QSAR Toolbox functionalities.

Clustering

Laboratory of Mathematical Chemistry
University “Prof. As. Zlatarov” Bourgas, Bulgaria

June 2019

Clustering - overview

- Cluster analysis (clustering) is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [1].

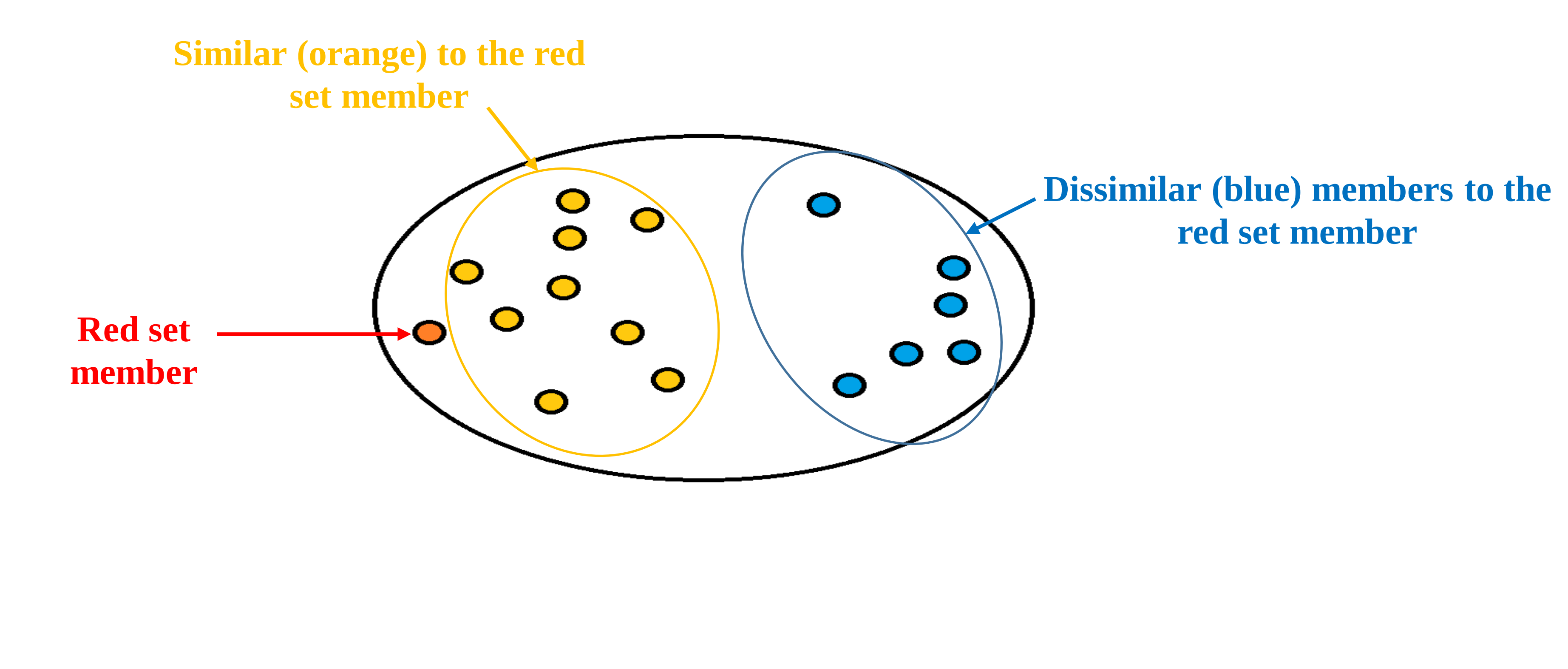
- Cluster analysis is based on similarities between each two chemicals of an initial set. Similarity values are compared to a selected threshold in a repeated procedure.

- Resulting clusters include only substances which are similar to each other above the required threshold.

[1] - http://en.wikipedia.org/wiki/Cluster_analysis

Clustering – how the clusters are formed?

1) The system prepares a matrix which contains the similarities between each two items of the initial set. Another matrix is also prepared that contains only flags if each two items are considered "similar" or "dissimilar". The latter values are calculated by comparing the similarity values from the first matrix to a given threshold.



Similar (orange) to the red set member

Dissimilar (blue) members to the red set member

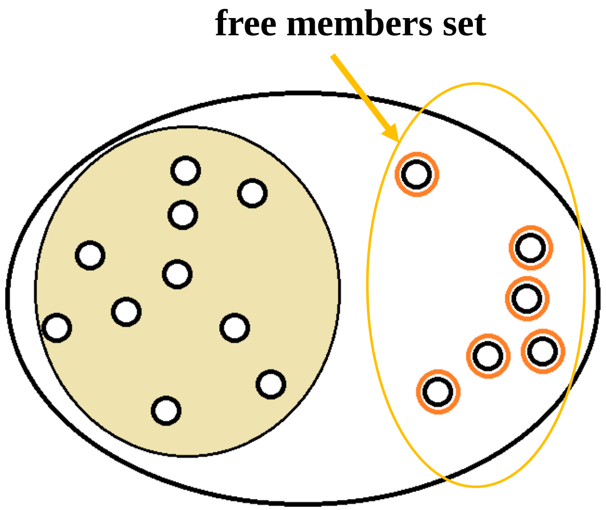
Red set member

Clustering – how the clusters are formed?

2) After creating a cluster, if there are free members set (not yet included in the existing clusters) still remaining, the process is repeated again.

Initially each free member* breeds a new cluster. Then, every cluster grows as much as possible, including only members that are similar to the ones it already included.

Depending on the option “***Allow overlapping***” the cluster may or may not include members that are present in already existing clusters.



*Always checked is the nearest free available member. The distance between a cluster and a free member is measured through the distances between this free member and the members already included in the new cluster. Different options are available here – ***taking the minimal***, ***taking the maximal***, ***taking the average***, ***taking the median***.

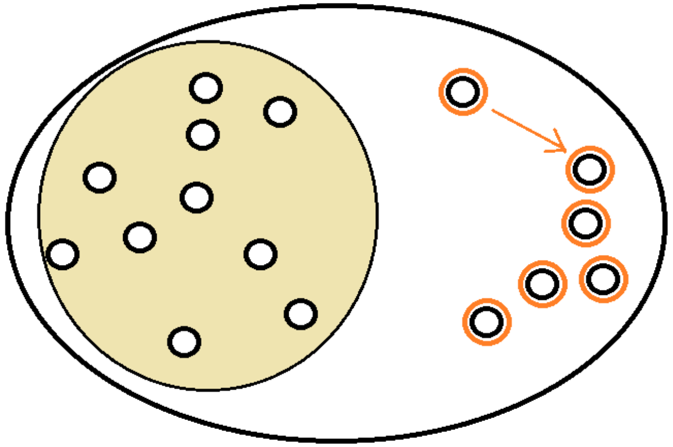
Clustering – how the clusters are formed?

2) After creating a cluster, if there are free members set (not yet included in the existing clusters) still remaining, the process is repeated again.

Initially each free member* breeds a new cluster. Then, every cluster grows as much as possible, including only members that are similar to the ones it already included.

Depending on the option “***Allow overlapping***” the cluster may or may not include members that are present in already existing clusters.

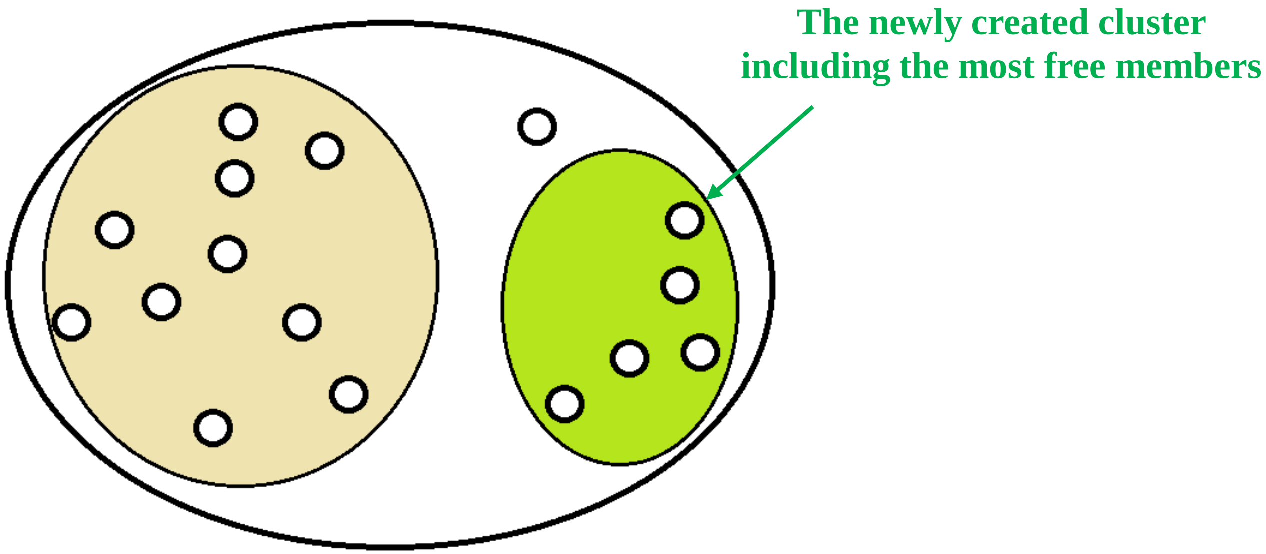
Every free member breeds a new cluster; the cluster grows to the nearest member



*Always checked is the nearest free available member. The distance between a cluster and a free member is measured through the distances between this free member and the members already included in the new cluster. Different options are available here – ***taking the minimal***, ***taking the maximal***, ***taking the average***, ***taking the median***.

Clustering – how the clusters are formed?

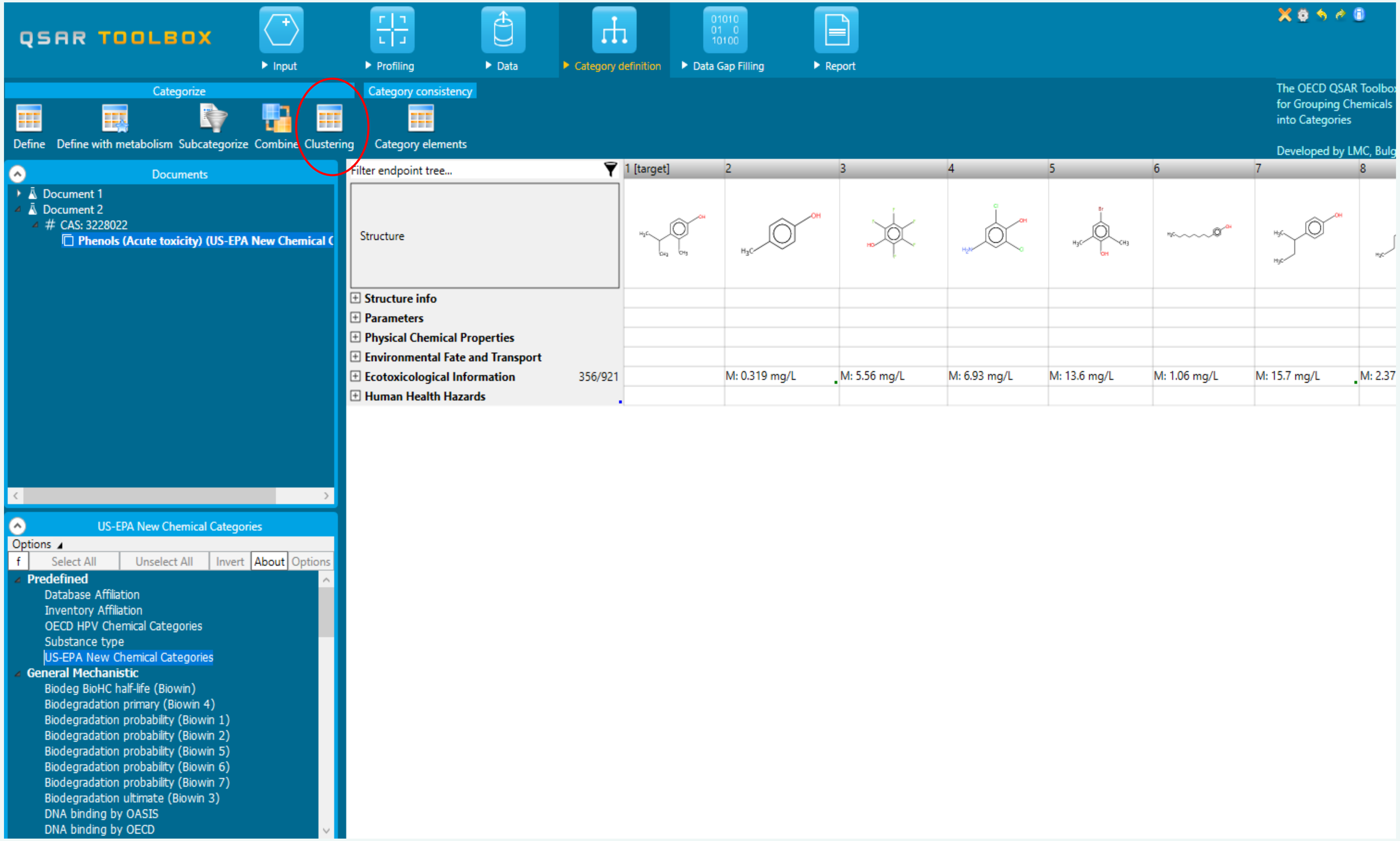
3) At the end, selected is the cluster which had included most free members



The newly created cluster
including the most free members

Clustering – implementation in QSAR Toolbox

The clustering functionality in Toolbox is available within the Category definition module.



Documents
- Document 1
- Document 2
  - CAS: 3228022
    - Phenols (Acute toxicity) (US-EPA New Chemical C

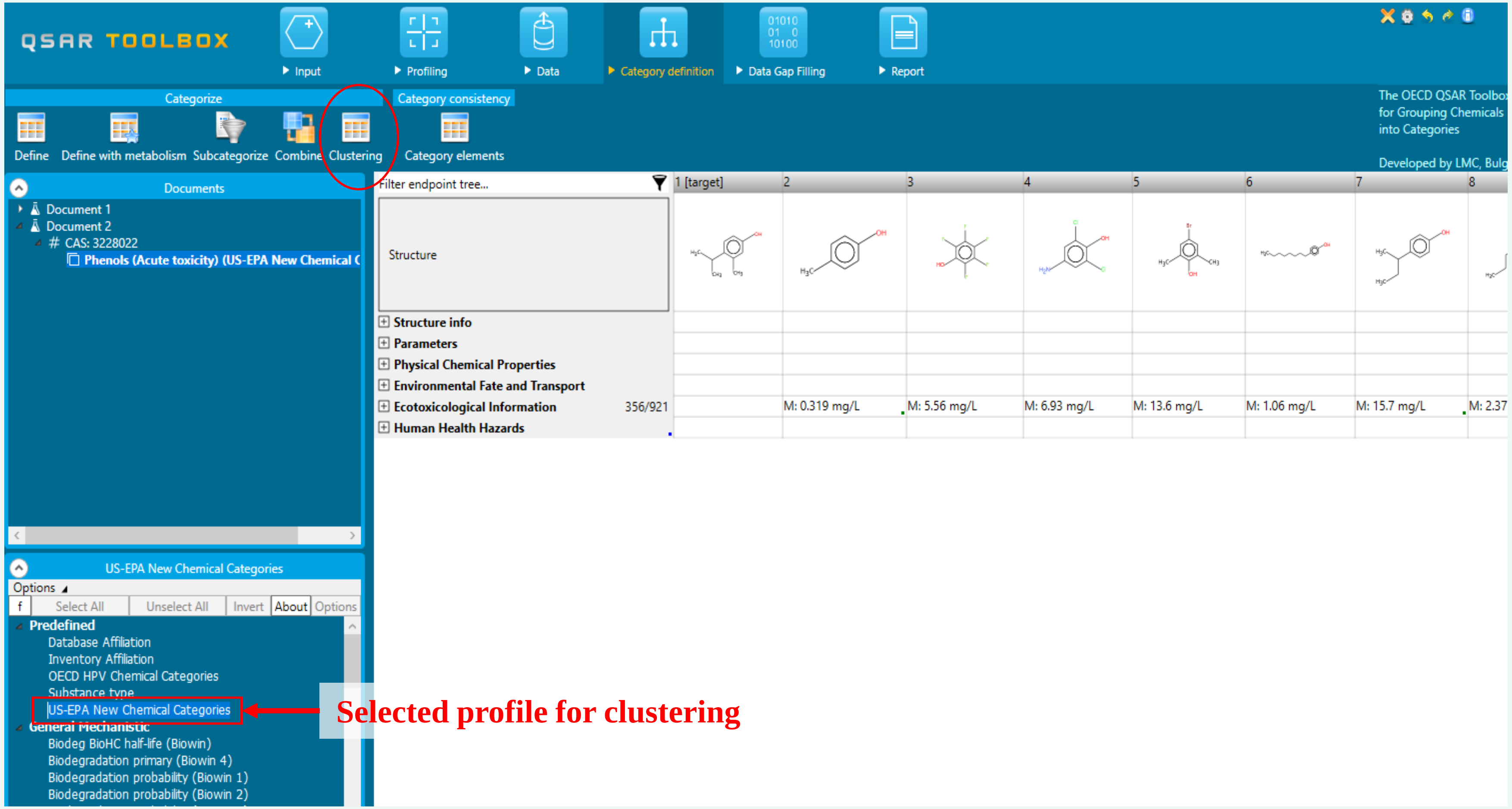
Predefined
- Database Affiliation
- Inventory Affiliation
- OECD HPV Chemical Categories
- Substance type
- US-EPA New Chemical Categories

General Mechanistic
- Biodeg BioHC half-life (Biowin)
- Biodegradation primary (Biowin 4)
- Biodegradation probability (Biowin 1)
- Biodegradation probability (Biowin 2)
- Biodegradation probability (Biowin 5)
- Biodegradation probability (Biowin 6)
- Biodegradation probability (Biowin 7)
- Biodegradation ultimate (Biowin 3)
- DNA binding by OASIS
- DNA binding by OECD

		1 [target]	2	3	4	5	6	7	8
Structure									
Structure info									
Parameters									
Physical Chemical Properties									
Environmental Fate and Transport									
Ecotoxicological Information	356/921		M: 0.319 mg/L	M: 5.56 mg/L	M: 6.93 mg/L	M: 13.6 mg/L	M: 1.06 mg/L	M: 15.7 mg/L	M: 2.37
Human Health Hazards									

Clustering – implementation in QSAR Toolbox

Allows distributing defined category into clusters. Cluster is presented as a sub-category of general category that includes chemicals with unique combination of profiling results or similar structural characteristics*.



- Document 1
- Document 2
  - # CAS: 3228022
    - Phenols (Acute toxicity) (US-EPA New Chemical C

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

General Mechanistic
- Biodeg BioHC half-life (Biowin)
- Biodegradation primary (Biowin 4)
- Biodegradation probability (Biowin 1)
- Biodegradation probability (Biowin 2)

Filter endpoint tree...		1 [target]	2	3	4	5	6	7	8
Structure									
⊞ Structure info									
⊞ Parameters									
⊞ Physical Chemical Properties									
⊞ Environmental Fate and Transport									
⊞ Ecotoxicological Information	356/921		M: 0.319 mg/L	M: 5.56 mg/L	M: 6.93 mg/L	M: 13.6 mg/L	M: 1.06 mg/L	M: 15.7 mg/L	M: 2.37
⊞ Human Health Hazards									

Selected profile for clustering

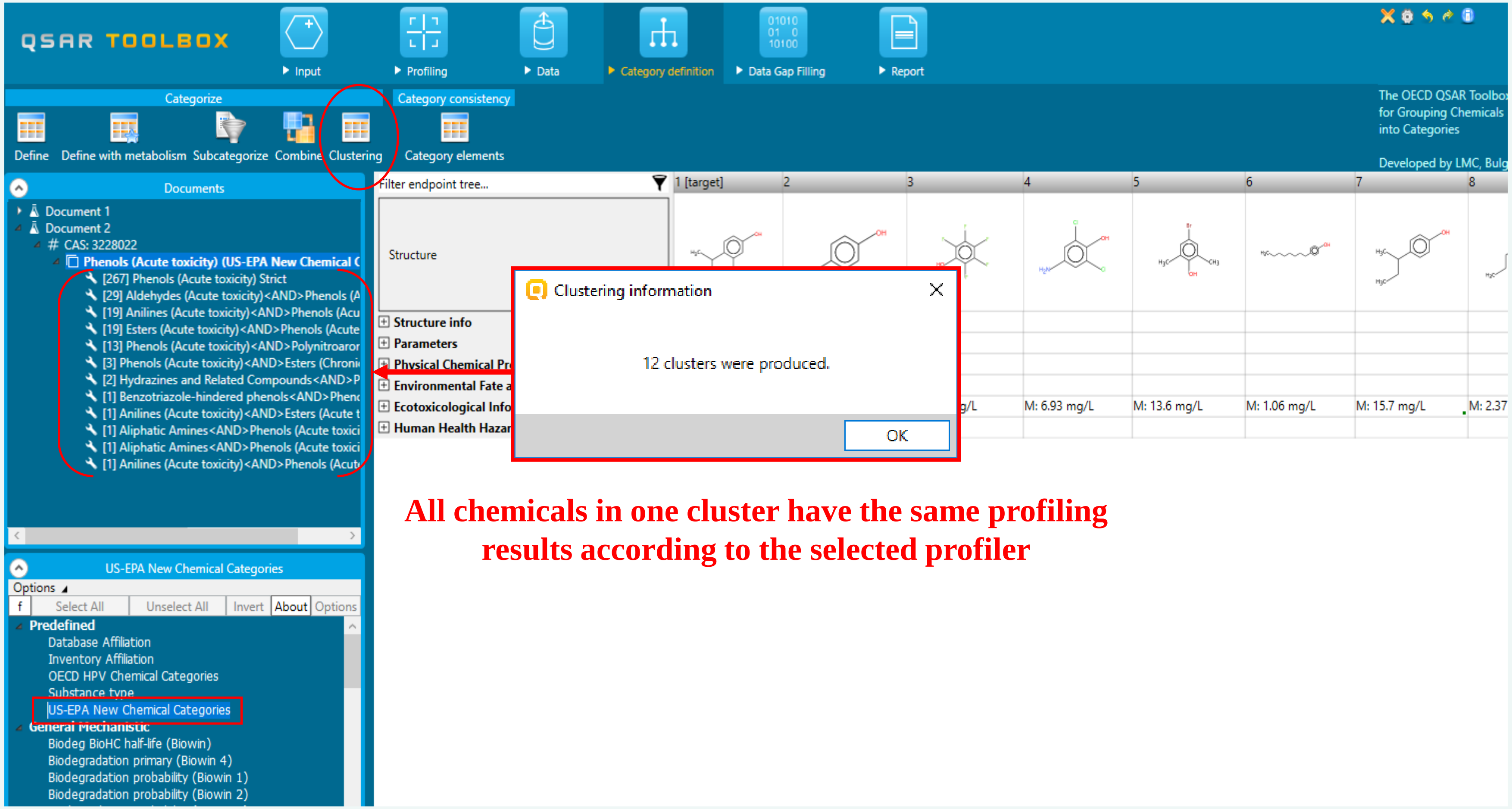
*Note: When *Structure similarity* profiler is selected to cluster a category, the sub-categories depends on the defined threshold and target chemical availability.

1) If there is a defined target chemical, then the category is splitted into clusters based on the structural similarity of the chemicals with respect to the target

2) If there is not a defined target chemical, then the category is splitted into clusters based on the structural similarity between each to each chemical

Clustering – implementation in QSAR Toolbox

Allows distributing defined category into clusters. Cluster is presented as a sub-category of general category that includes chemicals with unique combination of profiling results or similar structural characteristics.



Document 1
Document 2
CAS: 3228022
Phenols (Acute toxicity) (US-EPA New Chemical C
[267] Phenols (Acute toxicity) Strict
[29] Aldehydes (Acute toxicity)<AND>Phenols (A
[19] Anilines (Acute toxicity)<AND>Phenols (Acu
[19] Esters (Acute toxicity)<AND>Phenols (Acute
[13] Phenols (Acute toxicity)<AND>Polynitroaror
[3] Phenols (Acute toxicity)<AND>Esters (Chroni
[2] Hydrazines and Related Compounds<AND>P
[1] Benzotriazole-hindered phenols<AND>Phenc
[1] Anilines (Acute toxicity)<AND>Esters (Acute t
[1] Aliphatic Amines<AND>Phenols (Acute toxici
[1] Aliphatic Amines<AND>Phenols (Acute toxici
[1] Anilines (Acute toxicity)<AND>Phenols (Acute

g/L | M: 6.93 mg/L | M: 13.6 mg/L | M: 1.06 mg/L | M: 15.7 mg/L | M: 2.37

Clustering information

12 clusters were produced.

OK

Predefined
Database Affiliation
Inventory Affiliation
OECD HPV Chemical Categories
Substance type
US-EPA New Chemical Categories
General Mechanistic
Biodeg BioHC half-life (Biowin)
Biodegradation primary (Biowin 4)
Biodegradation probability (Biowin 1)
Biodegradation probability (Biowin 2)

All chemicals in one cluster have the same profiling results according to the selected profiler