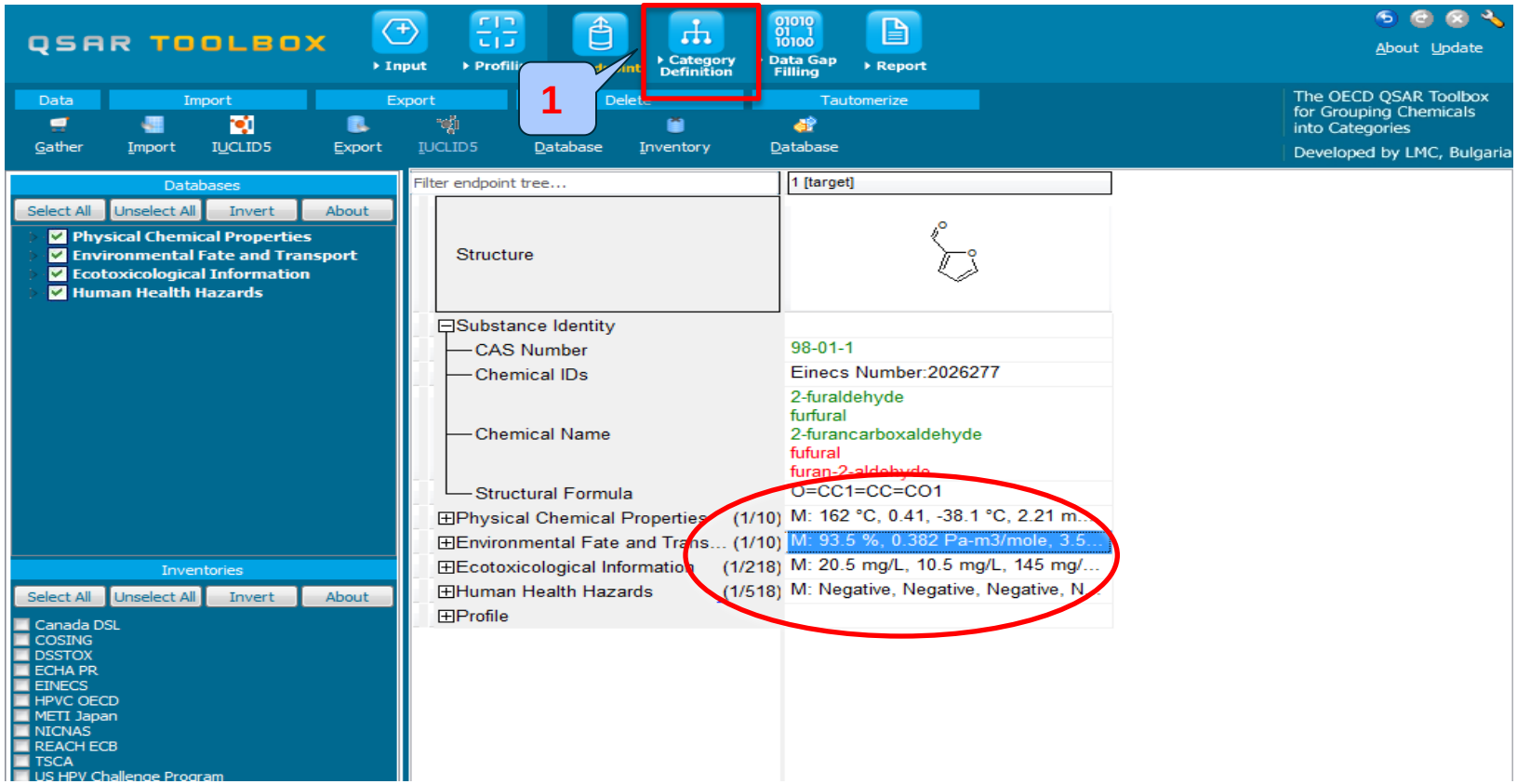
Endpoints
Read data for analogues

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



Endpoints

Inserting data for target in data matrix



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

Outlook

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

Category definition
Target endpoint

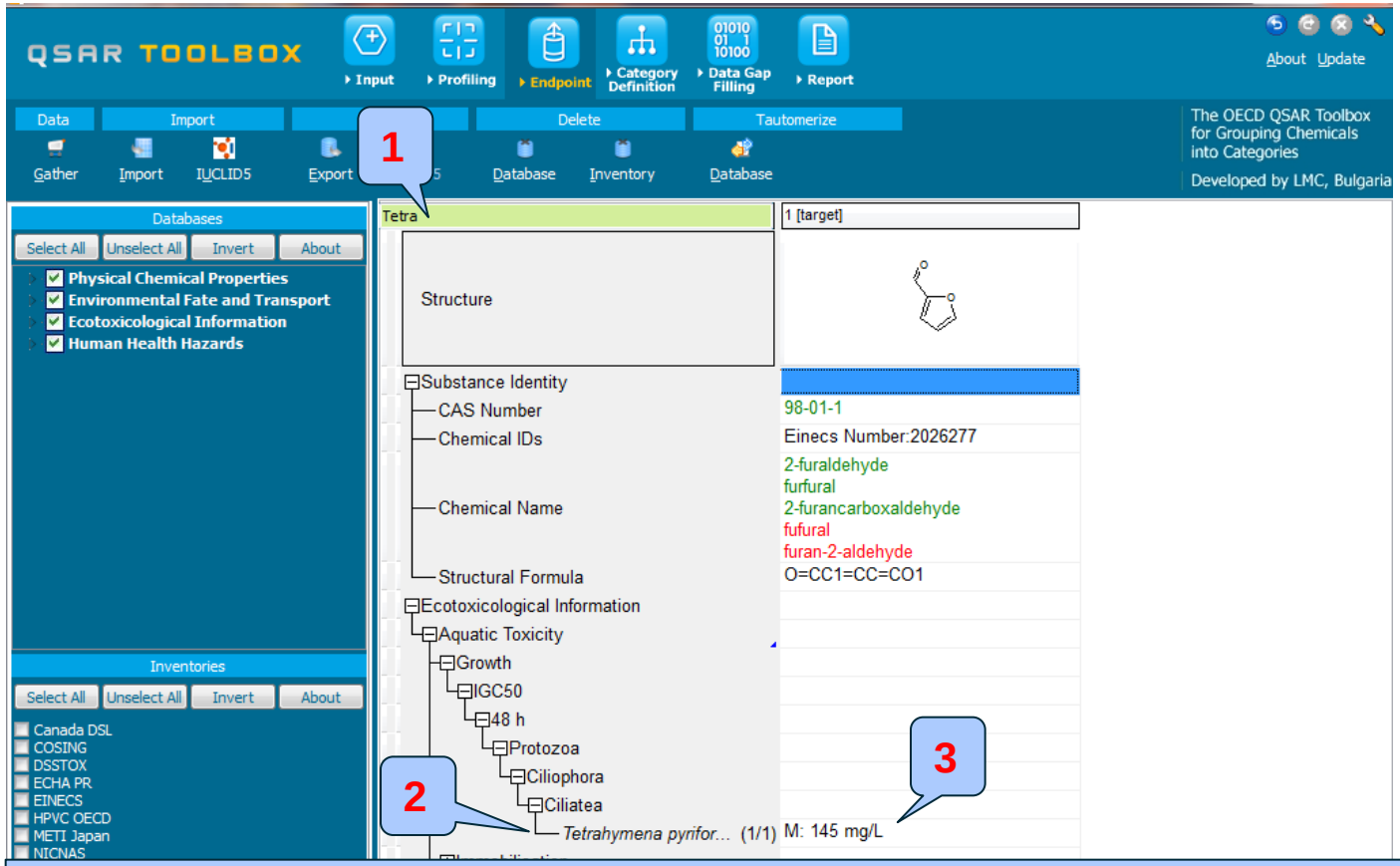
- In this exercise we will build a QSAR model to estimate the following endpoint :

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

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

Category definition

Navigate to the target endpoint



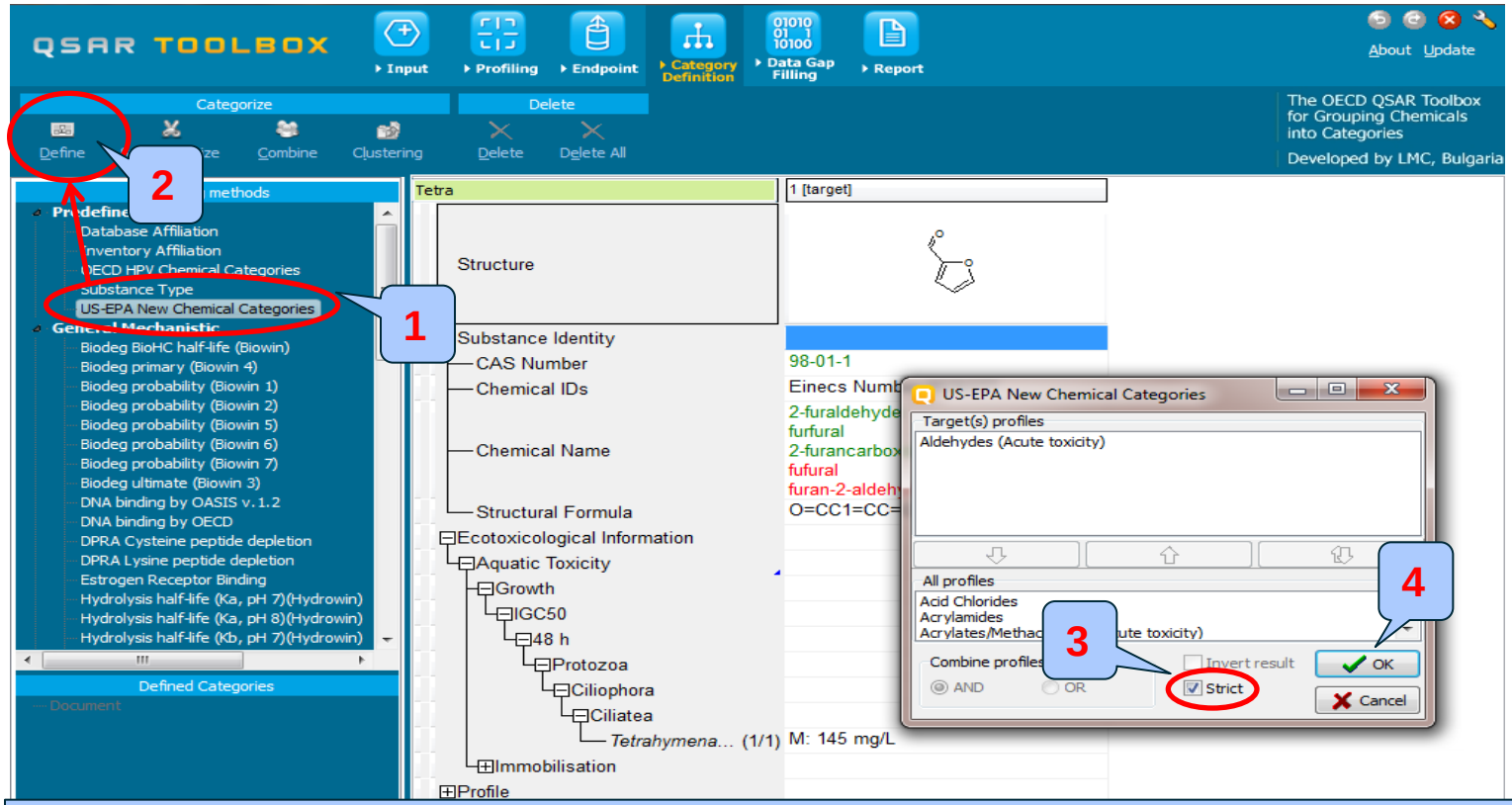
1. **Type** "Tetra" in the empty filter field; 2. **Open** the nodes to target endpoint; 3. **Highlight** the cell that will be filled in (in this case we will reproduce the observed data).

Category definition
Defining US-EPA category

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

Category definition

Defining US-EPA category



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

Category definition

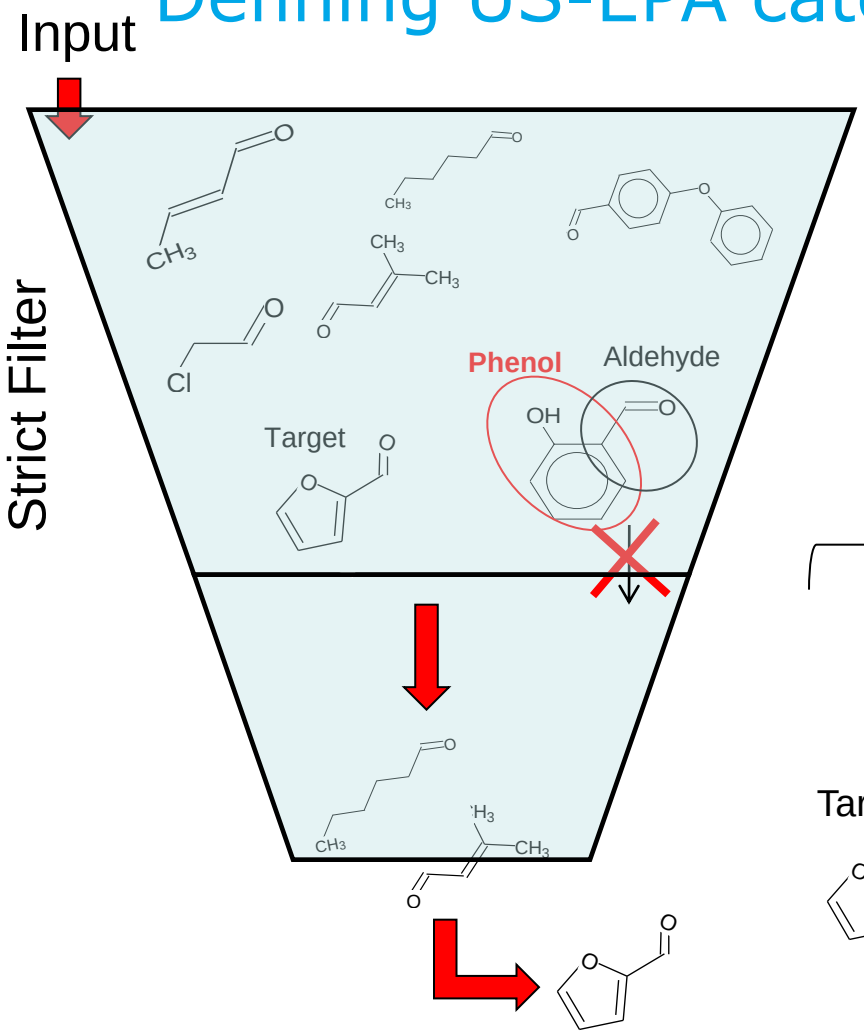
Defining US-EPA category strict functionality

- The **Strict** functionality means that the software will group analogues having **ONLY** the categories of the target and will exclude the analogues having any other categories according to the profiler used in the grouping method.

- For example, if the profiling for the target results in *Aldehydes(Acute toxicity)* **ONLY** according to US-EPA category, the group of analogues will include *Aldehydes(Acute toxicity)* **ONLY**.(See next screen shot)

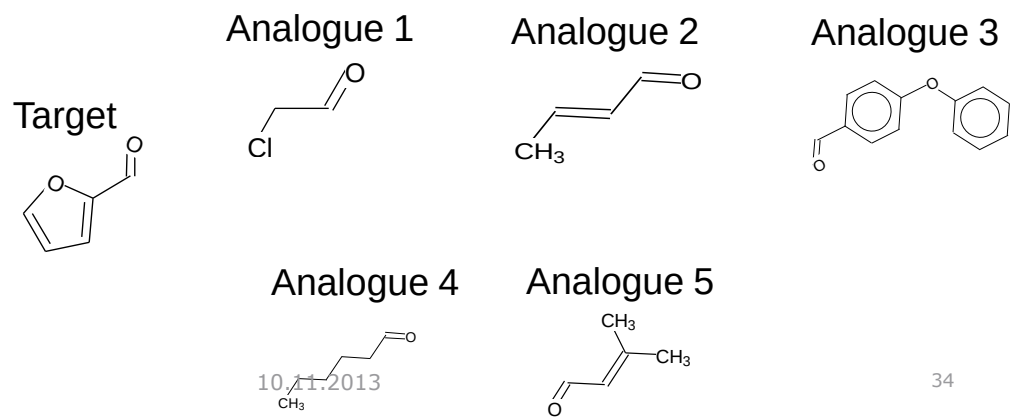
Category definition

Defining US-EPA category strict functionality



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

Defined Category



Category definition

Defining US-EPA category

QSAR TOOLBOX

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

About Update

Categorize: Define, Subcategorize, Combine, Clustering

Delete: Delete, Delete All

The OECD QSAR Toolbox for Grouping Chemicals into Categories

Developed by LMC, Bulgaria

Grouping methods

- Predefined
  - Database Affiliation
  - Inventory Affiliation
  - OECD HPV Chemical Categories
  - Substance Type
  - US-EPA New Chemical Categories
- General Mechanistic
  - Biodeg BioHC half-life (Biowin)
  - Biodeg primary (Biowin 4)
  - Biodeg probability (Biowin 1)
  - Biodeg probability (Biowin 2)
  - Biodeg probability (Biowin 5)
  - Biodeg probability (Biowin 6)
  - Biodeg probability (Biowin 7)
  - Biodeg ultimate (Biowin 3)
  - DNA binding by OASIS v.1.2
  - 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 (Ka, pH 8)(Hydrowin)
  - Hydrolysis half-life (Kb, pH 7)(Hydrowin)

Defined Categories

Document

Filter endpoint tree...

1 [target]

Structure

Substance Identity

CAS Number: 98-01-1

Chemical IDs: Einecs Number:2026277

Define category name

Category name (513 chemicals): toxicity) Strict (US-EPA New Chemical Categories)

OK | Cancel

1

Ecotoxicological Information

Aquatic Toxicity

Behavior	(1/3)	M: 29 mg/L, 13 mg/L, 51 mg/L
Behaviour and Morphology...	(1/2)	M: 0.33 mg/L, 0.47 mg/L
Cell Multiplication Inhibition	(1/1)	M: 16 mg/L
Cell Number	(1/2)	M: 2.7 mg/L, 31 mg/L
Development	(1/3)	M: 0.915 mg/L, 0.624 mg/L, 0.426 ...

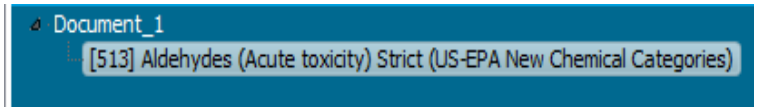
1. **Click** OK to confirm the name of the category

Category definition
Analogues

- The Toolbox now identifies all chemicals corresponding to *Aldehydes(Acute toxicity)* by US-EPA listed in the databases selected under "Endpoints".

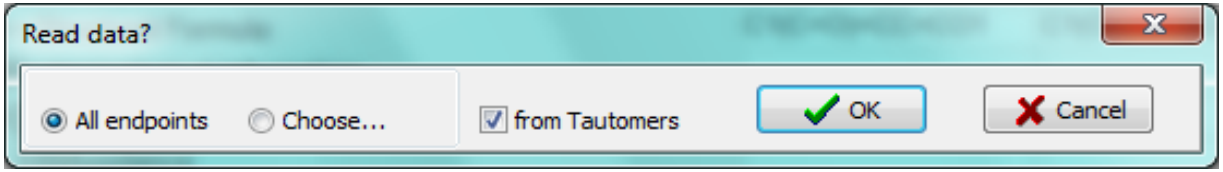
- 513 analogues including the target chemical are identified; they form a mechanistic category "**Aldehydes (Acute toxicity)**", which will be used for gap filling.

- The name of the analogues and name of the category appear in the "Defined Categories" window.



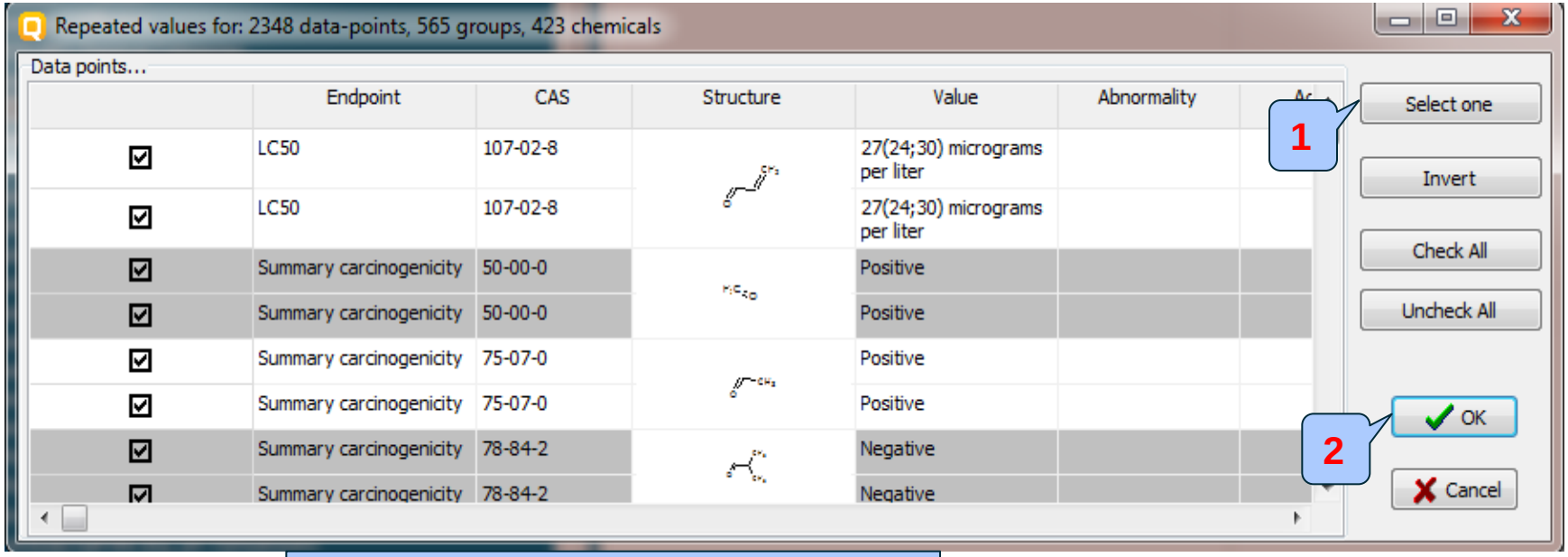
Category definition
Reading data for Analogues

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



Category definition
Reading data for Analogues

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



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

Category definition
Summary information for Analogues

The experimental results for the analogues are inserted into the matrix

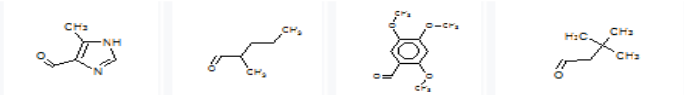
QSAR TOOLBOX
Input | Profiling | Endpoint | Category Definition | Data Gap Filling | Report
About Update

Categorize: Define, Subcategorize, Combine, Clustering
Delete: Delete, Delete All

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

Grouping methods
- Predefined
  - Database Affiliation
  - Inventory Affiliation
  - OECD HPV Chemical Categor
  - Substance Type
  - US-EPA New Chemical Categ
- General Mechanistic
  - Biodeg BioHC half-life (Biowin
  - Biodeg primary (Biowin 4)
  - Biodeg probability (Biowin 1)
  - Biodeg probability (Biowin 2)
  - Biodeg probability (Biowin 5)
  - Biodeg probability (Biowin 6)
  - Biodeg probability (Biowin 7)
  - Biodeg ultimate (Biowin 3)
  - DNA binding by OASIS v.1.2
  - DNA binding by OECD
  - DPRA Cysteine peptide deple
  - DPRA Lysine peptide depletic
  - Estrogen Receptor Binding
  - Hydrolysis half-life (Ka, pH 7
  - Hydrolysis half-life (Ka, pH 8
  - Hydrolysis half-life (Kb, pH 7

Defined Categories
- Document_1
  - [513] Aldehydes (Acute toxicity)
  - [513] sorted descending(targets

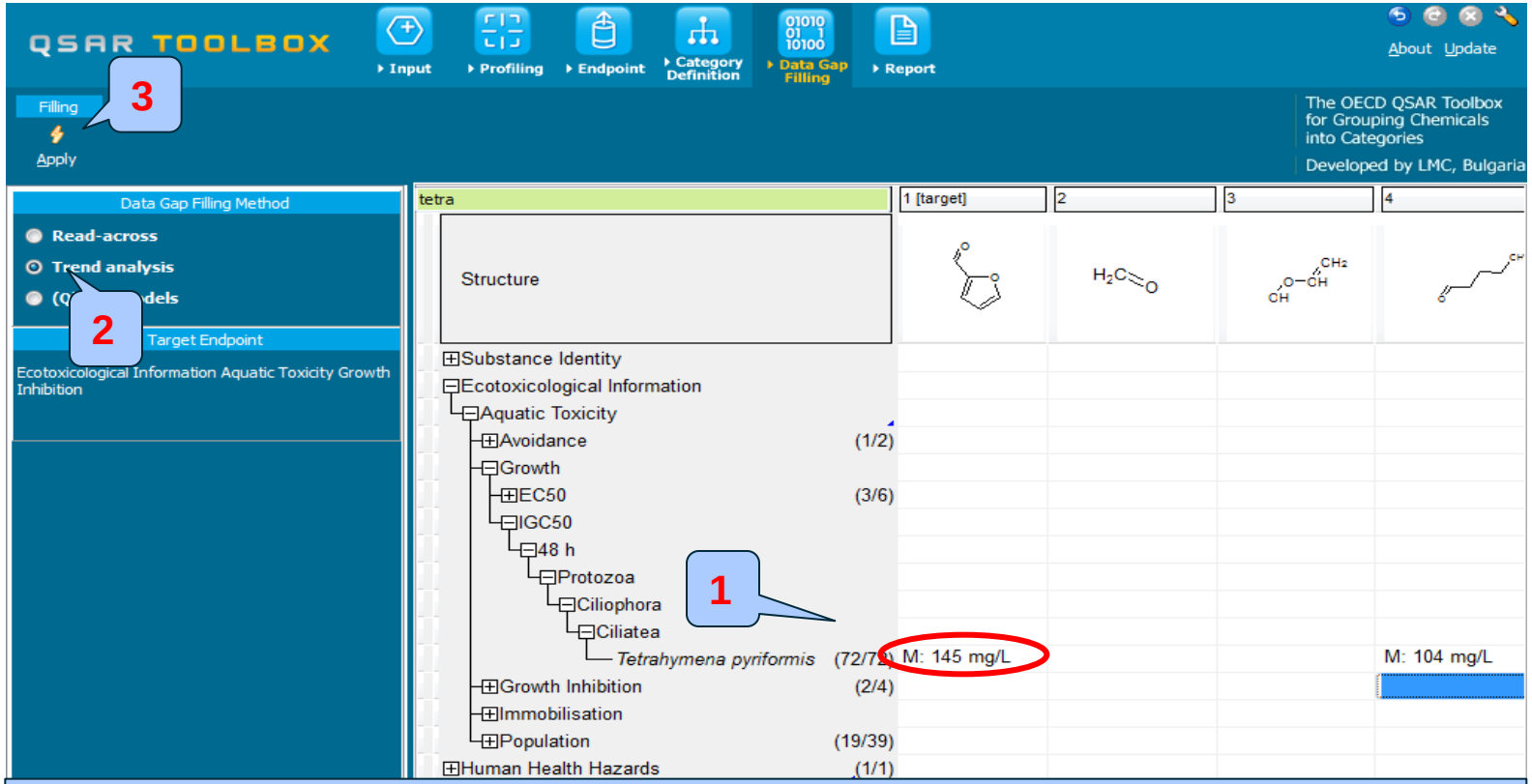
tetra	2	3	4	5	6
Structure					
Substance Identity					
Ecotoxicological Information					
Aquatic Toxicity					
Avoidance (1/2)					
Growth					
EC50 (3/6)					
IGC50					
48 h					
Protozoa					
Ciliophora					
Ciliatea					
Tetrahymena pyriformis (72/73)	M: 937 mg/L	M: 296 mg/L	M: 247 mg/L	M: 235 mg/L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

Outlook

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

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

Apply Trend analysis



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

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

QSAR TOOLBOX

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

About Update

Filling

Apply

The OECD QSAR Toolbox for Grouping Chemicals into Categories

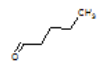
Developed by LMC, Bulgaria

Data Gap Filling Method

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

Target Endpoint

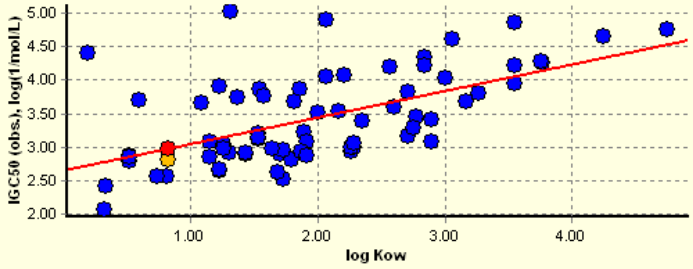
Ecotoxicological Information Aquatic Toxicity Growth IGC50 48 h Protozoa Ciliophora Ciliatea Tetrahymena pyriformis

tetra	1 [target]	4	10	13
Structure				
Tetrahymena pyriformis (72/72)	M: 145 mg/L	M: 104 mg/L	M: 2.18 mg/L	M: 134 mg/L

Descriptors | Prediction | Adequacy | Cumul. freq. | Statistics | Residuals

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

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



Descriptor X: log Kow

Accept prediction

Return to matrix

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

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

Interpreting dots on the graph

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

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

An accurate analysis of data set

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

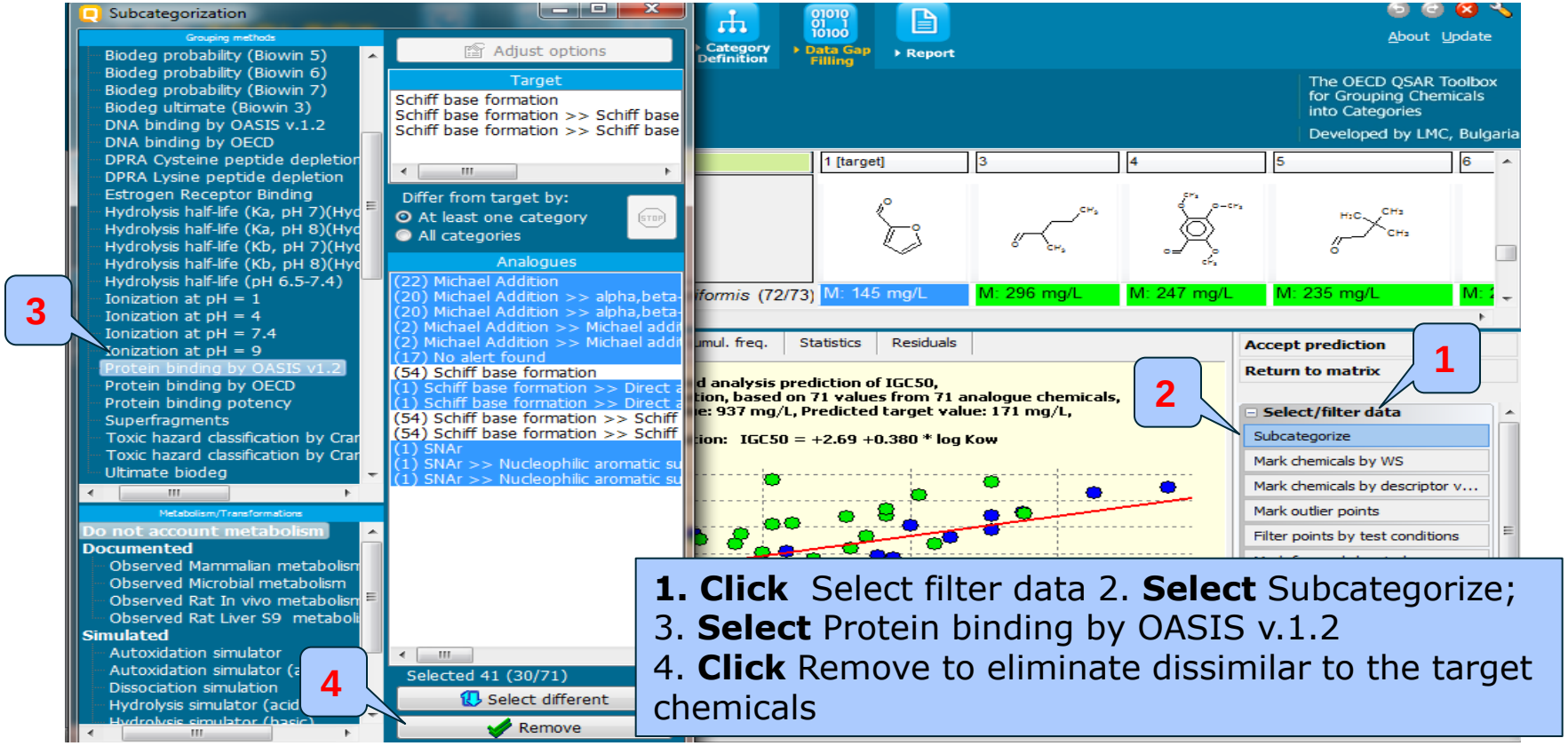
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

Subcategorisation by Protein binding by OASIS v.1.2

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

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

Subcategorisation by Protein binding by OASIS v.1.2



1. Click Select filter data 2. **Select** Subcategorize;
3. **Select** Protein binding by OASIS v.1.2
4. **Click** Remove to eliminate dissimilar to the target chemicals

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

Results after subcategorisation

QSAR TOOLBOX

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

About Update

Filling

Apply

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

tetra	1 [target]	4	14	16
Structure				
Tetrahymena pyriformis (72/72)	M: 145 mg/L	M: 104 mg/L	M: 194 mg/L	M: 152 mg/L

Descriptors | Prediction | Adequacy | Cumul. freq. | Statistics | Residuals

Trend analysis prediction of IGC50,
making a linear approximation, based on 30 values from 30 analogue chemicals,
Observed target value: 145 mg/L, Predicted target value: 275 mg/L,

Model equation: IGC50 = +2.10 +0.539 * log Kow

IGC50 (obs.), log(1/mol/L): 4.50, 4.00, 3.50, 3.00, 2.50, 2.00

log Kow: 1.00, 2.00, 3.00, 4.00

Descriptor X: log Kow

Accept prediction

Return to matrix

Select/filter data

- Subcategorize
- Mark chemicals by WS
- Mark chemicals by descriptor v...
- Mark outlier points
- Filter points by test conditions
- Mark focused chemical
- Mark focused points
- Remove marked chemicals/points
- Clear existing marks

Selection navigation

Gap filling approach

Descriptors/data

Model/(Q)SAR

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

An accurate trend analysis of data set

- The chemicals which differ from the target according to Protein binding by OASIS v1.2 are:
  - Michael addition<<alpha, beta – unsaturated carbonyl compounds<< alpha, beta-unsaturated aldehydes (20);
  - Michael addition<< Michael addition on conjugated systems with electron withdrawing group<< alpha, beta-Carbonyl compounds with polarized (2);
  - No alert found (17);
  - SNAr<<Nucleophilic aromatic substitution on activated halogens<<Activated aryl and hetetoalyl compounds(1).

- Another way for refining the data set is to ask what makes the obvious outliers different from the target.
- Click on Selection navigation then, click Back (see next screen shot).

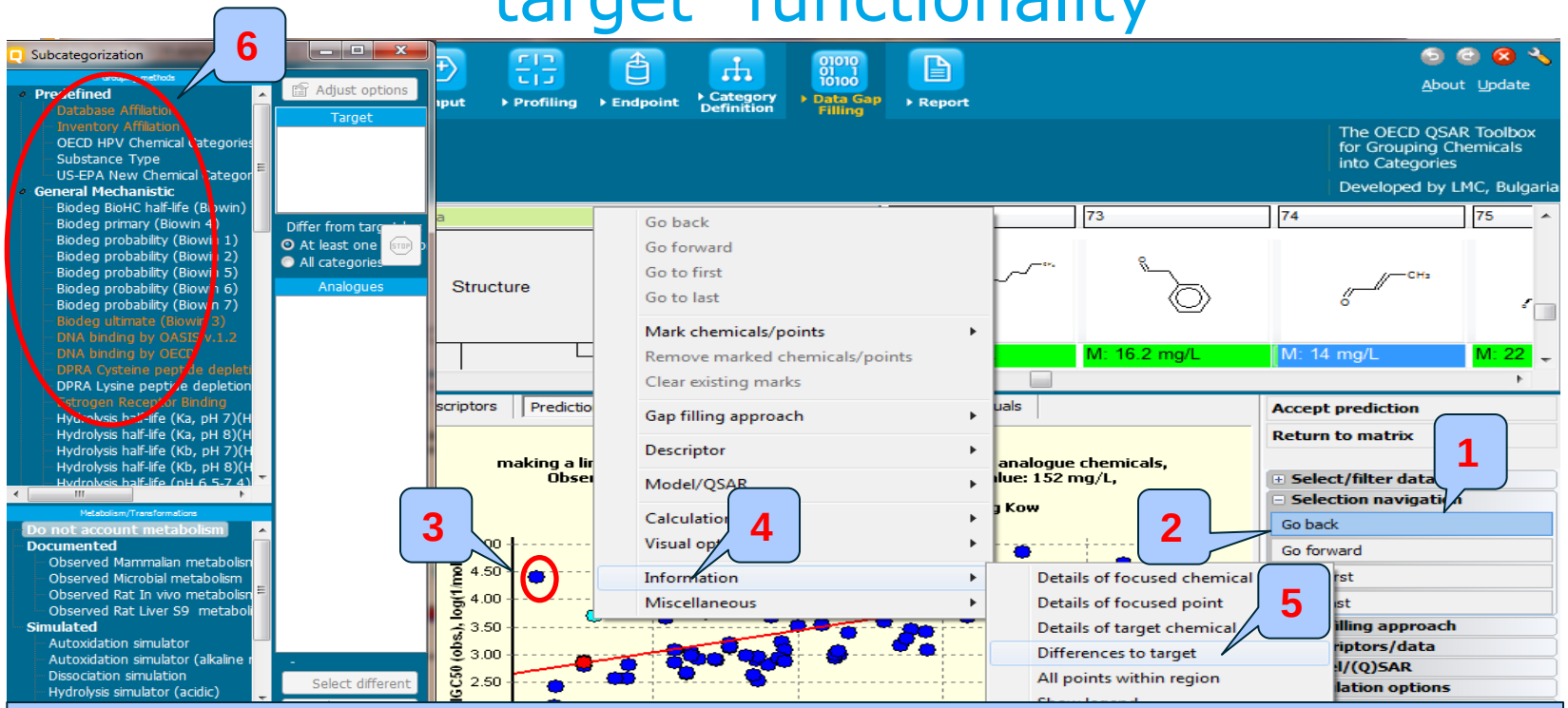
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

Subcategorisation by using "Difference to target" functionality

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

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

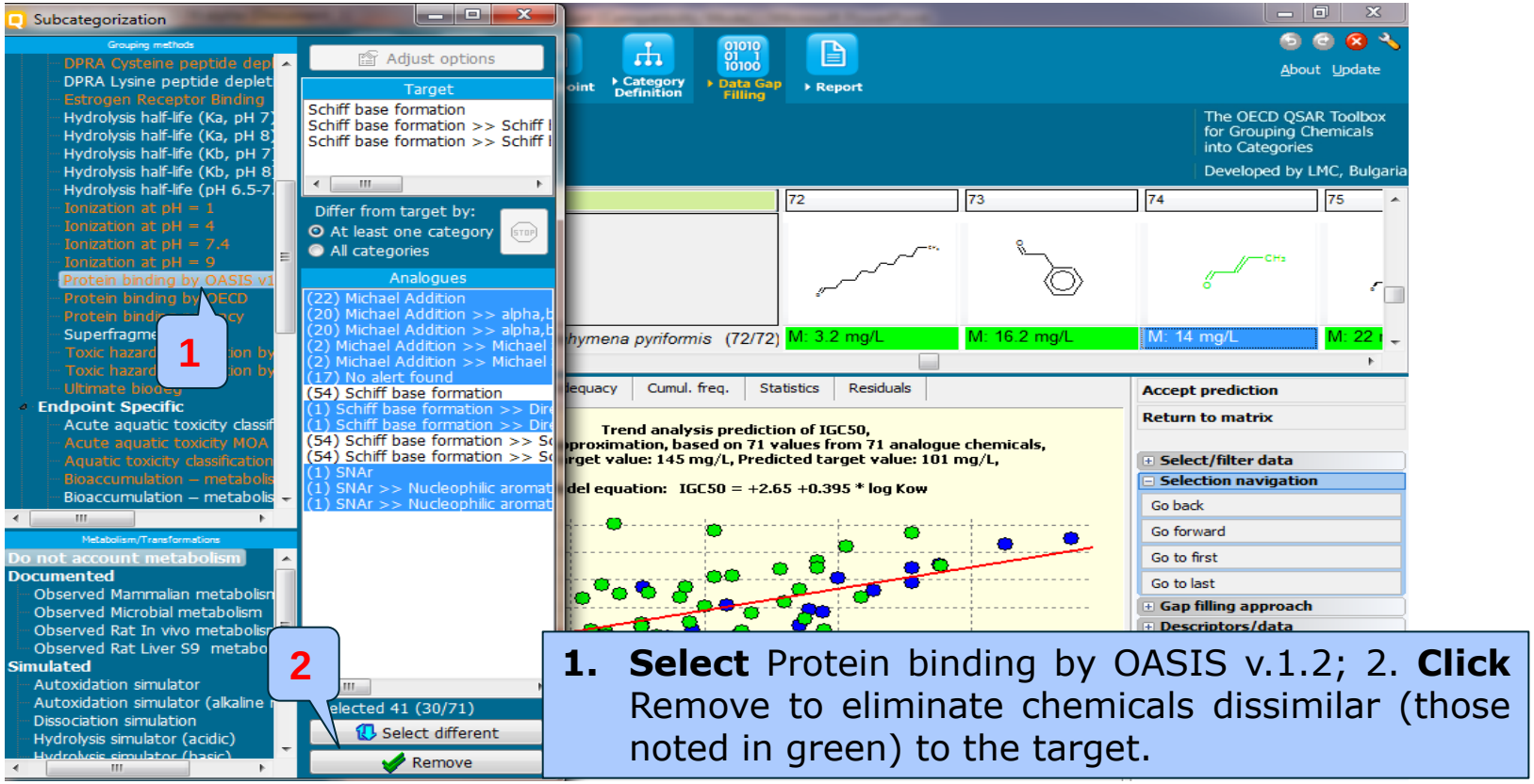
Subcategorisation by using "Difference to target" functionality



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

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

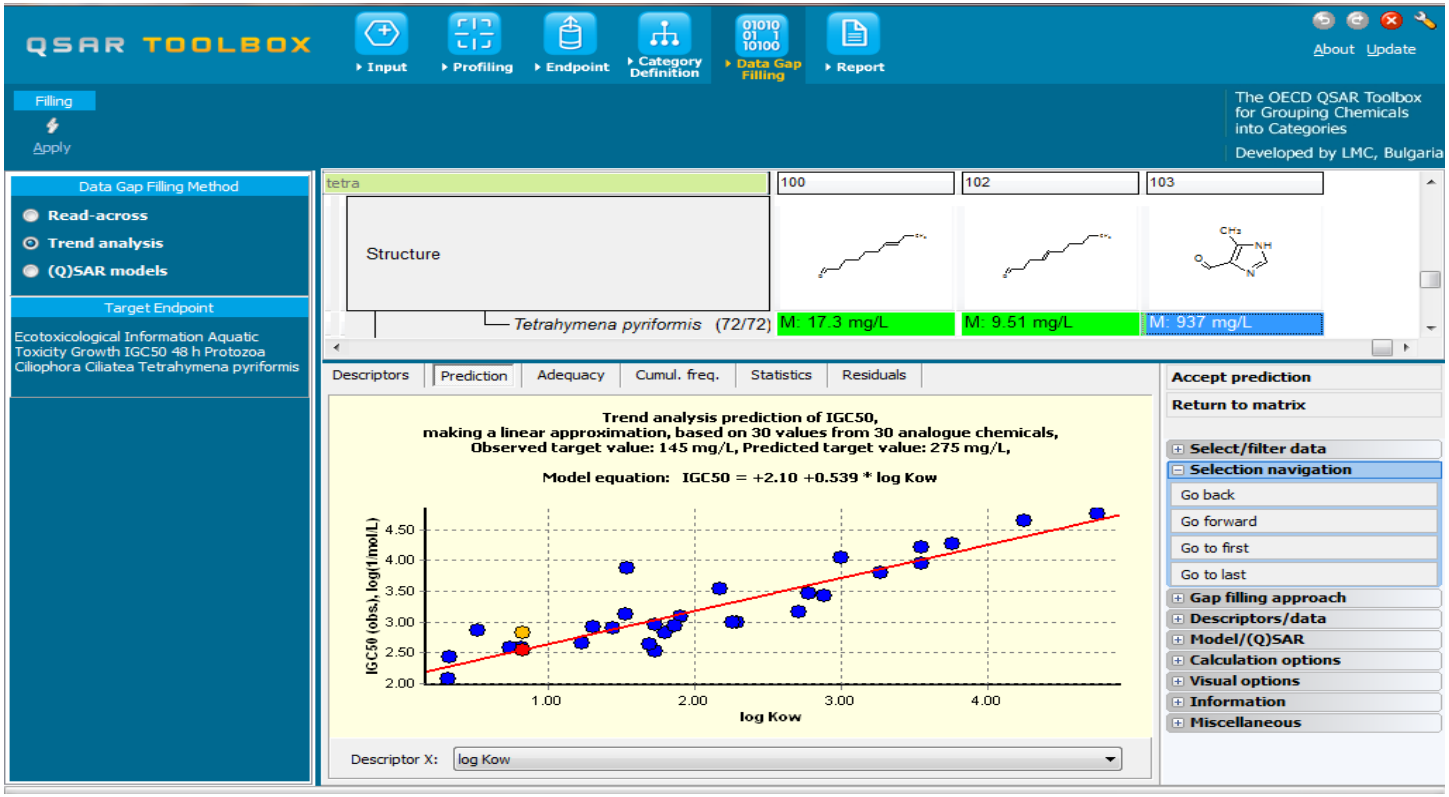
Subcategorisation by using "Difference to target" functionality



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

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

Subcategorisation by using "Difference to target" functionality



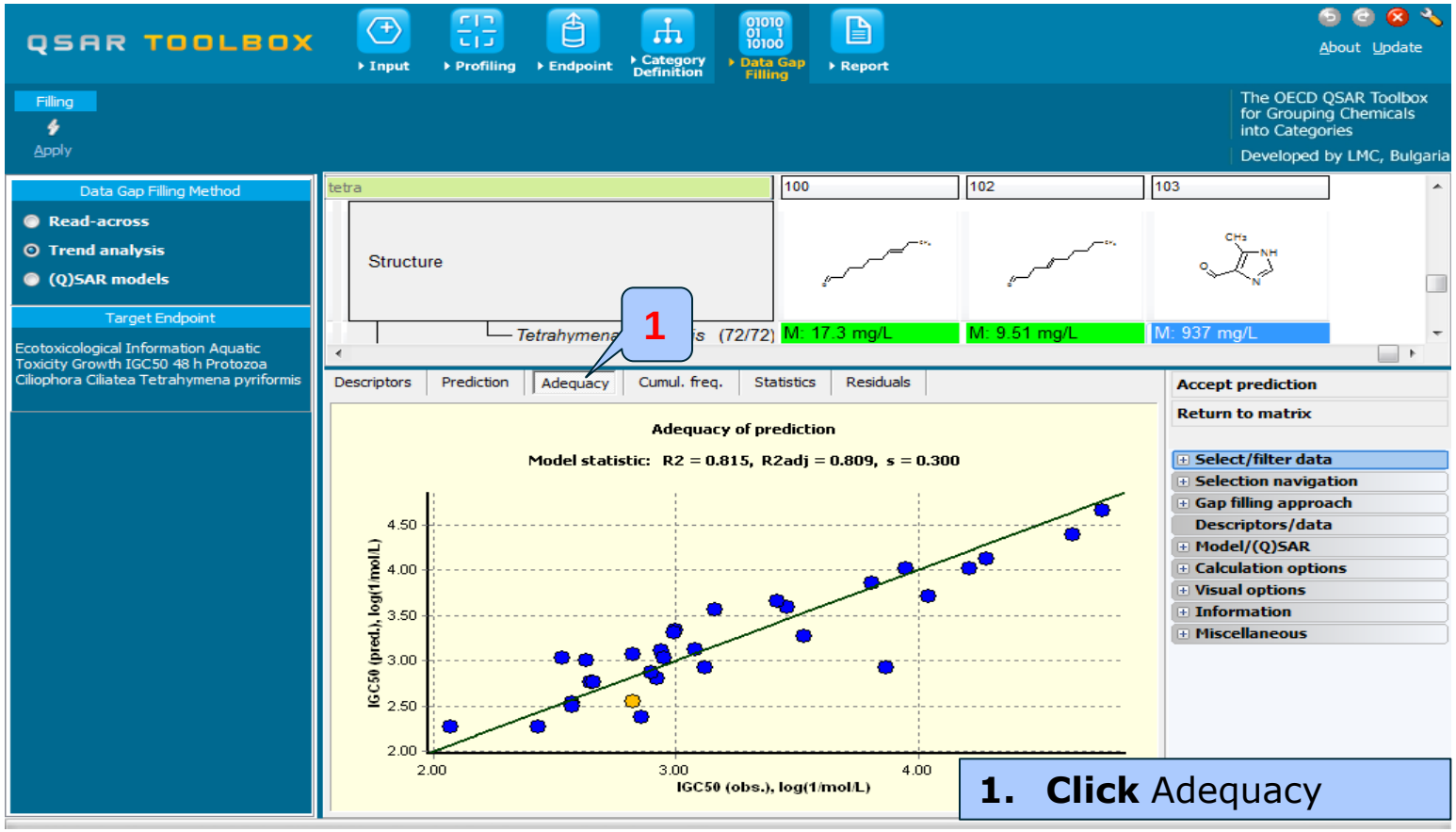
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

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

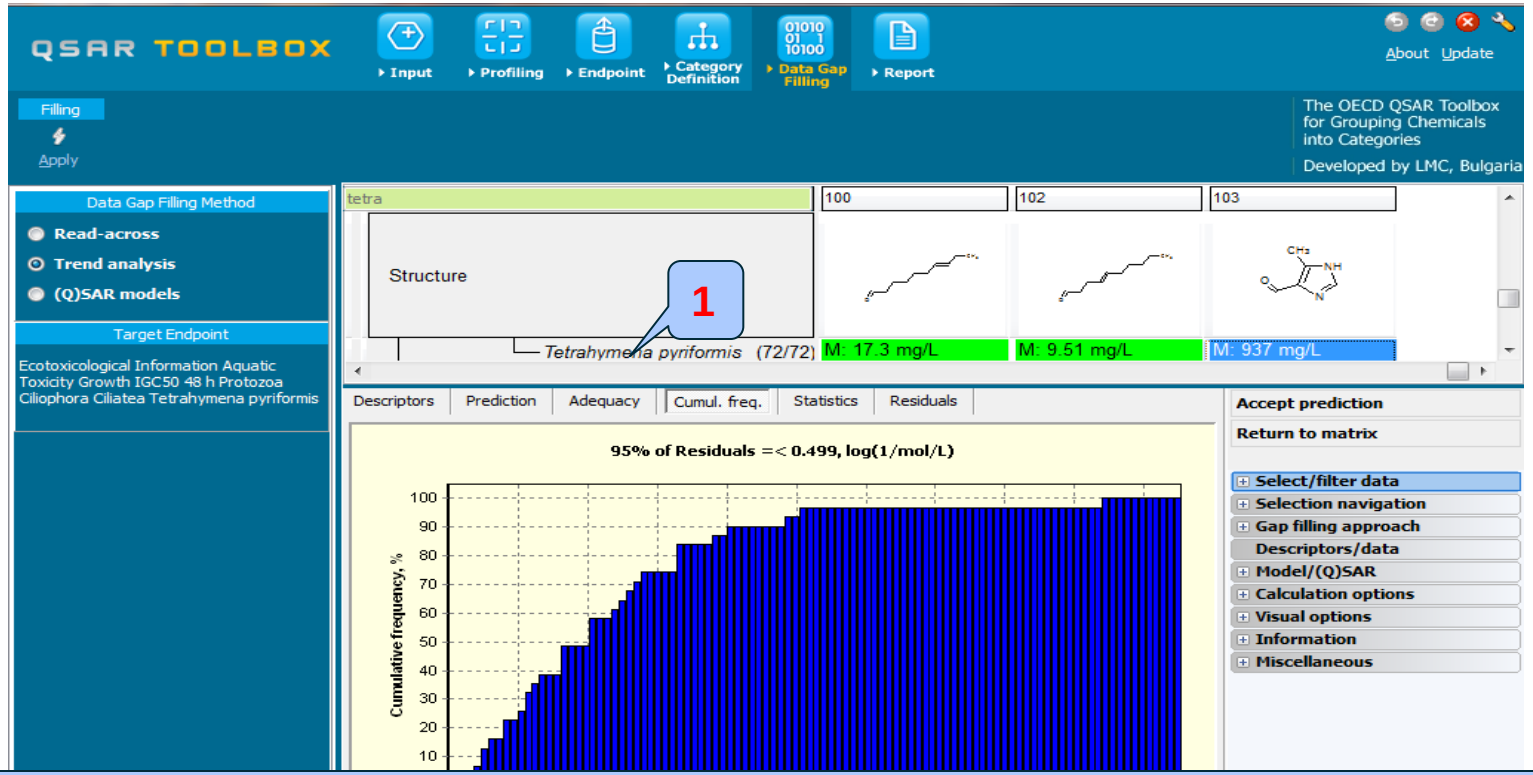
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

Evaluation of the model



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

Evaluation of the model cumulative frequency



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

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

Evaluation of the model statistics

QSAR TOOLBOX

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

About Update

Filling

Apply

The OECD QSAR Toolbox for Grouping Chemicals into Categories

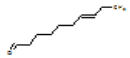
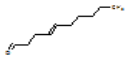
Developed by LMC, Bulgaria

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

tetra	100	102	103
Structure			
Tetrahymena pyriformis (72/72)	M: 17.3	M: 9.51 mg/L	M: 937 mg/L

1

Descriptors | Prediction | Adequacy | Cumul. freq. | Statistics | Residuals

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

Accept prediction

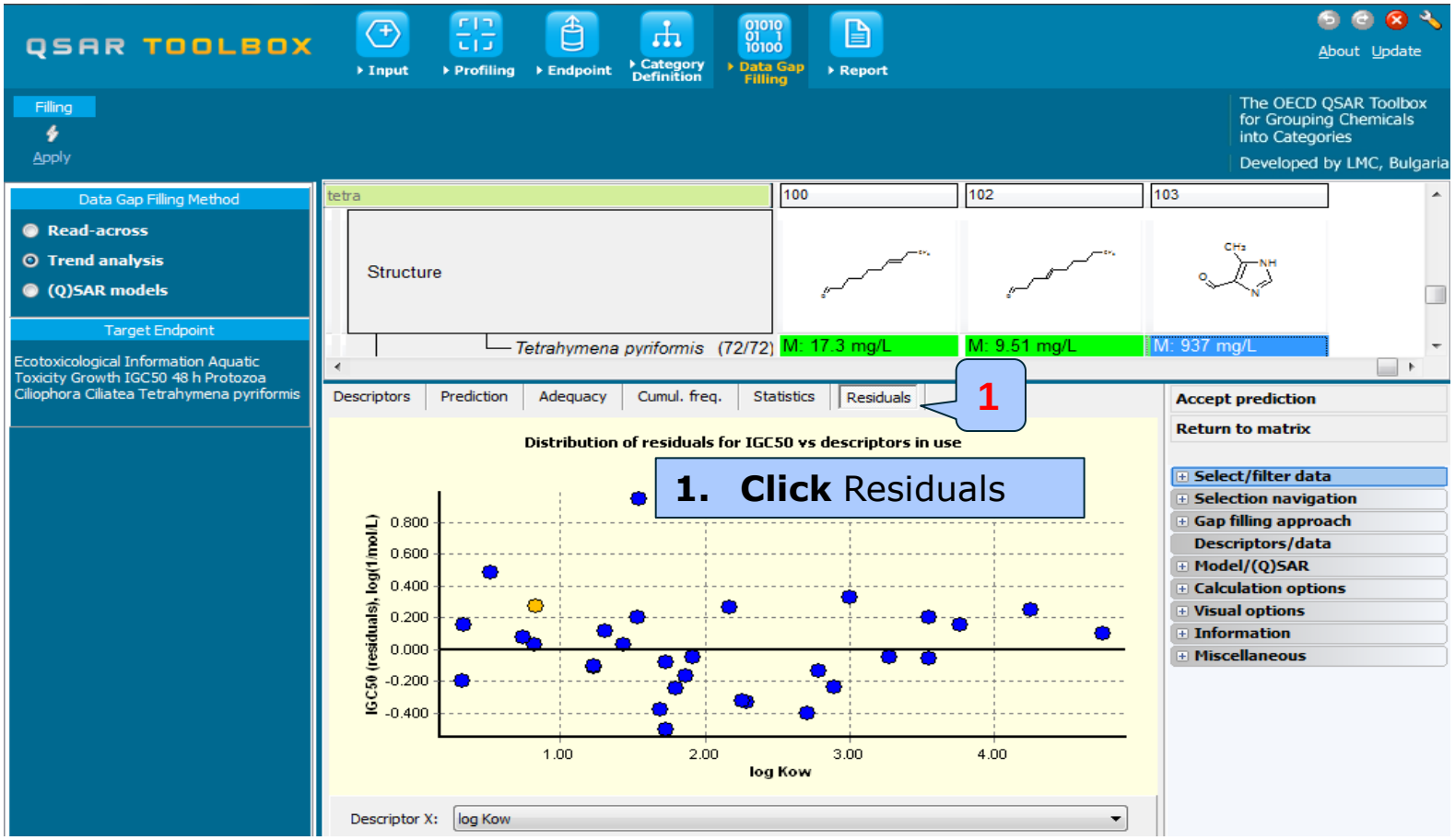
Return to matrix

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

1. **Click** Statistics

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

Evaluation of the model statistics



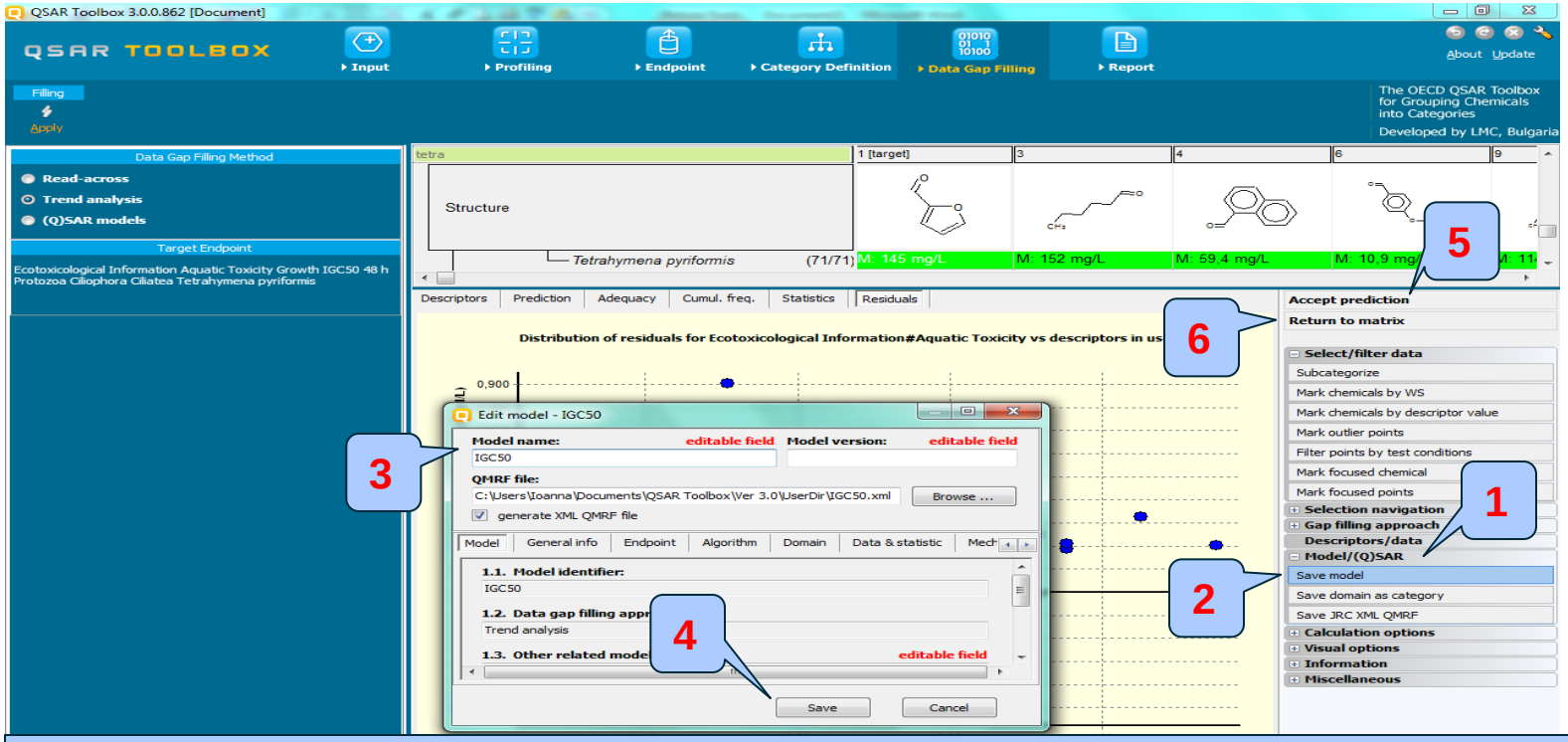
Data Gap Filling (IGC 50 48h of *T. pyriformis*)

Save the derived QSAR model

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

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

Save the derived QSAR model



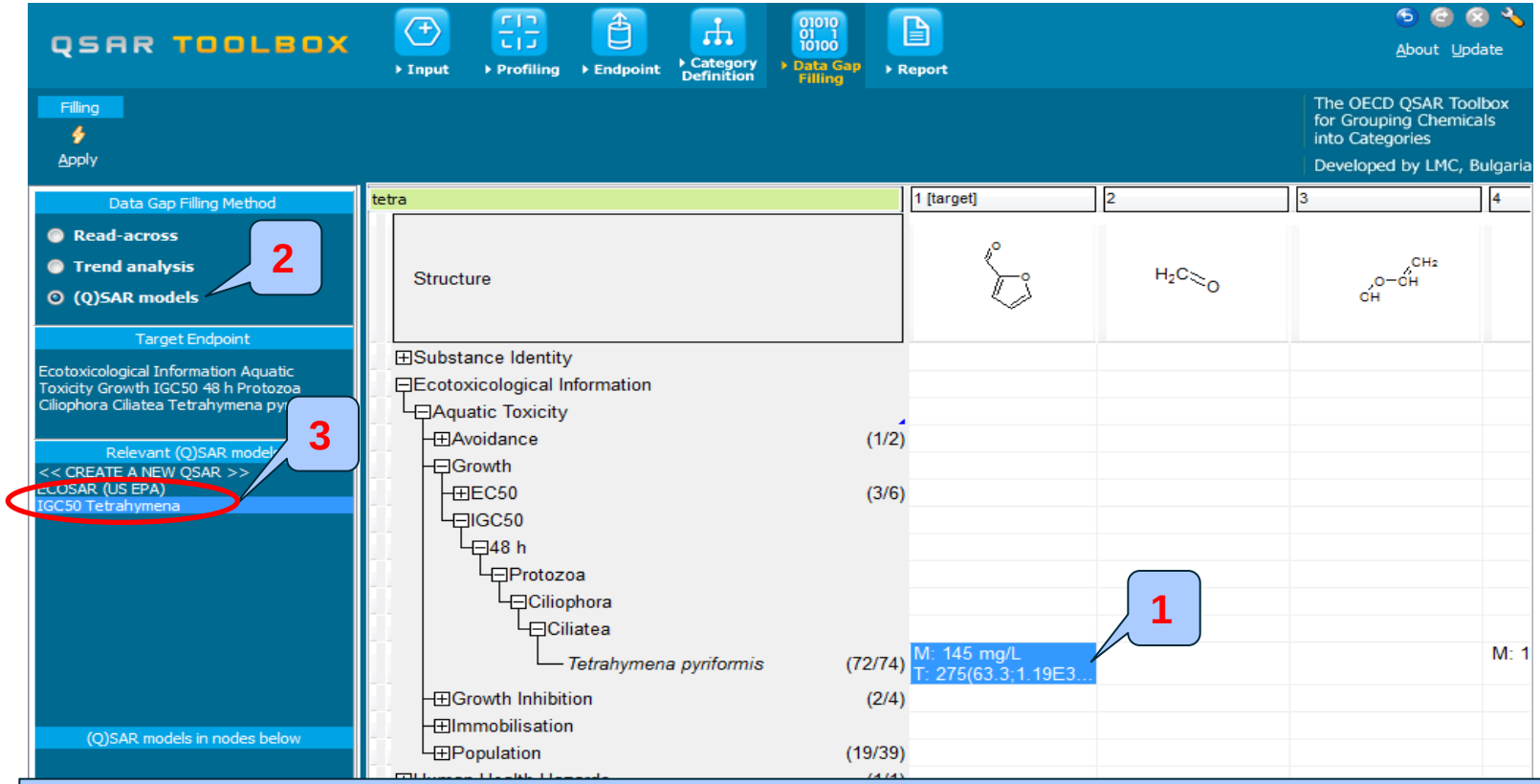
1. Click Model (Q)SAR; **2. Select** Save model; 3. **Type** Name of the model and fill fields if necessary; 4. **Click** Save; 5. **Click** Accept prediction; 6. **Select** Return to the matrix

Outlook

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

Data Gap Filling

How to see the derived QSAR?



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

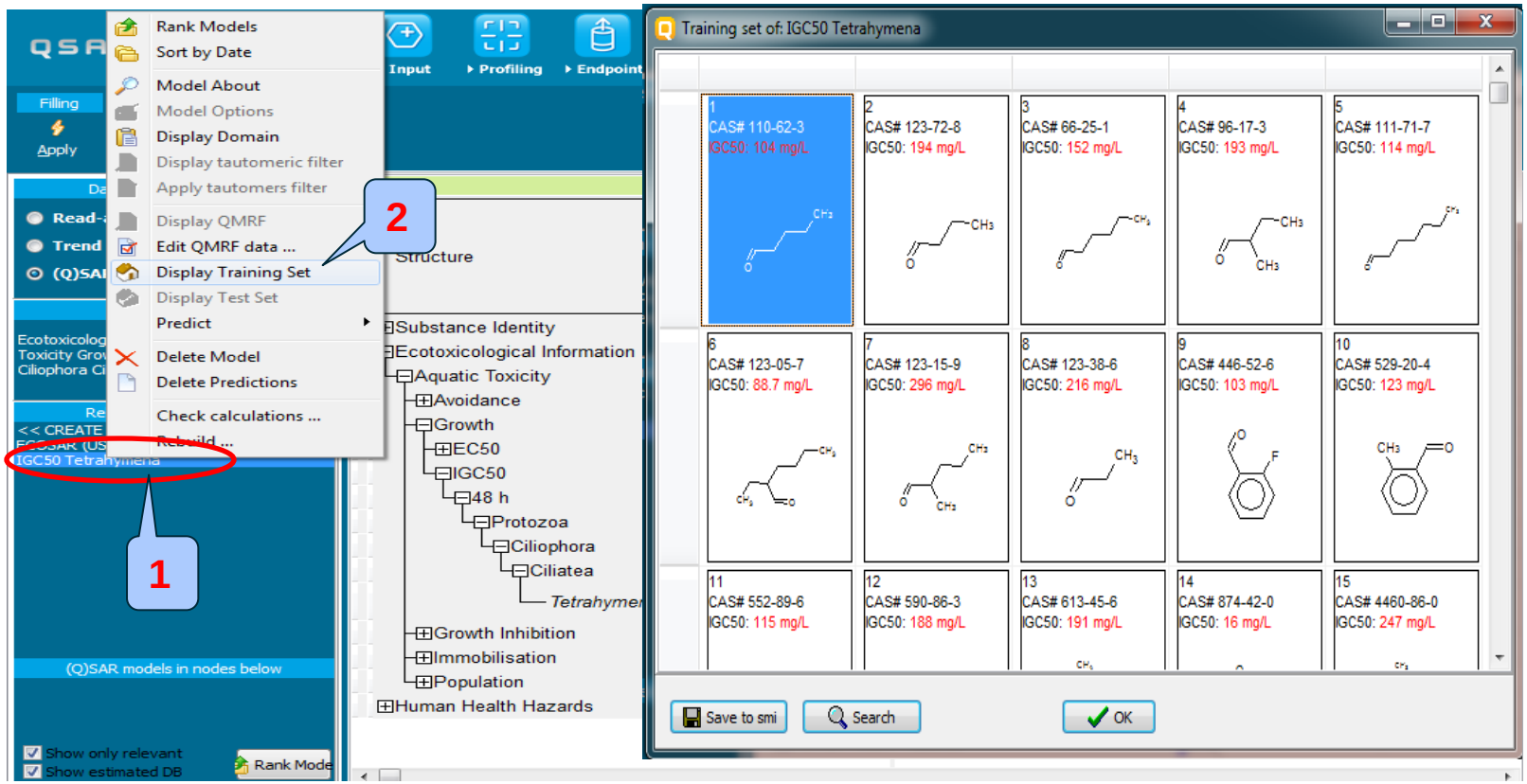
Data Gap Filling
How to see the derived QSAR?

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

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

Data Gap Filling

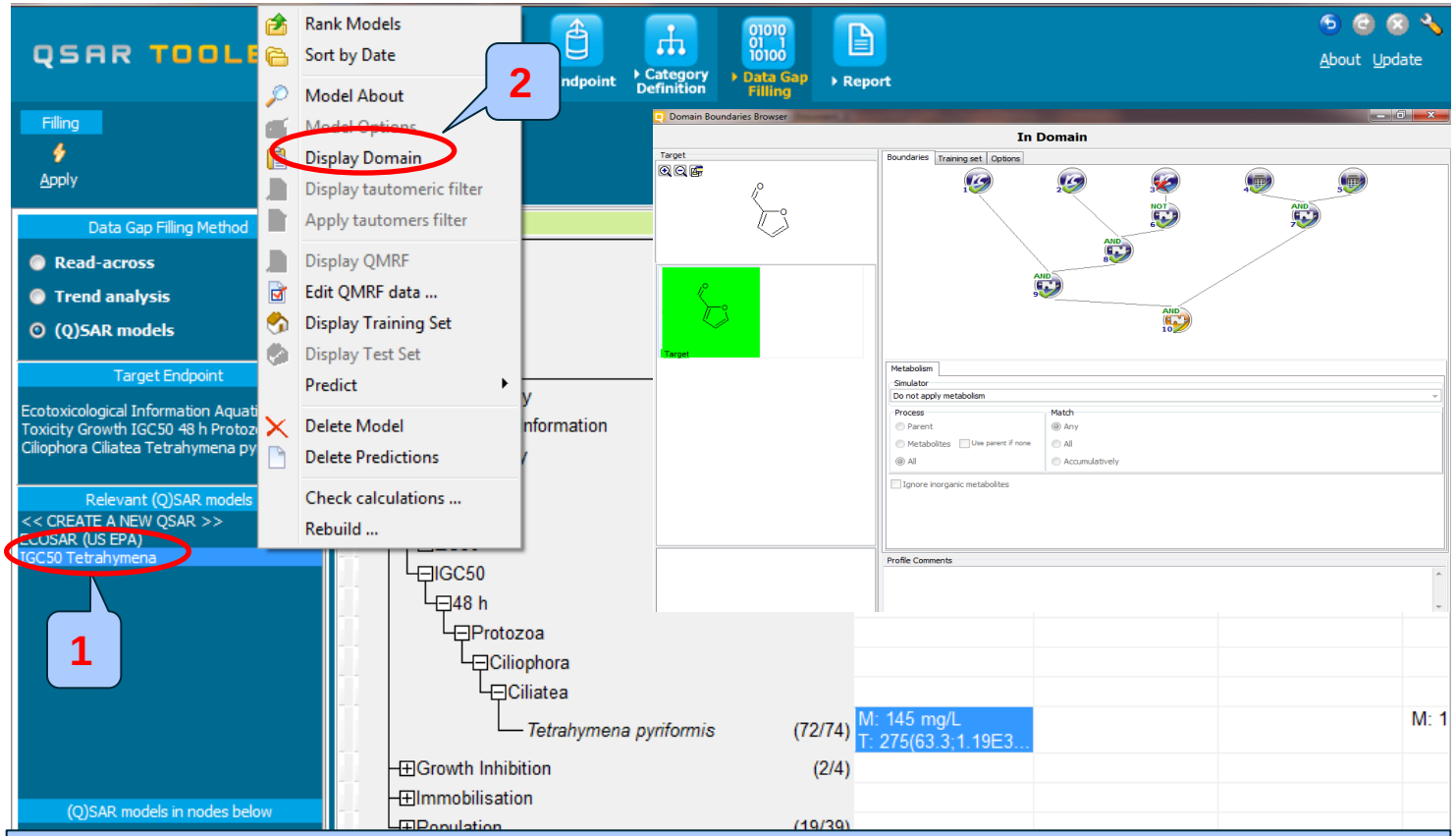
Visualisation of the training set



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

Data Gap Filling

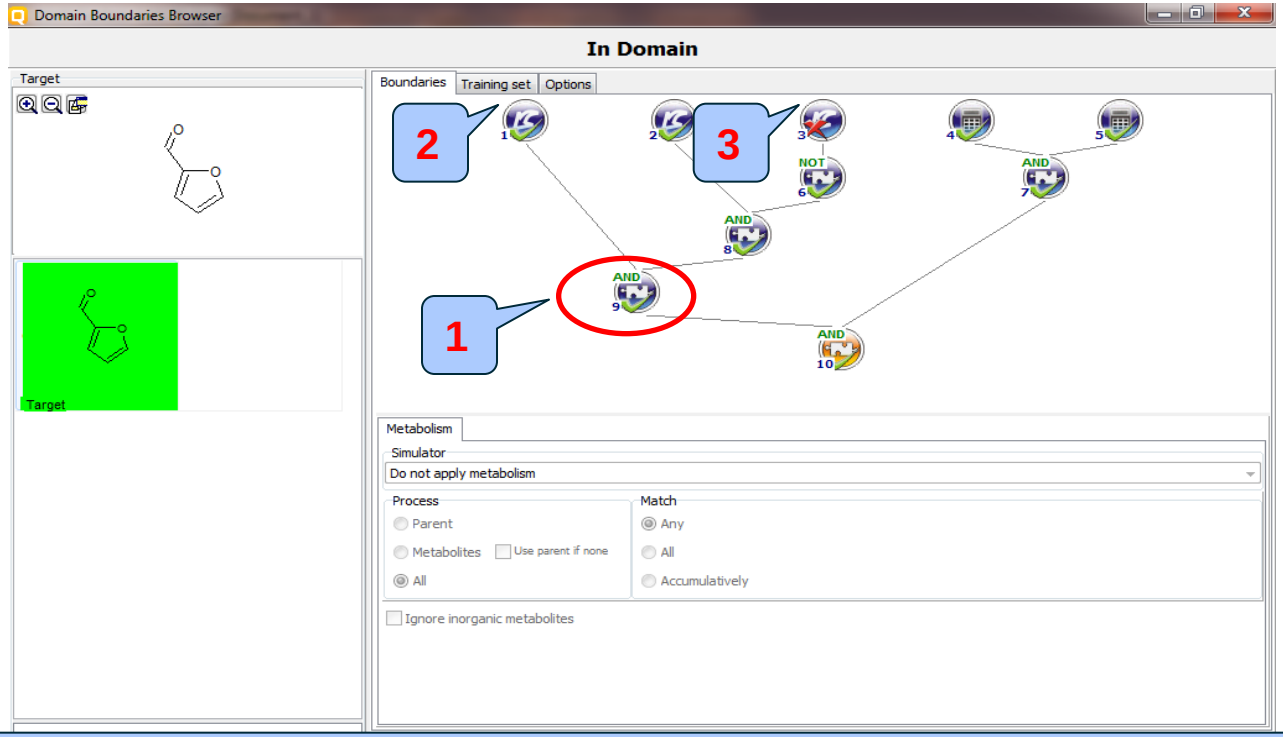
Visualisation of model domain



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

Data Gap Filling

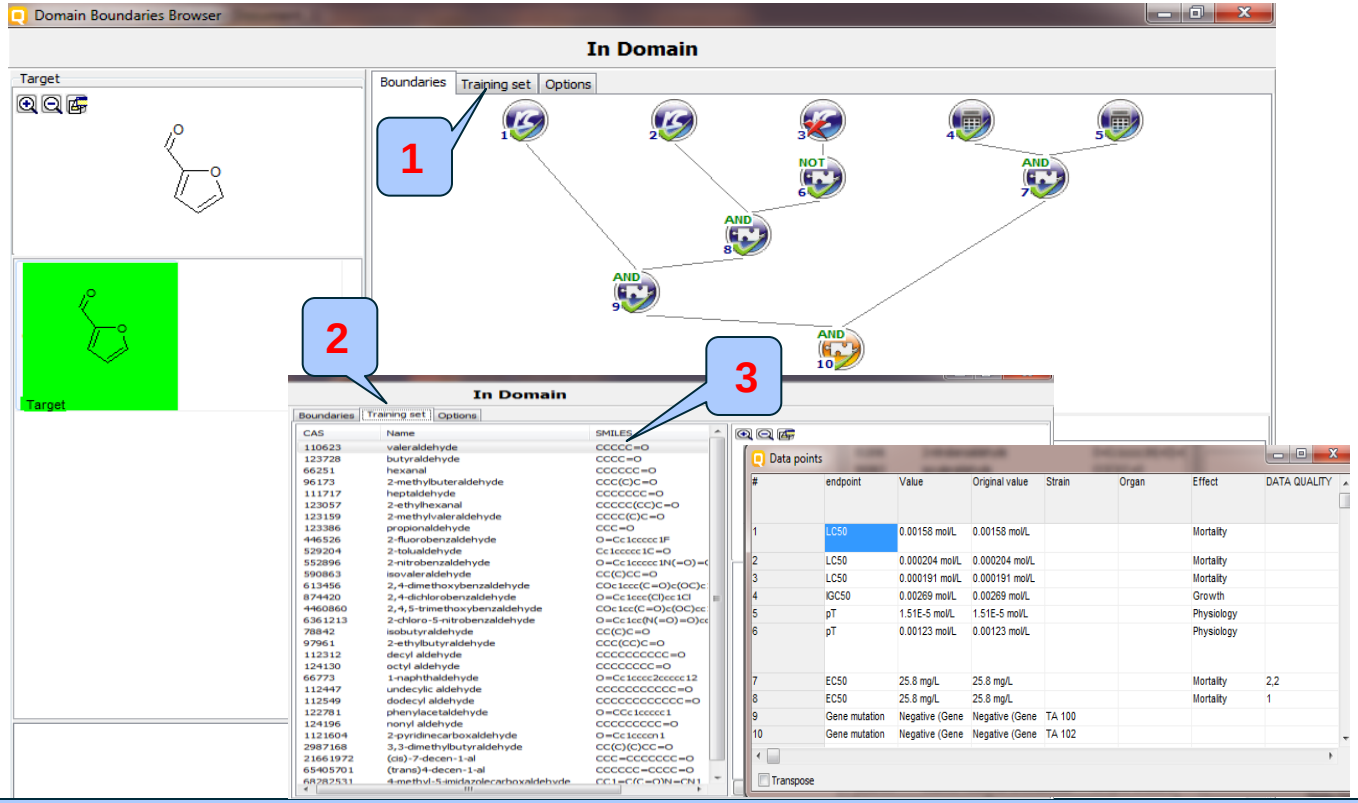
Visualisation of model domain



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

Data Gap Filling

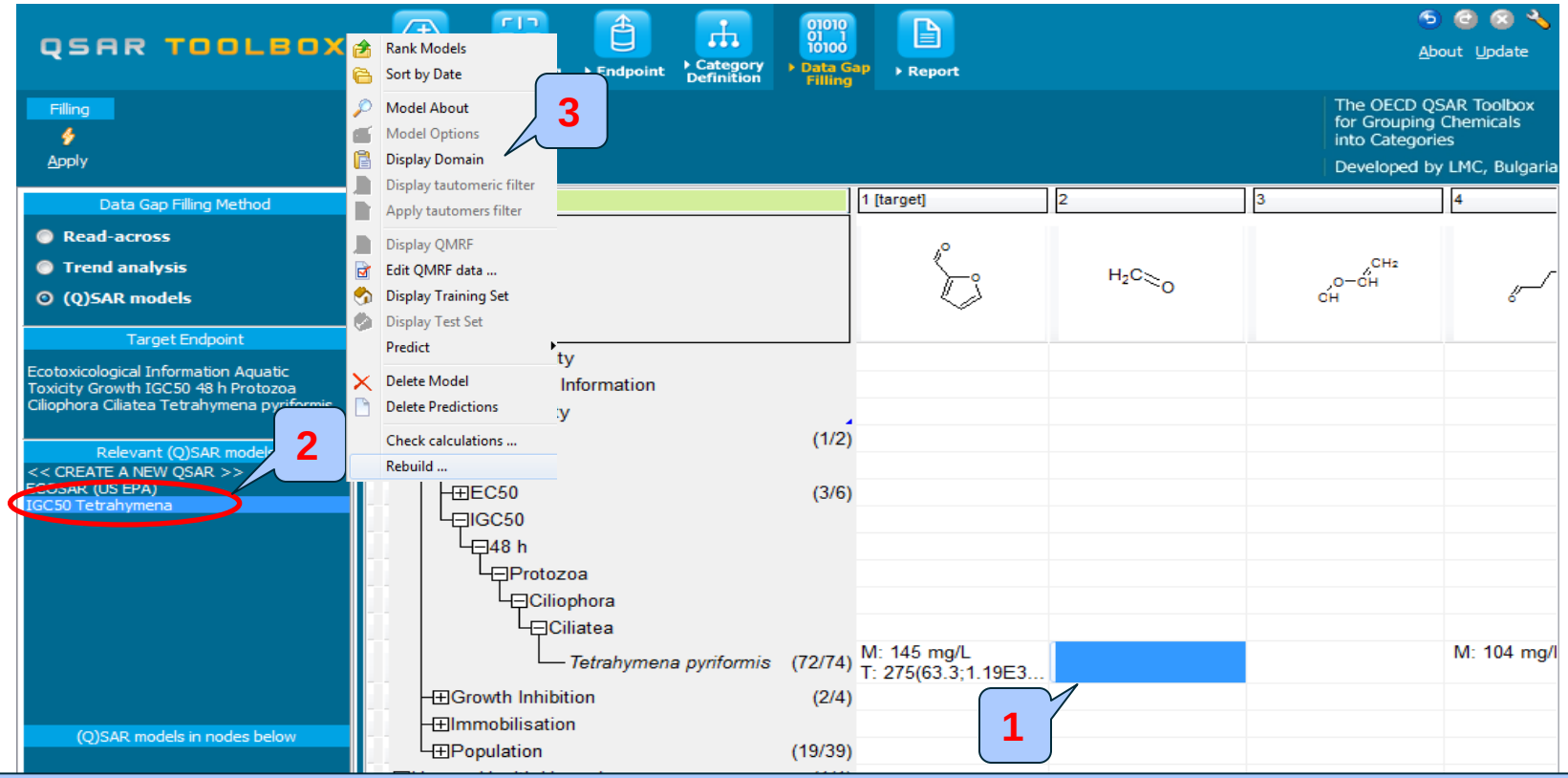
Visualisation of the training set of the model



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

Data Gap Filling

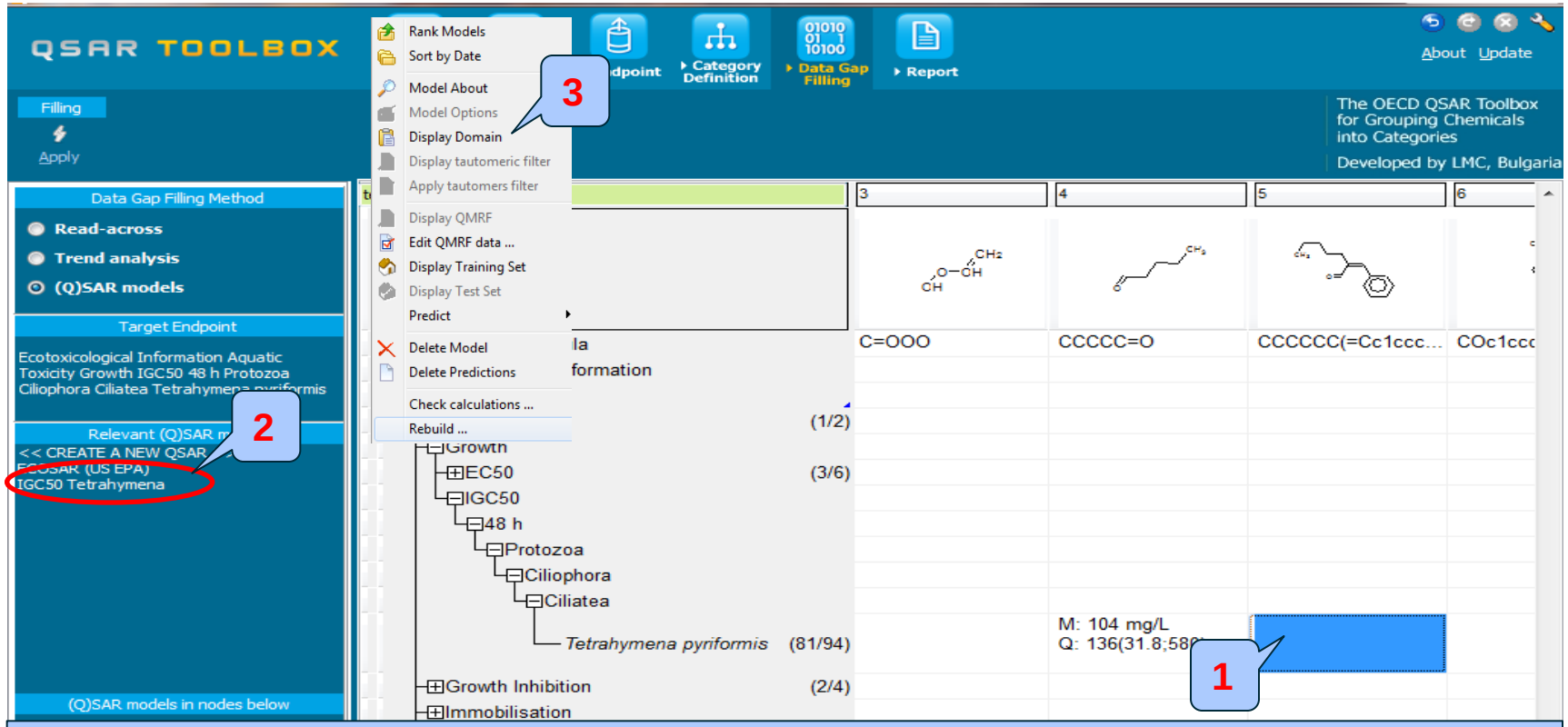
Visualisation whether a chemical is in the domain of the model



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

Data Gap Filling

Visualisation whether a chemical is in the domain of the model



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

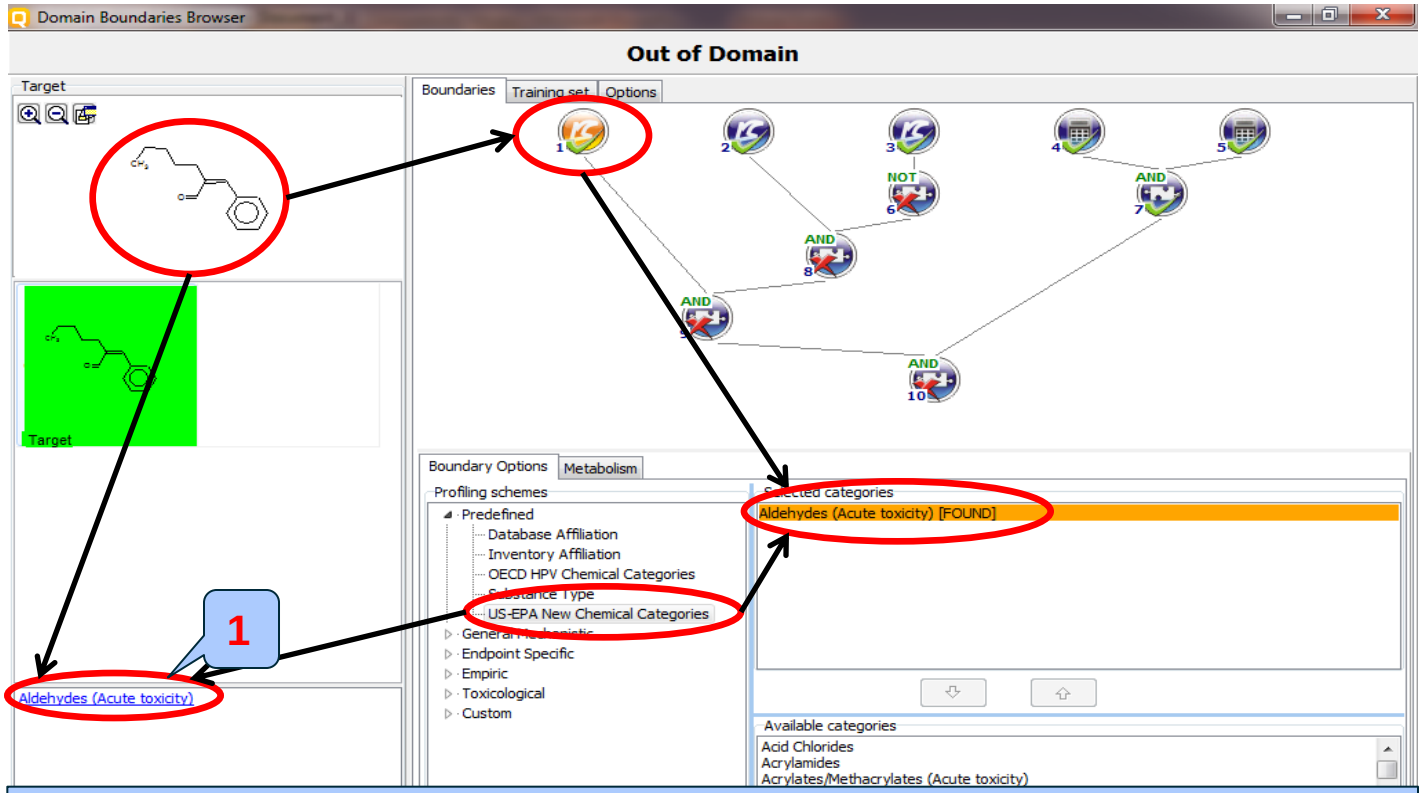
Data Gap Filling

Visualisation whether a chemical is in the domain of the model

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

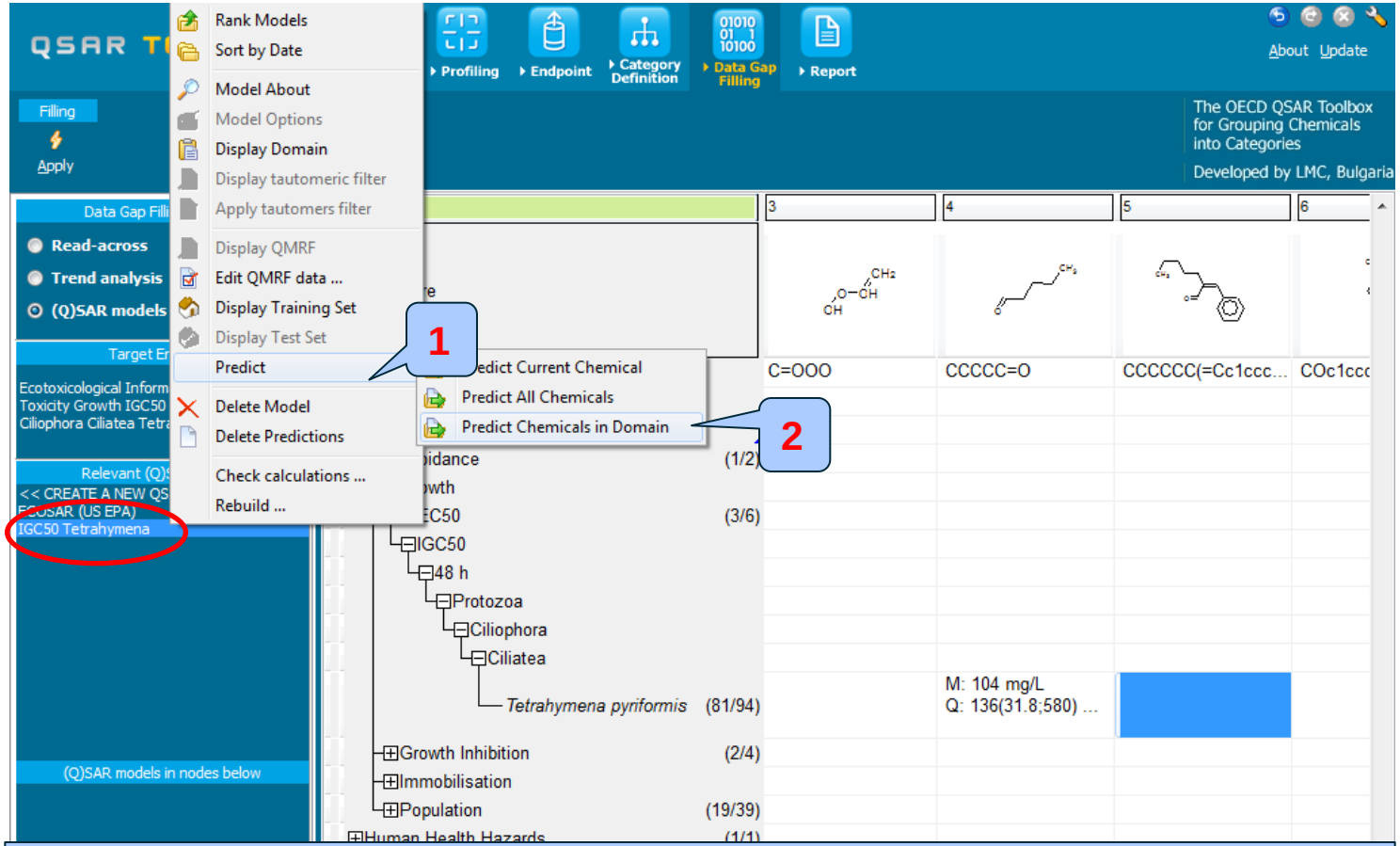
Data Gap Filling

Visualisation whether a chemical is in the domain of the model



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

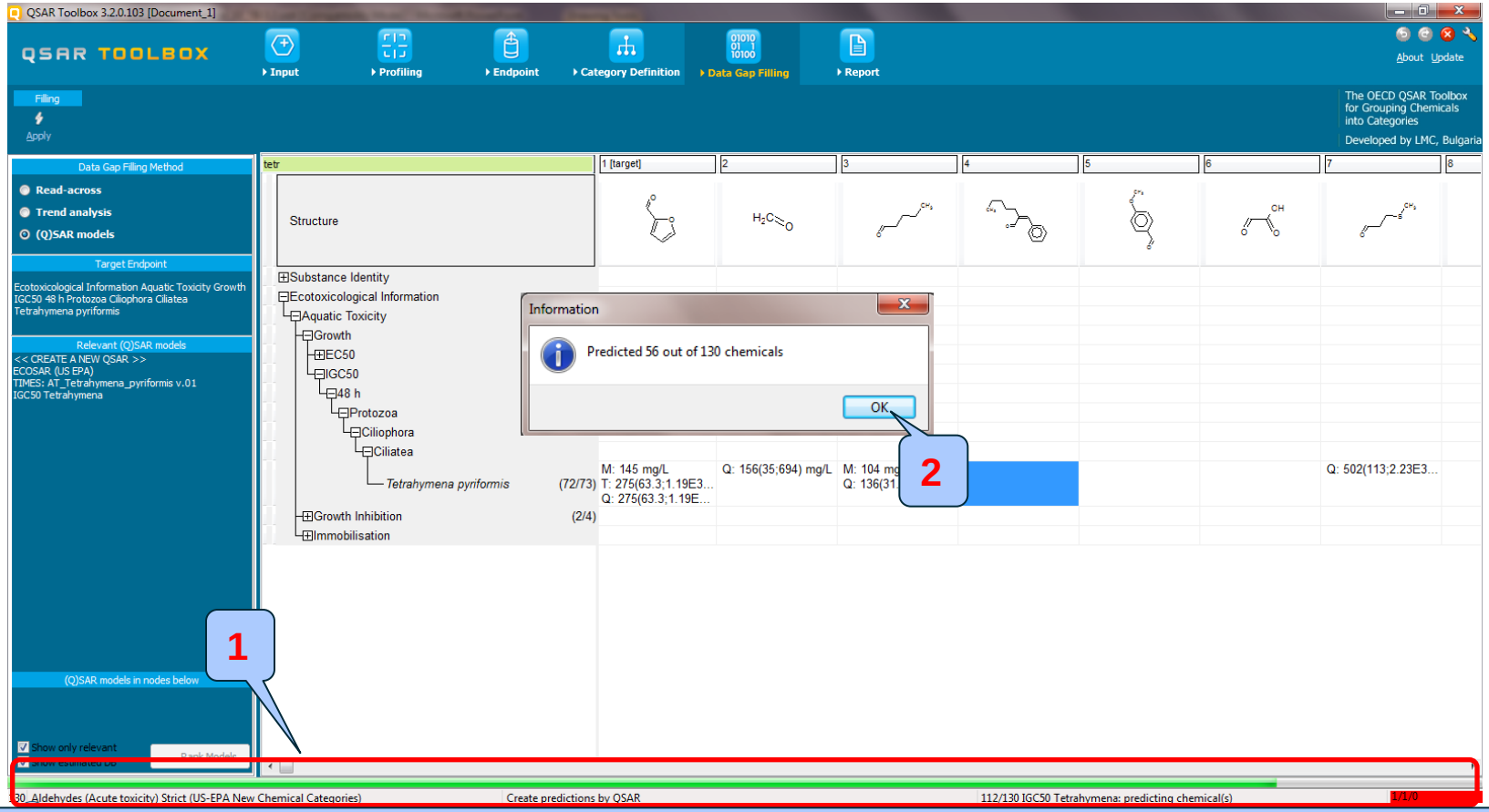
Data Gap Filling
Edit QMRF data



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

Data Gap Filling

Perform prediction



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

Outlook

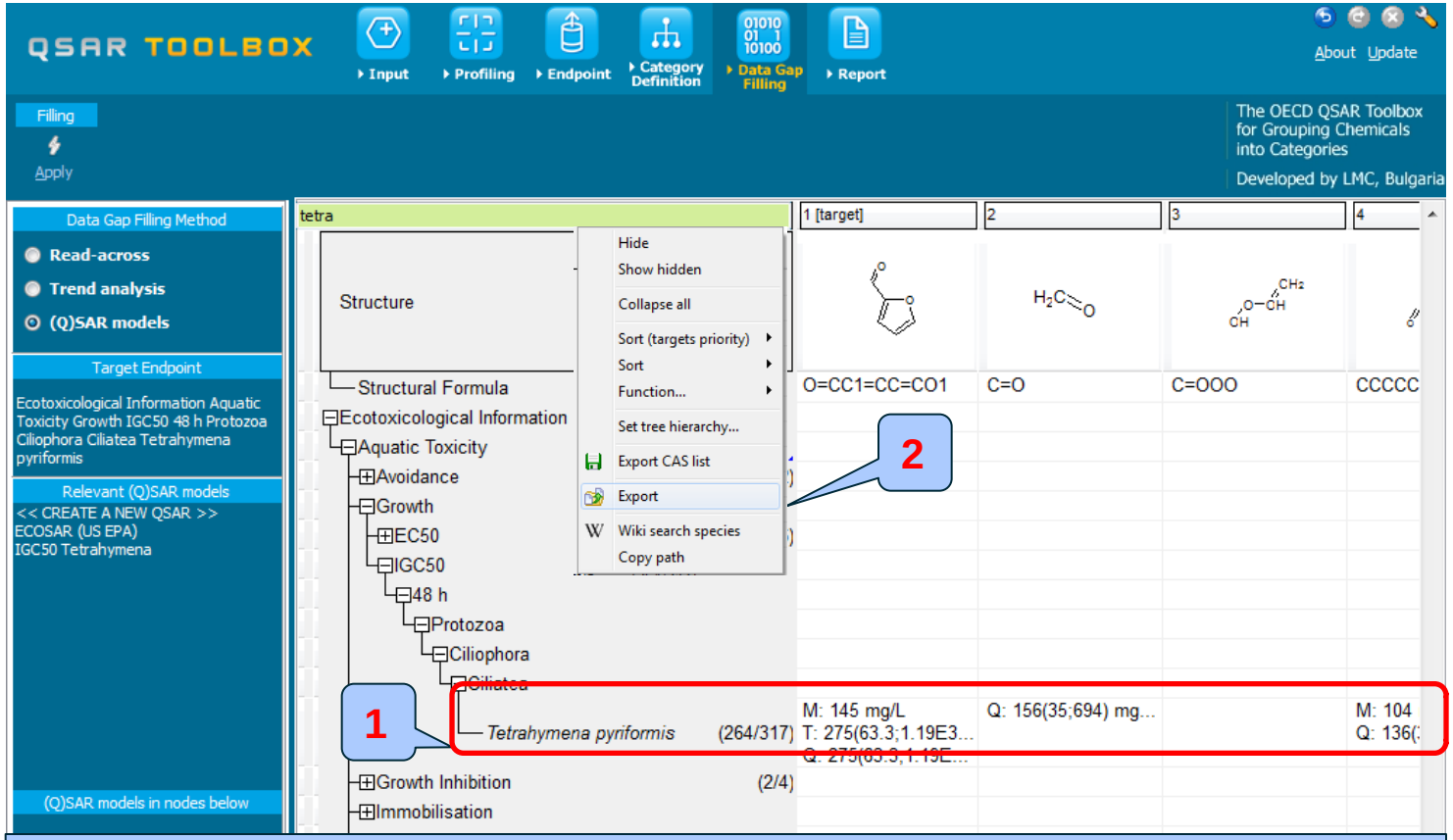
- Background
- Objectives
- The exercise
- **Workflow of the exercise**
  - Chemical Input
  - Profiling
  - Endpoints
  - Category definition
  - Data gap filling
    - QSAR model
  - **Export QSAR prediction**

Export QSAR results

- The predictions for the chemicals in the matrix can be exported into text file.

- In the data tree right-click on Tetrahymena pyriformis (for the endpoint IGC50 48h for Tetrahymena pyriformis) and select Export from the context menu (see next three screen shots).

Export QSAR results



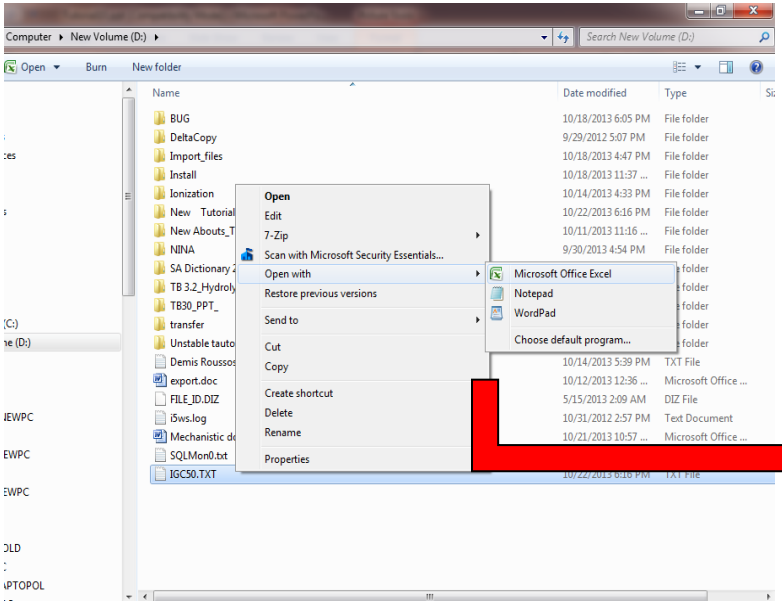
1. **Right click** on the row of endpoint tree associated with predictions from the QSAR model; 2. **Select** Export (see next screen shot).

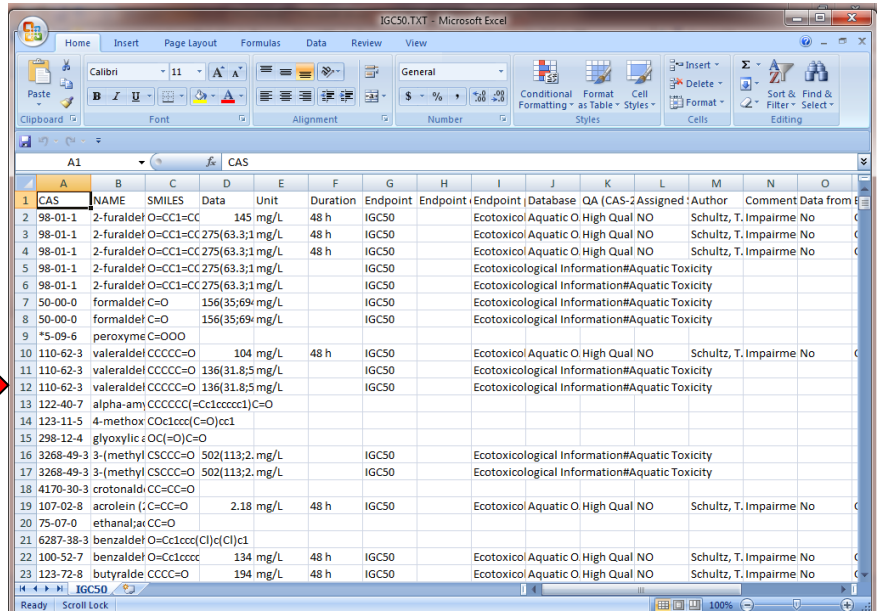
Export QSAR results

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

Export QSAR results

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





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.