

## OECD QSAR Toolbox v.4.1

Tutorial illustrating new options of the  
structure similarity

# Outlook

- **Background**
- Aims
- PubChem features
- The exercise
- Workflow

# Background

- This presentation is designed to familiarize the Toolbox user with the new structure similarity features;
- Structure similarity options in TB 4.0 have been expanded including PubChem substructure features.

# Outlook

- Background
- **Aims**
- PubChem features
- The exercise
- Workflow

## Aims

- To show to the Toolbox user how to compare two chemicals with respect to PubChem substructure similarity features;
- To show to the Toolbox user how to compare a chemical with list of chemicals with respect to PubChem substructure similarity features.

# Outlook

- Background
- Aims
- **PubChem features**
- The exercise
- Workflow

# PubChem features

## Overview

The *PubChem* generates a binary substructure fingerprint for each chemical structure. These fingerprints are used by PubChem for similarity neighboring and similarity searching.

A substructure is a fragment of chemical structure. A fingerprint is an ordered list of binary (1/0) bits. Each bit represent a Boolean determination of specific atom or test features used further for similarity neighboring and similarity searching. Seven groups of PubChem features are defined and used:

- Hierarchical element counts;
- Rings;
- Simple atom pairs;
- Simple atom nearest neighbors;
- Detailed atom neighbors;
- Simple SMARTS patterns;
- Complex SMARTS patterns.

# PubChem features Overview

Numbers in brackets show how many are the common substructure features between the two compared structures out of all features found in a single chemical.

- Green colored features are common for both structures;
- Red colored features are unique.

Each of the *PubChem* features has **bit position** (1) which correspond to a **bit substructure** (2).

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
<p><b>PubChem features (101/104)</b></p> <ul style="list-style-type: none"> <li>#001 - "&gt;= 4 H" (1/1)</li> <li>#010 - "&gt;= 2 C" (1/1)</li> <li>#011 - "&gt;= 4 C" (1/1)</li> <li>#015 - "&gt;= 1 N" (1/1)</li> <li>#019 - "&gt;= 1 O" (1/1)</li> <li>#020 - "&gt;= 2 O" (1/1)</li> <li>#038 - "&gt;= 1 Cl" (1/1)</li> <li>#179 - "&gt;= 1 any ring size 6" (1/1)</li> <li>#180 - "&gt;= 1 saturated or aromatic carbon-only ring size 6" (1/1)</li> <li>#256 - "&gt;= 1 aromatic ring" (1/1)</li> <li>#284 - "C-H" (1/1)</li> <li>#285 - "C-C" (1/1)</li> <li>#286 - "C-N" (1/1)</li> <li>#287 - "C-O" (1/1)</li> <li>#295 - "C-Cl" (1/1)</li> <li>#302 - "N-O" (1/1)</li> <li>#333 - "C(-C)(-C)" (1/1)</li> <li>#334 - "C(-C)(-C)(-C)" (1/1)</li> <li>#341 - "C(-C)(-C)(-N)" (1/1)</li> <li>#343 - "C(-C)(-Cl)" (1/1)</li> <li>#345 - "C(-C)(-H)" (1/1)</li> <li>#352 - "C(-C)(-N)" (1/1)</li> <li>#353 - "C(-C)(-O)" (1/1)</li> <li>#356 - "C(-C)(C)" (1/1)</li> <li>#357 - "C(-C)(C)(C)" (1/1)</li> <li>#371 - "C(-H)(C)" (1/1)</li> <li>#372 - "C(-H)(C)(C)" (1/1)</li> <li>#377 - "C(-N)(C)" (1/1)</li> </ul>	<ul style="list-style-type: none"> <li>#604 - "C-C-C-C" (1/1)</li> <li>#608 - "N-C-C-C" (1/1)</li> <li>#609 - "C-C-C-C" (1/1)</li> <li>#619 - "C-C-C-C" (1/1)</li> <li>#624 - "O=C-C-C" (1/1)</li> <li>#634 - "N-C-C-C" (1/1)</li> <li>#635 - "C-C-C-C" (1/1)</li> <li>#641 - "C=C-C-C" (1/1)</li> <li>#661 - "C-C-C-C" (1/1)</li> <li>#665 - "C=C=C=C" (1/1)</li> <li>#666 - "N-C-C-C" (1/1)</li> <li>#669 - "C-C-C-C" (1/1)</li> <li>#672 - "O=C=C=C" (1/1)</li> <li>#673 - "O=C=C=C" (1/1)</li> <li>#678 - "C-C-C-C" (1/1)</li> <li>#679 - "C-C=C-C" (1/1)</li> <li>#680 - "C-C-C-C" (1/1)</li> <li>#684 - "N-C-C-C" (1/1)</li> <li>#685 - "O=C-C-C" (1/1)</li> <li>#689 - "C-C-C-C" (1/1)</li> <li>#693 - "O=C-C-C" (1/1)</li> <li>#696 - "O=C-C-C-N" (0/1)</li> <li>#705 - "O=C-C-C-C" (1/1)</li> <li>#708 - "O=C-C-C(N)-C" (0/1)</li> <li>#709 - "C-C-C-C" (1/1)</li> <li>#710 - "C-C-C-C" (1/1)</li> <li>#711 - "C-C-C-C" (1/1)</li> <li>#717 - "C1c(N)cc1" (0/1)</li> <li>#780 - "C1CC(N)CC1" (0/1)</li> </ul>



# PubChem features

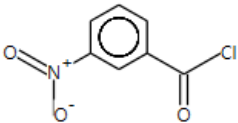
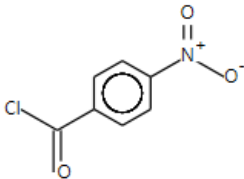
## Hierarchical element counts

- **Hierarchical element counts** - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.

They include bit positions from 001 to 115.

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
	
<p>PubChem features (101/104)</p> <ul style="list-style-type: none"> <li>#001 - "&gt;= 4 H" (1/1)</li> <li>#010 - "&gt;= 2 C" (1/1)</li> <li>#011 - "&gt;= 4 C" (1/1)</li> <li>#015 - "&gt;= 1 N" (1/1)</li> <li>#019 - "&gt;= 1 O" (1/1)</li> <li>#020 - "&gt;= 2 O" (1/1)</li> <li>#038 - "&gt;= 1 Cl" (1/1)</li> <li>#179 - "&gt;= 1 any ring size 6" (1/1)</li> <li>#180 - "&gt;= 1 saturated or aromatic" (1/1)</li> <li>#256 - "&gt;= 1 aromatic ring" (1/1)</li> <li>#284 - "C-H" (1/1)</li> <li>#285 - "C-C" (1/1)</li> <li>#286 - "C-N" (1/1)</li> </ul>	<p>PubChem features (101/105)</p> <ul style="list-style-type: none"> <li>#001 - "&gt;= 4 H" (1/1)</li> <li>#010 - "&gt;= 2 C" (1/1)</li> <li>#011 - "&gt;= 4 C" (1/1)</li> <li>#015 - "&gt;= 1 N" (1/1)</li> <li>#019 - "&gt;= 1 O" (1/1)</li> <li>#020 - "&gt;= 2 O" (1/1)</li> <li>#038 - "&gt;= 1 Cl" (1/1)</li> <li>#179 - "&gt;= 1 any ring size 6" (1/1)</li> <li>#180 - "&gt;= 1 saturated or aromatic" (1/1)</li> <li>#256 - "&gt;= 1 aromatic ring" (1/1)</li> <li>#284 - "C-H" (1/1)</li> <li>#285 - "C-C" (1/1)</li> <li>#286 - "C-N" (1/1)</li> </ul>

# PubChem features

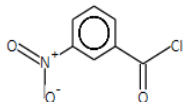
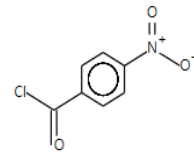
## Rings in a canonic ESSSR ring set

- Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set** - These bits test for the presence or count of the described chemical system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.

They include bit positions from 116 to 263.

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
	
<b>PubChem features (101/104)</b> #001 - ">= 4 H" (1/1) #010 - ">= 2 C" (1/1) #011 - ">= 4 C" (1/1) #015 - ">= 1 N" (1/1) #019 - ">= 1 O" (1/1) #020 - ">= 2 O" (1/1) #038 - ">= 1 Cl" (1/1) <b>#179 - "&gt;= 1 any ring size 6" (1/1)</b> <b>#180 - "&gt;= 1 saturated or aromatic carbon-only ring size 6" (1/1)</b> <b>#256 - "&gt;= 1 aromatic ring" (1/1)</b> #284 - "C-Cl" (1/1) #285 - "C-C" (1/1) #286 - "C-N" (1/1)	<b>PubChem features (101/105)</b> #001 - ">= 4 H" (1/1) #010 - ">= 2 C" (1/1) #011 - ">= 4 C" (1/1) #015 - ">= 1 N" (1/1) #019 - ">= 1 O" (1/1) #020 - ">= 2 O" (1/1) #038 - ">= 1 Cl" (1/1) <b>#179 - "&gt;= 1 any ring size 6" (1/1)</b> <b>#180 - "&gt;= 1 saturated or aromatic carbon-only ring size 6" (1/1)</b> <b>#256 - "&gt;= 1 aromatic ring" (1/1)</b> #284 - "C-Cl" (1/1) #285 - "C-C" (1/1) #286 - "C-N" (1/1)

# PubChem features

## Simple atom pairs

- **Simple atom pairs** – These bits test for the presence of patterns of bonded atom pairs, regardless of bond order or count.

They include bit positions from 264 to 327.

Target chemical	Query chemical
<p>Structure similarity: 96.65%</p> <p>PubChem features (101/104)</p> <ul style="list-style-type: none"> <li>#001 - "&gt;= 4 H" (1/1)</li> <li>#010 - "&gt;= 2 C" (1/1)</li> <li>#011 - "&gt;= 4 C" (1/1)</li> <li>#015 - "&gt;= 1 N" (1/1)</li> <li>#019 - "&gt;= 1 O" (1/1)</li> <li>#020 - "&gt;= 2 O" (1/1)</li> <li>#038 - "&gt;= 1 Cl" (1/1)</li> <li>#179 - "&gt;= 1 any ring size 6" (1/1)</li> <li>#180 - "&gt;= 1 saturated or aromatic carbon-only r</li> <li>#256 - "&gt;= 1 aromatic ring" (1/1)</li> <li>#284 - "C-H" (1/1)</li> <li>#285 - "C-C" (1/1)</li> <li>#286 - "C-N" (1/1)</li> <li>#287 - "C-O" (1/1)</li> <li>#295 - "C-Cl" (1/1)</li> <li>#302 - "N-O" (1/1)</li> <li>#333 - "C(-C)(-C)" (1/1)</li> <li>#334 - "C(-C)(-C)(-C)" (1/1)</li> </ul>	<p>PubChem features (101/105)</p> <ul style="list-style-type: none"> <li>#001 - "&gt;= 4 H" (1/1)</li> <li>#010 - "&gt;= 2 C" (1/1)</li> <li>#011 - "&gt;= 4 C" (1/1)</li> <li>#015 - "&gt;= 1 N" (1/1)</li> <li>#019 - "&gt;= 1 O" (1/1)</li> <li>#020 - "&gt;= 2 O" (1/1)</li> <li>#038 - "&gt;= 1 Cl" (1/1)</li> <li>#179 - "&gt;= 1 any ring size 6" (1/1)</li> <li>#180 - "&gt;= 1 saturated or aromatic carbon-only r</li> <li>#256 - "&gt;= 1 aromatic ring" (1/1)</li> <li>#284 - "C-H" (1/1)</li> <li>#285 - "C-C" (1/1)</li> <li>#286 - "C-N" (1/1)</li> <li>#287 - "C-O" (1/1)</li> <li>#295 - "C-Cl" (1/1)</li> <li>#302 - "N-O" (1/1)</li> <li>#333 - "C(-C)(-C)" (1/1)</li> <li>#334 - "C(-C)(-C)(-C)" (1/1)</li> </ul>

# PubChem features

## Simple atom nearest neighbors

- **Simple atom nearest neighbors** – These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") (1) or count, but where bond aromaticity (denoted by ":") (2) is significant.

They include bit positions from 328 to 416.

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
<ul style="list-style-type: none"> <li>#302 - "N-O" (1/1)</li> <li>#333 - "C(-C)(-C)" (1/1)</li> <li>#334 - "C(-C)(-C)(-C)" (1/1)</li> <li>#341 - "C(-C)(-C)(-N)" (1/1)</li> <li>#343 - "C(-C)(-Cl)" (1/1)</li> <li>#345 - "C(-C)(-H)" (1/1)</li> <li>#352 - "C(-C)(-N)" (1/1)</li> <li>#353 - "C(-C)(-O)" (1/1)</li> <li>#356 - "C(-C)(C)" (1/1)</li> <li>#357 - "C(-C)(C)(C)" (1/1)</li> <li>#371 - "C(-H)(C)" (1/1)</li> <li>#372 - "C(-H)(C)(C)" (1/1)</li> <li>#377 - "C(-N)(C)" (1/1)</li> <li>#378 - "C(-N)(C)(C)" (1/1)</li> <li>#385 - "C(C)(C)" (1/1)</li> <li>#390 - "N(-C)(-O)" (1/1)</li> <li>#402 - "N(-O)(-O)" (1/1)</li> <li>#417 - "C=C" (1/1)</li> <li>#421 - "C=O" (1/1)</li> </ul>	<ul style="list-style-type: none"> <li>#302 - "N-O" (1/1)</li> <li>#333 - "C(-C)(-C)" (1/1)</li> <li>#334 - "C(-C)(-C)(-C)" (1/1)</li> <li>#341 - "C(-C)(-C)(-N)" (1/1)</li> <li>#343 - "C(-C)(-Cl)" (1/1)</li> <li>#345 - "C(-C)(-H)" (1/1)</li> <li>#352 - "C(-C)(-N)" (1/1)</li> <li>#353 - "C(-C)(-O)" (1/1)</li> <li>#356 - "C(-C)(C)" (1/1)</li> <li>#357 - "C(-C)(C)(C)" (1/1)</li> <li>#371 - "C(-H)(C)" (1/1)</li> <li>#372 - "C(-H)(C)(C)" (1/1)</li> <li>#377 - "C(-N)(C)" (1/1)</li> <li>#378 - "C(-N)(C)(C)" (1/1)</li> <li>#385 - "C(C)(C)" (1/1)</li> <li>#396 - "N(-C)(-O)" (1/1)</li> <li>#402 - "N(-O)(-O)" (1/1)</li> <li>#417 - "C=C" (1/1)</li> <li>#421 - "C=O" (1/1)</li> </ul>

Callout 1 points to: #378 - "C(-N)(C)(C)" (1/1)

Callout 2 points to: #402 - "N(-O)(-O)" (1/1)

# PubChem features

## Detailed atom neighborhoods

- Detailed atom neighborhoods** – These bits test for the presence of detailed atom neighborhood patterns, regardless of count, but where bond orders are specific, bond aromaticity matches both single and double bonds, and where "-", "=", and "#" matches a single bond, double bond, and triple bond order, respectively.

They include bit positions from 417 to 460.

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
#402 - "N(-O)(-O)" (1/1) #417 - "C=C" (1/1) #421 - "C=O" (1/1) #424 - "N=O" (1/1) #431 - "C(-C)(-C)(=O)" (1/1) <b>#434 - "C(-C)(-Cl)(=O)" (1/1)</b> #435 - "C(-C)(-H)(=O)" (1/1) #438 - "C(-C)(-N)(=O)" (1/1) #442 - "C(-C)(=O)" (1/1) #444 - "C(-C)(=O)" (1/1) #445 - "C(-Cl)(=O)" (1/1) #447 - "C(-H)(=O)" (1/1) #450 - "C(-N)(=O)" (1/1) #455 - "N(-C)(=O)" (1/1) #456 - "N(-O)(=O)" (1/1) #465 - "N-C-C-C" (1/1) #471 - "C-C-C=C" (1/1)	#402 - "N(-O)(-O)" (1/1) #417 - "C=C" (1/1) #421 - "C=O" (1/1) #424 - "N=O" (1/1) #431 - "C(-C)(-C)(=O)" (1/1) <b>#434 - "C(-C)(-Cl)(=O)" (1/1)</b> #435 - "C(-C)(-H)(=O)" (1/1) #438 - "C(-C)(-N)(=O)" (1/1) #442 - "C(-C)(=O)" (1/1) #444 - "C(-C)(=O)" (1/1) #445 - "C(-Cl)(=O)" (1/1) #447 - "C(-H)(=O)" (1/1) #450 - "C(-N)(=O)" (1/1) #455 - "N(-C)(=O)" (1/1) #456 - "N(-O)(=O)" (1/1) #465 - "N-C-C-C" (1/1) #471 - "C-C-C=C" (1/1)

# PubChem features

## Simple SMARTS patterns

- **Simple SMARTS patterns** – These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.

They include bit positions from 461 to 713.

The screenshot shows a 'Similarity form' window with a 'Structure similarity: 96.65%' label. It is divided into two main columns: 'Target chemical' and 'Query chemical'. Each column contains a chemical structure and a list of SMARTS patterns. The patterns are numbered and include counts in parentheses. Two red boxes highlight specific lists of patterns in both columns.

Target chemical	Query chemical
<chem>CC1=CC=C(C=C1)C(=O)N=[O-]</chem>	<chem>CC1=CC=C(C=C1)C(=O)N=[O+]</chem>
<ul style="list-style-type: none"> <li>#465 - "N-C-C=C" (1/1)</li> <li>#471 - "C-C-C=C" (1/1)</li> <li>#491 - "C-C-C=C" (1/1)</li> <li>#494 - "O=C-C=C" (1/1)</li> <li>#503 - "N-C-C-[#1]" (1/1)</li> <li>#515 - "O-N-C-C" (1/1)</li> <li>#517 - "[#1]-C=C-[#1]" (1/1)</li> <li>#521 - "C=C-C-C" (1/1)</li> <li>#525 - "C-C=C-C" (1/1)</li> <li>#536 - "O=C-C-C" (1/1)</li> <li>#546 - "N-C-C-C" (1/1)</li> <li>#547 - "N-C-C-[#1]" (1/1)</li> <li>#550 - "N-C-C-C" (1/1)</li> <li>#551 - "Cl-C-C-C" (1/1)</li> <li>#553 - "C-C-C-C" (1/1)</li> <li>#554 - "O=C-C=C" (1/1)</li> <li>#557 - "C=C-C-C" (1/1)</li> <li>#559 - "O=N-G-C" (1/1)</li> <li>#565 - "C=C-C=C" (1/1)</li> <li>#571 - "C-C-C-C" (1/1)</li> <li>#579 - "C-C-C-C" (1/1)</li> <li>#580 - "O=C-C-C-C" (1/1)</li> <li>#583 - "C-C-C-C-C" (1/1)</li> <li>#585 - "C-C-C-C-C" (1/1)</li> <li>#592 - "Cl-C-C-C-C" (1/1)</li> <li>#593 - "N-C-C-C-C" (1/1)</li> <li>#596 - "C-C-C-C-C" (1/1)</li> <li>#598 - "O=C-C-C-C" (1/1)</li> <li>#600 - "[#1]-C-C=C-[#1]" (1/1)</li> <li>#601 - "N-C-C-C" (1/1)</li> </ul>	<ul style="list-style-type: none"> <li>#601 - "N-C-C-C" (1/1)</li> <li>#604 - "C-C-C-C" (1/1)</li> <li>#608 - "N-C-C-C" (1/1)</li> <li>#609 - "C-C-C-C" (1/1)</li> <li>#619 - "C-C-C-C" (1/1)</li> <li>#624 - "O=C-C-C" (1/1)</li> <li>#634 - "N-C-C-C" (1/1)</li> <li>#635 - "C-C-C-C" (1/1)</li> <li>#641 - "C=C-C-C" (1/1)</li> <li>#661 - "C-C=C-C" (1/1)</li> <li>#665 - "C-C=C-C" (1/1)</li> <li>#666 - "N-C-C-C" (1/1)</li> <li>#669 - "C-C-C-C" (1/1)</li> <li>#672 - "O=C-C=C" (1/1)</li> <li>#673 - "O=C-C=C-[#1]" (1/1)</li> <li>#678 - "C-C=C-C" (1/1)</li> <li>#679 - "C-C=C-C" (1/1)</li> <li>#680 - "C-C-C-C-C" (1/1)</li> <li>#684 - "N-C-C-C-C" (1/1)</li> <li>#685 - "O=C-C-C-C" (1/1)</li> <li>#689 - "C-C-C-C-C" (1/1)</li> <li>#693 - "O=C-C-C-C-C" (1/1)</li> <li>#696 - "O=C-C-C-C-C-N" (0/1)</li> <li>#705 - "O=C-C-C-C-C-C" (1/1)</li> <li>#708 - "O=C-C-C-C-C(N)-C" (0/1)</li> <li>#709 - "C-C(C)-C" (1/1)</li> <li>#710 - "C-C(C)-C-C" (1/1)</li> <li>#711 - "C-C(C)(C)-C" (1/1)</li> <li>#717 - "C-C(C)(N)CC" (0/1)</li> <li>#780 - "CC1CCC(N)CC1" (0/1)</li> </ul>

# PubChem features

## Complex SMARTS patterns

- **Complex SMARTS patterns** – These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.

They include bit positions from 714 to 881.

Similarity form

Structure similarity: 96.65%

Target chemical	Query chemical
<p>#665 - "C-C=C-C=C" (1/1)</p> <p>#666 - "N-C-C-C-C" (1/1)</p> <p>#669 - "C-C:C-C-C" (1/1)</p> <p>#672 - "O=C-C=C-C" (1/1)</p> <p>#673 - "O=C-C=C-[#1]" (1/1)</p> <p>#678 - "C-C=C-C-C" (1/1)</p> <p>#679 - "C-C-C=C-C" (1/1)</p> <p>#680 - "C-C-C-C-C" (1/1)</p> <p>#684 - "N-C-C-C-C" (1/1)</p> <p>#685 - "O=C-C-C-C" (1/1)</p> <p>#686 - "O=C-C-C-C-N" (0/1)</p> <p>#689 - "C-C-C-C-C-C" (1/1)</p> <p>#693 - "O=C-C-C-C-C" (1/1)</p> <p>#696 - "O=C-C-C-C-C-N" (0/1)</p> <p>#705 - "O=C-C-C-C-C-C" (1/1)</p> <p>#709 - "C-C(O)-C-C" (1/1)</p> <p>#710 - "C-C(O)-C-C" (1/1)</p> <p>#711 - "C-C-C(O)-C-C" (1/1)</p> <p>#738 - "Cc1cc(N)ccc1" (0/1)</p> <p>#801 - "CC1CC(N)CCC1" (0/1)</p>	<p>#666 - "N-C:C-C-C" (1/1)</p> <p>#669 - "C-C:C-C-C" (1/1)</p> <p>#672 - "O=C-C=C-C" (1/1)</p> <p>#673 - "O=C-C=C-[#1]" (1/1)</p> <p>#678 - "C-C=C-C-C" (1/1)</p> <p>#679 - "C-C-C=C-C" (1/1)</p> <p>#680 - "C-C-C-C-C" (1/1)</p> <p>#684 - "N-C-C-C-C" (1/1)</p> <p>#685 - "O=C-C-C-C" (1/1)</p> <p>#689 - "C-C-C-C-C-C" (1/1)</p> <p>#693 - "O=C-C-C-C-C" (1/1)</p> <p>#696 - "O=C-C-C-C-C-N" (0/1)</p> <p>#705 - "O=C-C-C-C-C-C" (1/1)</p> <p>#708 - "O=C-C-C-C-C(N)-C" (0/1)</p> <p>#709 - "C-C(C)-C-C" (1/1)</p> <p>#710 - "C-C(C)-C-C" (1/1)</p> <p>#711 - "C-C-C(C)-C-C" (1/1)</p> <p>#717 - "Cc1ccc(N)cc1" (0/1)</p> <p>#780 - "CC1CCC(N)CC1" (0/1)</p>

# Outlook

- Background
- Aims
- PubChem features
- **The exercise**
- Workflow



# The Exercise

- In this exercise we will compare:
  1. two chemicals with respect to *PubChem* substructure similarity features (we will use *m-Chloroaniline* and *benzoic acid*);
  2. One chemical with a list of chemicals with respect to *PubChem* substructure similarity features (we will use *m-Chloroaniline* and Skin sensitization ECETOC database ).

# Outlook

- Background
- Aims
- PubChem features
- The exercise
- **Workflow**
  - **Substructure similarity between two chemicals**

# Workflow

## Substructure similarity between two chemicals

The screenshot shows the QSAR Toolbox software interface. The top navigation bar includes 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling', and 'Report'. The 'Profiling' module is active, showing a list of profiling methods. The 'Structure similarity' method is selected, and its 'Options' dialog box is open. The dialog box has four numbered callouts: 1 points to the 'Profiling' button in the top bar; 2 points to the 'Options' button in the 'Structure similarity' method's context menu; 3 points to the 'Molecular features' section where 'PubChem features' is checked; 4 points to a chemical structure in the 'Example' section.

1. Go to *Profiling* module; 2. Right click over the *Structure similarity* profiler and select **Options**;
3. Uncheck all molecular features and select only *PubChem* features. The additional similarity options (e.g. *Calculation* and *Atom characteristics*) do not have influence to the *PubChem* features.;
4. Double click on the structure in left which will be our target. *2D editor* window appears.

# Workflow

## Substructure similarity between two chemicals

The image shows two screenshots of the 2D Editor software interface. The left screenshot shows the 'Blank page' button (a red square icon) in the top toolbar, highlighted with a callout box labeled '1'. A 'Clear All' dialog box is open in the center, asking 'Want to clear everything?' with 'Yes' and 'No' buttons. The 'Yes' button is highlighted with a callout box labeled '2'. The right screenshot shows the 'm-chloroaniline' structure drawn on the canvas, highlighted with a red dashed box and a callout box labeled '3'. The 'OK' button at the bottom right of the window is highlighted with a callout box labeled '4'.

1. Click on the **Blank page** button; 2. Confirm that you want to clear everything by click on **Yes**;  
 3. Draw *m-chloroaniline* structure; 4. Click on **OK**

# Workflow

## Substructure similarity between two chemicals

**Similarity options**

**Measure**

- Tanimoto (Jaccard)
- Dice
- Kulczynski-2
- Ochiai(Cosine)
- Yule

**Molecular features**

- Atom pairs
- Topologic torsions
- Atom centered fragments
- Path
- Cycles
- PubChem features

**Calculation**

- Fingerprint
- Hologram

**Formula**

$$\frac{c}{0.5 [(a + b) + (b + c)]}$$

**Description**

The **PubChem System** generates a binary substructure fingerprint for chemical structures. A substructure is a fragment of a chemical structure. A fingerprint is an ordered list of binary (1/0) bits. Each bit represents a boolean determination of, or test for, the presence of, for example, an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure.

**Atom characteristics**

- Atom type
- Count H attached
- Count heavy atoms attached
- Hybridization
- Incident pi-bonds
- Valency
- Charge
- Cyclic

**Structure**

NCCN Define

**Example**

A	B	C
9	38	67

Similarity = 74.033% Details

**1** Nc1ccc(Cl)cc1 **2** ClC(=O)c1ccc([N+](=O)[O-])cc1

<-> Default Help OK Cancel

The structure which you have drawn appears (1). Double click on the structure in right (2) to draw the second chemical.

# Workflow

## Substructure similarity between two chemicals

The image shows two windows of the 2D Editor software. The left window displays a complex chemical structure with a nitro group and a chlorine atom. The right window shows the resulting benzoic acid structure after editing. Numbered callouts (1-5) indicate the steps in the workflow.

Click on the **Eraser** button (1) and remove the nitro group (2). Select the oxygen symbol (3) and click over the chlorine atom (4). Now you are ready with drawing of benzoic acid and have to click on **OK** (5)

# Workflow

## Substructure similarity between two chemicals

**Similarity options**

Measure

- Tanimoto (Jaccard)
- Dice
- Kulczynski-2
- Ochiai(Cosine)
- Yule

Molecular features

- Atom pairs
- Topologic torsions
- Atom centered fragments
- Path
- Cycles
- PubChem features

Options

Formula

$$\frac{c}{0.5 [(a + b) + (b + c)]}$$

Description

Structure

NCCN

Example

**Similarity form**

Structure similarity: 53.75%

Target chemical	Query chemical

PubChem features (43/76)

- #001 - ">= 4 H" (1/1)
- #010 - ">= 2 C" (1/1)
- #011 - ">= 4 C" (1/1)
- #015 - ">= 1 N" (0/1)
- #038 - ">= 1 Cl" (0/1)
- #179 - ">= 1 any ring size 6" (1/1)
- #180 - ">= 1 saturated or aromatic carbon-only ring size 6" (1/1)
- #256 - ">= 1 aromatic ring" (1/1)
- #284 - "C-H" (1/1)
- #285 - "C-C" (1/1)
- #286 - "C-N" (0/1)
- #295 - "C-Cl" (0/1)
- #300 - "N-H" (0/1)
- #333 - "C(-C)(-C)" (1/1)
- #341 - "C(-C)(-C)(-N)" (0/1)
- #343 - "C(-C)(-Cl)" (0/1)
- #345 - "C(-C)(-H)" (1/1)
- #352 - "C(-C)(-N)" (0/1)
- #356 - "C(-C)(C)" (1/1)
- #363 - "C(-C)(C)" (0/1)
- #371 - "C(-H)(C)" (1/1)
- #372 - "C(-H)(C)(C)" (1/1)
- #377 - "C(-N)(C)" (0/1)
- #378 - "C(-N)(C)(C)" (0/1)
- #385 - "C(C)(C)" (1/1)
- #394 - "N(-C)(-H)" (0/1)
- #417 - "C=C" (1/1)

PubChem features (43/84)

- #001 - ">= 4 H" (1/1)
- #010 - ">= 2 C" (1/1)
- #011 - ">= 4 C" (1/1)
- #019 - ">= 1 O" (0/1)
- #020 - ">= 2 O" (0/1)
- #179 - ">= 1 any ring size 6" (1/1)
- #180 - ">= 1 saturated or aromatic carbon-only ring size 6" (1/1)
- #256 - ">= 1 aromatic ring" (1/1)
- #284 - "C-H" (1/1)
- #285 - "C-C" (1/1)
- #287 - "C-O" (0/1)
- #309 - "O-H" (0/1)
- #333 - "C(-C)(-C)" (1/1)
- #334 - "C(-C)(-C)(-C)" (0/1)
- #345 - "C(-C)(-H)" (1/1)
- #353 - "C(-C)(-O)" (0/1)
- #356 - "C(-C)(C)" (1/1)
- #357 - "C(-C)(C)(C)" (0/1)
- #371 - "C(-H)(C)" (1/1)
- #372 - "C(-H)(C)(C)" (1/1)
- #381 - "C(-O)(-O)" (0/1)
- #385 - "C(C)(C)" (1/1)
- #407 - "O(-C)(-H)" (0/1)
- #417 - "C=C" (1/1)
- #421 - "C=O" (0/1)
- #431 - "C(-C)(-C)(=C)" (0/1)
- #435 - "C(-C)(-H)(=C)" (1/1)

Similarity between the two structures is calculated automatically with respect to the *PubChem* features (1). Click on **Details** button (2) to get more information about the common and unique structural features.

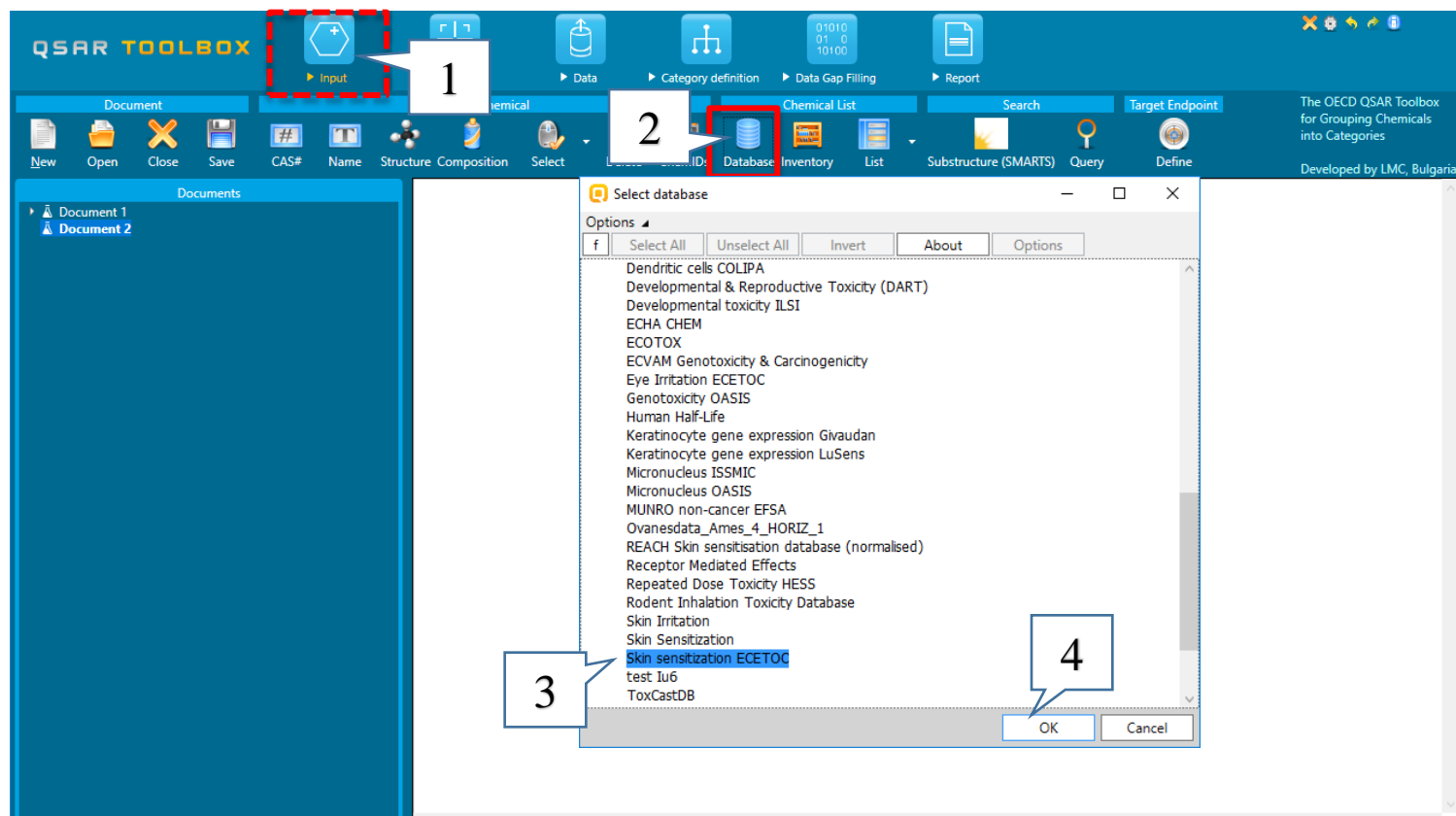
# Outlook

- Background
- Aims
- PubChem features
- The exercise
- **Workflow**
  - Substructure similarity between of two chemicals
  - **Substructure similarity between single chemical and chemical list**



# Workflow

## Substructure similarity between single chemical and chemical list



1. Go to *Input* module; 2. Click on *Chemical List* > **Database**; 3. Select **Skin sensitization ECETOC** database; 4. Click on **OK**.

# Workflow

## Substructure similarity between single chemical and chemical list

The screenshot shows the QSAR Toolbox software interface. The top navigation bar includes 'Input', 'Profiling', 'Category definition', 'Data Gap Filling', and 'Report'. The main window displays a list of profiling methods on the left, with 'Structure similarity' selected. A red box highlights the 'Options' button for 'Structure similarity'. A dialog box titled 'Similarity options' is open, showing various settings for the similarity calculation. The dialog box includes sections for 'Measure' (Dice is selected), 'Molecular features' (PubChem features is checked), 'Calculation' (Hologram is selected), and 'Atom characteristics' (Atom type, Count H attached, and Hybridization are checked). A Venn diagram shows the overlap of features between two sets, and a chemical structure of *m*-chloroaniline is shown. The dialog box also displays a similarity score of 74.033% and buttons for 'Default', 'Help', 'OK', and 'Cancel'.

1. Go to *Profiling* module; 2. Right click over the Structure similarity profile and select **Options**;
3. Select only *PubChem* features; 4. Click on the **Define** button and draw *m*-chloroaniline (see slide 20); 5. Click on **OK**.

# Workflow

## Substructure similarity between single chemical and chemical list

The screenshot shows the QSAR Toolbox interface. On the left, a list of endpoints is visible, with 'Structure similarity' checked (indicated by a red arrow and callout '1'). In the top-left corner, the 'Apply' button is highlighted with a red box and callout '2'. The main window displays a 'Filter endpoint tree...' dialog with 'Structure' selected. Below this, a table shows chemical structures in columns 1 through 7. A callout box in the center of the table reads: 'Check the box in front of the *Structure similarity* profile (1) and click on **Apply** (2) .'. The bottom status bar shows '33/39' and 'Profiling'.

! Keep in mind that the chemical list which you are insert have no a target structure. Otherwise the structure similarity will be calculated based on the target structure in the data matrix.

# Workflow

## Substructure similarity between single chemical and chemical list

The screenshot illustrates the workflow for substructure similarity analysis in the QSAR Toolbox. The main interface shows the 'Filter endpoint tree...' with 'Structure similarity' selected. A 'Similarity form' window is open, displaying a comparison between a 'Target chemical' (4-chloroaniline) and a 'Query chemical' (4-chloro-2-nitrophenol). The similarity percentage is 84.34%. The 'Similarity form' also displays two columns of 'PubChem features (70/76)' for each molecule, with a red dashed box highlighting the differences. A red box labeled '1' points to the 'Explain' button in the 'Structure similarity' section of the main window. A red box labeled '2' points to the 'Structure similarity: 84.34%' text in the 'Similarity form' window. A red box labeled '3' points to the 'PubChem features' lists in the 'Similarity form' window.

You can see more details about the structure similarity by right click over the result and selection of **Explain** (1). In the *Similarity form* you can see the exact percentage of structure similarity (2) as well as all similar and different features (3) between the two molecules.

# Congratulation

- You have now been familiarized with *PubChem* substructure similarity features;
- You have compared: 1) two chemicals and 2) single chemical and chemical list with respect to PubChem substructure similarity;
- Note proficiency comes with practice.