

## *Classifying Compounds in Public Databases*

ACS San Diego, March 15, 2016

OntoChem IT Solutions GmbH  
Blücherstr. 24  
06120 Halle (Saale)  
Germany  
Tel. +49 345 4780470  
Fax: +49 345 4780471  
mail: [info@ontochem.com](mailto:info@ontochem.com)

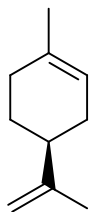


## Ontology...

### Why in chemistry ?

Organizing knowledge

e.g. terpenes are made of x times 5 carbon isoprene units



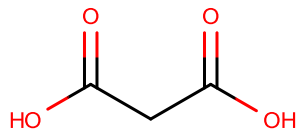
## Ontology...

### Why in chemistry ?

Organizing knowledge

Conserving old knowledge

what are dicarboxylic acids?



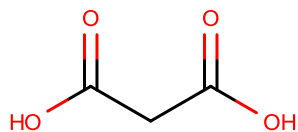
## Ontology...

### Why in chemistry ?

Organizing knowledge

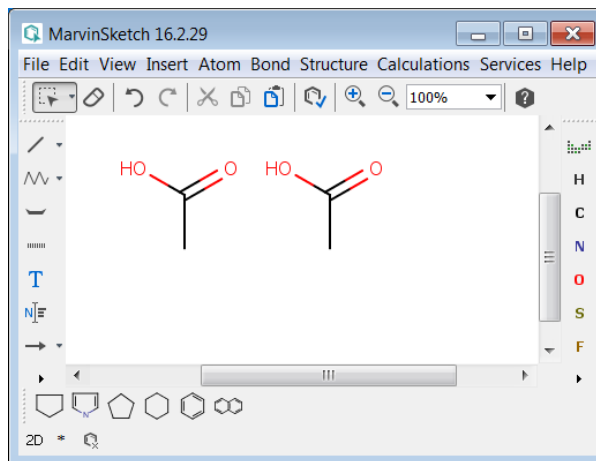
Conserving old knowledge

what are dicarboxylic acids?



*is\_a*: dicarboxylic acid

**ontology**



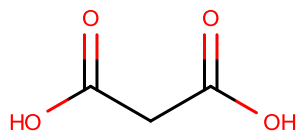
## Ontology...

### Why in chemistry ?

Organizing knowledge

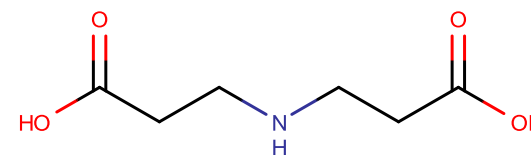
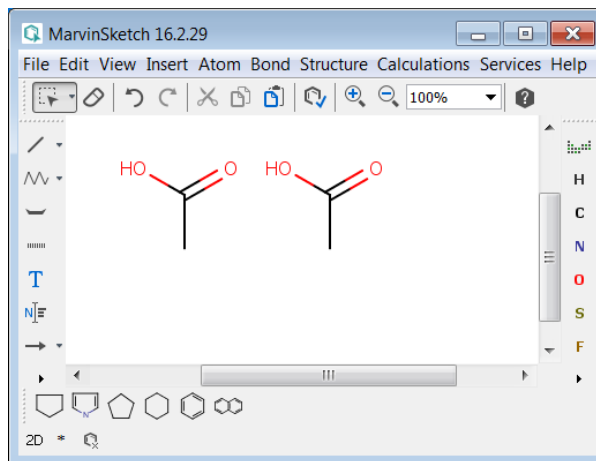
Conserving old knowledge

what are dicarboxylic acids?



*is\_a*: dicarboxylic acid

**ontology**



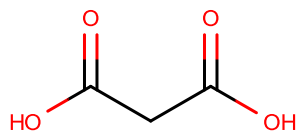
## Ontology...

### Why in chemistry ?

Organizing knowledge

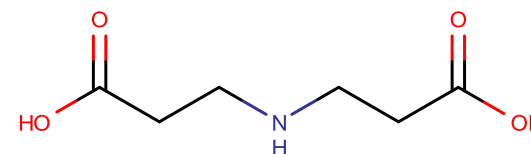
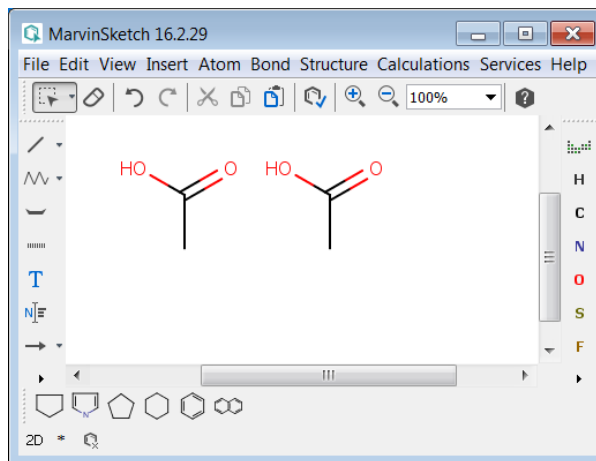
Conserving old knowledge

what are dicarboxylic acids?



*is\_a*: dicarboxylic acid

ontology



*is\_part*: 2 carboxylic acid moieties

„partology“ (Bolton, Hastings)  
SSS based

## Ontology...

### Why in chemistry ?

Organizing knowledge

Conserving old knowledge

Transporting our knowledge

e.g. teaching humans and computers

Using our knowledge

e.g. data mining, property predictions, knowledge inference,  
... creating new products



# ...need to build on common rules

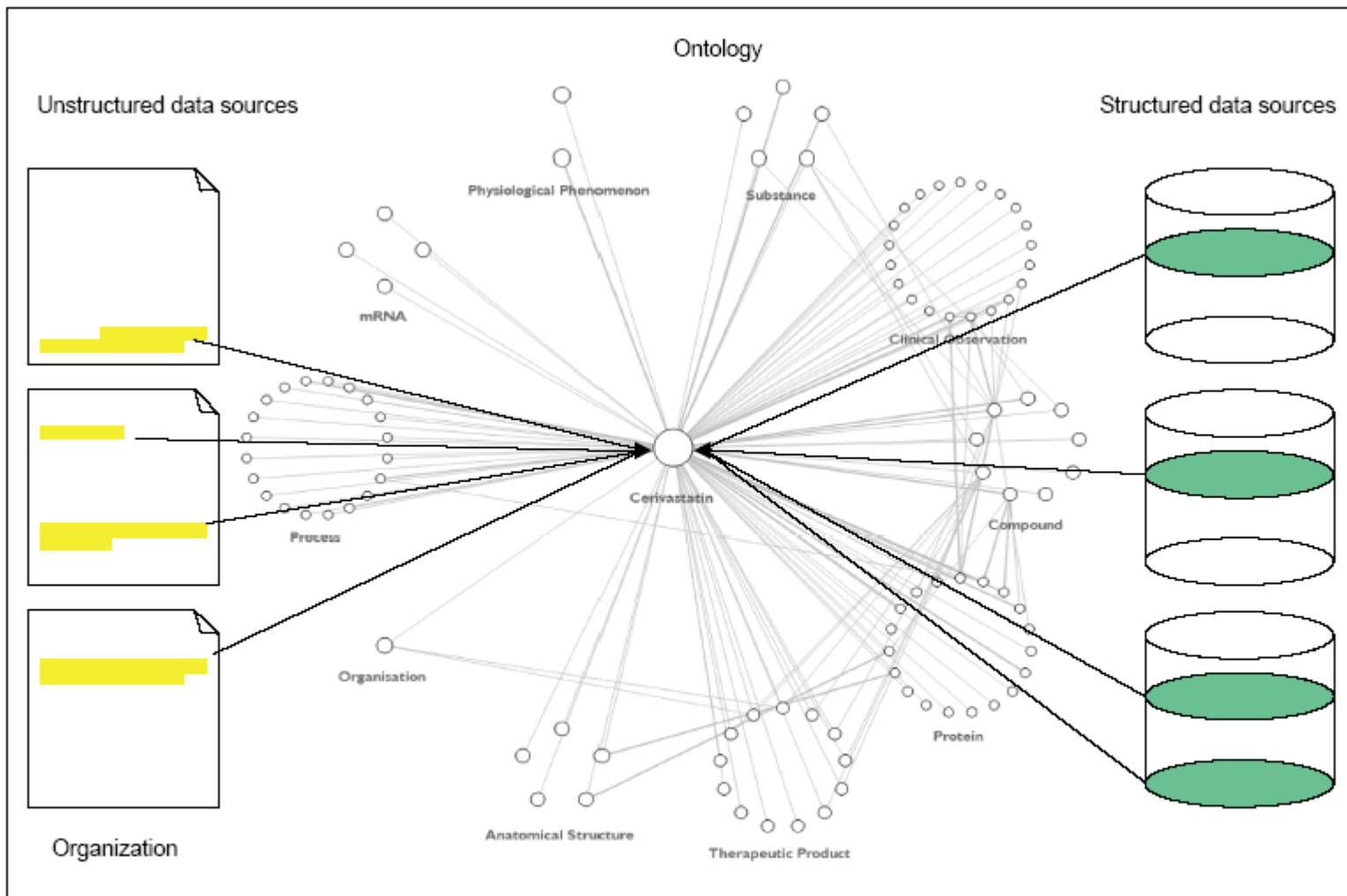


FIGURE 2

**Granular semantic data integration using an ontology.** An ontology can be used as a semantic middle layer to map references semantically to the same concepts among multiple data sources. In the example shown, the concept 'Cerivastatin' in the ontology is used as a consistent marker to connect data from multiple structured and unstructured data sources that refer to 'Cerivastatin', 'Baycol' or other equivalent terms.

## Assignment of organic compounds in structure files, databases, documents

Organic Chemistry:

7,916 class concepts

32,469 synonyms

4,698 SMARTS concepts

The screenshot displays the SODIAC software interface for ontology management. The main window is titled "SODIAC - oc\_chem\_compound\_classes\_2016-02-18.obo". The interface is divided into several panels:

- Tree View:** A hierarchical tree on the left side showing the ontology structure. The "pyrimidines" class is highlighted under the "6-membered heterocycles" category.
- Details:** A panel on the right showing the details for the selected "pyrimidines" class, including its ID (150000003230) and name.
- Compound Assignment:** A central panel showing a list of compounds (1, 2, 3, 4) and their details. The "Structure" field displays a chemical structure and its SMILES string: CS(=O)(=O)Cc1cccc(Nc2nccc(Oc3ccc(NC(=O)C4(C)C(=O)Nc4ccc(Br)cc4)cc3F)n2)c1.
- Assigned Ontology Classes:** A list of classes assigned to the compound, including "pyrimidines".
- Single Structure Matching:** A panel at the bottom right showing a chemical structure with a green highlight on the pyrimidine ring, indicating a match with the ontology class.

## Assignment of compounds in structure files, databases, documents

SMARTS based

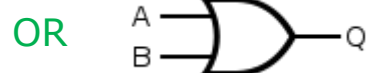
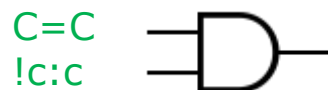
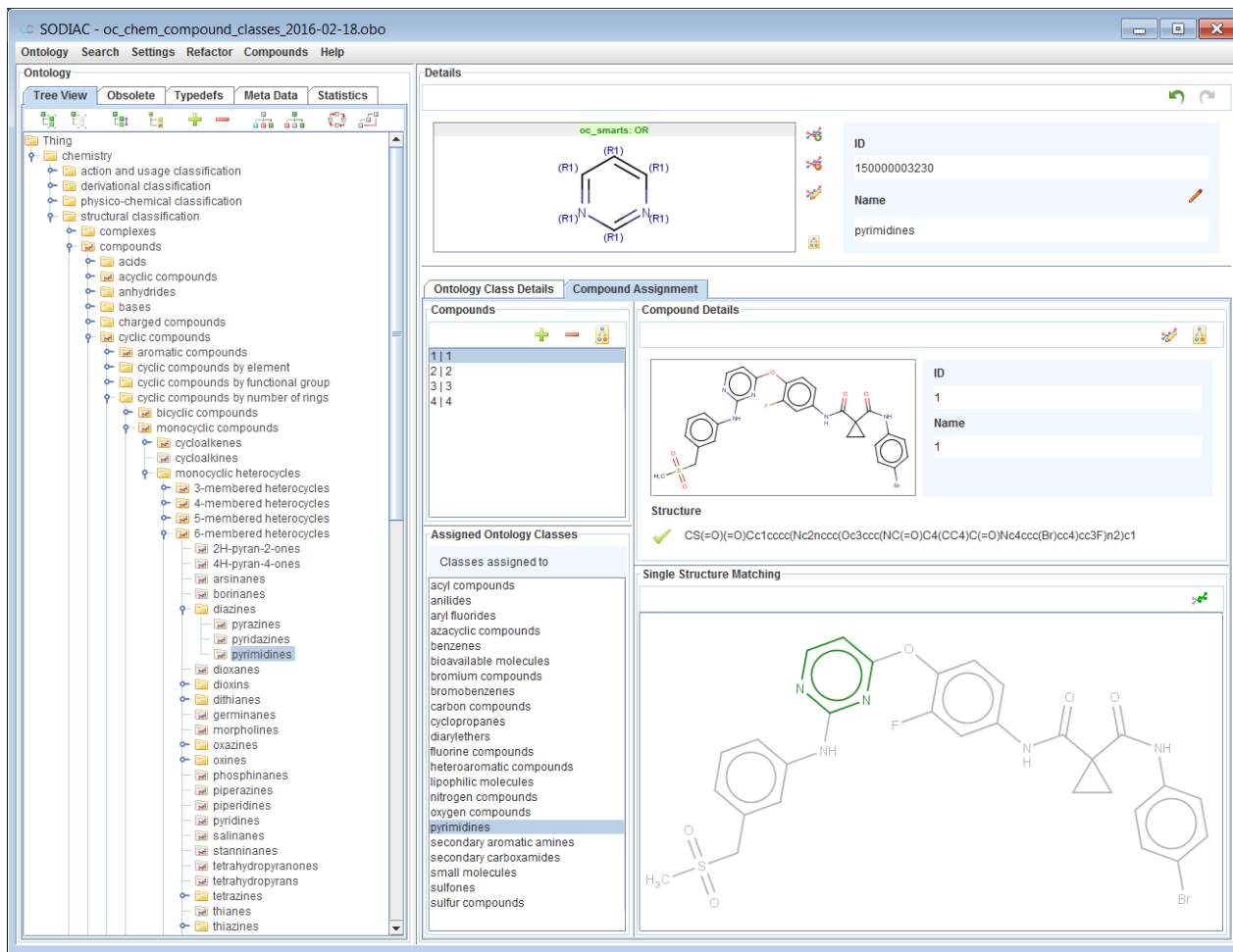
e.g. alkylbromides

[C;!\$(C=[O,S,N,P]))Br

D. Weininger

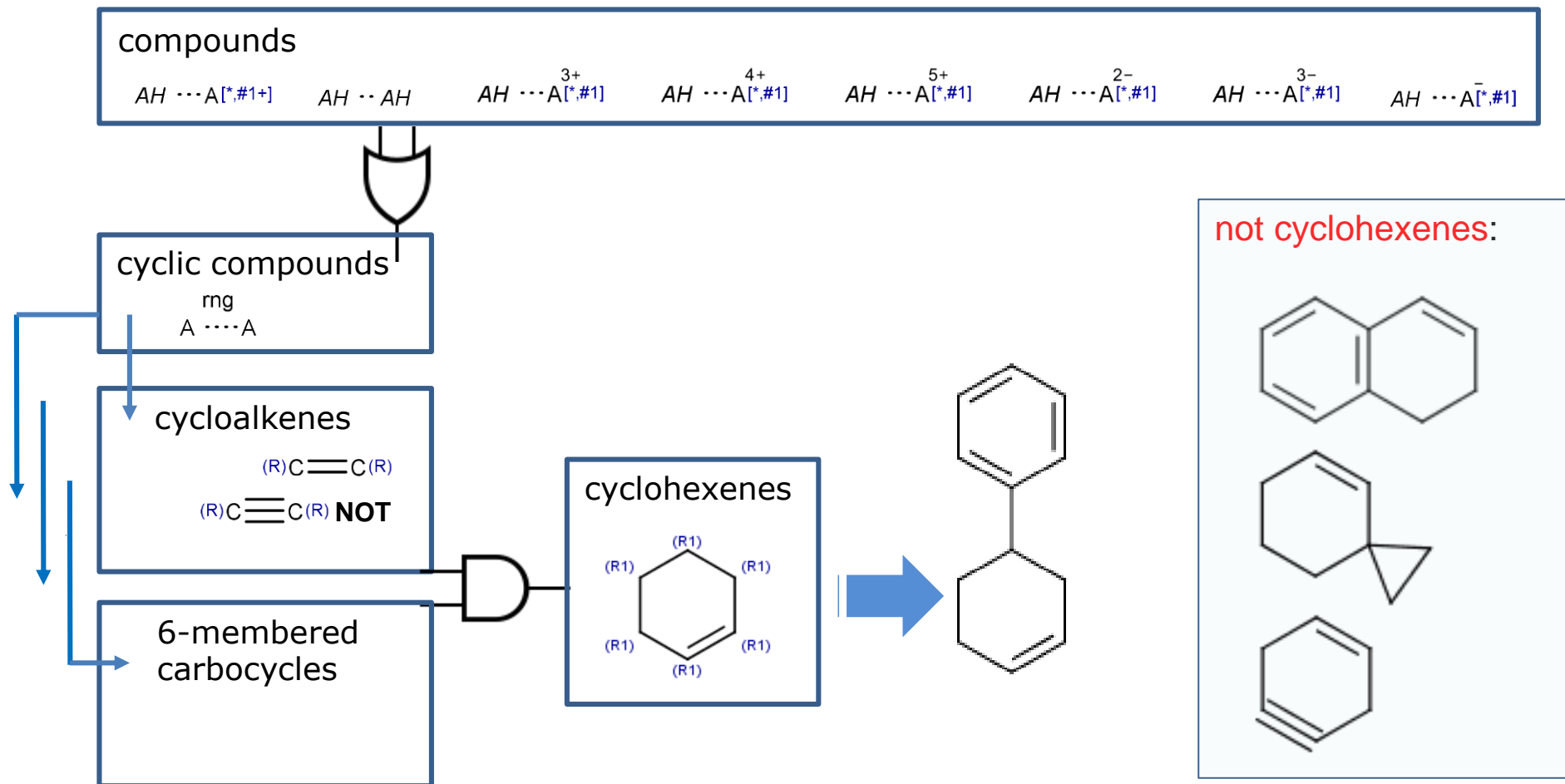
AND, OR, NOT logic

e.g. AND NOT for alkenes:

The screenshot displays the SODIAC software interface. On the left is a hierarchical ontology tree under 'chemistry', including categories like 'cyclic compounds', 'aromatic compounds', and 'heterocycles'. The 'pyrimidines' class is highlighted. The main window is divided into several panels: 'Details' at the top right shows a pyrimidine structure and its ID (150000003230) and name (pyrimidines). Below this is the 'Compound Assignment' panel, which lists assigned ontology classes such as 'acyl compounds', 'anilides', 'aryl fluorides', etc., with 'pyrimidines' selected. The 'Single Structure Matching' panel at the bottom right shows a complex chemical structure with a pyrimidine ring highlighted in green, indicating a successful match.

using SMARTS filters with AND, OR & NOT – a *hierarchical state machine* on DB



Testing **Sodiac** by annotating public databases:

- **ChEBI** (from E. Bolton, PubChem, March 1, 2016): 42,055 compounds
- **MeSH** (from E. Bolton, PubChem, March 1, 2016): 116,542 compounds
- **ChEMBL** (from Steve Boyer, IBM, February 2, 2016): 1,456,019 compounds
- **PubChem** (downloaded Sept. 2015, PubChem): 68,420,615 compounds

performance: takes 6 hours to load db, takes 4h to assign classes

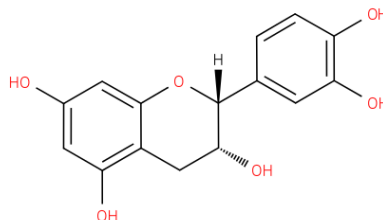
- **Patent** full text documents: >90 million full text documents

## ChEBI

- Manually curated, can we do it automatically ?
- Can we recognize the correct ChEBI parent ID for a compound ?
  - 61,904 ChEBI concepts in total, 3,584 ChEBI concepts contain a \* in a SMILES
  - 45,748 ChEBI concepts are compounds with SMILES, 44,355 after cleaning

## Assignment

1,164 ChEBI compounds are present as scaffold concepts in SODIAC as well,  
e.g. CHEBI:90 (-)-epicatechin



## Assignment of ChEBI concepts

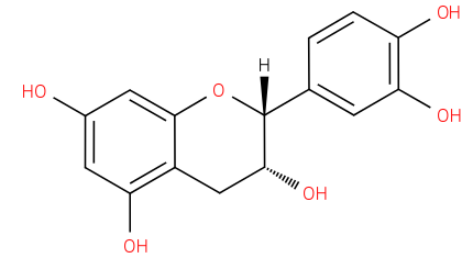
**CHEBI:90** (-)-epicatechin

**is\_a** catechin (CHEBI:23053)

**is\_a** polyphenol (CHEBI:26195)

**has\_role** antioxidant (CHEBI:22586)

**is\_enantiomer\_of** (+)-epicatechin (CHEBI:76125)



**SODIAC:150000000071** (-)-epicatechin

**is\_a** (-)-epicatechin derivatives; alkylarylethers; aromatic compounds; carbon compounds; chromanes; oxacyclic compounds; oxygen compounds; polyols; **pyrocatechols**; secondary alcohols; triols;

## ChEBI – SODIAC (ChEBI) matches:

### ChEBI

1,888 ChEBI parent concepts have been assigned to the **Test Set** of **1,164 ChEBI** compounds (avg 1.62 parent classes / compound)

### SODIAC

1,056 ChEBI parents were assigned by SODIAC from 1,888 **is\_a** ChEBI parents = 56%

832 ChEBI concepts not found as they are not represented in SODIAC

29,833 SODIAC classes assigned to 1,164 ChEBI compounds (avg. 26 classes/compound)

**14,476 ChEBI concepts found by SODIAC that were not assigned as parents**



**ChEBI – SODIAC (ChEBI)** matches: CHEBI:583099 ent-zingiberene

ChEBI: *is\_a* sesquiterpene (CHEBI:35189)

SODIAC: 16 new proposals:

CHEBI:33654 **alicyclic compound**

CHEBI:33653 **aliphatic compound**

CHEBI:32878

CHEBI:36480

CHEBI:33598

CHEBI:33595

CHEBI:33643

CHEBI:33646 **alkadiene**

CHEBI:24632

CHEBI:18059

CHEBI:33661

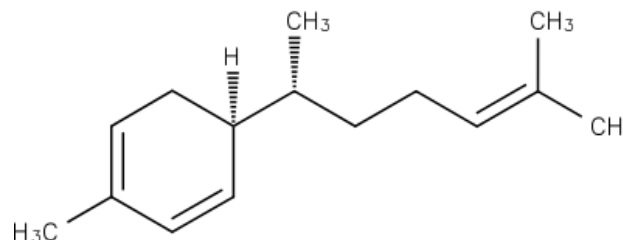
CHEBI:35187

CHEBI:50073 **p-menthadiene**

CHEBI:25826

CHEBI:50034

CHEBI:35186



## ChEBI – SODIAC (ChEBI)

cyclobutane (CHEBI:30377) *is\_a* cycloalkane (CHEBI:23453)

cyclohexane (CHEBI:29005) *is\_a* cycloalkane (CHEBI:23453)

cyclopentane (CHEBI:23492) *is\_a* cycloalkane (CHEBI:23453)

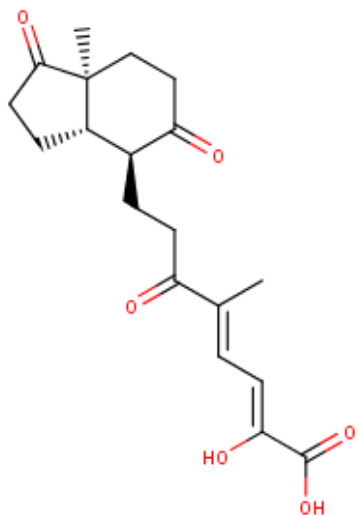
cyclopentane (CHEBI:23492) *is\_a* cyclopentanes (CHEBI:23493)

**.... automation could help to eliminate human inconsistencies**

**SODIAC** coverage from 7,916 compound class concepts in common:

**ChEBI:** 3,468 (of 16,156) **MeSH:** 2,570 (of 3,943) **ClassyFire:** 483 (of 4,832)

Next steps: complete overlap with ChEBI and MeSH



pubchem\_ID  
25244777

SODIAC parents	SODIAC_xrefs	ClassyFire parents	ClassyFire_xref
4,5:9,10-diseco-androstenes; acyclic ketones; alkene; bicyclic compounds; carbocycles; carbon compounds; <b>carboxylic acids</b> ; <b>cyclic ketones</b> ; dienes; oxygen compounds;	CHEBI:32878; MeSH:D000475; CHEBI:33636; CHEBI:33575; MeSH:D002264; CHEBI:3992; CHEBI:33646; MeSH:D000466; MeSH:D017601;	Medium-chain keto acids and derivatives; Hydroxy fatty acids; Branched fatty acids; Unsaturated fatty acids; Alpha-branched alpha,beta-unsaturated ketones; Enones; Acryloyl compounds; <b>Cyclic ketones</b> ; Monocarboxylic acids and derivatives; Enols; <b>Carboxylic acids</b> ; Organic oxides; Hydrocarbon derivatives	CHEBI:25754; CHEBI:24654; CHEBI:35819; CHEBI:27208; CHEBI:51689; CHEBI:51689; CHEBI:78840; CHEBI:3992; CHEBI:25384; CHEBI:33823; CHEBI:33575; CHEBI:25701; CHEBI:72695

## SODIAC / ClassyFire top-bottom statistics

SODIAC_ID	SODIAC_name	count	ClassyFire_ID	ClassyFire_name	count	difference
OC:150000001336	alkene	8781	CHEMONTID:0002501	Alkenes	0	8781
OC:150000001344	alkene derivatives	4910	CHEMONTID:0002501	Alkenes	0	4910
OC:150000002891	organic anions	7219	CHEMONTID:0003608	Organic anions	4481	2738
OC:150000003997	cyclic ketones	3945	CHEMONTID:0003487	Cyclic ketones	1433	2512
OC:150000003142	primary aromatic amines	2606	CHEMONTID:0002453	Primary aromatic amines	101	2505
OC:150000001993	dienes	2487	CHEMONTID:0001019	Alkadienes	0	2487
OC:150000003333	secondary alcohols	17455	CHEMONTID:0001661	Secondary alcohols	15293	2162
OC:150000002989	phenols	1714	CHEMONTID:0000134	Phenols	256	1458
OC:150000003335	secondary amines	1389	CHEMONTID:0002451	Secondary amines	99	1290
OC:150000002484	inorganic compounds	1251	CHEMONTID:0000001	Inorganic compounds	0	1251
OC:150000003568	thioethers	1132	CHEMONTID:0001202	Thioethers	0	1132
OC:150000003231	pyrocatechols	1642	CHEMONTID:0000135	Catechols	713	929
OC:159000000021	acetate salts	8	CHEMONTID:0003919	Acetate salts	1459	-1451
OC:150000001734	carboxylic acid salts	1090	CHEMONTID:0001166	Carboxylic acid salts	3141	-2051
OC:150000002341	glycosylamines	41	CHEMONTID:0002203	Glycosylamines	2402	-2361
OC:150000003527	tetrahydropyrans	4079	CHEMONTID:0002012	Oxanes	6583	-2504
OC:150000003141	primary amines	45	CHEMONTID:0002450	Primary amines	2980	-2935
OC:150000001241	acetals	1588	CHEMONTID:0001656	Acetals	4947	-3359
OC:150000001735	carboxylic acids	5969	CHEMONTID:0001205	Carboxylic acids	11204	-5235
OC:150000004030	azacyclic compounds	6174	CHEMONTID:0004139	Azacyclic compounds	11970	-5796
OC:150000001728	carbonyl compounds	1335	CHEMONTID:0001831	Carbonyl compounds	18235	-16900

OntViewer

File Search View Content Options Help

- organic compounds
  - alkene derivatives
  - alkine derivatives
  - alkoxides
  - ammonium compounds
  - arylidenes
  - carbocycles
  - carbonyl compounds
  - esters
  - ethers
  - heterocycles
  - hydrocarbons
  - hydroxy compounds
  - organic compound scaffolds
    - alkaloid derivatives
    - alpha-amino acid derivatives
    - alpha-amino acids
    - carbohydrate derivatives
    - cephams
    - diarylethenes
    - lipids
      - fatty acyls
      - glycerolipids
      - phospholipids
      - prenol lipids
        - hydroquinone lipids
        - polyprenols
        - quinone lipids
      - terpenes
        - C35 terpenes
        - diterpene derivatives
        - monoterpenes
          - bicyclic monoterpenes**
          - monocyclic monoterpenes
          - non cyclic monoterpenes
        - sesquiterpene derivatives
        - sesterterpene derivatives

preferred name: bicyclic monoterpenes

id: 150000001604

synonym: bicyclic monoterpene [scope=EXACT type=SYNONYM xref=OC {date=2011-09-14, lang=eng}]  
bicyclic monoterpenes [scope=EXACT type=SYNONYM xref=OC {date=2012-03-04, lang=eng, prefname=1}]

is-a: 150000002773 (monoterpenes)

pubchem: 86728

pubchem (oc\_chem\_compound\_classes\_150910.obo)

- chemistry
  - structural classification
    - compounds
      - organic compounds
        - organic compound scaffolds
          - lipids
            - prenol lipids
              - terpenes
                - monoterpenes
                  - bicyclic monoterpenes

100GB ontology file

2 0

**Statistics:** find **missing/wrong** definitions / wrong compounds?

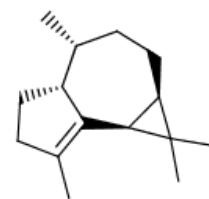
| Concept                             | Aggregated Children | Direct Children | Concept                             |
|-------------------------------------|---------------------|-----------------|-------------------------------------|
| chemistry                           | 1,242,661,248       |                 | chemistry                           |
| structural classification           | 1,240,832,811       |                 | structural classification           |
| compounds                           | 1,227,157,075       |                 | compounds                           |
| organic compounds                   | 574,930,108         | <b>52,160</b>   | organic compounds                   |
| cyclic compounds                    | 379,604,259         | <b>221,327</b>  | cyclic compounds                    |
| element compounds                   | 248,665,892         |                 | element compounds                   |
| organic nitrogen compounds          | 175,349,613         | 1,083,882       | organic nitrogen compounds          |
| carbonyl compounds                  | 169,293,274         | 1,342,539       | carbonyl compounds                  |
| cyclic compounds by element         | 135,330,213         |                 | cyclic compounds by element         |
| cyclic compounds by ring size       | 96,080,495          |                 | cyclic compounds by ring size       |
| 5-7-membered cyclic compounds       | 91,681,812          |                 | 5-7-membered cyclic compounds       |
| chalcogen compounds                 | 84,914,021          |                 | chalcogen compounds                 |
| organic halogen compounds           | 83,842,178          | 21,610          | organic halogen compounds           |
| cyclic compounds by number of rings | 83,377,605          |                 | cyclic compounds by number of rings |
| heterocycles                        | 74,645,912          | 60,802          | heterocycles                        |

e.g. organic compound:

2-Benzofuranmethyl [CH2]c1occ2ccccc12 <sup>^1:0</sup>

radicals?

(-)- $\alpha$ -gurjunene  
missing scaffold



## Statistics: Find missing / wrong definitions

| Concept                               | Aggregated Children | Direct Children | Concept                               |
|---------------------------------------|---------------------|-----------------|---------------------------------------|
| iron cyano complex                    | 0                   |                 | iron cyano complex                    |
| iron nitroso complex                  | 0                   |                 | iron nitroso complex                  |
| 2-exo-hydroxy-1,4-cineole derivatives | 0                   |                 | 2-exo-hydroxy-1,4-cineole derivatives |
| Eucarvone derivatives                 | 0                   |                 | Eucarvone derivatives                 |
| pinocarvone derivatives               | 0                   |                 | pinocarvone derivatives               |
| ocimene derivatives                   | 0                   |                 | ocimene derivatives                   |
| $\beta$ -ocimene derivatives          | 0                   |                 | $\beta$ -ocimene derivatives          |
| (Z)- $\beta$ -ocimene derivatives     | 0                   |                 | (Z)- $\beta$ -ocimene derivatives     |
| (E)- $\beta$ -ocimene derivatives     | 0                   |                 | (E)- $\beta$ -ocimene derivatives     |
| radium salts                          | 0                   |                 | radium salts                          |
| unclassified entities                 | 0                   |                 | unclassified entities                 |
| iron cyano complex                    | 0                   |                 | iron cyano complex                    |
| iron nitroso complex                  | 0                   |                 | iron nitroso complex                  |
| 2-exo-hydroxy-1,4-cineole derivatives | 0                   |                 | 2-exo-hydroxy-1,4-cineole derivatives |
| Eucarvone derivatives                 | 0                   |                 | Eucarvone derivatives                 |

## ChemAnalyser: searching "pinocarvone" in patent and non-patent literature:

303 patent hits in total  
(53 in SciFinder)

20 patent hits in 2009  
(1 in SciFinder)

Welcome LutzWeber! My ChemAnalyser Tips & tricks About Logout A- | A | A+

chemanalyser version 1.2.4.3 Search results

Search Expert search Structure search Search history Co-occurrences

infoapps (303) PubMed Central (37)

+OCID:"190000036336"

303 document(s) found.  Go to page

(1 of 31)

**1. A TOPICAL COMPOSITION**

Sem-IP.com

[OC DocID: 97830513] [Patent publ. date: 2015-08-13] [Patent ID: WO2015117957A1] [Patent classifications IPC: A61K31/365; A61K33/38; A61K8/19; A61K8/34; A61K8/43; A61K8/49; A61Q17/00; A61Q19/00; A61Q19/10] [First publ.: 2014-02-07] [Score: 6.711] [Normalized score: 1.000]

an essential oil active selected from thymol, eugenol, menthol, geraniol, vertenone, eucalyptol, **pinocarvone**, cedrol, anethol, carvacrol, hinokitiol, berberine, ferulic acid, cinnamic acid, methyl salicylate, ... Preferred essential oil actives are thymol, eugenol, menthol, geraniol, vertenone, eucalyptol, **pinocarvone**, cedrol, anethol, carvacrol, hinokitiol, berberine, ferulic acid, cinnamic acid, methyl salicylate,

**2. METHOD OF EXTRACTING ACTIVE MOLECULES FROM NATURAL RESINS AND USE THEREOF**

Sem-IP.com

[OC DocID: 90302044] [Patent publ. date: 2014-01-23] [Patent ID: US2014023721A1] [Patent classifications IPC: A23L1/30; A61K35/64; A61K36/28; A61K36/328; A61K36/45] [First publ.: 2010-12-10] [Score: 4.643] [Normalized score: 0.692]

14 Camphor 25000 ppm casnfora 15 E-chrystanthemyl acetate 23250 ppm E-cristantemil acetato 16 **Pinocarvone** 20 ppm **pinocarvone** 17 Borneol 90 ppm borneolo 18 Alpha terpineol 300 ppm alfa terpineolo 19 Linalool acetate 480 ppm linalolo acetato 20 Bornyl angelate 3750 ppm bornil angelate 21 **Pinocarvone** 20 ppm **pinocarvone** 22 Borneol 20 ppm borneolo 23 Citronellal hydrate 20 ppm citronellal idrato 24 Thymol 2450 ... 28000 32000 20000 E-chrystanthemyl acetate/ 23250 30000 420000 19500 E-cristantemil acetato **Pinocarvone**/unvaried 20 50 110 20 Borneol/borneolo 90 120 270 80 Alpha terpineol/alfa terpineolo 300 500 900

Export results list

Sort by  relevance  date

Filter by query concept distance (at least 2 query terms, no free text).

Max. distance: paragraph -  +

Filter by publication date

From  to

e.g. 2012 2013-12-31

Number of publications per year

Percentage of publications per year

Welcome LutzWeber! My ChemAnalyser Tips & tricks About Logout A- | A | A+

chemanalyser version 1.2.4.3 Search results

Search Expert search Structure search Search history Co-occurrences

infoapps (303) PubMed Central (37)

+OCID:"190000036336"

20 document(s) found.  Go to page

(1 of 2)

**1. Compositions and methods for treating hoof diseases**

Sem-IP.com

[OC DocID: 85671153] [Patent publ. date: 2013-03-05] [Patent ID: US8389581B2] [Patent classifications IPC: A61K31/11] [First publ.: 2009-11-25] [Score: 2.238] [Normalized score: 1.000]

selected from the group consisting of thymol, eugenol, menthol, geraniol, verbenone, eucalyptol, **pinocarvone**, cedrol, anethol, carvacrol, hinokitiol, berberine, ferulic acid, cinnamic acid, methyl salicylic ... in geranium and rose, citronella), verbenone (present for example in vervain), eucalyptol and **pinocarvone** (present in eucalyptus), cedrol (present for example in cedar), anethol (present for example in

**2. COMPOSITIONS AND METHODS FOR TREATING HOOF DISEASES**

Sem-IP.com

[OC DocID: 31534973] [Patent publ. date: 2010-06-03] [Patent ID: US2010137451A1] [Patent classifications IPC: A61K31/11; A61P31/00] [First publ.: 2009-11-25] [Score: 2.238] [Normalized score: 1.000]

selected from the group consisting of thymol, eugenol, menthol, geraniol, verbenone, eucalyptol, **pinocarvone**, cedrol, anethol, carvacrol, hinokitiol, berberine, ferulic acid, cinnamic acid, methyl salicylic ... in geranium and rose, citronella), verbenone (present for example in vervain), eucalyptol and **pinocarvone** (present in eucalyptus), cedrol (present for example in cedar), anethol (present for example in

**3. Hard surface cleaner compositions of sulfonated estolides and other derivatives of fatty acids and uses thereof**

Export results list

Sort by  relevance  date

Filter by query concept distance (at least 2 query terms, no free text).

Max. distance: paragraph -  +

Filter by publication date

From  to

2009 2009

Number of publications per year

Percentage of publications per year



**Search term "tomatidine"** (a class of natural products for food):

ChemAnalyser            664 patent hit documents

SciFinder                57 patent hit documents

**Search term "sesquiterpenes"** (a class of natural products for food):

ChemAnalyser            318,707 patent hit documents

SciFinder                1,837 patent hit documents

**Search terms "food additives" + "natural products" + "triterpenes"**

ChemAnalyser            362,406 patent hit documents

SciFinder                33 all hit documents

## Search term "neodymium iron boron magnet":

|              |                             |
|--------------|-----------------------------|
| ChemAnalyser | 25,616 patent hit documents |
| SciFinder    | 0 patent hit documents      |

## Search term "NdFeB":

|              |                             |
|--------------|-----------------------------|
| ChemAnalyser | 25,616 patent hit documents |
| SciFinder    | 6 patent hit documents      |

## Search term "304 stainless steel":

|              |                             |
|--------------|-----------------------------|
| ChemAnalyser | 14,578 patent hit documents |
| SciFinder    | 2,487 patent hit documents  |

## Search term "cosmetic formulation" + "glyzyrrhizin":

|              |  |
|--------------|--|
| ChemAnalyser | 51,024 patent hit documents  |
| SciFinder    | No References were found containing all of the concepts "cosmetic", "formulation" and "glyzyrrhizin" |

# Thanks!

## Special thanks to

**Stephen Boyer @IBM**

**Evan Bolton @NCBI**

**Yannick Djoumbou Feunang @ualberta.ca**

**Claudia Bobach & Anett Püschel @ontochem.com**

## please contact us for

- *working on your cognitive Big Data computing project*
- *test account to ChemAnalyser®*
- *supplying you with custom ontologies and tools*

visit [www.chemanalyser.com](http://www.chemanalyser.com)

