

OECD QSAR Toolbox v.3.4

Types of endpoint vs. endpoint correlations
using ToxCast and other endpoint data
applied in Toolbox 3.4

Outlook

- **Background**
- Objectives
- The exercise
- Workflow

Background

This presentation is designed to introduce the user with:

- ToxCast database is part of the Toolbox database
- Illustration of different types endpoint vs. endpoint correlations using:
 - ToxCast and other Estrogen receptor data
 - LLNA and GPMT skin sensitization data
 - DPRA and LLNA skin sensitization data
 - Skin sensitization and Ames mutagenicity data

Outlook

- Background
- **Objectives**
- The exercise
- Workflow

Objectives

This presentation demonstrates a number of functionalities of the Toolbox:

- Illustration of endpoint vs. endpoint correlations using different type endpoint data

Outlook

- Background
- Objectives
- **The exercise**
- Workflow

The exercise

- Illustration of different endpoint data correlations:
 - AC50 vs. AC50 endpoints associated with different test type extracted from Toxcast database
 - AC50 vs. Estrogen receptor binding data
 - LLNA vs. GPMT skin sensitization data
 - DPRA (reactivity) vs. LLNA (skin sensitization) data
 - GPMT (skin sensitization) vs. Ames mutagenicity data

Outlook

- Background
- Objectives
- The exercise
- **Workflow**

Workflow

- **The Toolbox has six modules which are typically used in a workflow:**
 - Chemical Input
 - Profiling
 - Endpoints
 - Category Definition
 - Filling Data Gaps
 - Report
- **In this example we will use the modules in a different order, tailored to the aims of the example.**

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - **Load ToxCast database**

ToxCast database

Loading database

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Database', and 'Report'. The 'Database' menu is open, and the 'DB' button is highlighted with a red box and a callout '1'. A 'Select database' dialog box is open, showing a list of databases with 'ToxCast DB' highlighted and a callout '2'. The 'OK' button in the dialog is highlighted with a red box and a callout '3'. The main window shows a 'Datamatrix' table with 8 columns, each labeled '1 (target)', '2 (target)', etc., and containing chemical structures. A red box highlights the first row of the table, and a callout '4' points to it.

1. **Click** on "DB" button;
loaded on datamatrix

2. **Select** "ToxCast DB";

3. **Click** "OK";

4. Chemicals are

ToxCast database Data gathering

The screenshot shows the QSAR Toolbox software interface. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The 'Endpoint' button is highlighted with a red circle and the number 1. The 'Gather' button is highlighted with a red circle and the number 3. The 'Inventories' panel on the left shows a list of databases, with 'ToxCastDB' selected and highlighted with a red circle and the number 2. The 'Filter endpoint tree...' panel in the center shows a tree structure of endpoints, with 'ToxCast' selected and highlighted with a red circle and the number 4. The main window displays a datamatrix table with columns for various endpoints and rows for different chemical substances.

Substance	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]	8 [target]
ACEA	(600/660)	M: 21.2 mg/L	M: 0.0039 mg/L	M: 8.08 mg/L, 0.00...	M: 0.000504 mg/L			
Apredica	(425/2642)			M: 23.4 mg/L, 26.9...	M: 0.0962 mg/L, 0...			
Attagene	(1374/6568)	M: 12.5 mg/L, 12.6...		M: 0.00268 mg/L, ...	M: 0.689 mg/L, 1.3...		M: 3.87 mg/L, 3.42...	
BioSeek	(971/21906)			M: 6.74 mg/L, 6.25...	M: 0.338 mg/L, 0.3...			
NCGC	(1475/6890)	M: 0.00436 mg/L, ...	M: 0.000106 mg/L, ...	M: 12.4 mg/L, 9.25...	M: 0.000262 mg/L, ...	M: 0.66 mg/L, 0.05...	M: 0.219 mg/L, 1.0...	M: 5.84E-5 mg/L, ...
Novascreen	(975/8054)	M: 3.59 mg/L, 0.02...		M: 0.00646 mg/L, ...	M: 0.236 mg/L, 0.1...	M: 2.67 mg/L		
Odyssey Thera	(969/2794)	M: 19.8 mg/L, 4.56...	M: 2.46 mg/L	M: 5.79 mg/L	M: 0.00676 mg/L, ...	M: 3.52 mg/L, 2.19...		
Undefined Assay Provider	(2/2)							

1. Go to "Endpoint"; 2. Select "ToxCastDB"; 3. Click "Gather"; 4. The data appears on datamatrix separated in a new node called "ToxCast"

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - **ToxCast database - overview**

ToxCast database

Background

- A major part of EPA's CompTox research is the ToxCast™ project. ToxCast is a multi-year project launched in 2007 that uses automated chemical screening technologies (called "high-throughput screening assays") to expose living cells or isolated proteins to chemicals. The cells or proteins are then screened for changes in biological activity that may suggest potential toxic effects. These innovative methods have the potential to limit the number of required laboratory animal-based toxicity tests while quickly and efficiently screening large numbers of chemicals.
- ToxCast has evaluated over 2,000 chemicals from a broad range of sources including: industrial and consumer products, food additives, and potentially "green" chemicals that could be safer alternatives to existing chemicals. Chemicals were evaluated in over 700 high-throughput assays that cover a range of high-level cell responses and approximately 300 signaling pathways.
- ToxCast results are contributed to the federal agency collaboration called Toxicity Testing in the 21st Century (Tox21). Tox21 pools chemical research, data and screening tools from multiple federal agencies including the National Toxicology Program. So far, Tox21 has compiled high-throughput screening data on nearly ten thousand chemicals.

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - **Correlation of data - background**

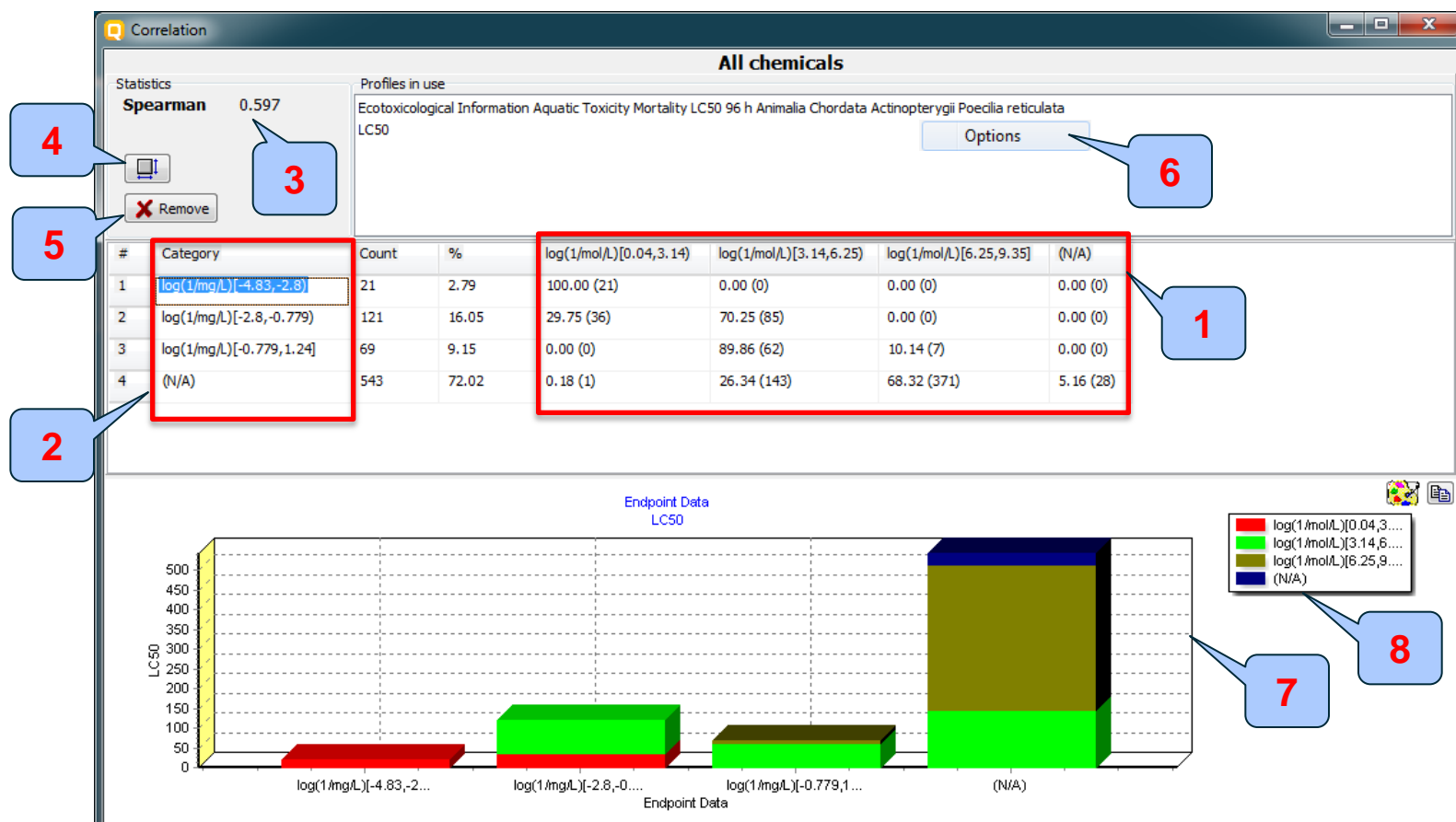
Correlation of endpoint data

Background

- This functionality introduce the user with opportunity to analyze correlations between selected gap filling endpoint (endpoint used for prediction) and other endpoint data.
- It is applicable for correlation analysis of data presented in ordinary, interval or ratio scale.
- If correlated data are measured in interval or ratio scale they are transformed in ordinary scale and the strength of the correlation is estimated by Spearman correlation coefficient.
- Basically, this functionality provides a correlation between target endpoint (this is the initial endpoint selected by the user) displayed on ordinate axis (Y-axis) and other endpoint data displayed on abscissa (X-axis). Illustration is provided on next slide.
- By default, the interval or ratio data, associated with initial endpoint and displayed on Y-axis of the graph is distributed into three bins (illustrated on the graph shown on next slide). The users are able to change the number of bins or their magnitudes.

Correlation of endpoint data

Graphical illustration of "Correlation" window



1. Columns with initial endpoint data displayed on Y axis;
2. Column with endpoint data placed on X-axis;
3. Spearman correlation index
4. Button for changing position of X and Y axis;
5. Button, which removes range(s) from the contingency table;
6. Option functionality allowing to change settings of the selected endpoint (evoked by right click).
7. Bar graph of the obtained correlation;
8. Color legend

Correlation of endpoint data

Spearman coefficient factor

- Spearman's rank correlation coefficient is a nonparametric rank statistic proposed by Charles Spearman as a measure of the strength of an association between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.
- Spearman correlation coefficient could be used for exploring the covary between:
 - two ranked variables
 - one measurement variable and one ranked variable (in this case, the measurement variable need to be to converted to ranks)
- Spearman correlation varies from -1 to +1 and the interpretation of the coefficient factor is provided below:
 - 0.00 – 0.19 – very weak correlation
 - 0.20 – 0.39 – weak correlation
 - 0.40 – 0.59 – moderate correlation
 - 0.60 – 0.79 – strong correlation
 - 0.80 – 1.0 – very strong

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - Correlation of data – background
 - **Types endpoint correlations**

Types endpoint correlations

Types endpoint correlations are as follows:

- Continuous vs. continuous
- Categorical vs. categorical:
 - ✓ Categorical vs. categorical
 - ✓ Categorized continuous vs. categorical
 - ✓ Categorized continuous vs. categorized continuous

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - Correlation of data – background
 - **Types endpoint correlations**
 - Continuous vs. continuous

Types endpoint correlations

Continuous vs. continuous

- The aim of this type correlation is to illustrate how continuous type endpoint data or so called ratio data correlates each other (e.g. LC50 vs. EC50 data)
- In this example we will illustrate how AC50 data associated with two different test assays extracted from ToxCast DB correlates each other:
 - NCGC Reporter Gene Assay ERα Agonist, Estrogen receptor 1 (assay 1)
 - Tox21_Era_BLA_Agonist_ch2 (assay 2)
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
 - *Gather experimental data (step 1)*
 - *Define target endpoint (step 2)*
 - *Enter Gap filling (step 3)*
 - *Change default X-descriptor (logKow) with other AC50 data (step 5)*

Types endpoint correlations

Continuous vs. continuous

Gather experimental data – step 1

The screenshot shows the QSAR TOOLBOX software interface. The 'Endpoint' tab is selected in the top menu bar, indicated by a red circle with the number '1'. In the left sidebar, under the 'Inventories' section, the 'ToxCastDB' checkbox is checked, indicated by a red circle with the number '2'. The 'Gather' button in the top left corner is highlighted with a red circle and the number '3'. The main window displays a table with 8 columns, each labeled '1 [target]' through '8 [target]', and a row for 'Structure' with chemical structures. Below the table, there are sections for 'Substance Identity', 'Physical Chemical Properties', 'Environmental Fate and Transport', 'Ecotoxicological Information', and 'Human Health Hazards'.

1. Go to "Endpoint" 2. Select "ToxCast" DB 3. Click "Gather"

Types endpoint correlations

Continuous vs. continuous
Gather experimental data – step 1

The screenshot displays the QSAR Toolbox software interface. The main window shows the 'Endpoint' tab, which includes a 'Filter endpoint tree...' panel on the left and a grid of chemical structures on the right. The 'Filter endpoint tree...' panel lists various endpoints, including 'AC50'. The grid on the right shows chemical structures for different targets, with the first target being 'AC50'. Two dialog boxes are overlaid on the main window. The first dialog, 'Read data?', has 'All endpoints' selected and 'from Tautomers' checked. The second dialog, 'Repeated values for', shows a table of data points with 'AC50' as the endpoint and various CAS numbers. Red callouts with numbers 1, 2, and 3 point to the 'OK' button in the first dialog, the 'Select one' button in the second dialog, and the 'OK' button in the second dialog respectively.

1. Click "OK"

2. Click "Select one" button

3. Click "OK"

Types endpoint correlations

Continuous vs. continuous
Gather experimental data – step 1

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Gather Import Import IUCID5 Export IUCID5 Database Inventory Database

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

Databases: Select All Unselect All Invert About

- ☐ Carcinogenicity & mutagenicity ISSCAN
- ☐ Cell Transformation Assay ISSCTA
- ☐ Dendritic cells COLIPA
- ☐ Developmental & Reproductive Toxicity
- ☐ Developmental toxicity ILSI
- ☐ ECHA CHEM
- ☐ ECOTOX
- ☐ Estrogen Receptor Binding Affinity OA
- ☐ Eye Irritation ECETOC
- ☐ Genotoxicity OASIS
- ☐ Human Half-Life
- ☐ Keratinocyte gene expression Givaudan
- ☐ Micronucleus ISSMIC
- ☐ Micronucleus OASIS
- ☐ MUNRO non-cancer EFSA
- ☐ Rep Dose Tox Fraunhofer ITEM
- ☐ Repeated Dose Toxicity HESS
- ☐ Rodent Inhalation Toxicity Database
- ☐ Skin Irritation
- ☐ Skin Sensitization
- ☐ Skin sensitization ECETOC
- ☒ ToxCastDB
- ☐ Toxicity Japan MHLW
- ☐ ToxRefDB US-EPA
- ☐ Yeast estrogen assay database

Inventories: Select All Unselect All Invert About

- ☐ Canada DSL
- ☐ COSING
- ☐ DSSTOX
- ☐ ECHA PR
- ☐ EINECS
- ☐ HPV OCED
- ☐ METI Japan
- ☐ NICNAS
- ☐ REACH ECB
- ☐ TSCA
- ☐ US HPV Challenge Program

Filter endpoint tree...

Structure

- Environmental Fate and Transport
- Ecotoxicological Information
- Human Health Hazards
 - Acute Toxicity
 - Bioaccumulation
 - Carcinogenicity
 - Developmental Toxicity / Teratogenicity
 - Genetic Toxicity
 - Immunotoxicity
 - Irritation / Corrosion
 - Neurotoxicity
 - Photoinduced Toxicity
 - Repeated Dose Toxicity
 - Sensitisation
 - ToxCast
 - ACEA (600/660) M: 21.2 mg/L
 - Apredica (425/2642) M: 0.0039 mg/L
 - Attagene (1374/6568) M: 47.2 mg/L, 22.3...
 - BioSeek (971/21906) M: 23.4 mg/L, 26.9...
 - NCGC (1475/6890) M: 8.76 mg/L, 29.1...
 - Novascreen (975/8054) M: 0.0962 mg/L, 0...
 - Odyssey Thera (969/2794) M: 3.87 mg/L, 3.42...
 - Undefined Assay Provider (2/2)
 - Toxicity to Reproduction
 - Toxicokinetics, Metabolism and Distribution

1 [target] 2 [target] 3 [target] 4 [target] 5 [target] 6 [target] 7 [target] 8 [target]

1

Table with 8 columns (1 [target] to 8 [target]) and 10 rows of chemical data. The first row shows chemical structures. The second row shows molecular weights (M) and concentrations (mg/L) for various endpoints.

1. **ToxCast data** has been loaded on datamatrix in a separate "Endpoint tree" node

Types endpoint correlations

Continuous vs. continuous

Define target endpoint – step 2

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Filter endpoint tree...

Structure

Photoinduced Toxicity

Repeated Dose Toxicity

Sensitisation

ToxCast

ACEA (600/660) M: 21.2 mg/L

Apredica (425/2642) M: 0.0039 mg/L

Attagene (1374/6568) M: 12.5 mg/L, 12.6...

BioSeek (971/21906) M: 8.08 mg/L, 0.00...

NCGC

NCGC Reporter Gene Assay ERa Agonist

Homo sapiens

Estrogen Receptor 1 (374/505) M: 0.000224 mg/L, ...

NCGC Reporter Gene Assay ERa Antagonist (487/559) M: 19.1 mg/L

Tox21_AhR (237/237) M: 0.66 mg/L

Tox21_AhR_viability (319/319) M: 0.000106 mg/L

Tox21_AR_BLA_Agonist_ch1 (439/439) M: 0.00436 mg/L

Tox21_AR_BLA_Agonist_ch2 (67/67) M: 0.000991 mg/L

Tox21_AR_BLA_Agonist_ratio (89/89) M: 10.6 mg/L

Tox21_AR_BLA_Antagonist_ratio (150/150) M: 0.219 mg/L

Tox21_AR_BLA_Antagonist_viability (207/207) M: 14.4 mg/L

1

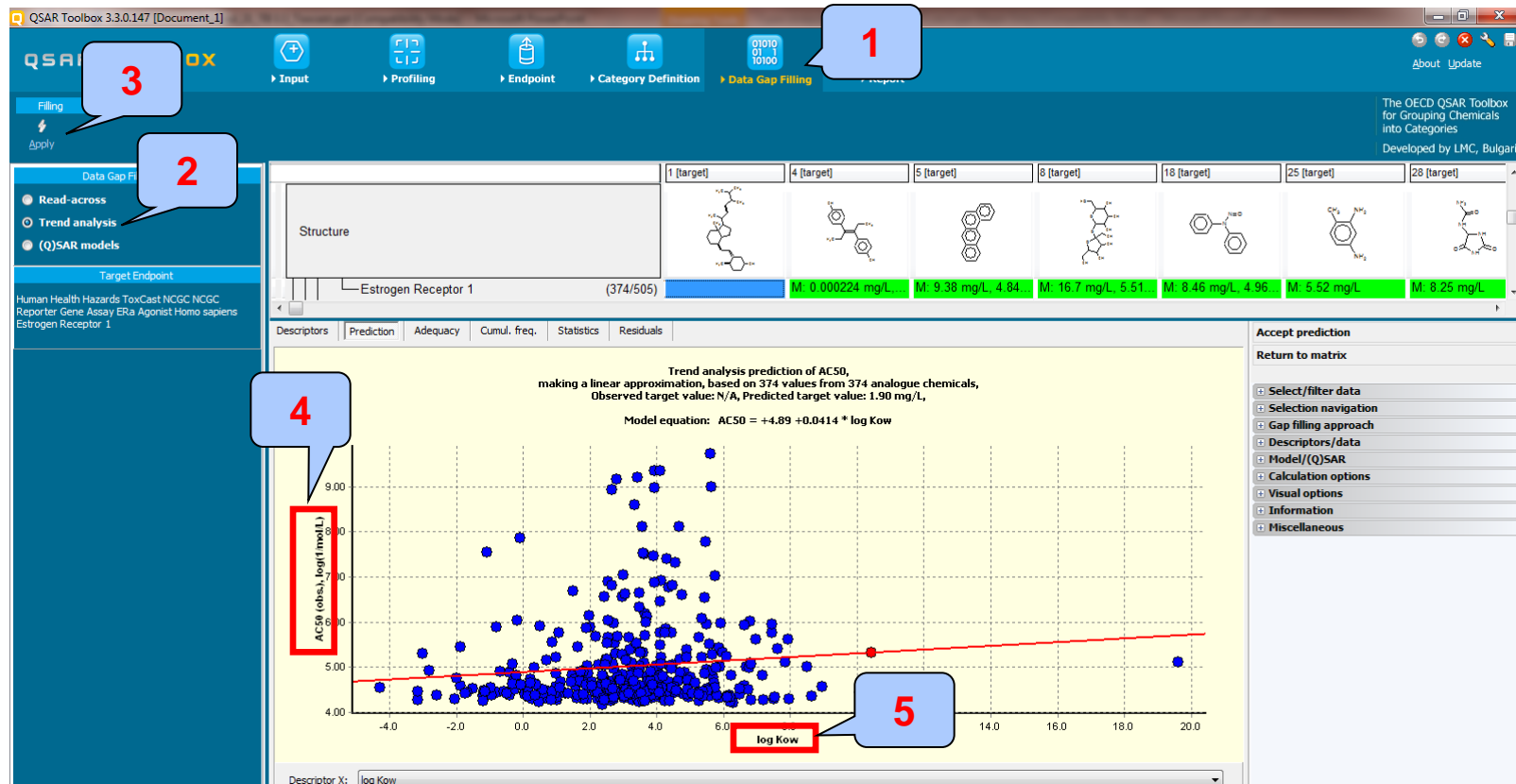
The target endpoint is A50 associated with assay "NCGC Reporter Gene Assay ERa Agonist"

1. **Click** on the cell related to the investigated endpoint, below the first chemical of datamatrix

Types endpoint correlations

Continuous vs. continuous

Enter Gap filling – step 3



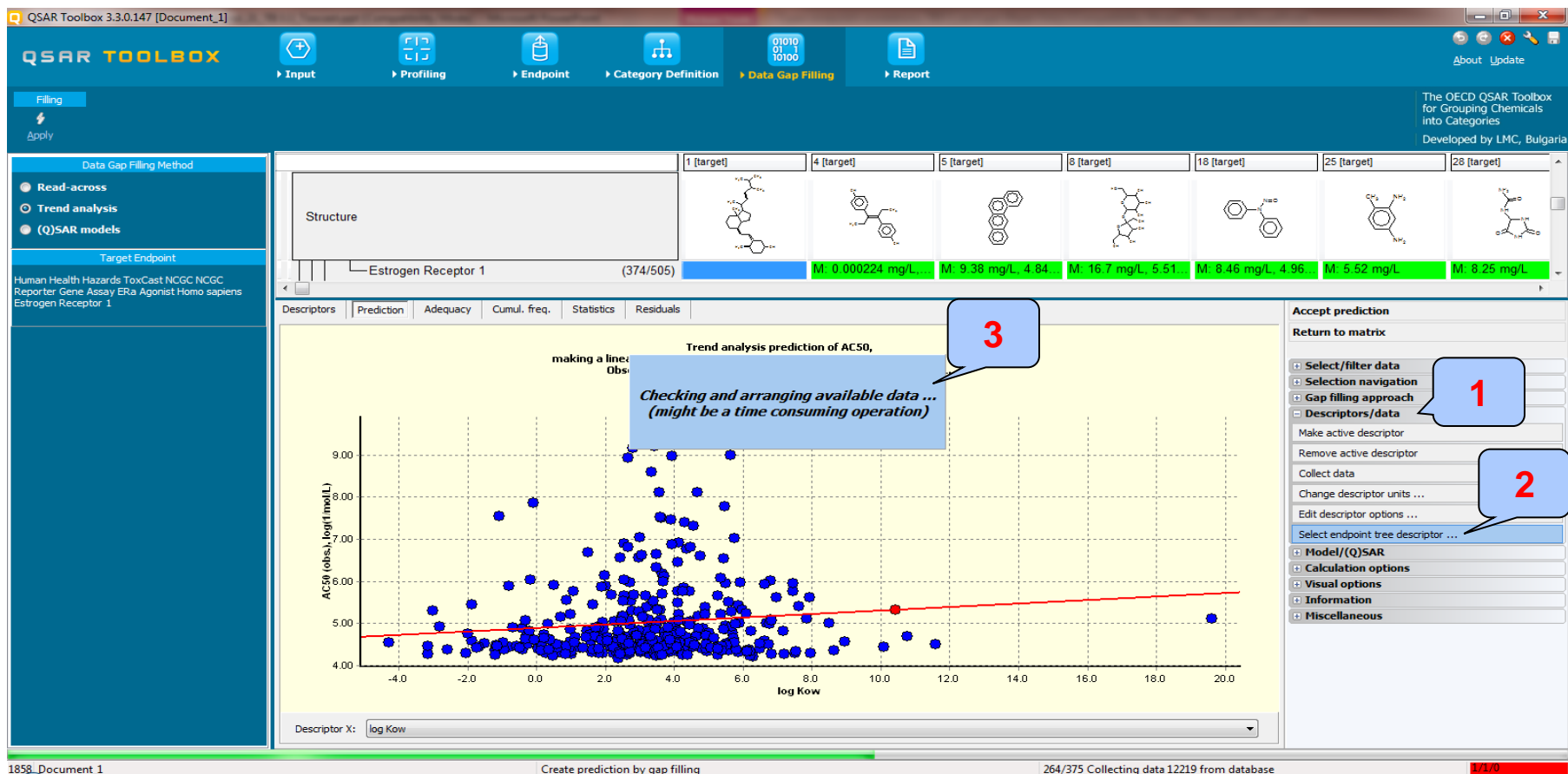
Enter Gap filling applying trend analysis. Trend analysis is applied because the target endpoint is in continues range of data and there is enough data to build a linear regression.

1. **Go** to "Data Gap filling"
2. **Select** "Trend analysis"
3. **Click** "Apply"
4. AC50 endpoint related to ER enzyme assay
5. Pay attention that default descriptor displayed on X-axis is log Kow.

Types endpoint correlations

Continuous vs. continuous

Replacement of default X-descriptor (logKow) with other AC50 data – step 4



1. Click on "Descriptors/Data" menu item
2. Click on "Select endpoint tree descriptor..." menu item
3. Message informing for checking and arranging data appears

Types endpoint correlations

Continuous vs. continuous

Replacement of default X-descriptor (logKow) with other AC50 data – step 4

The screenshot displays the QSAR Toolbox 3.3.0.147 interface. The main window shows a scatter plot of AC50 (obs.) log (mol/L) versus log Kow. A dialog box titled "Select an endpoint tree node as gap filling descriptor" is open, showing a tree structure of endpoints. The "Human Health Hazards" node is selected, and a list of sub-nodes is shown. A blue callout bubble with the number "1" points to the dialog box. The background shows a scatter plot of AC50 (obs.) log (mol/L) versus log Kow, and a sidebar with various tool options.

Endpoint Data Tree:

- Human Health Hazards (375/17480)
 - ToxCast (375/17480)
 - ACEA (200/242)
 - Apredica (118/882)
 - Attagene (305/3060)
 - BioSeek (220/7038)
 - NCGC (375/2744)
 - Novascreen (247/2360)
 - Odyssey Thera (251/1154)

Scale/Units: uM, 375/17480

Descriptor X: log Kow

1. A window with arranged "Endpoint data tree" appears

Types endpoint correlations

Continuous vs. continuous

Replacement of default X-descriptor (logKow) with other AC50 data – step 4

QSAR Toolbox 3.3.0.147 [Document_1]

Input Profile

Select an endpoint tree node as gap filling descriptor

1

2

3

Scale/Units Available data
uM 220/220

AC50 (obs.) log10 mol/L

Descriptors Prediction

Structure

Estrogen R

Target Endpoint

Human Health Hazards ToxCast NCGC NCGC Reporter Gene Assay Era Agonist Homo sapiens Estrogen Receptor 1

Assay provider
Assay name
Test organisms (species)
Entrez gene name

target] 25 [target] 28 [target]

8.46 mg/L 4.96... M: 5.52 mg/L M: 8.25 mg/L

Accept prediction

Return to matrix

Select/filter data

Selection navigation

Gap filling approach

Descriptors/data

Make active descriptor

Remove active descriptor

Collect data

Change descriptor units ...

Edit descriptor options ...

Select endpoint tree descriptor ...

Model/(Q)SAR

Calculation options

Visual options

Information

Miscellaneous

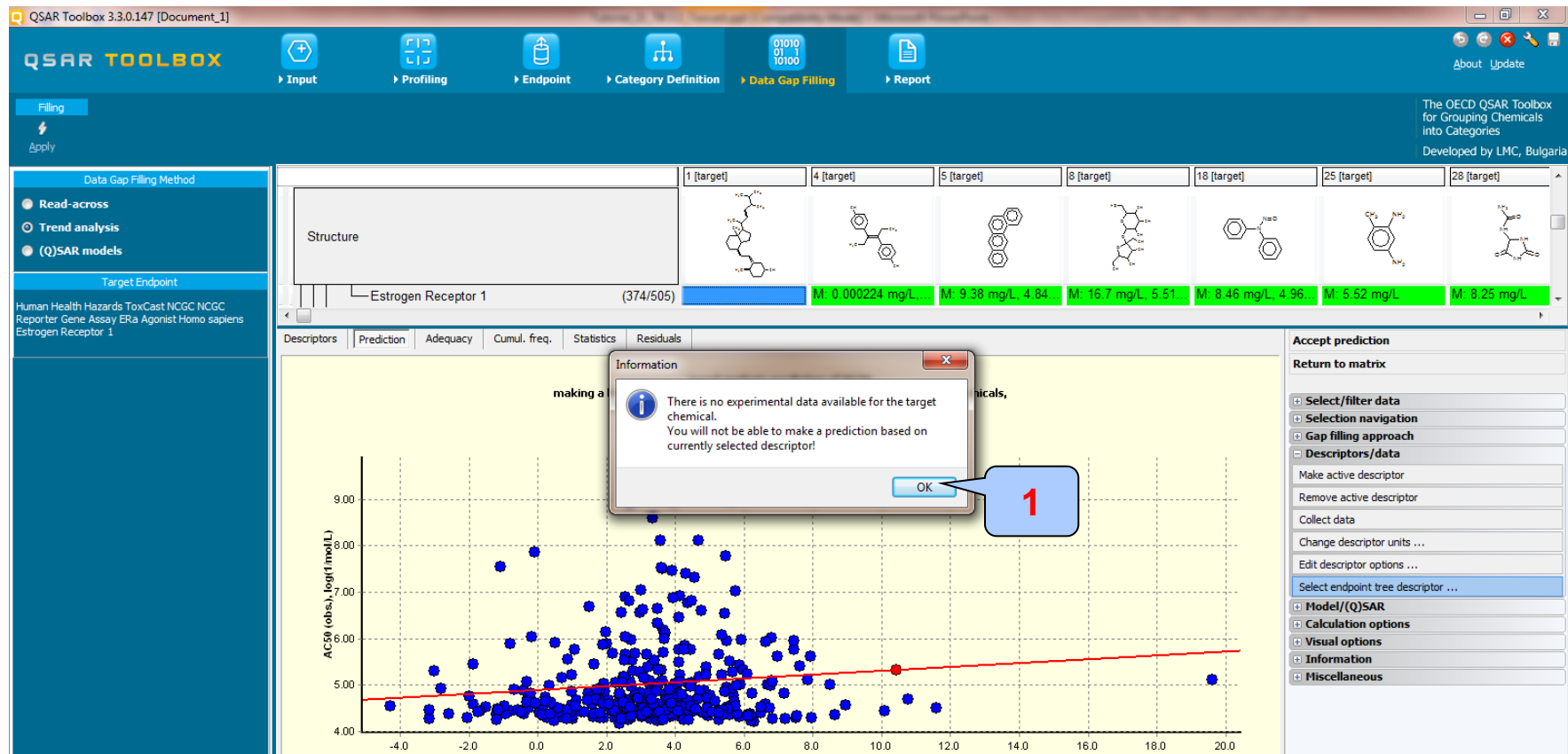
1. **Open** nodes under "NCGC" node;
X-axis circled in red box;

2. **Select** second endpoint, which will be placed on
3. **Click** "OK" button

Types endpoint correlations

Continuous vs. continuous

Replacement of default X-descriptor (logKow) with other AC50 data – step 4



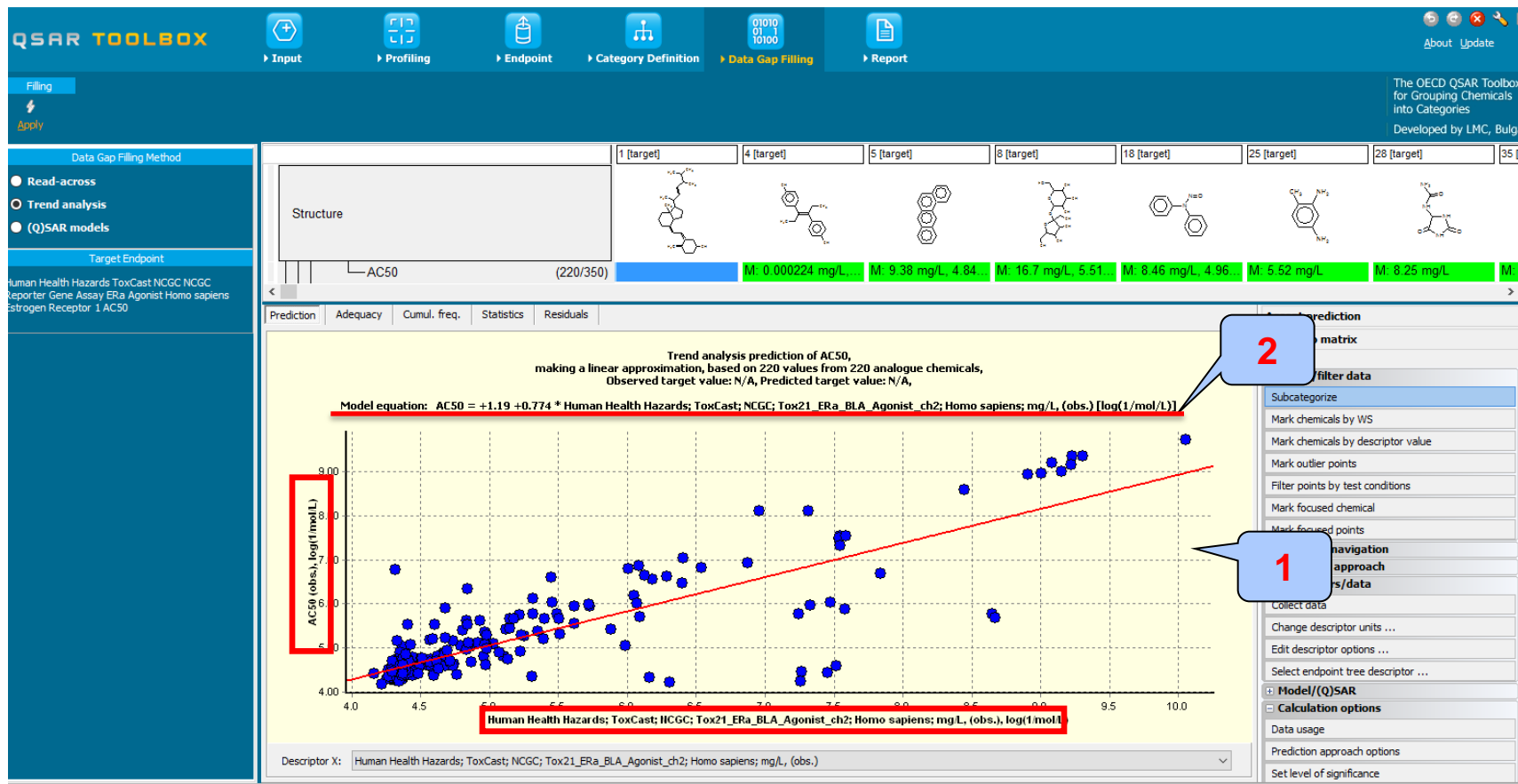
1. **Click** "OK" on the message informing that there is no experimental data for the target chemical.

The aim of this example is to see how the data correlates, so we ignore this message.

Types endpoint correlations

Continuous vs. continuous

Replacement of default X-descriptor (logKow) with other AC50 data – step 4



1. The graph obtained after replacing logKow with other ToxCast endpoint is visualized
2. The equation including endpoint data is rebuild

Types endpoint correlations

Continuous vs. continuous

Interpretation of correlation results

- In this example, we have correlated two AC50 endpoints associated with different type assay
- As seen from the graph, a linear relationship between two endpoints has been observed
- In order to assess only the chemicals having positive estrogen activity we remove the “Non-binders” chemicals based on subcategorization by “Estrogen receptor binding by OASIS” profiler (illustrated on next slide)

Types endpoint correlations

Continuous vs. continuous

Subcategorization by Estrogen receptor binding profiler

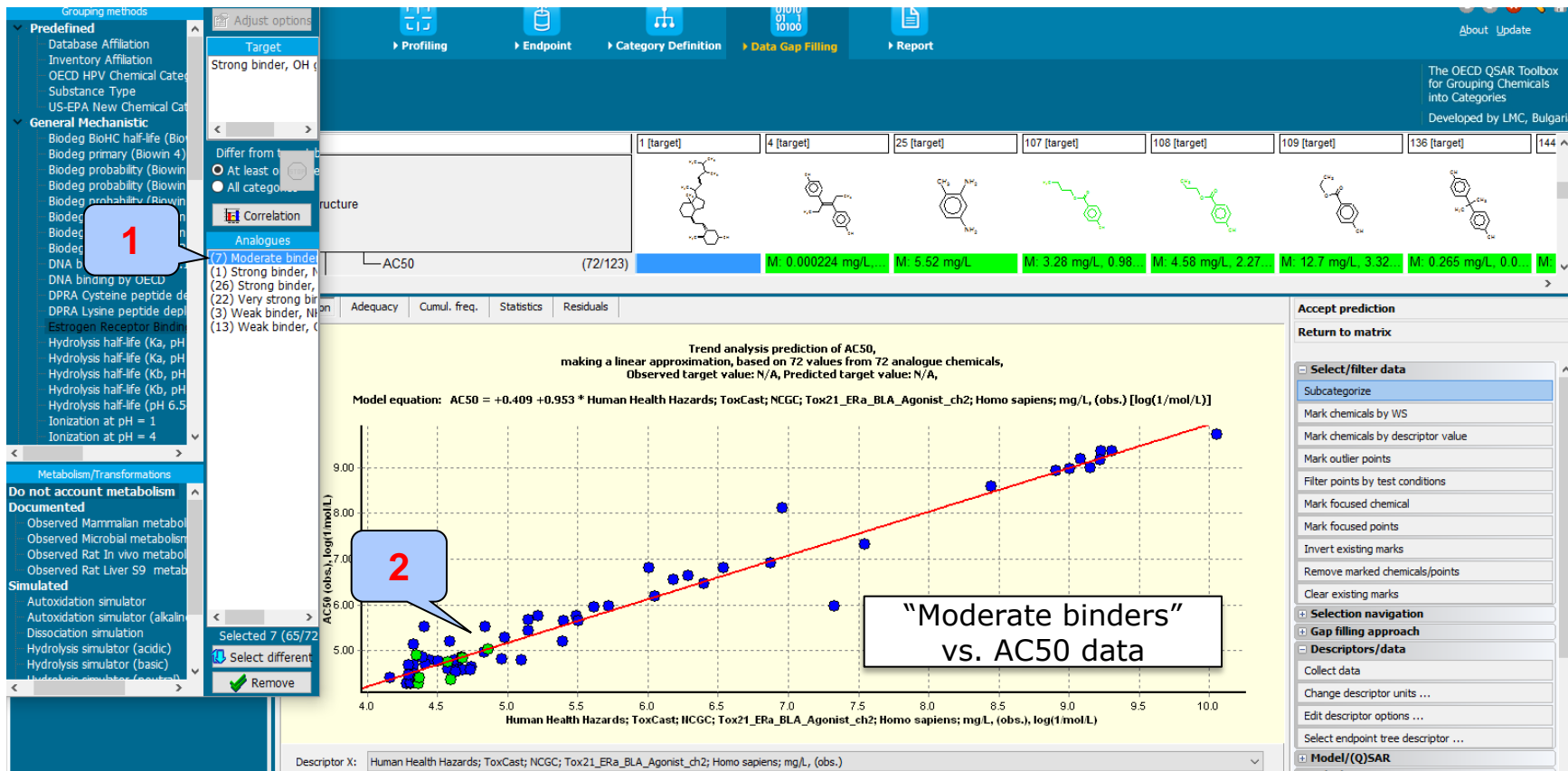
The screenshot displays the QSAR Toolbox software interface. On the left, the 'Endpoint Specific' section is expanded, showing 'Estrogen Receptor Binding' as the selected endpoint. In the center, the 'Correlation' section is active, showing a list of chemical categories. Callout 3 points to the selection of 'Non binder, impaired OH or NH2 group' and 'Non binder, MW>500'. Below this, a scatter plot titled 'Trend analysis prediction of AC50, making a linear approximation, based on 220 values from 220 analogue chemicals' is shown. The plot displays a positive correlation between 'Human Health Hazards; ToxCast; NCGC; Tox21_ERa_BLA_Agonist_ch2; Homo sapiens; mg/L, (obs.) [log(1/mol/L)]' on the x-axis and 'AC50 = +1.19 + 0.774 * Human Health Hazards; ToxCast; NCGC; Tox21_ERa_BLA_Agonist_ch2; Homo sapiens; mg/L, (obs.) [log(1/mol/L)]' on the y-axis. On the right, the 'Select/filter data' menu is open, and callout 1 points to the 'Subcategorize' option. At the bottom, callout 4 points to the 'Remove' button.

1. **Open** "Select/filter data" menu item, then **click** "Subcategorize";
2. **Select** "Estrogen receptor binding" profiler;
3. **Select** only Non binder categories by **left mouse click** and **hold** "Ctrl" button
4. **Click** "Remove" button

Types endpoint correlations

Continuous vs. continuous

Correlation of active Estrogen receptor categories vs.AC50 endpoint

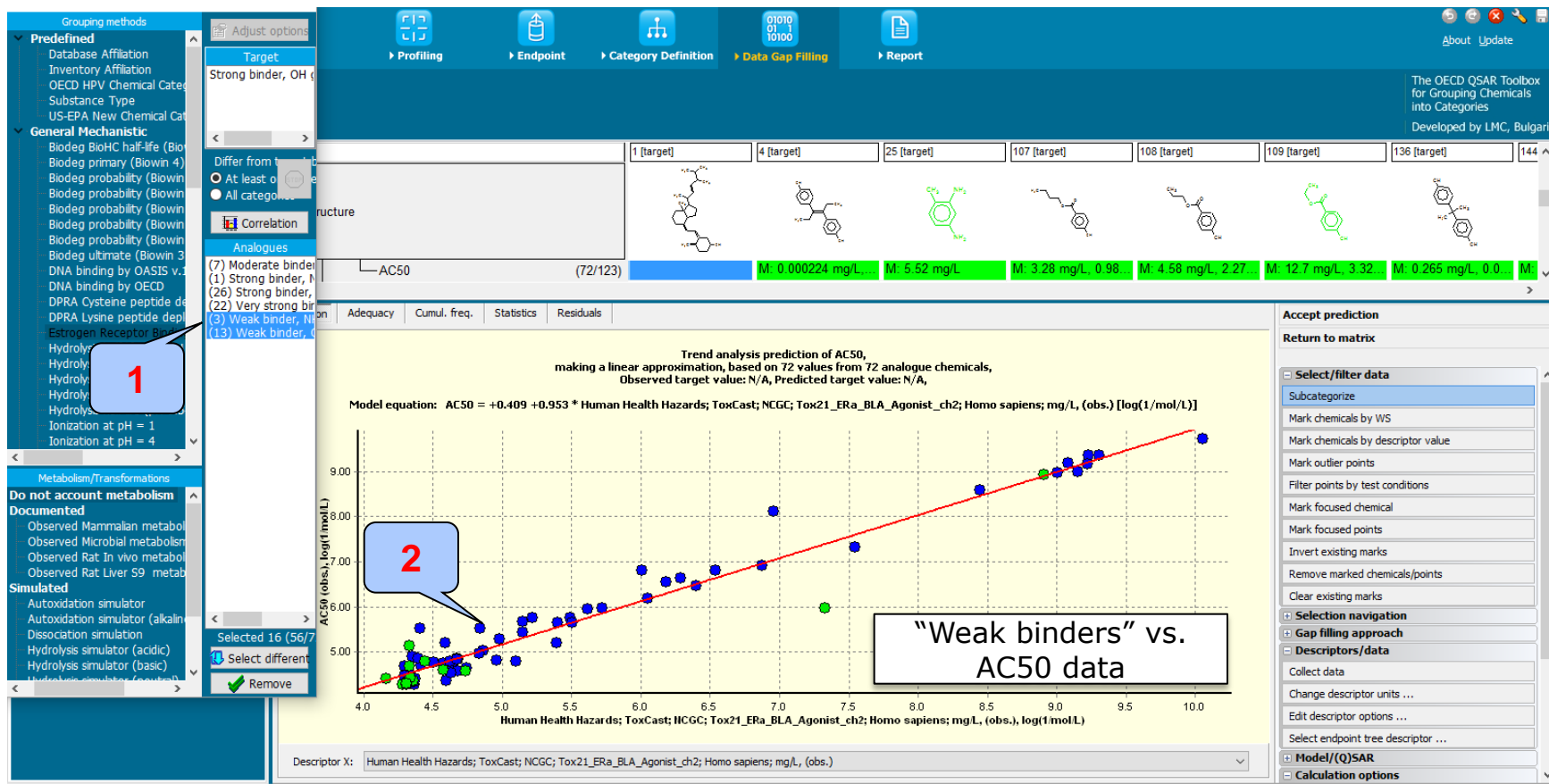


1. **Select** "Moderate binder" categories (left mouse click and hold "Ctrl" button)
2. The chemicals corresponding to the selected categories are highlighted in green

Types endpoint correlations

Continuous vs. continuous

Correlation of active Estrogen receptor categories vs.AC50 endpoint

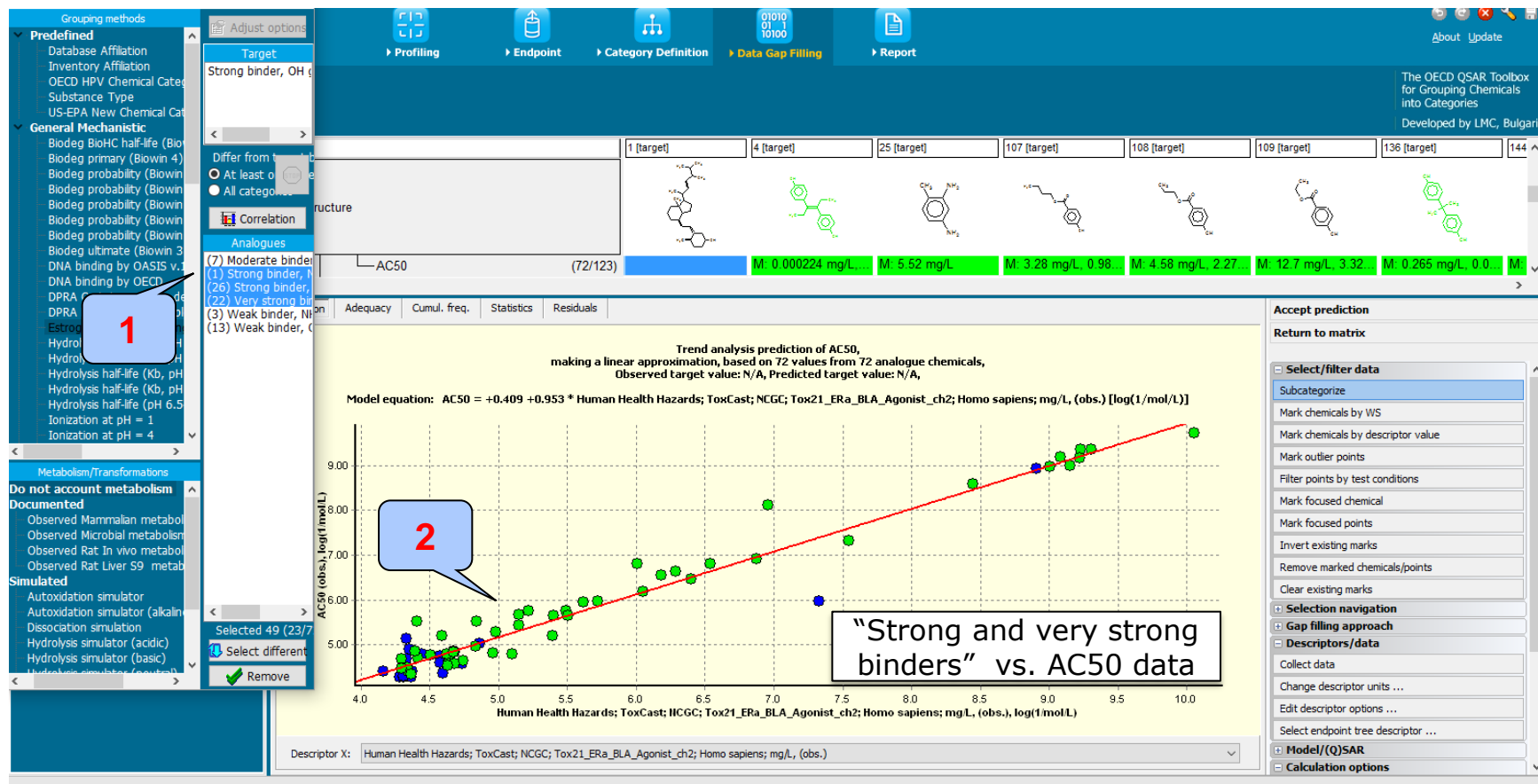


1. **Select** "Weak binder" categories (left mouse click and hold "Ctrl" button)
2. The chemicals corresponding to the selected categories are highlighted in green

Types endpoint correlations

Continuous vs. continuous

Correlation of active Estrogen receptor categories vs.AC50 endpoint



Types endpoint correlations

Continuous vs. continuous

Correlation results

- The two AC50 endpoints associated with different type assay have been correlated each other
- Non binders according to Estrogen receptor binding profiler have been eliminated from the correlation
- User can analyse the distribution of remaining ER binders (Very strong, Strong, Moderate and Weak) across selected AC50 endpoint

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - Correlation of data – background
 - **Types endpoint correlations**
 - Continuous vs. continuous
 - Categorical vs. categorical

Types endpoint correlations

Categorical vs. categorical

- The aim of this type correlation is to illustrate how categorical type data correlates each other.
- Categorical type data is the statistical data type consisting of categorical variables or of data that has been converted into that form. Such data is binary Ames data (dichotomic type): positive, negative or polytomic type data such as GPMT data: strong, weak and negative.
- Two examples illustrating this type correlation will be demonstrated:
 - Example 1: Correlation of two types skin sensitization data
 - LLNA (Strongly positive, Weakly positive, Negative) vs. GPMT (Strong, Moderate, Weak and Non)
 - Example 2: Correlation of skin sensitization and Ames mutagenicity data
 - GPMT (Strong, Moderate, Weak and Non) vs. AMES (Positive, Equivocal, Negative)
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
 - *Load Skin sensitization database (step 1)*
 - *Gather experimental data (step 2)*
 - *Define target endpoint (step 3)*
 - *Enter Gap filling (step 4)*
 - *Perform correlation between endpoints (step 5).*

Types endpoint correlations

Categorical vs. categorical

Load Skin sensitization database – step 1

Example 1: Correlation of LLNA and GPMT data

The screenshot shows the QSAR Toolbox software interface. The 'Input' menu is highlighted. The 'DB' button in the toolbar is highlighted with a red box. A 'Select database' dialog box is open, showing a list of databases. 'Skin Sensitization' is selected in the list. The 'OK' button is highlighted. The main workspace shows a datamatrix with chemical structures loaded from the database.

1. **Go** to "Input";
2. **Click** "DB" button;
3. **Select** "Skin sensitization" database;
4. **Click** OK;
5. The chemicals from database have been loaded on datamatrix

Types endpoint correlations

Categorical vs. categorical
Gather experimental data – step 2

Example 1: Correlation of LLNA and GPMT data

The screenshot shows the QSAR Toolbox 3.4.0.17 interface. The 'Endpoint' tab is selected in the top menu bar. In the left sidebar, 'Skin sensitization' is checked under the 'Inventories' section. The 'Gather' button is highlighted with a red box. A 'Read data?' dialog box is open, showing 'All endpoints' selected and 'from Tautomers' checked. A 'Repeated values' dialog box is also open, showing a table of data points. A status window at the bottom right indicates '1684 data points gathered across 1208 chemicals.'.

Data points...	Endpoint	CAS	Structure	Value	Assay
<input checked="" type="checkbox"/>	SWAN	56-81-5	<chem>CC(C)O</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	56-81-5	<chem>CC(C)O</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	56-81-5	<chem>CC(C)O</chem>	Not sensitising	Miscellaneous
<input checked="" type="checkbox"/>	SWAN	67-68-5	<chem>CS(=O)(=O)C</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	67-68-5	<chem>CS(=O)(=O)C</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	67-68-5	<chem>CS(=O)(=O)C</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	67-68-5	<chem>CS(=O)(=O)C</chem>	Not sensitising	Miscellaneous
<input type="checkbox"/>	SWAN	67-68-5	<chem>CS(=O)(=O)C</chem>	Not sensitising	Miscellaneous

1. Go to "Endpoint";
2. Select "Skin sensitization";
3. Click "Gather"
4. Click "OK"
5. Click "Select one";
6. Click "OK";
7. Click "OK"

Types endpoint correlations

Categorical vs. categorical
Gather experimental data – step 2

Example 1: Correlation of LLNA and GPMT data

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Data Import Export Delete Tautomerize

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

Databases

Select All Unselect All Invert About

- Developmental & Reproductive Toxicity
- Developmental toxicity ILSI
- ECHA CHEM
- ECOTOX
- Estrogen Receptor Binding Affinity OA
- Eye Irritation ECETOC
- Genotoxicity OASIS
- Human Half-Life
- Keratinocyte gene expression Givauda
- Keratinocyte gene expression LuSens
- Micronucleus ISSMIC
- Micronucleus OASIS
- MUNRO non-cancer EFSA
- Rep Dose Tox Fraunhofer ITEM
- Repeated Dose Toxicity HESS
- Rodent Inhalation Toxicity Database
- Skin Irritation
- ☒ Skin sensitization
- Skin sensitization ECETOC
- ToxCastDB
- Toxicity Japan MHLW
- ToxRefDB US-EPA
- Yeast estrogen assay database
- ZEBET database

Filter endpoint tree...

Structure

- Physical Chemical Properties
- Environmental Fate and Transport
- Ecotoxicological Information
- Human Health Hazards
 - Acute Toxicity
 - Bioaccumulation
 - Carcinogenicity
 - Developmental Toxicity / Teratogenicity
 - Genetic Toxicity
 - Immunotoxicity
 - Irritation / Corrosion
 - Neurotoxicity
 - Photoinduced Toxicity
 - Repeated Dose Toxicity
 - Sensitisation
 - Skin
 - In Chemico
 - In Vitro
 - In Vivo
 - GPMT (334/335) M: Negative
 - HRIPT (116/165)
 - LLNA (617/674) M: Negative
 - Miscellaneous (421/509)
 - Undefined Assay (1/1)

1

2

All skin sensitization data has been converted into positive/negative data based on implemented scale conversion.
Note: A reminder slide illustrating what is scale and scale conversion is provided on next click.

1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]	8 [target]
<chem>CC(C)O</chem>	<chem>c1ccc2ccccc2c1</chem>	<chem>CCCCCCCCCCCCCCCC</chem>	<chem>CC(C)O</chem>	<chem>O=C(O)c1ccccc1</chem>	<chem>Nc1ccccc1</chem>	<chem>Nc1ccccc1</chem>	<chem>Nc1ccccc1</chem>
(334/335) M: Negative							
(116/165)							
(617/674) M: Negative	M: Positive			M: Negative		M: Negative	
(421/509)		M: Positive		M: Negative, not c...	M: Negative	M: Positive	
(1/1)							M:

1. Skin sensitization data appeared on data matrix.
2. Data associated with different type assay (e.g LLNA, GPMT) are distributed in separate nodes

What is “scale” and “scale conversion” ?

Reminder slide

- Skin sensitisation as an example is a “qualitative” endpoint for which the results are presented with categorical type of data (for example: positive; negative; weak sensitizer; strong sensitizer, etc).
- Skin sensitisation potential of the chemicals came from different authors coded with different names (for example: data from John Moores University of Liverpool are: *Strongly sensitizing*, *Moderately sensitizing* etc.; data from European centre for Ecotoxicology and Toxicology of chemicals are: *Positive*, *Negative*, and *Equivocal*).
- The main purpose of the scales is to unify all data available in the Toolbox databases for a certain endpoint.
- “Scale conversion” is the TB instrument to create conversions between scales. More reasonable is to convert more informative to less informative scale.
- The default scale for Skin Sensitisation data is “Skin Sensitisation ECETOC”. It converts all skin sensitization data into: Positive and Negative. This allows skin sensitization data to be used as much as possible for gap filling purposes.

Types endpoint correlations

Categorical vs. categorical

Define target endpoint – step 3

Example 1: Correlation of LLNA and GPMT data

The screenshot shows the QSAR Toolbox interface. On the left, the 'Filter endpoint tree...' panel lists various endpoints. The 'GPMT' endpoint is selected, and its sub-endpoint 'LLNA' is highlighted with a red box. A blue callout bubble with the number '1' points to the 'LLNA' entry. The main table displays data for various targets, with the 'LLNA' row showing 'M: Negative' for target 1 and 'M: Positive' for target 2.

Endpoint	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]	8 [target]
GPMT	(334/335) M: Negative							
LLNA	(617/674) M: Negative	M: Positive						

The target endpoint is EC3 data associated with LLNA assay
1. Click on the cell associated with target endpoint

Types endpoint correlations

Categorical vs. categorical

Enter Gap filling – step 4

Example 1: Correlation of LLNA and GPMT data

Note: By default EC3 data has been converted into binary categories: positive/negative based on scale "Skin sensitization II (ECETOC)". For the purpose of this exercise, Skin sensitization I (OASIS) will be used. This scale converts EC3 data into three categories: Strongly positive (EC3 0-10%), Weakly positive (EC3 10-50%) and Negative (EC3>50%).

Enter Gap filling and apply read across. Read across is applied because a categorical type data is analyzed. Follow the steps:

1. **Go** to "Data Gap filling";
2. **Select** "Read-across";
3. **Click** "Apply";
4. **Select** "Skin sensitization I (OASIS)" scale (see Note);
5. **Click** "OK"

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation of LLNA and GPMT data

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

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

Human Health Hazards Sensitisation Skin In Vivo LLNA EC3

EC3

Structure

EC3 (617/674)

1 [target] 2 [target] 4 [target] 6 [target] 9 [target] 12 [target] 15 [target] 16 [target]

M. Negative M. Positive M. Negative M. Negative M. Negative M. Negative M. Positive M. Negative

Descriptors Prediction

Read across prediction of EC3, taking the highest mode from the nearest 5 neighbours, based on 5 values from 5 neighbour chemicals, Observed target value: 'Negative', Predicted target value: 'Strongly positive'

EC3 (obs)

Strongly positive

Weakly positive

Accept prediction

Return to matrix

Select/filter data

Selection navigation

Gap filling approach

Descriptors/data

Model/(Q)SAR

Calculation options

Data usage

Prediction approach options

Use target data for prediction

Visual options

Set usage of data

All

Minimal

Maximal

Lower median

Higher median

Mode(s)

Lowest mode

Highest mode

OK Cancel

1

2

3

4

Correlation assumes a single value per chemical to be used. In this respect the default calculation settings should be changed from "All" to something different. In our case study we play a worst case scenario, thus an option "All" is changed to "Maximal" values. Follow the steps:

1. **Open** "Calculation options";
2. **Click** on "Data usage" menu item;
3. **Select** Maximal;
4. **Click** "OK"

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation of LLNA and GPMT data

1. Click "Subcategorize" ; 2. Click "Endpoint data" node; 3. Click "Adjust options"; 4. "Endpoint data grouper" window appears. More details about this window are provided on next slide.

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Endpoint correlation options...

1. Select descriptor button

2. descriptor: Human Health Hazards Sensitisation Skin In Vivo GPMT S M W N

3. Default number of ratio bins: 3

4. Single category per chemical

5. Scale/Unit: Skin sensitisation IV (GPMT)

6. Data usage: highest value

7. highest value

8. Recreate bins

9. Units and Scales: Skin sensitisation IV (GPMT)

10. Bin constraints: Non sensitizer (Skin sensitisation IV (GPMT)), Weak sensitizer (Skin sensitisation IV (GPMT)), Strong sensitizer (Skin sensitisation IV (GPMT)), Moderate sensitizer (Skin sensitisation IV (GPMT))

11. Resulting categories list... Strong sensitizer (Skin sensitisation IV (GPMT)), Moderate sensitizer (Skin sensitisation IV (GPMT)), Weak sensitizer (Skin sensitisation IV (GPMT)), Non sensitizer (Skin sensitisation IV (GPMT))

Select descriptor...

1. Select descriptor button

2. Selected descriptor: Human Health Hazards Sensitisation Skin In Vivo GPMT S M W N

3. Select descriptor button

4. Structure: CC(=O)C, c1ccc2ccccc2c1, CC(C)O

5. GPMT: S ... (84/85) M: Negative

6. LLNA: ... (617/674) M: Positive

7. Mis... (88/133) M: Negative, not conver...

8. Undefined A

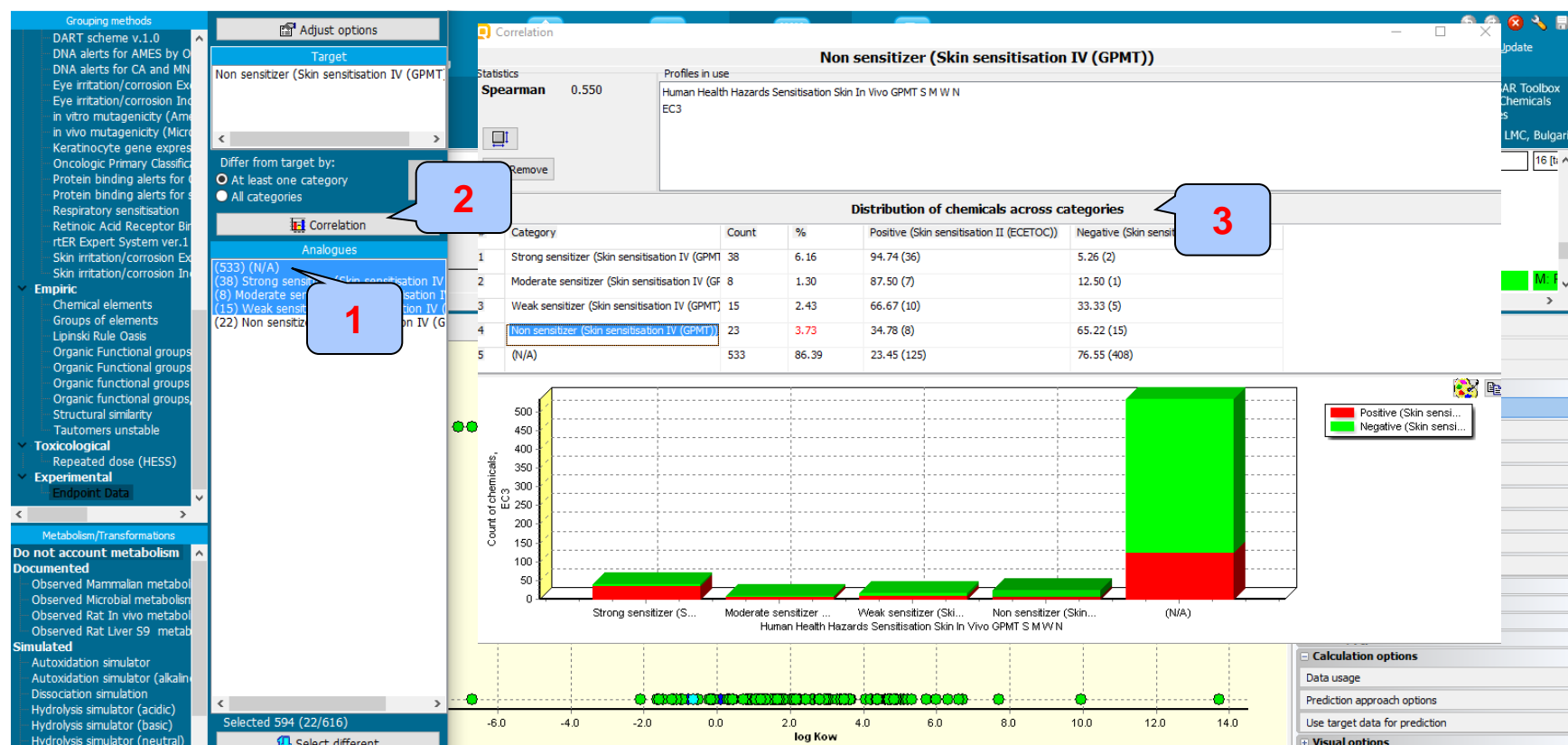
1. "Select descriptor" button allows the user to select second endpoint which will be used in the correlation. **Click on** the button. Additional window Appears;
2. **Click** on the row associated with "S M W N" endpoint;
3. **Click** "Select descriptor";
4. By default the program separates data into 3 bins, number of bins could be changed from here;
5. "Single category per chemical" produces a single value per chemical whenever multiple values of single unit/scale are present;
6. List of scales used in the correlation;
7. Highest mode are used in this case, because worst case scenario is played;
8. **Click** "Recreate bins" to finish the procedure of selecting endpoint;
9. Units and scales used in the correlation;
10. A panel with bins used in the correlation;
11. **Click** OK to finish the correlation settings;

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation of LLNA and GPMT data



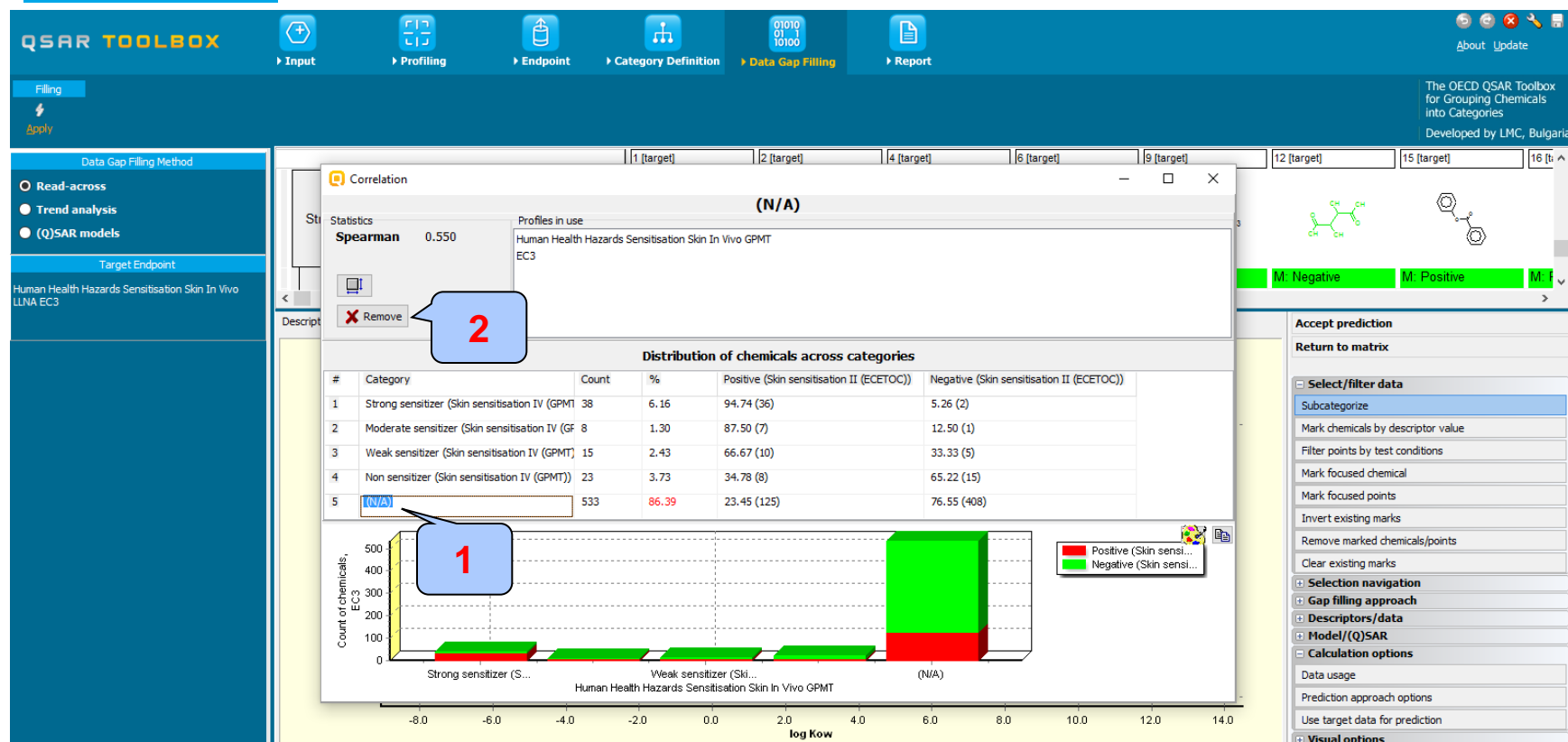
1. After the settings are configured all the analogues are distributed in 5 bins depending on GPMT data: Strong, Moderate, Weak, Non sensitizer and N/A. Analogues, which do not have GPMT data are marked as N/A (533 in this case).
2. Click "Correlation" button;
3. A window with contingency table appears.

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation of LLNA and GPMT data



Analogue with no GPMT data (N/A bin) could be removed from the table. This will not affect the value of correlation coefficient.

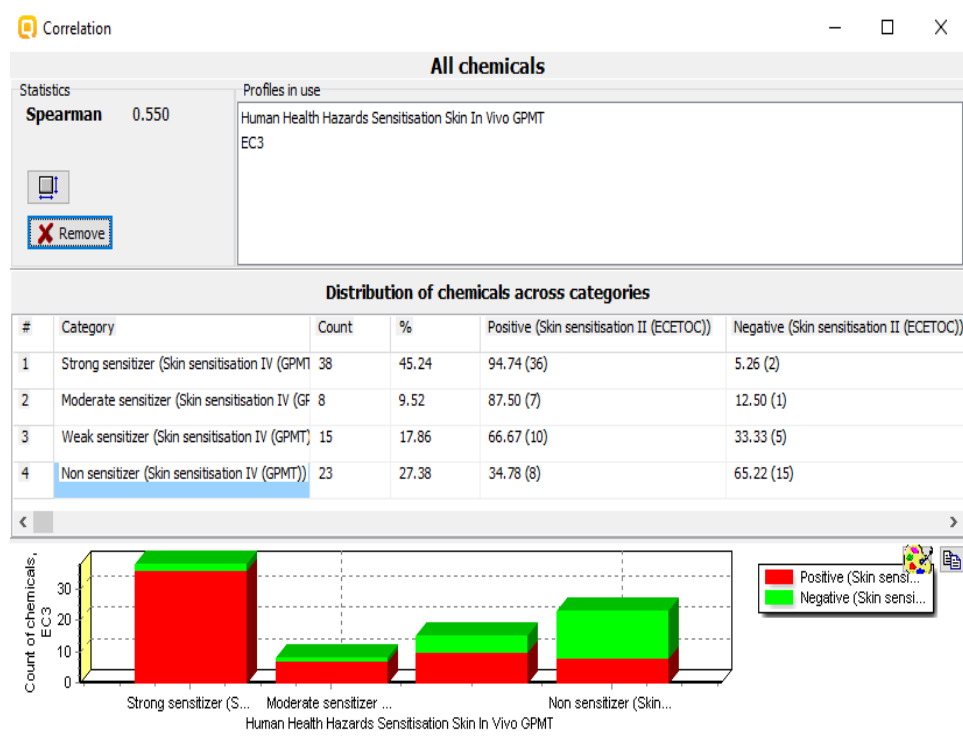
1. Click on the row with N/A
2. Click on "Remove" button

Types endpoint correlations

Categorical vs. categorical

Interpretation of correlation results (LLNA vs. GPMT)

- Correlation analysis between two categorical type skin sensitization data (LLNA and GPMT) shows strong endpoint correlation (Spearman coefficient is 0.55, see slide 17 for details).



Types endpoint correlations

Categorical vs. categorical

- The second example illustrating categorical vs. categorical type correlation is:
 - Example 1: Correlation of Skin sensitization data
 - LLNA (Strongly positive, Weakly positive, Negative)
 - GPMT (Strong, Moderate, Weak and Non)
 - **Example 2: Correlation of Skin sensitization and Ames mutagenicity data**
 - **GPMT (Strong, Moderate, Weak and Non)**
 - **AMES (Positive, Equivocal, Negative)**
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
 - *Load Skin sensitization database (step 1) – skipped, because this database is already loaded on data matrix*
 - *Gather experimental data (step 2)*
 - *Define target endpoint (step 3)*
 - *Enter Gap filling (step 4)*
 - *Perform correlation between endpoints (step 5)*

Types endpoint correlations

Categorical vs. categorical
Gather experimental data – step 2

Example 2: Correlation of GPMT and AMES data

1 Go to "Endpoint"; including Ames data;

2 Skin sensitization DB is already selected;

3 Select the databases

4 Click "Gather"

5 The data appeared on datamatrix

Note that the correlation between endpoints is possible when data is gathered and available on data matrix. One should be aware of the data values that would be using during the data gap filling and gather the data for the corresponding endpoint during the "Endpoint" stage of the workflow, prior to entering the "Data gap filling" module

Types endpoint correlations

Categorical vs. categorical

Define target endpoint – step 3

Example 2: Correlation of GPMT and AMES data

The screenshot shows the QSAR Toolbox software interface. The top bar includes the 'QSAR TOOLBOX' logo and navigation icons. Below this are tabs for 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The left sidebar contains 'Databases' (Human Health Hazards, Environmental Hazards, etc.) and 'Inventories' (Canada DSL, COSING, etc.). The main area displays a 'Filter endpoint tree...' on the left and a table of results on the right. The table has columns for 'Structure' and '1 [target]' through '8 [target]'. A red box highlights the 'S M W N' entry in the 'Structure' column, which is associated with 'M: Negative' in the '1 [target]' column. A blue callout bubble with the number '1' points to this entry.

The target endpoint is skin sensitization GPMT

1. **Click** on the cell associated with target endpoint

Types endpoint correlations

Categorical vs. categorical

Enter Gap filling – step 4

Example 2: Correlation of GPMT and AMES data

1

2

3

4

5

Converted data

Original data

#	Endpoint	Value	Original value	Organ	Reference source	Phylum (common name)	Phylum	Test method / Data	Type of method	Year	Test organism (species)	Title	Kind of organism
1	S M W N	Negative (Skin sensitizer sensitization (Skin on II (ECETOC) on IV	Non sensitizer sensitization (Skin sensitizer sensitization (Skin on II (ECETOC) on IV	Skin	SAR QSAR Environ. Res. 2(3): 159-179	Vertebrates	Chordata	GPMT	In Vivo	1994	guinea pig	Multivariate Analysis of QSAR data for skin sensitization	Mammal

Enter Gap filling applying read across. Read across is applied because a categorical type data is analyzed.

1. **Go** to "Data Gap filling";
2. **Select** "Read-across";
3. **Click** "Apply";
4. **Select** "Skin sensitization IV (GPMT)" scale;
5. **Click** "OK"

Types endpoint correlations

Categorical vs. categorical

Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of GPMT and AMES data

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

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

Read-across
Trend analysis
(Q)SAR models

Target Endpoint
Human Health Hazards Sensitisation Skin In Vivo
GPMT S M W N

Structure

1 [target] 6 [target] 7 [target] 8 [target] 9 [target] 11 [target] 14 [target] 15 [target]

S M W N (334/335) M: Negative M: Positive M: Negative M: Negative M: Negative M: Positive M: Positive M: F

Descriptors Prediction

Read across prediction of S M W N, taking the highest mode from the nearest 5 neighbours, based on 5 values from 5 neighbour chemicals, Observed target value: 'Non sensitizer', Predicted target value: 'Non sensitizer'

Strong sensitizer
Moderate sensitizer
Weak sensitizer
Non sensitizer

log Kow

Accept prediction
Return to matrix

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

Data usage

Prediction approach options
Use target data for prediction

Visual options

Set usage of d

All
Minimal
Maximal
Lower median
Higher median
Mode(s)
Lowest
Highest

OK Cancel

1. Open "Calculation" options. 2. Click on "Data usage" 3. Select "Maximal" 4. Click "OK" (refer to slide 53 for details)

Types endpoint correlations

Categorical vs. categorical

Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of GPMT and AMES data

The screenshot displays the QSAR Toolbox software interface during a correlation analysis. The left sidebar contains a tree view of grouping methods and endpoint data. The main window shows the 'Adjust options' dialog box, which is used to select the target descriptor. The 'Select descriptor' dialog box is also visible, showing the selection of the 'Gene Mutation' descriptor under the 'Salmonella typhimurium' category. The 'Endpoint data' section on the left shows the selection of 'With S9' data. The 'Accept prediction' dialog box on the right shows the 'Subcategorize' option selected. Numbered callouts 1 through 6 highlight the following steps:

- Click on "Subcategorize" in the "Accept prediction" dialog.
- Click on "Endpoint data" node in the left sidebar.
- Click on "Adjust options" button in the main window.
- Click "Select descriptor" button in the "Adjust options" dialog.
- Click on "With S9" under In Vitro|Bacterial Reverse Mutation Assay (e.g. Ames Test)|Gene Mutation| Salmonella typhimurium in the "Select descriptor" dialog.
- Click on "Select descriptor" button in the "Select descriptor" dialog.

1. **Open** "Subcategorize";
2. **Click** on "Endpoint data" node;
3. **Click** on "Adjust options" button;
4. **Click** "Select descriptor" button;
5. **Click** on "With S9" under In Vitro|Bacterial Reverse Mutation Assay (e.g. Ames Test)|Gene Mutation| Salmonella typhimurium;
6. **Click** on "Select descriptor" button

Types endpoint correlations

Categorical vs. categorical

Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of GPMT and AMES data

The screenshot shows the 'Endpoint data grouper options...' dialog box in the OECD QSAR Toolbox. The dialog is used to define categories for correlation. The following steps are highlighted with numbered callouts:

1. Select descriptor: Human Health Hazards Genetic Toxicity In Vitro Bacterial Reverse Mutation
2. Select "Single category per chemical"
3. Select scale "Gene mutation I"
4. Select "highest value" (worst case)
5. Click "OK"

The background shows a scatter plot of log Kow vs. log Kow, with data points clustered around the origin. The x-axis ranges from -8.0 to 16.0, and the y-axis ranges from -8.0 to 16.0. The plot is titled 'log Kow'.

1. 3 bins are set by default;
2. **Select** "Single category per chemical";
3. **Select** scale "Gene mutation I"
4. **Select** "highest value" (worst case);
5. **Click** "OK"

Types endpoint correlations

Categorical vs. categorical

Perform correlation between GPMT and AMES data – step 5

Example 2: Correlation of GPMT and AMES data

The screenshot shows the OECD QSAR Toolbox interface. The main window displays the 'Correlation' results for 'All chemicals'. The 'Statistics' section shows a Spearman correlation of 0.2873. The 'Distribution of chemicals across categories' table shows the following data:

#	Category	Count	%	Strong sensitizer (Skin sensitisation IV (GPMT))	Moderate sensitizer (Skin sensitisation IV (GPMT))	Weak sensitizer (Skin sensitisation IV (GPMT))
1	Positive (Gene mutation I)	24	7.19	58.33 (14)	12.50 (3)	12.50 (3)
2	Equivocal (Gene mutation I)	4	1.20	50.00 (2)	0.00 (0)	25.00 (1)
3	Negative (Gene mutation I)	92	27.54	25.00 (23)	15.22 (14)	16.30 (15)
4	(N/A)	214	64.07	9.35 (20)	29.44 (63)	41.59 (89)

The bar chart below the table shows the distribution of chemicals across categories, with the x-axis representing log Kow and the y-axis representing the count of chemicals. The legend indicates: Strong sensitizer (S...), Moderate sensitizer (S...), Weak sensitizer (Ski...), and Non sensitizer (Skin...).

Four numbered callouts highlight specific steps in the process:

1. Analogues are distributed in 4 bins
2. Click on "Correlation" button
3. Click on the row with N/A data
4. Click on "Remove"

1. Analogues are distributed in 4 bins
3. Click on the row with N/A data

2. Click on "Correlation" button
4. Click on "Remove"

Types endpoint correlations

Categorical vs. categorical

Interpretation of correlation results (GPMT vs. AMES)

- Correlation analysis between two categorical type data: GPMT and AMES shows weak correlation between two endpoints (Spearman coefficient is 0.3, see slide 22 for details).



Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - Correlation of data – background
 - **Types endpoint correlations**
 - Continuous vs. continuous
 - Categorical vs. categorical
 - Categorized continuous vs. categorical

Types endpoint correlations

Categorized continuous vs. categorical

- The aim of this type correlation is to illustrate how categorized continuous and categorical type data correlates each other.
- Categorized continuous data is the continuous type data (e.g LC50 or AC50, EC3, %) converted into categories.
- In this example we will illustrate how DPRA ratio data (%) correlates with LLNA data:
 - DPRA (ratio data expressed in % and converted in categories)
 - LLNA (categorical type: Strongly positive, Weakly positive, Negative)
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
 - *Load Skin sensitization database (step 1) – skipped, because this database has been already loaded on data matrix*
 - *Gather experimental data (step 2)*
 - *Define target endpoint (step 3)*
 - *Enter Gap filling (step 4)*
 - *Perform correlation between endpoints (step 5).*

Types endpoint correlations

Categorized continuous vs. categorical

Gather experimental data – step 2

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Go to "Endpoint";

2. Skin sensitization DB is already selected;

3. Select "Chemical reactivity (COLIPA)" database;

4. Click "Gather" button;

5. The data appeared on datamatrix

Types endpoint correlations

Categorized continuous vs. categorical

Define target endpoint – step 3

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

The screenshot shows the QSAR Toolbox software interface. The top navigation bar includes tabs for Input, Profiling, Endpoint, Category Definition, Data Gap Filling, and Report. Below this, there are sub-tabs for Data, Import, Export, Delete, and Tautomerize. The left sidebar contains a 'Databases' panel with a tree view of various toxicity endpoints, including 'Human Health Hazards' and 'Sensitisation'. The main workspace displays a 'Filter endpoint tree...' on the left and a table of results on the right. The table has columns for different target types (1 [target] to 8 [target]). The 'Sensitisation' endpoint is expanded, showing sub-endpoints like 'Skin' and 'In Vivo'. The 'LLNA' sub-endpoint is highlighted with a red box, and its result 'M: Negative' is highlighted with a blue box. A red circle with the number '1' is placed over the 'LLNA' result.

The target endpoint is EC3 skin sensitization data

1. **Click** on the cell associated with target endpoint and target chemical

Types endpoint correlations

Categorized continuous vs. categorical

Enter Gap filling – step 4

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

1. Go to "Data Gap filling"; 2. Select "Read-across"; 3. Click "Apply"; 4. Select "Skin sensitization I (OASIS)" scale; 5. Click "OK"

Converted data

Original data

#	Endpoint	Value	Original value	Organ	Reference source	Phylum (common name)	Test method / Data source
1	EC3	Negative (Skin sensitization I (ECETOC))	Negative (Skin sensitization I (Oasis))	Skin	Dermatitis, 16 (4): 1-46	Vertebrates	LLNA

Enter Gap filling and apply read across. Read across is applied because a categorical type data is analyzed.

1. **Go** to "Data Gap filling"; 2. **Select** "Read-across"; 3. **Click** "Apply"; 4. **Select** "Skin sensitization I (OASIS)" scale; 5. **Click** "OK"

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA (%) and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

QSAR TOOLBOX

Input Profiling Endpoint Category Definition Data Gap Filling Report

Target Endpoint: Human Health Hazards Sensitisation Skin In Vivo LLNA EC3

Structure: EC3 (617/674)

Descriptors: Prediction

Accept prediction: Return to matrix

1. Open "Calculation options"; 2. Click on "Data usage" (refer to slide 53); 3. Select "Maximal"; 4. Click "OK"

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

The screenshot shows the OECD QSAR Toolbox software interface. The sidebar on the left contains a list of endpoints under 'Endpoint Data'. The main window has two tabs: 'Adjust options' and 'Endpoint data grouper options...'. The 'Adjust options' tab is active, showing a list of endpoints. The 'Endpoint data grouper options...' tab is also visible, showing a 'Select descriptor' button. The 'Select descriptor...' dialog box is open, showing a list of descriptors. The 'DPRA' descriptor is selected. The 'Select descriptor...' dialog box also shows a table of descriptors and their correlation coefficients.

Descriptor	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]
DPRA							
% Depletion of Cystine (216/244)	M: -0.872 %	M: -3.78 %	M: 0 %	M: 3.8 %	M: 64.9 %		
% Depletion of Lysine (215/250)	M: 0.832 %	M: 2.1 %	M: 9.7 %, 2.38 %	M: 1.2 %	M: 2.23 %		
ITLC-MS (285/286)	M: -5.3 %	M: -5.83 %	M: 1 %	M: -5.98 %	M: 25.5 %		

1. Click on "Endpoint data"
2. Click on "Adjust options" button
3. Click on "Select descriptor" button
4. Click on the endpoint tree on the level of "DPRA". In this case we mixed DPRA lysine and Cysteine data
5. Click on "Select descriptor" button

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

The screenshot displays the 'Endpoint data grouper options...' dialog box within the OECD QSAR Toolbox. The dialog is used to configure the correlation between DPRA and LLNA data. Numbered callouts 1 through 5 highlight specific steps:

1. 'Default number of ratio bins' set to 3
2. 'Single category per chemical' checkbox checked
3. 'Scale/Unit' set to 'Skin sensitization DPRA (ratio)(%)'
4. 'Data usage' set to 'maximal value'
5. 'OK' button

The background shows a scatter plot of log Kow vs. Skin sensitization DPRA (ratio)(%) with data points categorized into bins. The plot shows a positive correlation between log Kow and Skin sensitization DPRA (ratio)(%).

1. 3 bins are set by default
2. **Select** "Single category per chemical"
3. **Select** scale "Skin sensitization DPRA (ratio)"
4. **Select** "maximal value" (worst case)
5. **Click** "OK"

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data

The screenshot shows the OECD QSAR Toolbox software interface. The main window displays the 'Correlation' analysis results for 'All chemicals'. The 'Statistics' section shows a Spearman correlation coefficient of -0.472. The 'Distribution of chemicals across categories' table is visible, showing the count and percentage of chemicals in different categories. A 3D bar chart below the table shows the distribution of chemicals across categories. Numbered callouts 1 through 4 highlight specific steps in the process:

- 1. Analogues are distributed in 4 bins depending on data
- 2. Click "Correlation" button
- 3. Click on the row with N/A data
- 4. Click "Remove"

1. Analogues are distributed in 4 bins depending on data
4. Click "Remove"

2. Click "Correlation" button

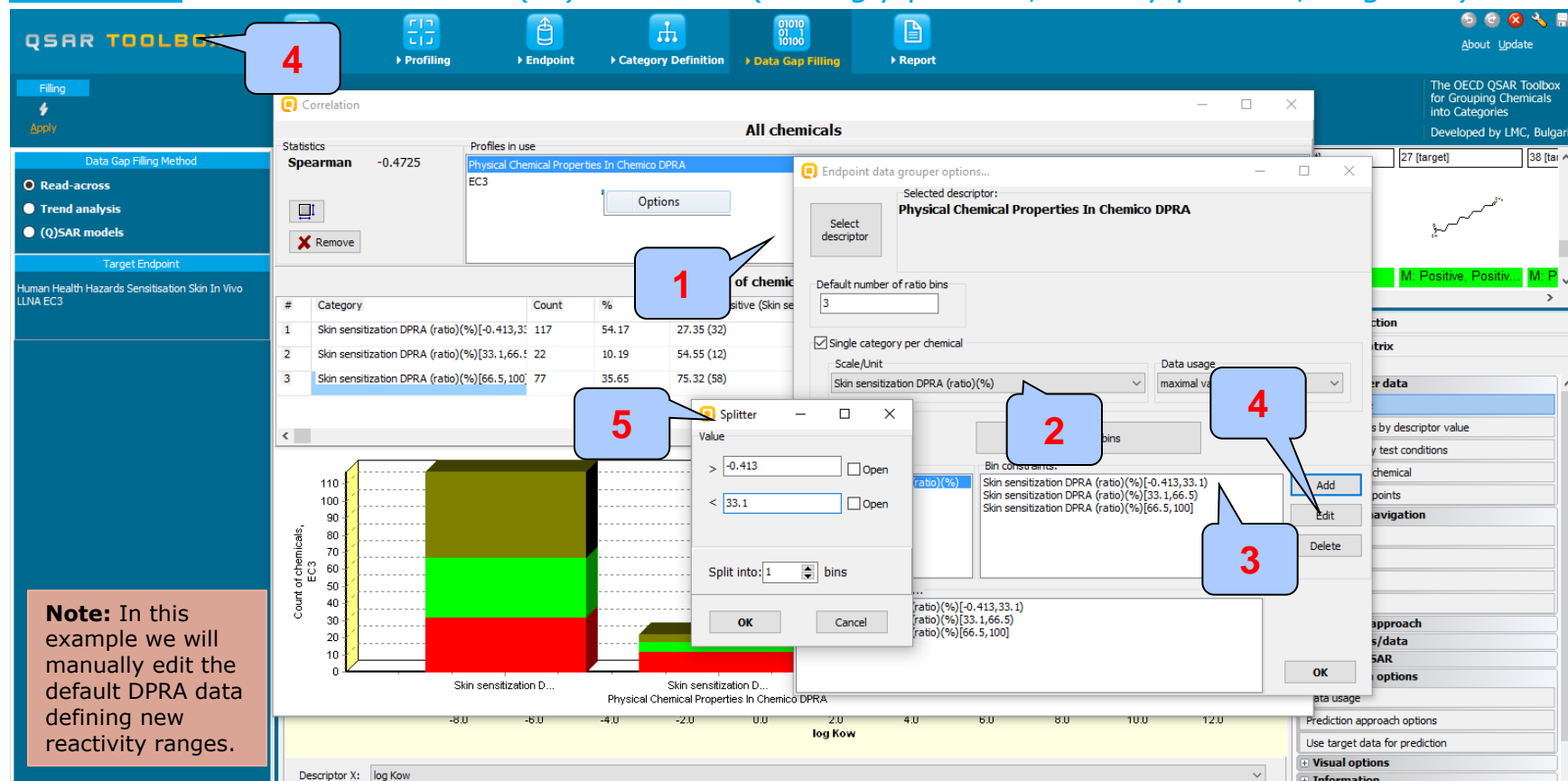
3. Click on the row with N/A

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data



1. Apply **right click** over DPRA descriptor and select "Options"
the first bin [-0.413 – 33.1]

4. **Click** on "Add" button

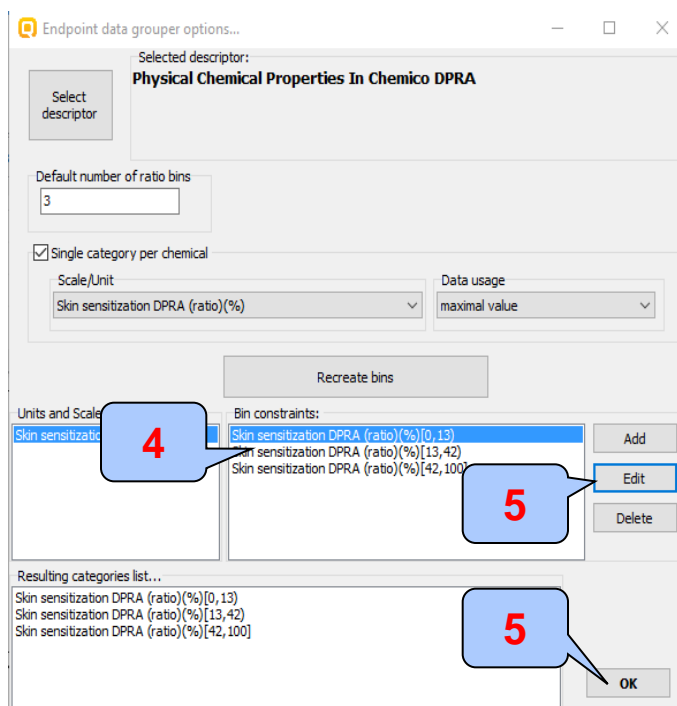
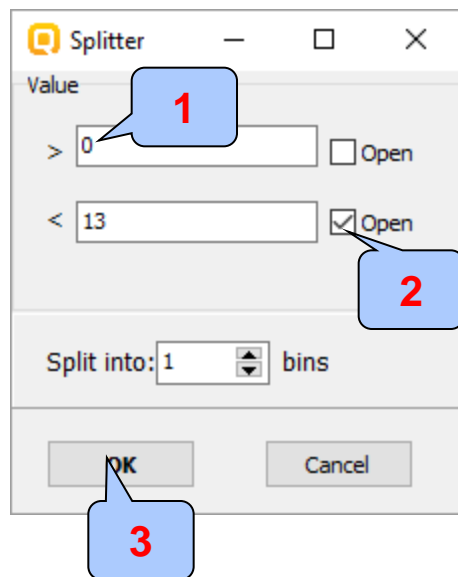
2. **Click** on "Skin sensitization DPRA (ratio)" scale
3. **Select**
5. Additional window appears

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between DPRA and LLNA data – step 5

Example: Correlation of DPRA (%) and LLNA (Strongly positive, Weakly positive, Negative) data



The following ranges have been configured:

- 0 – 13 %
- 13 – 42 %
- 42 – 100 %

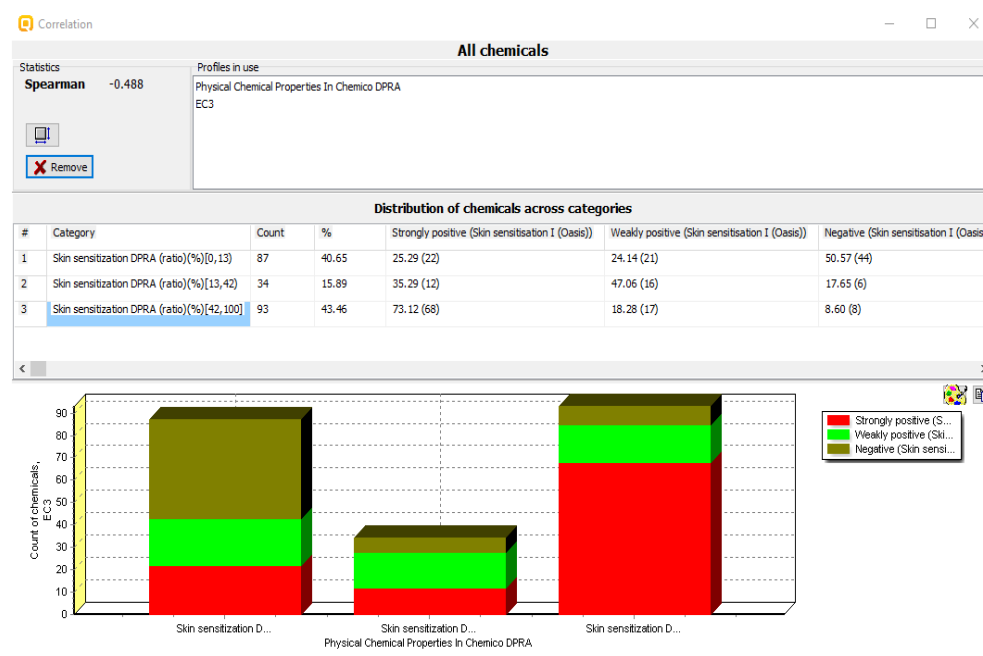
1. **Erase** the default lower value "-0.413" of the first range and type "0". The range is closed, that's why do not check the "open" box.
 2. **Set** "13" value for the upper value of the first range and **check** "open" box to set the range as open.
 3. **Click** "OK" button
 4. **Select** second bin
 5. **Click** "Edit" button and enter the lower and upper values of the second range (13 – 42%).
 6. **Click** "OK"
- Note** that the lower and upper values of the second range are opened. The lower value of the third range is open and the upper value is closed.

Types endpoint correlations

Categorized continuous vs. categorical

Interpretation of correlation results (DPRA vs. LLNA)

- In this example we have correlate continues DPRA (%) data distributed into 3 bins (0-13; 13-42; 42 - 100%) and categorical LLNA data (Strongly positive, Weakly positive, Negative)
- The high absolute value of Spearman coefficient (0.49) shows a good monotonic tendency in the data *.



*The absolute value of the Spearman coefficient shows how monotonic is the data, while the sign of the coefficient specifies the direction of the slope - positive or negative.

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
 - Load ToxCast database
 - ToxCast database – overview
 - Correlation of data – background
 - **Types endpoint correlations**
 - Continuous vs. continuous
 - Categorical vs. categorical
 - Categorized continuous vs. categorical
 - Categorized continuous vs. categorized continuous

Types endpoint correlations

Categorized continuous vs. categorized continuous

- The aim of this type correlation is to illustrate how two different categorized continuous endpoints correlates each other.
- Categorized continuous data is the continuous type data (e.g LC50 or AC50, EC3, %) converted in categories.
- In this example we will illustrated how AC50 ratio data (mol/L) correlates with Relative ERBA (%) data:
 - AC50 (mol/L) associated with assay "NCGC Reporter Gene Assay ER α Agonist" converted in 3 categories
 - Relative ERBA (ratio data expressed in %) converted in 5 categories
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
 - *Load ToxCast database (step 1)*
 - *Gather experimental data (step 2)*
 - *Define target endpoint (step 3)*
 - *Enter Gap filling (step 4)*
 - *Perform correlation between endpoints (step 5).*

Types endpoint correlations

Categorized continuous vs. categorized continuous

Load ToxCast database – step 1

1. Go to "Input"

2. Click "DB" button

3. Select "ToxCast" DB

4. Click "OK"

5. The chemicals from database have been loaded on datamatrix

Types endpoint correlations

Categorized continuous vs. categorized continuous
Gather experimental data – step 2

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

The screenshot shows the QSAR TOOLBOX software interface. The 'Endpoint' tab is selected. The left sidebar shows a tree view of databases, with 'Estrogen Receptor Binding Affinity OASIS' and 'ToxCast' selected. The central panel shows a list of endpoints, with 'AC50' and 'Relative ERBA' selected. The main data matrix table displays the gathered data for these endpoints across various test organisms.

Endpoint	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]
ACEA	(600/660)	M: 21.2 mg/L	M: 0.0039 mg/L	M: 8.08 mg/L, 0.00...	M: 0.000504 mg/L		
Apredica	(425/2642)			M: 23.4 mg/L, 26.9...	M: 8.76 mg/L, 29.1...	M: 0.0962 mg/L, 0...	
Attagene	(1374/6568)	M: 12.5 mg/L, 12.6...		M: 12.8 mg/L, 4.96...	M: 0.00268 mg/L, ...	M: 0.689 mg/L, 1.3...	M: 3.87 mg/L, 3.42...
BioSeek	(971/21906)			M: 4.62 mg/L, 4.96...	M: 6.74 mg/L, 6.25...	M: 0.338 mg/L, 0.3...	M: 0.288 mg/L
NCGC	(1475/6890)	M: 0.00436 mg/L, ...	M: 0.000106 mg/L, ...	M: 12.4 mg/L, 9.25...	M: 0.000262 mg/L, ...	M: 0.66 mg/L, 0.05...	M: 0.219 mg/L, 1.0...
Novascreen	(975/8054)	M: 3.59 mg/L, 0.02...		M: 0.00646 mg/L, ...	M: 0.236 mg/L, 0.1...	M: 2.67 mg/L	
Odyssey Thera	(969/2794)	M: 19.8 mg/L, 4.56...	M: 2.46 mg/L	M: 5.79 mg/L	M: 0.00676 mg/L, ...	M: 3.52 mg/L, 2.19...	
Undefined Assay Provider	(2/2)						
Androgen Binding Affinity							
Estrogen Receptor Binding							
Relative ERBA							
Human	(179/179)			M: 88.9 %	M: 0.0518 %		
Lamb	(1/2)						
Noassay	(13/13)			M: 219 %			
Rat	(110/111)			M: 398 %			
Trout	(65/65)			M: 179 %			
Undefined Test organisms (spec...							
Toxicokinetics, Metabolism and Dist...							

1. Go to "Endpoint"
2. Click on "Unselect all" button
3. Select "ToxCast"
4. Select "Estrogen Receptor Binding Affinity OASIS" DB
5. Click "Gather"
6. The data appears on datamatrix

Types endpoint correlations

Categorized continuous vs. categorized continuous

Define target endpoint – step 3

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

The screenshot shows the QSAR Toolbox software interface. The 'Databases' panel on the left lists various endpoints, with 'Estrogen Receptor 1' selected under 'Human Health Hazards'. The main panel displays a table of chemical structures and their associated data for various endpoints. A red box highlights the 'Estrogen Receptor 1' endpoint, and a blue callout with the number '1' points to the selected cell.

Structure	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]	8 [target]
Structure 1								
Structure 2								
Structure 3								
Structure 4								
Structure 5								
Structure 6								
Structure 7								
Structure 8								
Structure 9								
Structure 10								
Structure 11								
Structure 12								
Structure 13								
Structure 14								
Structure 15								
Structure 16								
Structure 17								
Structure 18								
Structure 19								
Structure 20								
Structure 21								
Structure 22								
Structure 23								
Structure 24								
Structure 25								
Structure 26								
Structure 27								
Structure 28								
Structure 29								
Structure 30								
Structure 31								
Structure 32								
Structure 33								
Structure 34								
Structure 35								
Structure 36								
Structure 37								
Structure 38								
Structure 39								
Structure 40								
Structure 41								
Structure 42								
Structure 43								
Structure 44								
Structure 45								
Structure 46								
Structure 47								
Structure 48								
Structure 49								
Structure 50								
Structure 51								
Structure 52								
Structure 53								
Structure 54								
Structure 55								
Structure 56								
Structure 57								
Structure 58								
Structure 59								
Structure 60								
Structure 61								
Structure 62								
Structure 63								
Structure 64								
Structure 65								
Structure 66								
Structure 67								
Structure 68								
Structure 69								
Structure 70								
Structure 71								
Structure 72								
Structure 73								
Structure 74								
Structure 75								
Structure 76								
Structure 77								
Structure 78								
Structure 79								
Structure 80								
Structure 81								
Structure 82								
Structure 83								
Structure 84								
Structure 85								
Structure 86								
Structure 87								
Structure 88								
Structure 89								
Structure 90								
Structure 91								
Structure 92								
Structure 93								
Structure 94								
Structure 95								
Structure 96								
Structure 97								
Structure 98								
Structure 99								
Structure 100								

The target endpoint in our case is "Estrogen Receptor 1"

1. **Click** on the cell associated with target endpoint and target chemical

Types endpoint correlations

Categorized continuous vs. categorized continuous

Enter Gap filling – step 4

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

The screenshot shows the QSAR Toolbox software interface during the 'Data Gap Filling' step. The top menu bar includes 'Input', 'Profiling', 'Endpoint', 'Category Definition', 'Data Gap Filling', and 'Report'. The left sidebar shows 'Read-across', 'Trend analysis', and '(Q)SAR models'. The main area displays a 'Filter endpoint tree...' on the left and a table of endpoints on the right. The 'Filter endpoint tree...' shows a hierarchy of endpoints, with 'ToxCast' and 'ACEA' selected. The table shows data for various endpoints, including 'ACEA', 'Apredica', 'Attagene', 'BioSeek', 'NCGC', 'NCGC Reporter Gene Assay ERA Agonist', 'Homo sapiens', 'Estrogen Receptor 1', 'Tox21_AhR', 'Tox21_AhR_viability', 'Tox21_AR_BLA_Agonist_ch1', 'Tox21_AR_BLA_Agonist_ch2', and 'Tox21_AR_BLA_Agonist_ratio'. The 'ACEA' row is highlighted in blue.

Endpoint	1 [target]	2 [target]	3 [target]	4 [target]	5 [target]	6 [target]	7 [target]
ACEA (600/660)	M: 21.2 mg/L	M: 0.0039 mg/L		M: 8.08 mg/L, 0.00...		M: 0.000504 mg/L	
Apredica (425/2642)			M: 47.2 mg/L, 22.3...	M: 23.4 mg/L, 26.9...	M: 8.76 mg/L, 29.1...	M: 0.0962 mg/L, 0...	
Attagene (1374/6568)	M: 12.5 mg/L, 12.6...		M: 12.8 mg/L, 4.96...	M: 0.00268 mg/L, ...	M: 0.689 mg/L, 1.3...		M: 3.87 m
BioSeek (971/21906)			M: 4.62 mg/L, 4.96...	M: 6.74 mg/L, 6.25...	M: 0.338 mg/L, 0.3...	M: 0.288 mg/L	
NCGC							
NCGC Reporter Gene Assay ERA Agonist							
Homo sapiens							
Estrogen Receptor 1 (374/505)				M: 0.000224 mg/L, ...	M: 9.38 mg/L, 4.84...		
NCGC Reporter Gene Assay ERA ... (487/559)	M: 19.1 mg/L	M: 0.000531 mg/L					M: 5.84E-4
Tox21_AhR (237/237)					M: 0.66 mg/L		
Tox21_AhR_viability (319/319)		M: 0.000106 mg/L	M: 12.4 mg/L		M: 0.0557 mg/L		M: 5.84E-4
Tox21_AR_BLA_Agonist_ch1 (439/439)	M: 0.00436 mg/L	M: 0.000991 mg/L					
Tox21_AR_BLA_Agonist_ch2 (67/67)					M: 10.6 mg/L	M: 0.219 mg/L	
Tox21_AR_BLA_Agonist_ratio (89/89)							

1. Go to "Data Gap filling" 2. Select "Trend analysis" 3. Click "Apply"

Types endpoint correlations

Categorized continuous vs. categorized continuous
Perform correlation between AC50 and Relative ERBA – step 5

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

1. Open "Subcategorize"
2. Click "Endpoint data"
3. Click on "Adjust options" button
4. Click "Select descriptor" button
5. Click on the row related to "Human" data. Human data has been selected in order to have consistency between two endpoints (AC50, Estrogen receptor 1 is associated with Homo sapience species)
6. Click on "Select descriptor" button

Types endpoint correlations

Categorized continuous vs. categorized continuous
Perform correlation between AC50 and Relative ERBA – step 5

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

The screenshot displays the 'Endpoint data grouper options...' dialog box in the QSAR Toolbox. The 'Selected descriptor' is 'Human Health Hazards Toxicity to Reproduction Relative ERBA Human'. The 'Default number' is set to 5. The 'Differ from target by' section has 'All categories' selected. The 'Correlation' method is chosen. The 'Scale/Unit' is set to '%'. The 'Data usage' is set to 'maximal value'. The 'Recreate bins' button is highlighted. The background shows a scatter plot of data points and a list of chemical structures.

Further workflow illustrates manually editing of first 3 bins. Follow the steps:

1. **Enter** 5 bins
2. **Check** "Single category per chemical"
3. **Select** "%"
4. **Select** "maximal value"
5. **Click** "Recreate bins".

Edit the automatically generated ranges of "Estrogen receptor binding" activity into following 5 ranges: 0 – 0.1; 0.1 – 1; 1– 10; 10 – 100; > 100 %. The procedure of manual editing of bins is illustrated on next slide.

Types endpoint correlations

Categorized continuous vs. categorized continuous
Perform correlation between AC50 and Relative ERBA – step 5

Example: Correlation of AC 50 (mol/L) and Relative ERBA (%) data

The following ranges have been configured:

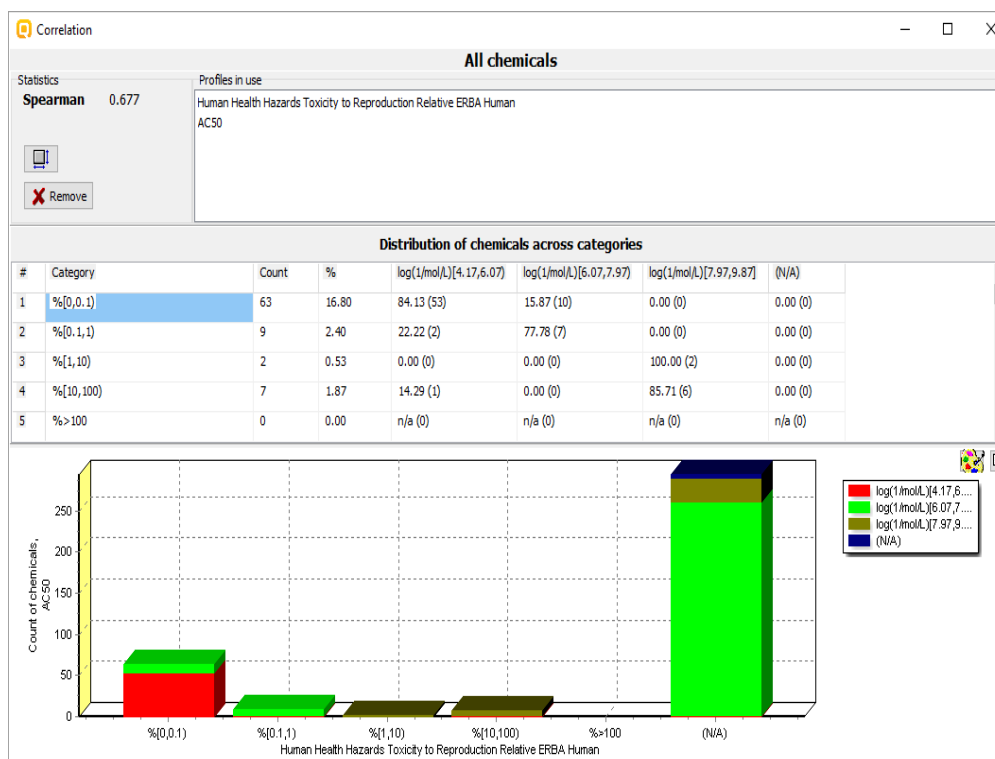
- 0 – 0.1 %
- 0.1 – 1 %
- 1 – 10 %
- 10 – 100%
- >100%

1. **Erase** the default upper value "13" of the range and type "0.1". The range is closed, that's why do not check the "open" box.
 2. **Click** "OK" button
 3. **Select** second bin
 4. **Click** "Edit" button and enter the lower and upper values of the second range (0.1 – 1%).
 5. **Click** "OK"
Note that the lower values of each range is opened. The upper values of each range is closed.

Types endpoint correlations

Categorized continuous vs. categorized continuous *Interpretation of correlation results (AC50 vs. Relative ERBA)*

- In this example we have correlate AC50 (mol/L) categorized continuous data distributed automatically in 3 default bins (categories) and another categorized continuous Relative ERBA data distributed manually into 5 bins (0-0.1; 0.1-1; 1 – 10; 10 – 100; >100 %)
- The high value of Spearman coefficient (0.68) shows a good monotonic tendency in the data*. The correlation is assumed as strong based on Spearman coefficient interpretation



*The absolute value of the Spearman coefficient shows how monotonic is the data, while the sign of the coefficient specifies the direction of the slope - positive or negative.

Summary

- Different type correlations have been illustrated in this tutorial based on type of endpoint data:
 - Continuous vs. continuous
 - Categorical vs. categorical:
 - Categorical vs. categorical
 - Categorized continuous vs. categorical
 - Categorized continuous vs. categorized continuous
- Correlation analysis has been evaluated by Spearman coefficient using a newly implemented functionality
- High endpoint correlations have been obtained for 3 out of 4 illustrated examples.