

Wikidata and Scholia as hub linking metabolite knowledge

Egon Willighagen

ORCID:0000-0001-7542-0286

chem-bla-ics.blogspot.com

T: @egonwillighagen

M: @egonw@scholar.social

BeNeLux Metabolomics Days

#NMCDays

Rotterdam, 2018-08-19

CC-BY 4.0 Int. (except slides with ©)



Maastricht University



Acknowledgements

- Finn Nielsen (Scholia inventor)
- Denise Slenter (BiGCaT PhD candidate)
- Others
 - Various Maastricht University research groups
 - EPA CompTox Dashboard: Tony Williams
 - MetaboLights team: Reza Salek and Chandu Venkata
 - ChEBI team: Christoph Steinbeck (now Jena), Gareth Owen
 - PubChem, WikiGenomes teams: Evan Bolton, Gang Fu, Sebastian Burgstaller, Andra Waagmeester (Micelio)
 - SPLASH team: Gert Wolgemuth, Sajjan Singh Mehta
 - Wikidata & WikiCite teams: Daniel Mietchen, Dario Tataborelli
 - Wikidata:WikiProject Chemistry
 - Reactome: Robin Haw, Henning Hermjakob



WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research

Denise N Slenker, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles ... [Show more](#)



[View Metrics](#)

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D661–D667, <https://doi.org/10.1093/nar/gkx1064>

Published: 10 November 2017 [Article history](#)

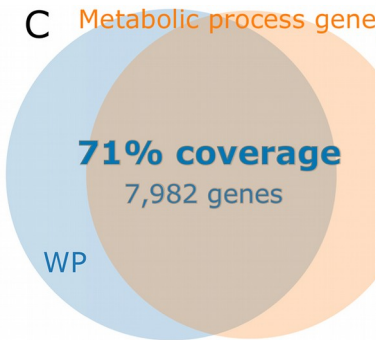
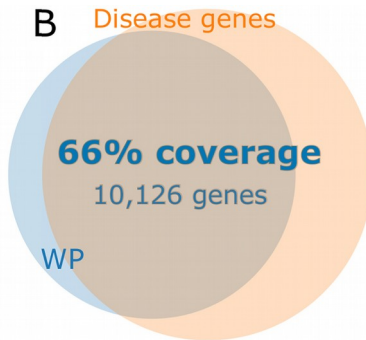
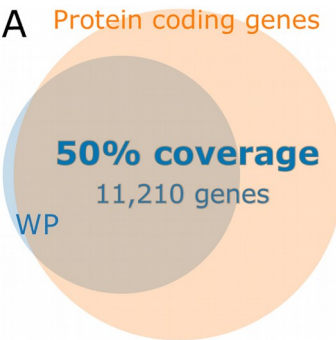
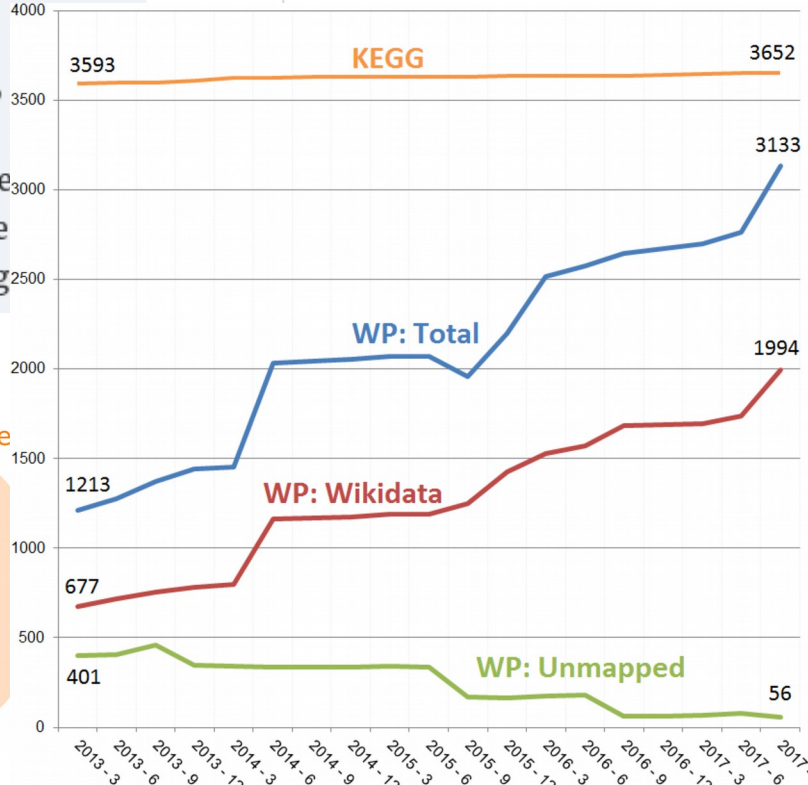
[Views](#) [PDF](#) [Cite](#) [Permissions](#) [Share](#)

Email alerts

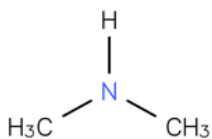
- [New issue alert](#)
- [Advance article alerts](#)
- [Article activity alert](#)

Abstract

WikiPathways (wikipathways.org) captures the collective knowledge represented in biological pathways. By providing a database in a curated, machine readable way, omics data analysis and visualization is enabled. WikiPathways and other pathway databases are used to analyze experimental data by research groups in many fields. Due to the open and collaborative nature of the WikiPathways platform, our content keeps growing and is g



One ID in the pathway, many in the popup



MOL SMILES InChIKey

WikiPathways KEGG Pathways Reactome Pathways

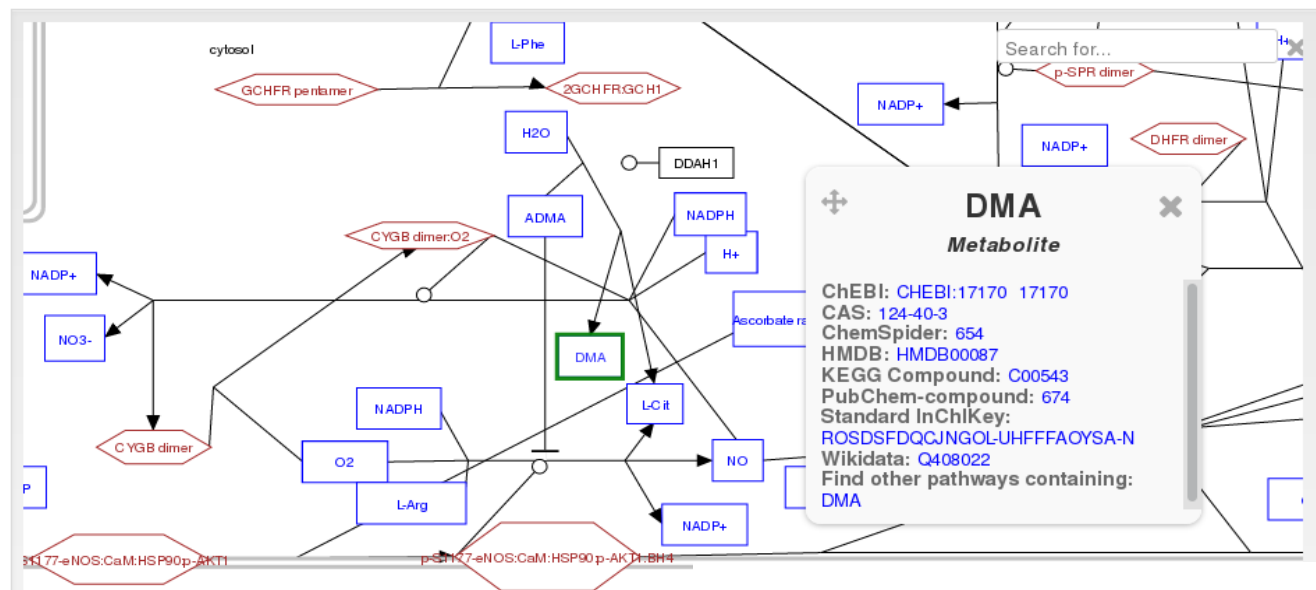
Select Species

Homo sapiens

Select Pathways

Metabolism of nitric oxide

This metabolite has been identified in the following MetaboLights studies. [MTBLS100](#) [MTBLS172](#)



Databases & identifiers

- HMDB: **H**uman **M**etabolome **D**atabase
- ChEBI: Database of **C**hemicals **E**ntities of **B**iological Interest
- ChemSpider, PubChem
- CAS: **C**hemical **A**bstracts **S**ervice
- InChI: **I**nternational **C**hemical **I**dentifier
 - UniChem, ...



EMBL-EBI
ChEBI
Home | Advanced Search | Browse | Documentation | Download | Tools | About ChEBI

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on biology.

Search for    only | All in ChEBI 

Example: COc1ccc(O)cc1, water
Advanced Search | About ChEBI

Documentation
Tutorial: An introduction to the ChEBI database and ontology, showing users how to search and browse the web/programmatic interface.

Downloads
SDF files ChEBI provides its chemical structures and additional data in structure-data file (SDF) format: CHEBI 4321
Ontology files ChEBI ontology is provided in



PubChem

BioAssay  Compound  Substance 

Go Limits Advanced

Try the new PubChem Search

 A revamped PubChem Compound Summary page is now released. [Read more...](#) more ... 

[Write to Helpdesk](#) | [Disclaimer](#) | [Privacy Statement](#) | [Accessibility](#) | [Data Citation Guidelines](#)
National Center for Biotechnology Information
NLM | NIH | HHS



- BioActivity Summary
- BioActivity Databale
- BioActivity SAR
- BioActivity DataCler
- Structure Search
- 3D Conformer Tools
- Structure Clustering
- Classification
- Upload
- Download
- PubChem FTP

So, what IDs are used in WikiPathways?

2017

datasource	numberEntries
ChEBI	1923
HMDB	623
CAS	299
KEGG Compound	251
PubChem-compound	245
Chempider	174
PubChem-substance	33
LIPID MAPS	10
Reactome	4
Wikidata	3
ChEMBL compound	2

2015

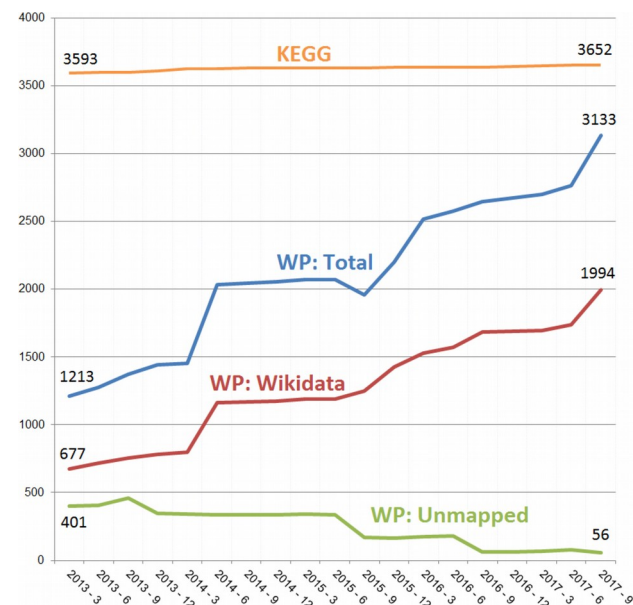
source	count
HMDB	569
ChEBI	496
KEGG Compound	408
CAS	293
PubChem-compound	217
Chempider	156
PubChem-substance	24
LIPID MAPS	11
Wikipedia	9
ChemIDplus	7
Reactome	4
ChEMBL compound	2
Other	1
CTD Chemical	1
ChemSpider	1

2012

source	count
HMDB	522
Kegg Compound	389
CAS	267
ChEBI	244
Entrez Gene	136
PubChem-compound	108
Chempider	15
Wikipedia	11
PubChem-substance	8
ChemIDplus	7
ChEMBL compound	2
3DMET	1
LIPID MAPS	1

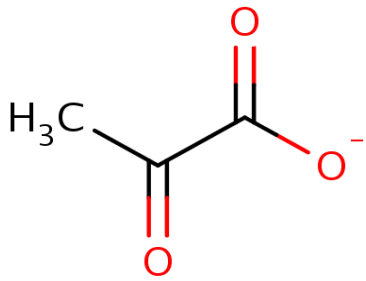
Curated subset

+ Reactome

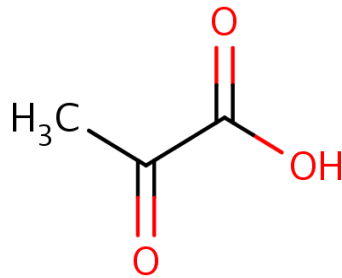


Expression chemistry in metabolic pathways

CHEBI:15361 (Pyruvate) -> Ce:CHEBI:32816 (conjugate) -> Ck:C00022 -> [WP2456 HIF1A and PPARG regulation of glycolysis, WP2453 TCA Cycle and PDHc]

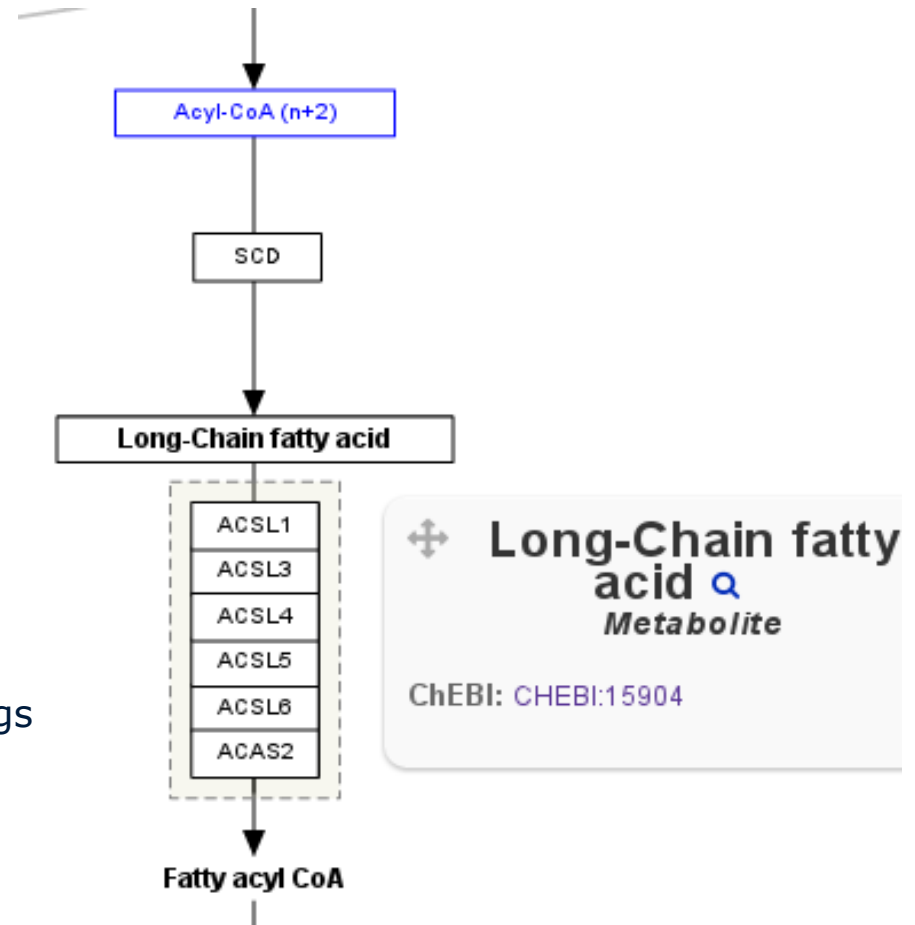


CHEBI:15361



CHEBI:32816

Brenninkmeijer, CYA, et al. "Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision." Proceedings of 2nd International Workshop on Linked Science. 2012.



What if we had more metabolite info?

- more CAS registry numbers?
- more physchem properties?
- more literature references?





Item [Discussion](#)

[Read](#) [View history](#) [☆](#) [More](#)



We need your opinion!
Help the Wikimedia Foundation determine directions to take in its work.

L-Leucine (Q483745)

amino acid

[edit](#)

Leu | L | (2S)-alpha-2-Amino-4-methylvaleric acid | (2S)-alpha-Leucine | L-Leucin | L-Leuzin | Leucine | (2S)-2-Amino-4-methylpentanoic acid | (S)-(+)-Leucine | (S)-Leucine | 2-Amino-4-methylvaleric acid

[In more languages](#)

Language	Label	Description	Also known as
English	L-Leucine	amino acid	Leu L (2S)-alpha-2-Amino-4-methylval... (2S)-alpha-Leucine L-Leucin L-Leuzin Leucine (2S)-2-Amino-4-methylpentanoic... (S)-(+)-Leucine (S)-Leucine 2-Amino-4-methylvaleric acid
German	Leucin	chemische Verbindung	Isobutylglycin Leuzin
French	Leucine	composé chimique	328-39-2 E641 L-Leucine
Dutch	Leucine	chemische stof	61-90-5 Leukine

Wikipedia (54 entries) [edit](#)

ar	لوسين	[ref]
be	Лейцын	[ref]
bg	Левцин	[ref]
bs	Leucin	[ref]
ca	Leucina	[ref]
cs	Leucin	[ref]
da	Leucin	[ref]
de	Leucin	[ref]
el	Λευκίνη	[ref]
en	Leucine	[ref]
eo	Leŭcino	[ref]
es	Leucina	[ref]
et	Leutsiin	[ref]
eu	Leuzina	[ref]
fa	لوسين (مواد)	[ref]
fi	Leusiini	[ref]

Mietchen, D. et al. Enabling open science: Wikidata for research (Wiki4R). Research Ideas and Outcomes 1, e7573+ (2015)

Wikidata: external (database) identifiers

Identifiers

Freebase ID [/m/0h1wg](#)
▶ 1 reference

ChEBI ID [57427](#)
▶ 1 reference

stated in	ChEMBL
ChEMBL ID	CHEMBL291962
language of work or name	English
title	L-Leucine (English)
retrieved	19 January 2016

ChEMBL ID [CHEMBL291962](#)
▶ 1 reference



[Main page](#)
[Community portal](#)
[Project chat](#)
[Create a new item](#)
[Item by title](#)
[Recent changes](#)
[Random item](#)
[Query Service](#)
[Nearby](#)
[Help](#)
[Donate](#)

Tools
[What links here](#)
[Related changes](#)

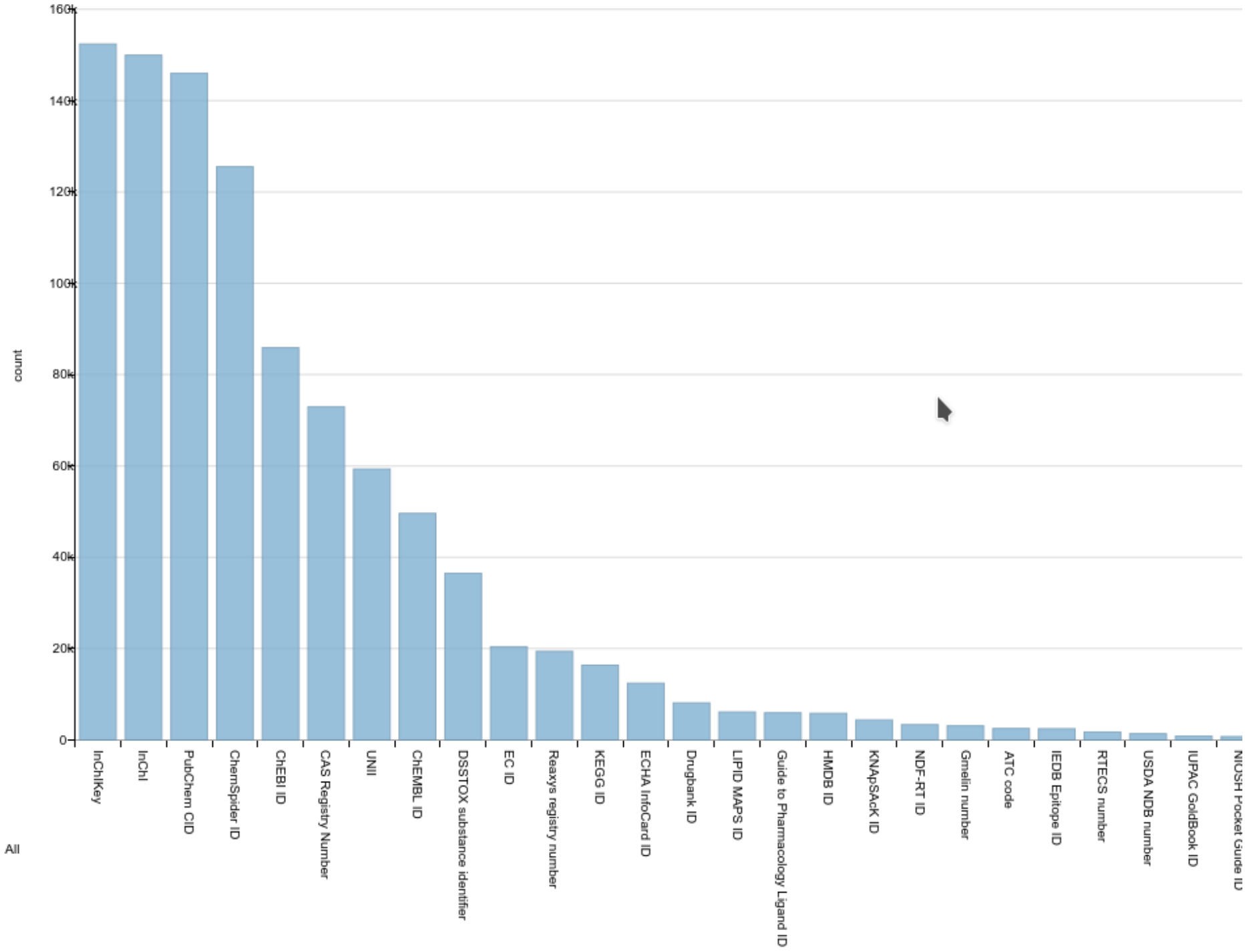
Property [Discussion](#)

DSSTOX substance identifier (P3117)

DSSTox substance identifier used in the Environmental Protection Agency (DTXSID)

▼ [In more languages](#)

Language	Label	Description
English	DSSTOX substance identifier	DSSTox substar Environmental F Dashboard
German	DSSTOX-Identifikator	No description c
French	No label defined	No description c
Dutch	No label defined	No description c
Swedish	No label defined	No description c



QuickStatements: scriptable adding of content

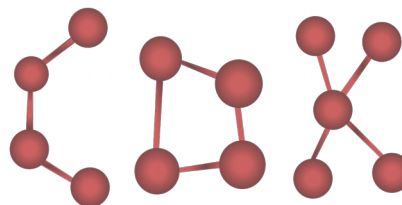
The screenshot shows the Bioclipse software interface. On the left is the Bioclipse Navigator with a tree view of project folders. The central pane contains a script editor with the following code:

```
1// critical input and output
2// smiFile = "/Wikidata/wp.smi"
3smiFile = "/Recon/MODEL1603150001.smi"
4// "/CompToxDash/Polychlorinated_biphenyls.smi"
5// "/Wikidata/input.smi"
6// smiFile = "/Wikidata/wp2356.smi"
7qsFile = "/Wikidata/output.quickstatements"
8reconFile = "/Wikidata/recon2wikidata.txt"
9
10// if all SMILES come from the same paper, enter the Wikidata item code
11// on the next line, e.g. paperQ = "Q22570477". It will be used as reference
12// to some of the information
13paperQ = null
14
15// a helper function
16def upgradeChemFormula(formula) {
17  formula = formula.replace("0","");
18  formula = formula.replace("1","");
19  formula = formula.replace("2","");
20  formula = formula.replace("3","");
21  formula = formula.replace("4","");
22  formula = formula.replace("5","");
23  formula = formula.replace("6","");
24  formula = formula.replace("7","");
25  formula = formula.replace("8","");
26  formula = formula.replace("9","");
27}
28def renewFile(file) {
29  if (ui.fileExists(file)) ui.remove(file)
30  ui.newFile(file)
31  return file
32}
33
34// reset the output (SD file + QuickStatements)
35renewFile(qsFile)
36mols = cdk.createMoleculeList()
```

On the right, a table titled 'wikipathways.sdf' displays chemical structures in a grid. The table has columns for '2D-structure', 'URL187_TI...', and 'URL212...'. The first six rows show different chemical structures, each with a corresponding number in the first column.

	2D-structure	URL187_TI...	URL212...
1			
2			
3			
4			
5			
6			

Spjuth, O. et al.,
2007, BMC
Bioinformatics



QuickStatements: scriptable adding of content, here, the SPLASH

convert2rdf.groovy

wikidataMashup.groovy

mona-identifier-table.csv

```
1 ID, SPLASH, InChIKey, SMILES
2 AU100601, splash10-0a4i-1900000000-d2bc1c887f6f99ed0f74, QKLPUVXBJHRFQZ-UHFFFAOYSA-N, C=1C=C(C=CC1N)S(N=C2C=NC=C(C1)N2)(=O)=O
3 AU100701, splash10-0a4i-1900000000-d2bc1c887f6f99ed0f74, XOXHILFPYWFOD-UHFFFAOYSA-N, C=1C=C(C=CC1N)S(NC=2C=CC(C1)=NN2)(=O)=O
4 AU100801, splash10-0pk9-2970000000-abb9e31dc053c1d21884, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
5 AU100802, splash10-0zi0-0590000000-2e8e1e943731aa842e16, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
6 AU100803, splash10-0kmi-0950000000-9e91e580b580594989a0, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
7 AU100804, splash10-00di-0920000000-216c25f077b257c6979d, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
8 AU100805, splash10-00di-0900000000-127b8b6e862f77ce8c13, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
9 AU100806, splash10-004i-0090000000-4841dadd912e95b736ce, ASWVTGNCAZCZNNR-UHFFFAOYSA-N, CC1=CC(C)=NC(=N1)NS(C2=CC=C(C=C2)N)(=O)=O
10 AU100902, splash10-0bt9-0905000000-78aa6da956c32ba36241, ZZORFUFYDOWNEF-UHFFFAOYSA-N, COC=1C=C(N=C(N1)OC)NS(C2=CC=C(C=C2)N)(=O)=O
11 AU100903, splash10-0a4i-0910000000-7b5873eb9276bcb413be, ZZORFUFYDOWNEF-UHFFFAOYSA-N, COC=1C=C(N=C(N1)OC)NS(C2=CC=C(C=C2)N)(=O)=O
12 AU100904, splash10-0k96-0910000000-24de656de5a8714caacb, ZZORFUFYDOWNEF-UHFFFAOYSA-N, COC=1C=C(N=C(N1)OC)NS(C2=CC=C(C=C2)N)(=O)=O
13 AU100905, splash10-0udl-0900000000-cf83e971f45b0faf66d2, ZZORFUFYDOWNEF-UHFFFAOYSA-N, COC=1C=C(N=C(N1)OC)NS(C2=CC=C(C=C2)N)(=O)=O
14 AU101001, splash10-0a4i-2901000000-8862dd6e93b57f07349b, PJSFRIWCGOHTNF-UHFFFAOYSA-N, COC1=C(N=CN=C1OC)NS(C2=CC=C(C=C2)N)(=O)=O
15 AU101101, splash10-0udi-0090000000-1bc83eabade65dc8953a, SEEPANYCNGTZFQ-UHFFFAOYSA-N, C1=CN=C(N=C1)NS(C2=CC=C(C=C2)N)(=O)=O
16 AU101102, splash10-0a4i-0910000000-bd163c77861c5a8afa49, SEEPANYCNGTZFQ-UHFFFAOYSA-N, C1=CN=C(N=C1)NS(C2=CC=C(C=C2)N)(=O)=O
17 AU101103, splash10-0a4i-0900000000-260a572b5c7f3c070c26, SEEPANYCNGTZFQ-UHFFFAOYSA-N, C1=CN=C(N=C1)NS(C2=CC=C(C=C2)N)(=O)=O
18 AU101104, splash10-052r-0900000000-580f983ba25c4cb2eb20, SEEPANYCNGTZFQ-UHFFFAOYSA-N, C1=CN=C(N=C1)NS(C2=CC=C(C=C2)N)(=O)=O
19 AU101201, splash10-0a4i-1900000000-95dc6d24a8
20 AU101302, splash10-0a4i-3900000000-23eed4bc89
21 AU101501, splash10-001i-0090000000-eef21c4c7e
```

QuickStatements: Magnus Manske, WT Sanger Institute in Cambridge

SPLASH: G. Wolhgemuth, 10.1038/nbt.3689

mona-identifier-table.csv

convert.groovy

mappings.txt

1	Q55167836	P4964	"splash10-0f89-0950000000-117ad116d5ac14a5be9c"
2	Q55167836	P4964	"splash10-001i-0900000000-507c3f3bafb08a0393f6"
3	Q55167836	P4964	"splash10-001i-0900000000-daa82fab29c3bc26425a"
4	Q55167836	P4964	"splash10-0g5i-0900000000-56c0d97c8ede7a0f5121"
5	Q4596780	P4964	"splash10-00di-4900000000-73bfbb6a2b2ba59e4f4f"
6	Q4596780	P4964	"splash10-00di-0900000000-2370c00b72994e6aa54f"
7	Q11751639	P4964	"splash10-00di-0910000000-74bd5c8746b39b3c0420"
8	Q11751639	P4964	"splash10-014i-0900000000-7ac1f98cc9b9752a76d6"
9	Q11751639	P4964	"splash10-014i-0900000000-f5c899766b4d912df5ce"
10	Q11751639	P4964	"splash10-014i-0900000000-5e5ee374be357b4e26a4"
11	Q55177068	P4964	"splash10-0059-0900000000-d1e136c596eda61fdad6"
12	Q55177068	P4964	"splash10-001i-1900000000-d9c4b6edb79ec7db500c"
13	Q55177068	P4964	"splash10-000i-9500000000-ec02387ba0107a9f05cb"
14	Q55177068	P4964	"splash10-000i-9100000000-c1e7be64cc9ef25c1291"
15	Q55177068	P4964	"splash10-0006-9000000000-d90115a5ff29ff776135"
16	Q11751639	P4964	"splash10-0gb9-0900000000-060dd1cb50610a20cf7a"



Maastricht University



More identifiers: EPA CompTox Dashboard, LIPID MAPS, PDB ligand identifiers

DSSTOX substance identifier

DTXSID30678817

▼ 1 reference

stated in

Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EPA CompTox Dashboard

PDB ligand ID

ACY



Egon Willigh@gen
@egonwillighagen



ok, we had 2333 @lipidmaps identifiers in @wikidata this morning... the current count is 6099 :) tinyurl.com/y8g3zwas #endocannabinoidSaturday

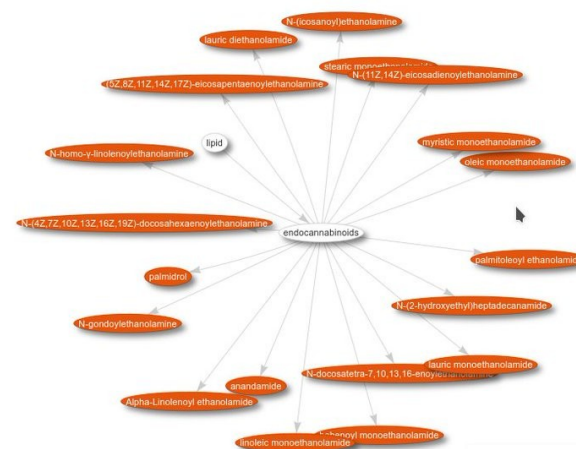
11:05 PM - Jun 30, 2018



See Egon Willigh@gen's other Tweets

endocannabinoids (Q55282178)

Class Hierarchy



Maastricht University



Scholia: visualizing data

[Scholia](#) [Author](#) [Work](#) [Organization](#) [Location](#) [Event](#) [Award](#) [Topic](#) [Tools](#) [Help](#)

Search

Search for a scientist, paper, organization, venue, event, topic, etc.

Examples

Profiles

[Denny Vrandečić](#)

View the researcher profile for the Semantic Web researcher Denny Vrandečić. It shows his papers, co-authors, etc.

[Technical University of Denmark](#)

View the profile for an organization: People associated with the organization, their publications, the co-author patterns, etc.

[NeuroImage](#)

View information about a venue, e.g., a scientific journal or scientific conference. Here, the *NeuroImage*

Comparisons

Scholia can show multiple items together.

[Technical University of Denmark and UCL](#)

Compare two or more organizations. Here a comparison between two universities with collaborating researchers, number of publications and citations.

[Tim Berners-Lee, James Hendler and Ruben Verborgh](#)

Compare three Semantic Web researchers.

Redirects

If you know the external identifier of a concept, then Scholia can make a lookup based on it:

[twitter/utafrih](#)

Look up by Twitter username @utafrih. This will identify the London-based researcher Uta Frith and redirect to her Scholia page.

[twitter/mitpress](#)

Redirect also works for organizations, here MIT Press

[orcid/0000-0001-7542-0286](#)

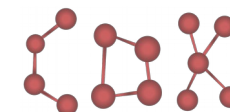
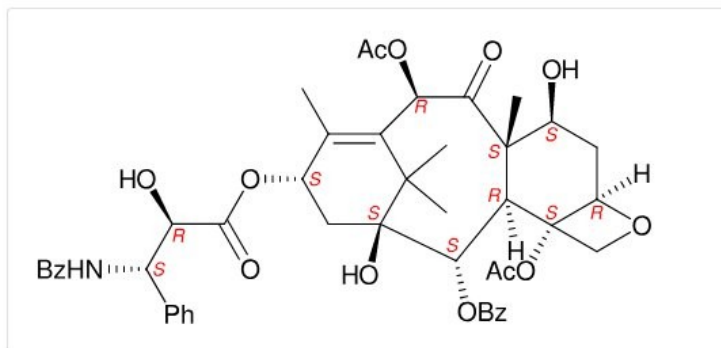
Lookup 0000-0001-7542-0286 that is



Scholia: compounds

paclitaxel (Q423762)

Paclitaxel (PTX), sold under the brand name Taxol among others, is a chemotherapy medication used to treat a number of types of cancer. This includes ovarian cancer, breast cancer, lung cancer, Kaposi sarcoma, cervical cancer, and pancreatic cancer. It is given by injection into a vein. ... (from the [English Wikipedia](#))



Identifiers

Show entries

Search:

IDpred	Id
ATC code	L01CD01
CAS Registry Number	33069-62-4

Show entries

Search:

Mol	InChIKey	CAS	ChemSpider	PubChem CID
acetic acid	QTBSBXVTEAMEQO-UHFFFAOYSA-N	64-19-7	171	176
deuterated acetic acid	QTBSBXVTEAMEQO-GUEYOVJQSA-N	1186-52-3	2006083	2723903
acetic acid c-14	QTBSBXVTEAMEQO-HQMMQRPSA-N	2845-03-6	144444	164769
acetic acid c-13	QTBSBXVTEAMEQO-VQEHIDDOSA-N	1563-79-7	8329490	10153982
acetic acid c-11	QTBSBXVTEAMEQO-JVVVGQRLSA-N	78887-71-5	396653	450349
acetate ion	QTBSBXVTEAMEQO-UHFFFAOYSA-M	71-50-1	170	175

[Edit on query.Wikidata.org](#)

Showing 1 to 6 of 6 entries

Scholia: visualizing data (compound classes)

Scholia Author Work Organization Venue Series Publisher Sponsor Award Topic Tools About

paraben ([Q410780](#))

Parabens are a class of widely used preservatives in cosmetic and pharmaceutical products. Chemically, they are a series of parahydroxybenzoates or esters of parahydroxybenzoic acid (also known as 4-hydroxybenzoic acid). Parabens are effective preservatives in many types of formulas. ... (from the English Wikipedia)

Class Hierarchy



Scholia: physical-chemical properties

Physchem Properties

Show entries

Search:

PropEntity	Value	Units	Qualifiers	Source	Doi
acid dissociation constant	4.74	1		Small Scale Determination of the pKa Values for Organic Acids	10.1021/ED071PA6
mass	60.021129	atomic mass unit		PubChem	
acid dissociation constant	4.756	1	temperature: 25	CRC Handbook of Chemistry and Physics (95th edition)	
boiling point	117.9	degrees Celsius	pressure: 101325	CRC Handbook of Chemistry and Physics (95th edition)	
density	1.0446	gram per cubic centimetre	temperature: 25	CRC Handbook of Chemistry and Physics (95th edition)	



Scholia: keeping up with literature

Recently published works on the chemical

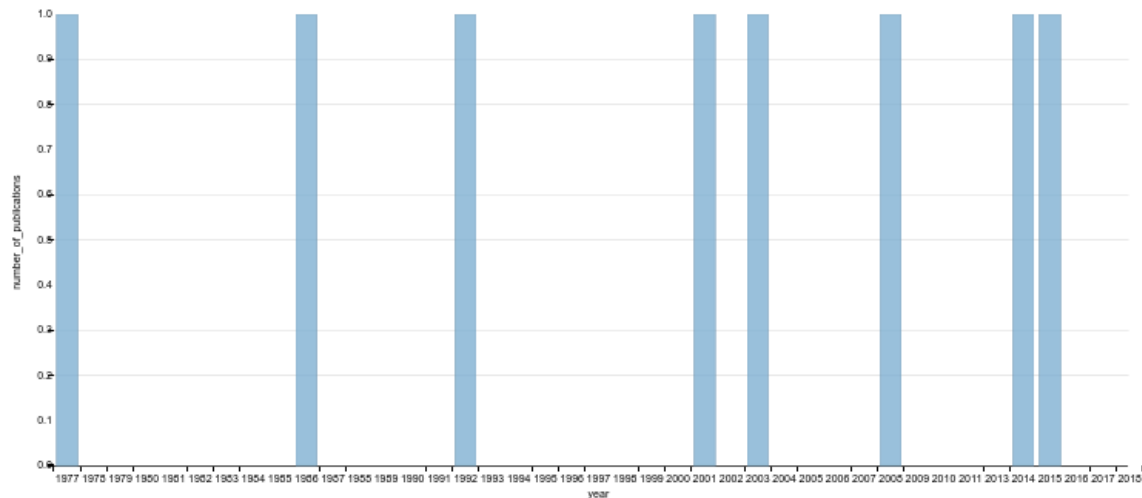
Search:

Date	Work	Type	Topics
2015-11-25	Metabolism and elimination of methyl, iso- and n-butyl paraben in human urine after single oral dosage.	scientific article	butylparaben // isobutylparaben // methylparaben // metabolism
2014-09-17	Overdosage of methylparaben induces cellular senescence in vitro and in vivo.	scientific article	drug overdose // methylparaben
2008-07-01	Paraben esters: review of recent studies of endocrine toxicity, absorption, esterase and human exposure, and discussion of potential human health risks.	scientific article	paraben // toxicity
2003-01-01	Propylparaben: physical characteristics.	scientific article	propylparaben
2001-06-01	Safety assessment of propyl paraben: a review of the published literature.	scientific article	propylparaben
1992-09-01	Methylparaben and propylparaben do not alter cerebral blood flow in humans.	scientific article	Homo sapiens // blood flow // propylparaben // methylparaben
1986-10-01	The toxicological implications of the interaction of butylated hydroxytoluene with other antioxidants and phenolic chemicals	scientific article	eugenol // ferulic acid // guaiacol // Butylated hydroxyanisole // arachidonic acid // methylparaben // butylated hydroxytoluene // hydrogen peroxide // vanillin
1977-05-01	Methylparaben--an overlooked cause of local anesthetic hypersensitivity.	scientific article	hypersensitivity // methylparaben

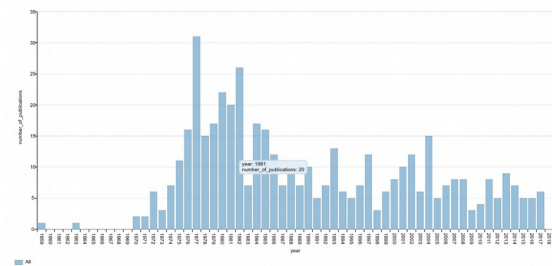
[Edit on query.Wikidata.org](#)

Showing 1 to 8 of 8 entries

Publications per year



Publications per year



Wikidata and Scholia as hub linking metabolite knowledge

Identifiers

Show entries

IDpred	Id
ATC code	G01AD02
ATC code	S02AA10
Beilstein Registry Number	506007
CAS Registry Number	64-19-7
ChEBI ID	15366
ChEMBL ID	CHEMBL539
ChemSpider ID	171
CosIng number	31572
DSSTOX substance identifier	DTXSID5024394
Drugbank ID	03166

[Edit on query.Wikidata.org](#)

Showing 1 to 10 of 30 entries

Redirecting

If you know the identifier then Scholia can make a lookup based on the identifier:

[cas/50-00-0](#)

Lookup CAS 50-00-0. This will identify formaldehyde and redirect to its Scholia page.

[inchikey/QTBSBXVTEAMEQO-UHFFFAOYSA-N](#)

Redirect also works for InChIKeys, here for acetic acid.

Previous 2 3 Next

Conclusions

- Wikidata + WikiCite + Scholia
 - Largest open data set that link CAS numbers to chemical structures
 - Literature is central
 - Integrates ontological relations with data
 - Large community around it
 - Powerful query service
 - FAIR by design

