Giving metabolites (and lipids) a chemical and biological context with open science

Egon Willighagen, BiGCaT team and collaborations
@egonw@social.edu.nl
0000-0001-7542-0286

NMetC, Turku 2024-08-26
doi:10.5281/zenodo.13373746

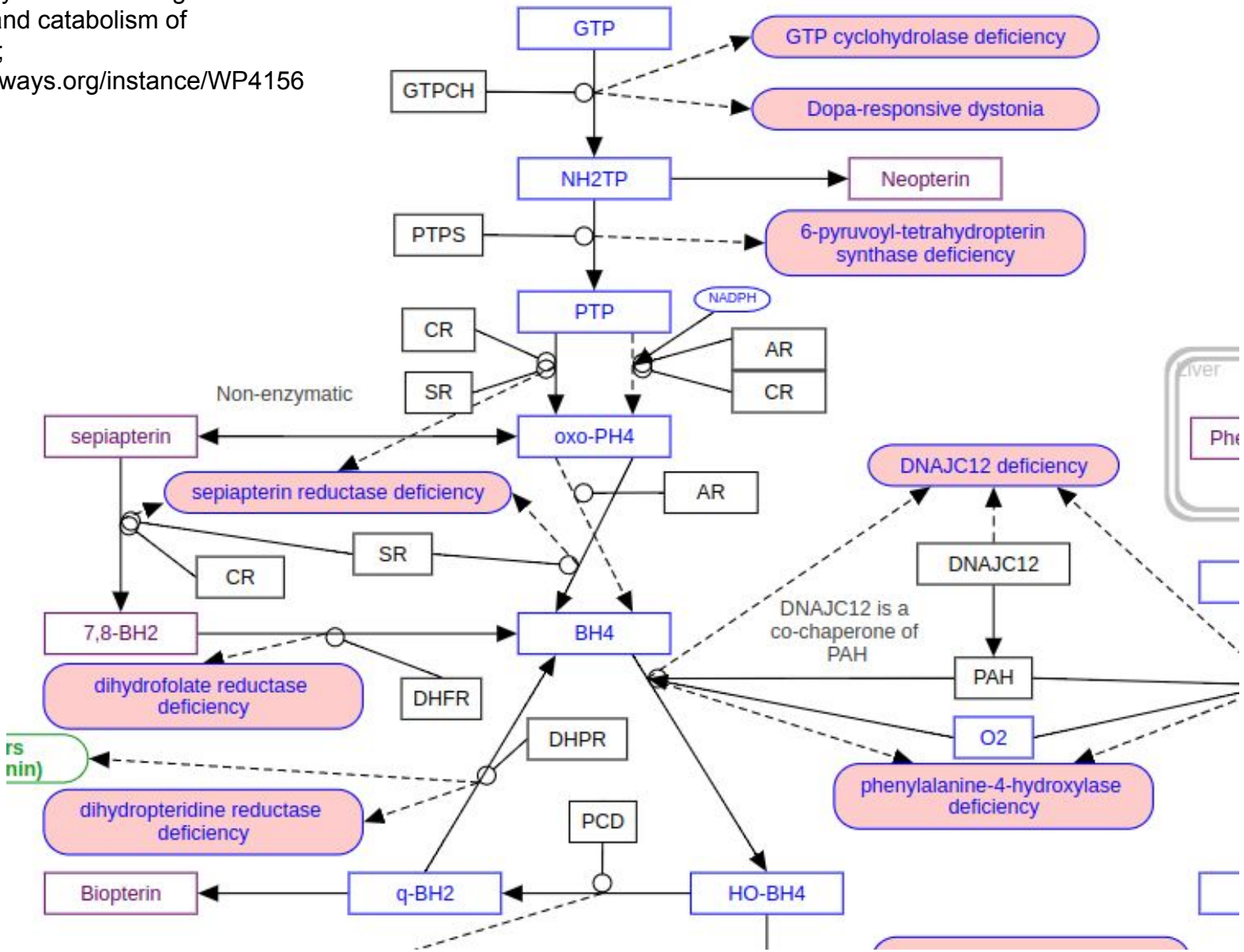
NUTRIM
School of Nutrition and Translational Research in Metabolism

Maastricht University | Maastricht UMC+

Abstract

A metabolite or lipid in vacuum is nice to look at, but when it comes into contact with other chemicals, things become interesting. However, finding and accessing knowledge about single chemicals is already not trivial, let alone that of multiple interacting chemicals. Maybe it was possible 100 years ago to know everything in your field, we now need to bridge many domains and many resources. This talk will sketch the roles open science is playing in giving metabolites and lipids their needed chemical and biological context. It will cover open source cheminformatics for linking metabolite and lipid databases (SMILES, InChI, CXSMILES), open standards for interoperability of knowledge (semantic web), open data to support metabolite and lipid identification (Wikidata and linked data), and linking knowledge via publications to their original research. Furthermore, it will explore mechanisms to describe the interactions these chemicals have in open biological pathway databases (WikiPathways).

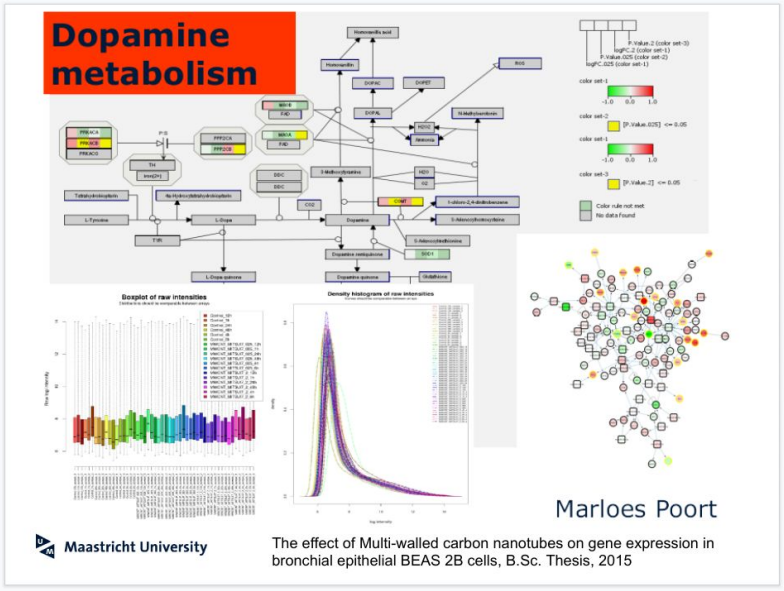
Slenter D, et al. Biosynthesis and regeneration of tetrahydrobiopterin and catabolism of phenylalanine. 2023; https://www.wikipathways.org/instance/WP4156



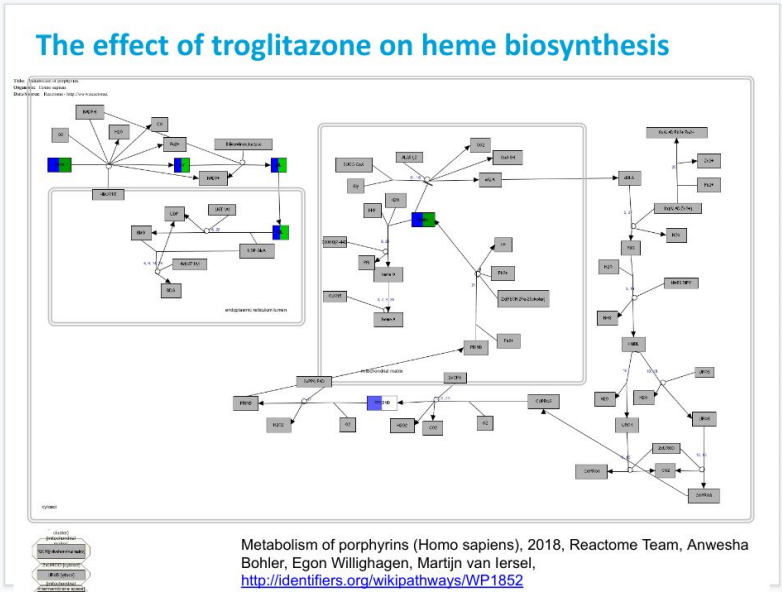
NUTRIM

Maastricht University | Maastricht UMC+

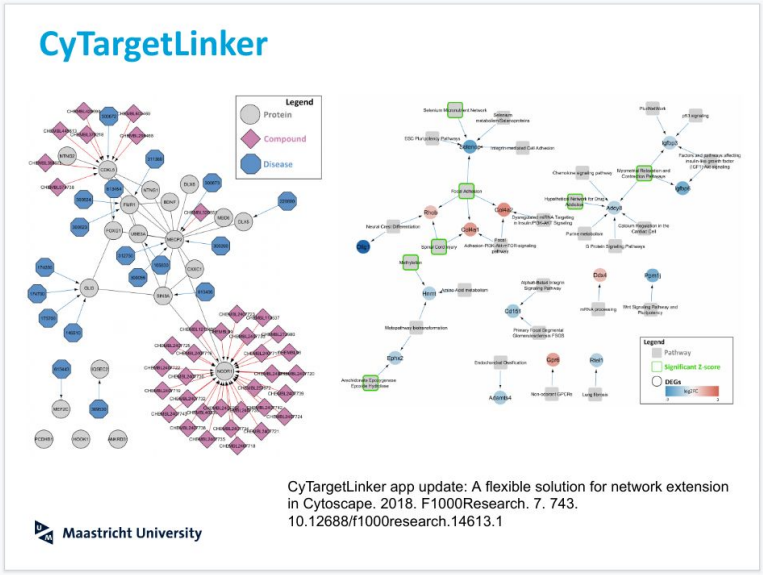
Integrative Systems Biology



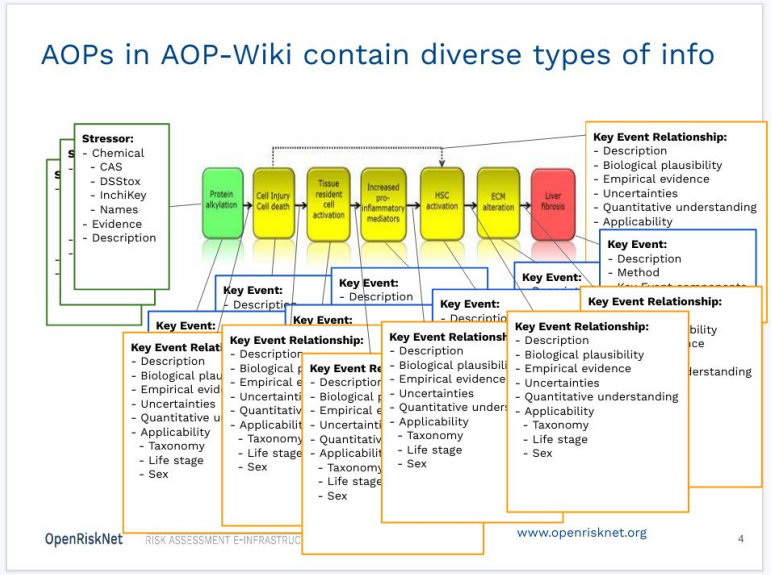
The effect of Multi-walled carbon nanotubes on gene expression in bronchial epithelial BEAS 2B cells, B.Sc. Thesis, 2015



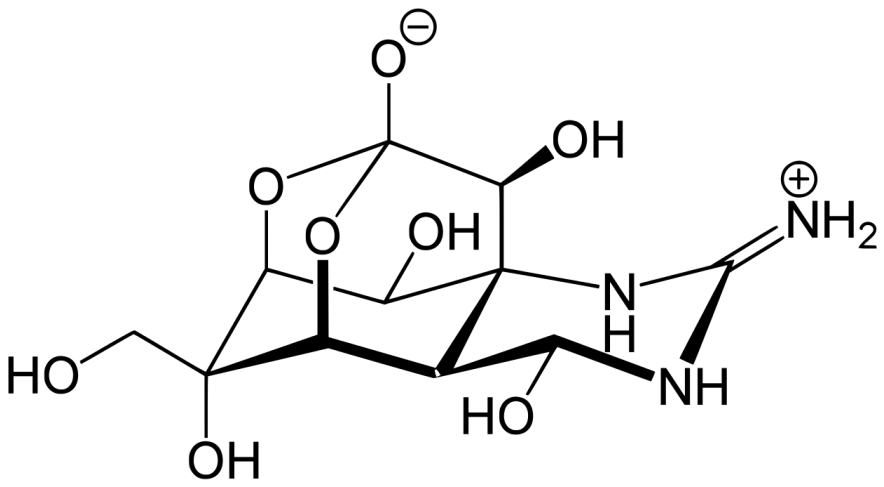
Metabolism of porphyrins (Homo sapiens), 2018, Reactome Team, Anwesha Bohler, Egon Willighagen, Martijn van Iersel, http://identifiers.org/wikipathways/WP1852



CyTargetLinker app update: A flexible solution for network extension in Cytoscape. 2018. F1000Research. 7. 743. 10.12688/f1000research.14613.1



A random compound…



- What exactly did we measure and what was it before the experiment started?
- Where do we find knowledge and/or can we make predictions?

NUTRIM

Maastricht University | Maastricht UMC+

Some concepts…

Theory behind modern data analysis and aspects essential to chemical and biological context

Semantics and Scientific Thinking

- **Meaning of concepts, of measured data**
  - Melting point
  - What is a “Metabolite”? Do we mean the same?
  - Implications on how to understand data and knowledge
- **Ontologies formalize meaning**
  - Name, definition, identifier
- **Representation**
  - Name, file format, data model, “shape”, …

Three types of computation

- **Statistics finds correlation (not causation)**
  - Same for machine learning, deep learning, AI
  - Numbers
  - Explainable AI?
- **Knowledge graphs**
  - Logic based: from facts to arguments
- **Algorithms (“Programming”)**
  - Solve physical equations
  - Converting representations
    - *Descriptors*, e.g. for predictive toxicology

Usable data and Explainable AI

- **Data exploration for *biological plausibility***
  - Multivariate statistics to find variance
    - Visualisation
  - Compare with existing knowledge
  - Find correlations, hopefully causation

- ***Garbage in, garbage out***
  - “Minimal reporting standards”
  - Experimental design
- ***Black boxes***

Open, FAIR, and used

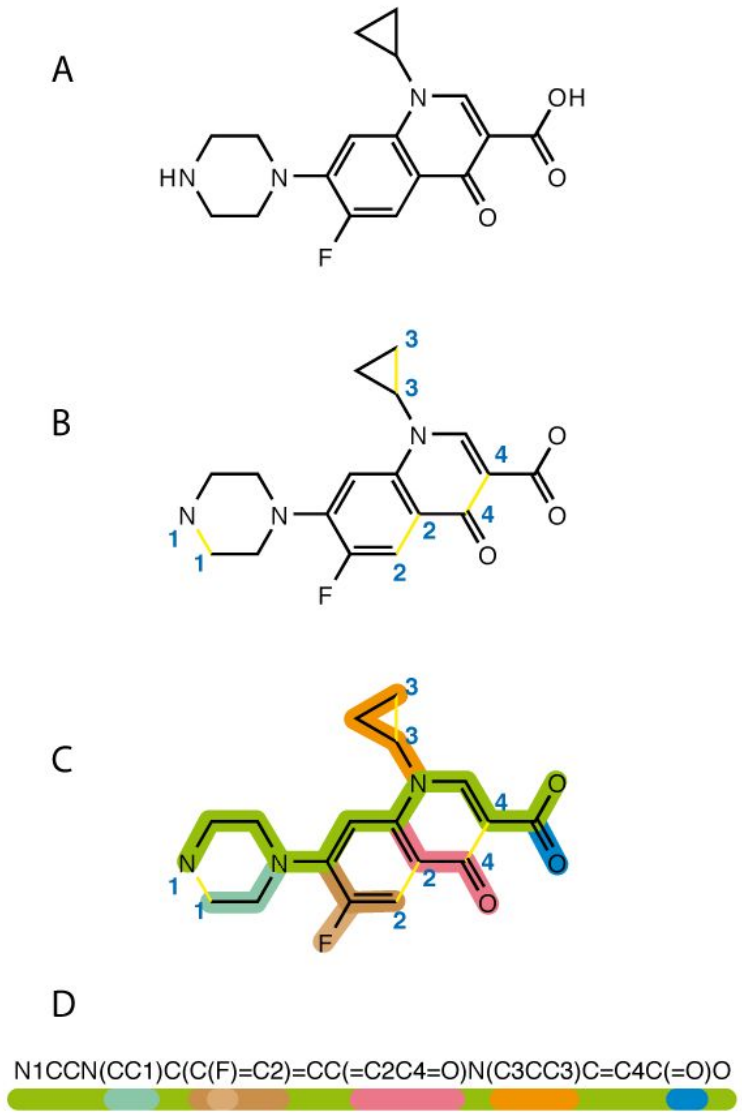
- Open
  - Not just theoretical reuse, real reuse
  - Needed rights: use, modify, reshare
- Findable, Accessible, Interoperable, Reusable
  - Interoperable: we have solutions for most
  - Reusable: it works. we need a lot more
- Used
  - How can we encourage reuse and resharing?

Being precise in your uncertainty

Cheminformatics

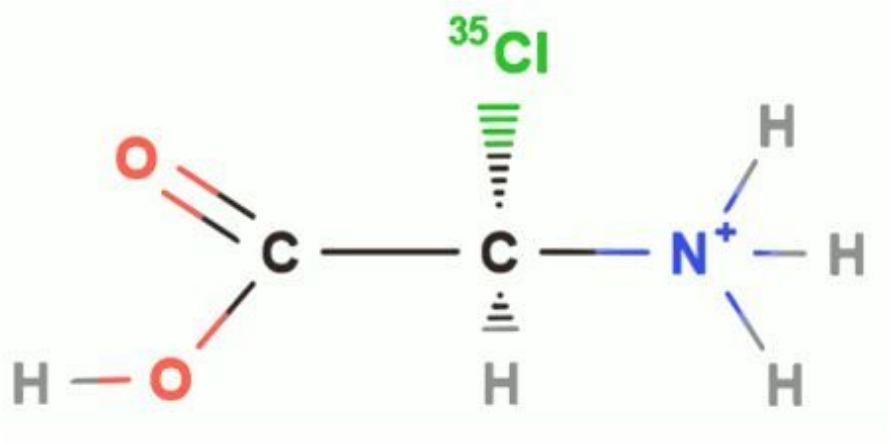
Identifiers: SMILES

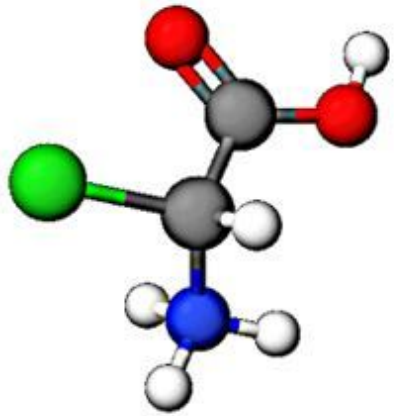
Simplified molecular-input line-entry system

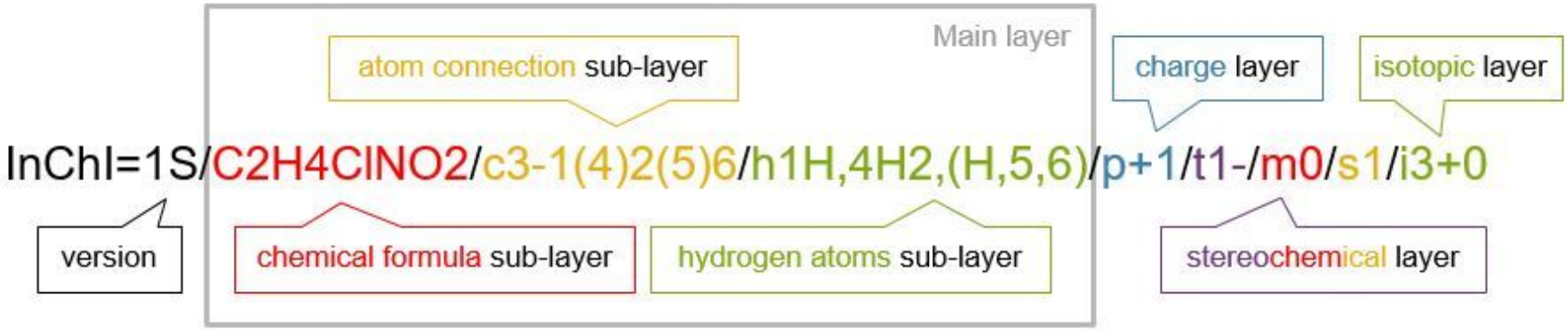


en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Identifiers: (Standard) InChI







https://chem.libretexts.org/Courses/Fordham_University/Chem1102%3A_Drug_Discovery_-_From_the_Laboratory_to_the_Clinic/05%3A_Organic_Molecules/5.08%3A_Line_Notation_(SMILES_and_InChI)

Journal of Cheminformatics

Home About Articles Submission Guidelines About The Editors Calls For Papers

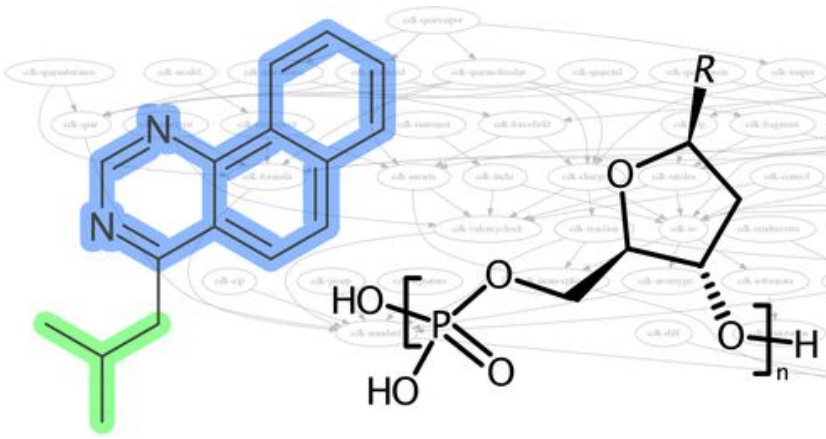
Software | Open Access | Published: 06 June 2017

The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching

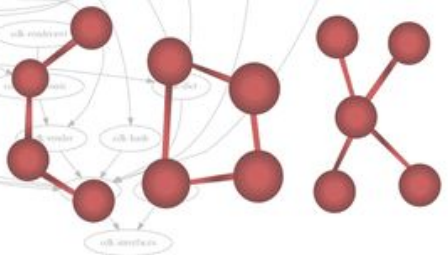
Egon L. Willighagen, John W. Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, Gilleain Torrance, Chris T. Evelo, Rajarshi Guha & Christoph Steinbeck

Journal of Cheminformatics 9, Article number: 33 (2017) | Download Citation

7825 Accesses | 50 Citations | 55 Altmetric | Metrics »



Toolkit	Time (s)
CDK 1.4.9	3,843
CDK 2.0	67



NUTRIM

Maastricht University | Maastricht UMC+

Identifiers: other identifiers

```
INFO: old database is HMDB-CHEBI-WIKIDATA HMDB5.0.20211102-CHEBI224-WIKIDATA20230807 [build: 20230807]
INFO: new database is HMDB-CHEBI-WIKIDATA HMDB5.0.20240416-CHEBI232-WIKIDATA20240416 [build: 20240416]
INFO: Number of ids in Ch [HMDB]: 297251 [unchanged]
INFO: Number of ids in Gpl [Guide to Pharmacology]: 7506 [0 added, 3 removed -> overall changed -0.0%]
INFO: Number of ids in Cpc [PubChem-compound]: 1273488 [50399 added, 22 removed -> overall changed +4.1%]
INFO: Number of ids in Ce [ChEBI]: 386946 [67162 added, 20 removed -> overall changed +21.0%]
INFO: Number of ids in Ck [KEGG Compound]: 16258 [46 added, 20 removed -> overall changed +0.2%]
INFO: Number of ids in Cl [ChEMBL compound]: 50331 [10 added, 0 removed -> overall changed +0.0%]
INFO: Number of ids in Cs [Chemspider]: 187969 [26207 added, 26 removed -> overall changed +16.2%]
INFO: Number of ids in Sl [SwissLipids]: 20341 [unchanged]
INFO: Number of ids in Wd [Wikidata]: 1405361 [4440 added, 3308 removed -> overall changed +0.1%]
INFO: Number of ids in Dr [DrugBank]: 12036 [25 added, 2 removed -> overall changed +0.2%]
INFO: Number of ids in Cks [KNApSAcK]: 5014 [36 added, 0 removed -> overall changed +0.7%]
INFO: Number of ids in Lm [LIPID MAPS]: 45139 [3150 added, 0 removed -> overall changed +7.5%]
INFO: Number of ids in Ect [EPA CompTox]: 850027 [29 added, 5 removed -> overall changed +0.0%]
INFO: Number of ids in Kd [KEGG Drug]: 4109 [14 added, 0 removed -> overall changed +0.3%]
INFO: Number of ids in Ca [CAS]: 944585 [515 added, 28 removed -> overall changed +0.1%]
INFO: Number of ids in Ik [InChIKey]: 1508499 [27550 added, 85 removed -> overall changed +1.9%]
```

Identifiers: InChIKey

Show 10 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-HQMMCQRPSA-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

Previous 1 Next

Visualization

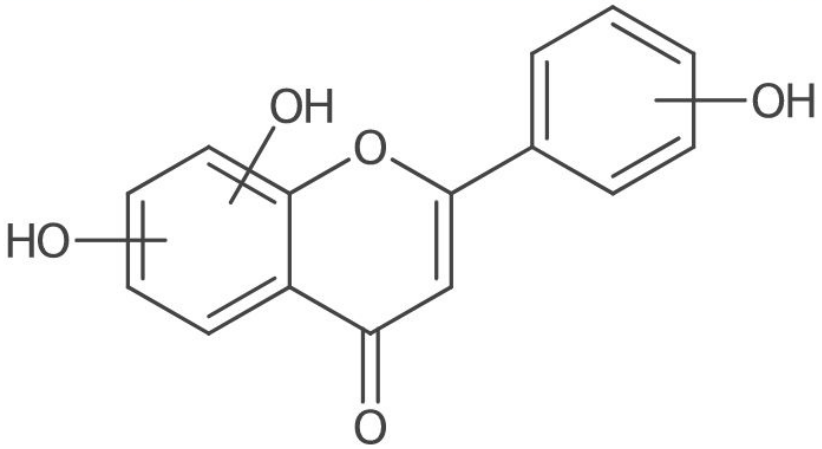
- Input: SMILES
- Output: drawing

E. Willighagen, A. Rutz, 2022, CDK CxSMILES, https://egonw.github.io/cdk-cxsmiles/

Flavonoid

Or for a flavonoid with known number of hydroxy groups on each rings, but with position uncertainty:

```
O*.O*.C1=C(C(=C2C(=C1)OC(=CC2=O)C3=CC(=C(C=C3))))).O* |m:3:4.5,m:1:8.9,m:22:18.19|
```



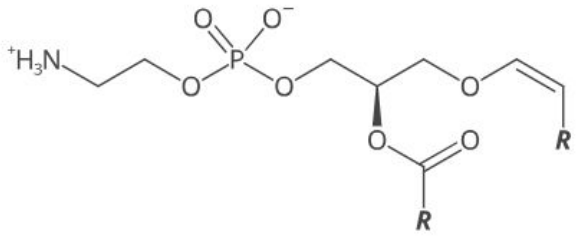
NUTRIM

Maastricht University | Maastricht UMC+

CXSMILES (a *de facto* standard)

R-groups

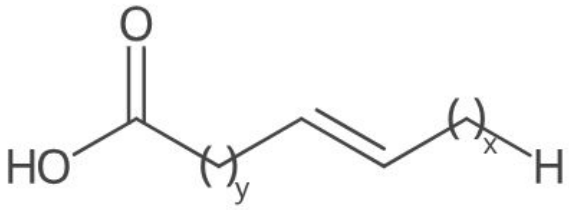
A common application of CXSMILES is the use of R-groups. This is done with the following CXSMILES pattern like `[NH3+]CCOP([O-])(=O)OC[C@@H](CO\C=C/[*])OC([*])=O |$;;;;;;;;;;;;;;R;;;R$|`:



Lipids

Lipids with a double bond somewhere in the tail

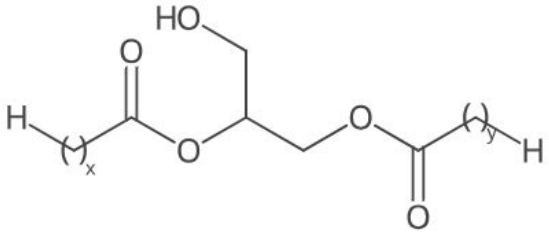
If we have a single tail lipid with `x+y+2` carbons in the tail but we do not know the location of the double bond, we can use a CxSMILES like `[H]C\\C=C\\CC(=O)O |Sg:n:1:x:ht,Sg:n:4:y:ht| x+y=15`:



x+y=15

Lipids with two tails

Sometimes experimental data does not provide enough information to decide how long the individual lipid tails are, but only provide the total length. Then a template like `OCC(OC(=O)C[H])COC(=O)C[H] |Sg:n:6:x:ht,Sg:n:12:y:ht| x+y=28` can be useful:



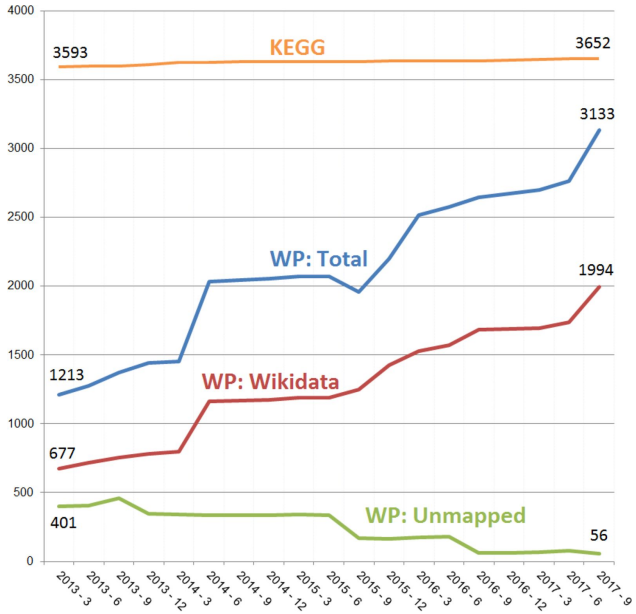
x+y=28

E. Willighagen, A. Rutz, 2022, CDK CxSMILES, https://egonw.github.io/cdk-cxsmiles/

WikiPathways

Biological context

WikiPathways
Pathways for the People

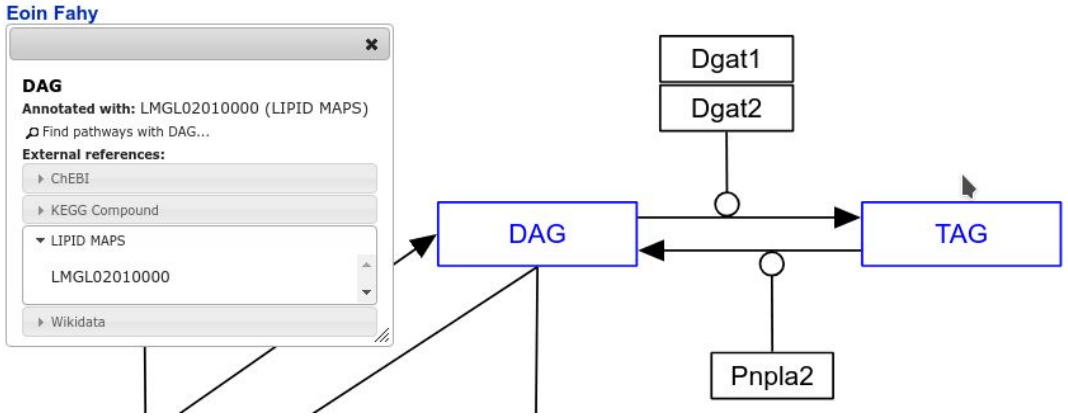


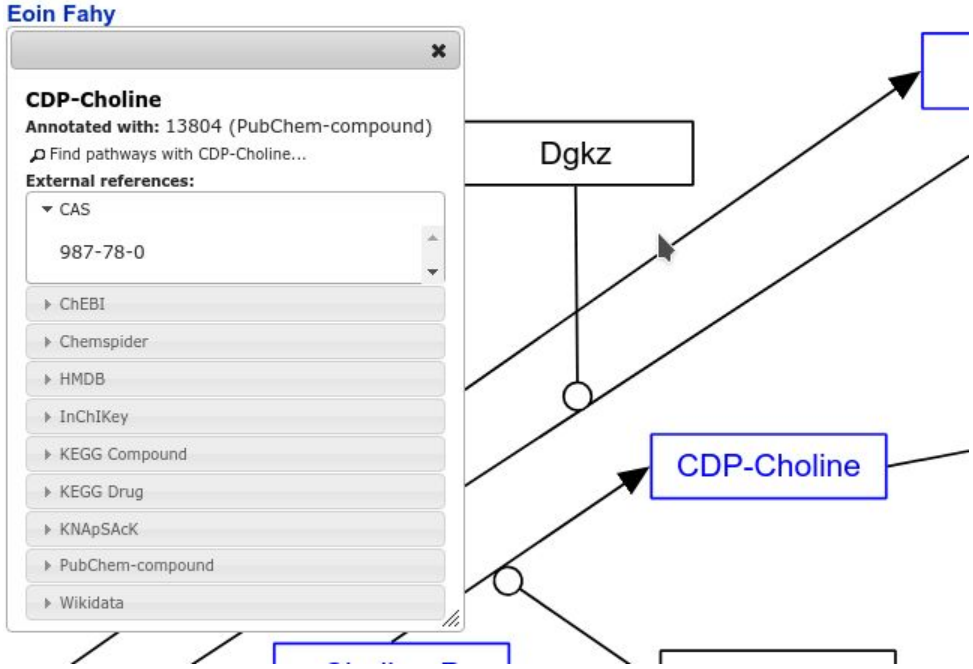
Agrawal *et al*, WikiPathways 2024: next generation pathway database, 10.1093/nar/gkx1064

NUTRIM
School of Nutrition and Translational Research in Metabolism

elixir Recommended Interoperability Resources

BridgeDb





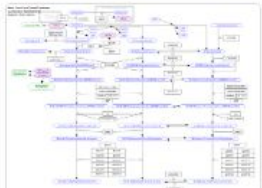
Lipid Pathways

Community Pathways

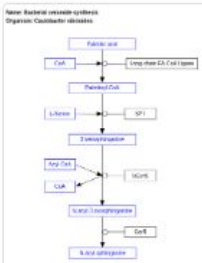
Gallery | List

Filters | Table

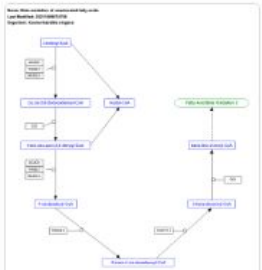
This community helps to curate 39 pathways:



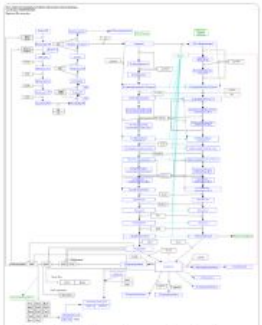
7-oxo-C and 7-beta-HC pathways *(Homo sapiens)*



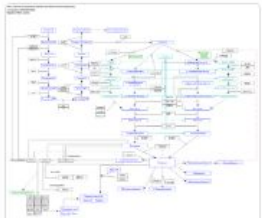
Bacterial ceramide synthesis *(Caulobacter vibrioides)*



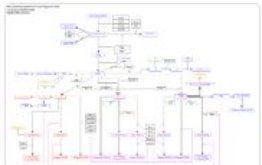
Beta-oxidation of unsaturated fatty acids *(Caenorhabditis elegans)*



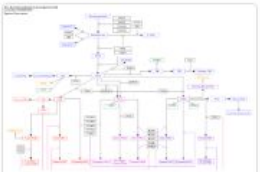
Cholesterol metabolism with Bloch and Kandutsch-Russell pathways *(Mus musculus)*



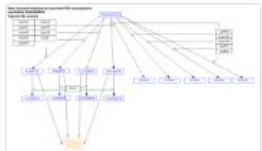
Cholesterol metabolism with Bloch and Kandutsch-Russell pathways *(Homo sapiens)*



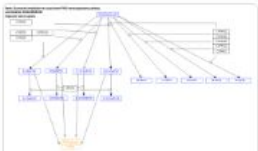
Eicosanoid metabolism via cyclooxygenases (COX) *(Mus musculus)*



Eicosanoid metabolism via cyclooxygenases (COX) *(Homo sapiens)*

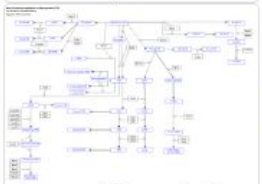


Eicosanoid metabolism via cytochrome P450 monooxygenases *(Mus musculus)*



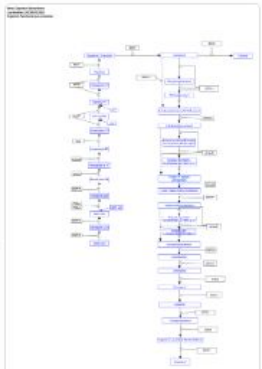
Eicosanoid metabolism via cytochrome P450 monooxygenases pathway *(Homo sapiens)*

Eicosanoid metabolism via lipooxygenases (LOX) *(Homo sapiens)*

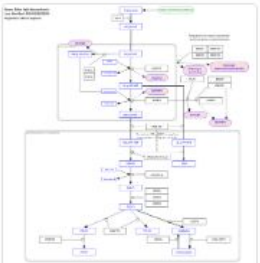


Eicosanoid metabolism via lipoxygenases (LOX) *(Mus musculus)*

Eicosanoid synthesis *(Homo sapiens)*



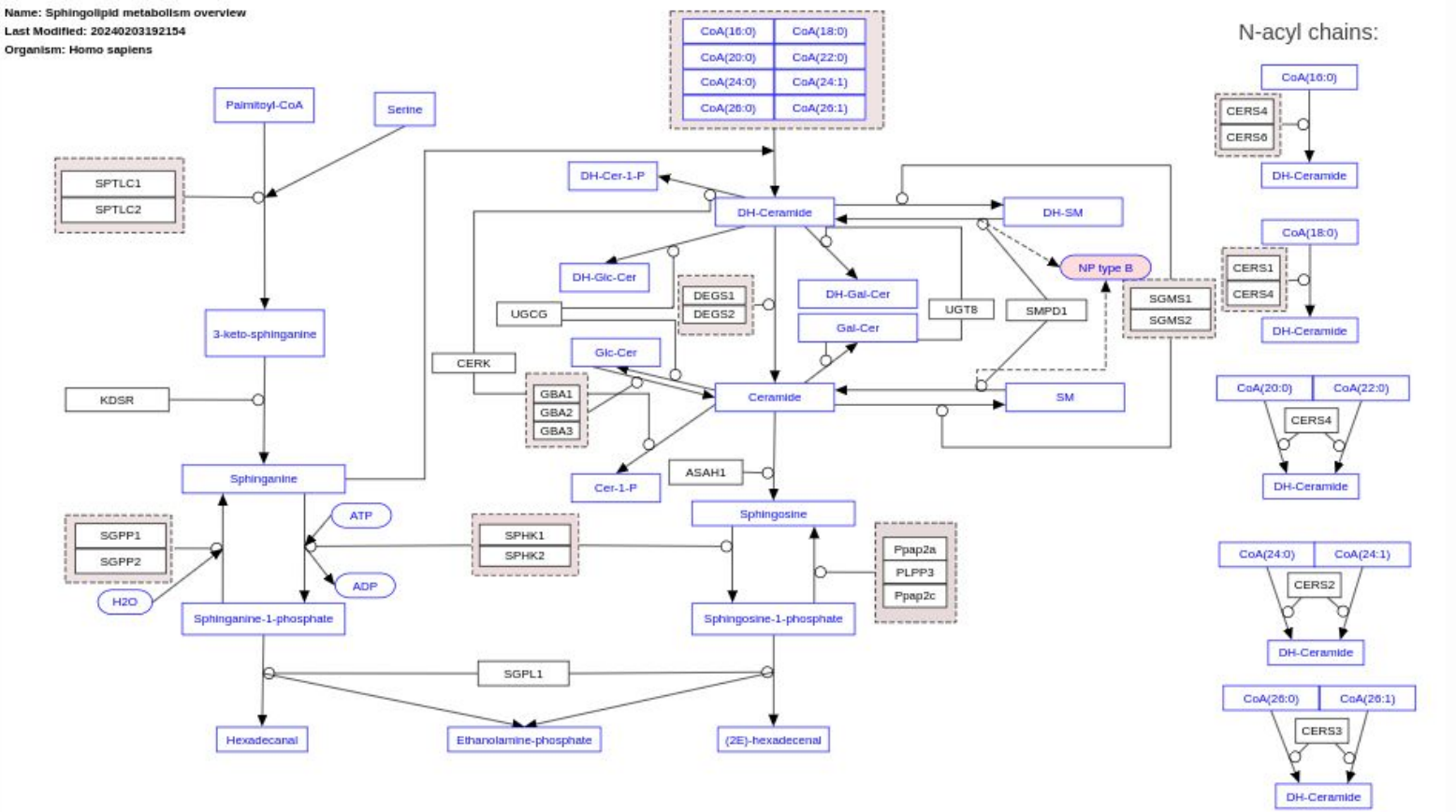
Ergosterol biosynthesis *(Saccharomyces cerevisiae)*



Ether lipid biosynthesis *(Homo sapiens)*

Sphingolipid metabolism overview (WP4725)

Homo sapiens



New PW, homology converted

Authors

Denise Slenter , Eric Weitz , Conroy lipids , and Ash Iyer

Links to other databases

BridgeDb

Participants

Label	Type	Compact URI	Comment
ATP	Metabolite	chebi:30616	
3-keto-sphinganine	Metabolite	lipidmaps:LMSP01020002	
Palmitoyl-CoA	Metabolite	lipidmaps:LMFA07050360	aka CoA(16:0)
Serine	Metabolite	pubchem.compound:5951	
Hexadecanal	Metabolite	lipidmaps:LMFA06000088	

more rows

References

1. Characterization of murine sphingosine-1-phosphate phosphohydrolase. Le Stunff H, Peterson C, Thornton R, Milstien S, Mandala SM, Spiegel S. J Biol Chem. 2002 Mar 15;277(11):8920–7. PubMed Europe PMC Scholia
2. De novo sphingolipid biosynthesis: a necessary, but dangerous, pathway. Merrill AH Jr. J Biol Chem. 2002 Jul 19;277(29):25843–6. PubMed Europe PMC Scholia
3. Identification of small subunits of mammalian serine palmitoyltransferase that confer distinct acyl-CoA substrate specificities. Han G, Gupta SD, Gable K, Niranjanakumari S, Moitra P, Eichler F, et al. Proc Natl Acad Sci U S A. 2009 May 19;106(20):8186–91. PubMed Europe PMC Scholia
4. A mouse macrophage lipidome. Dennis EA, Deems RA, Harkewicz R, Quehenberger O, Brown HA, Milne SB, et al. J Biol Chem. 2010 Dec 17;285(51):39976–85. PubMed Europe PMC Scholia

Citation.js

NUTRIM

Maastricht University | Maastricht UMC+

Pathway and Network Analysis

MetaboLights

Search

Examples: Alanine, Homo sapiens, Urine, MTBLS1

Home | Browse Studies | Browse Compounds | Browse Species | Download | Help | Give us feedback | About | Submit Study | Login

Status Public | Release Date 2016-07-13

MTBLS360: Metabolomics and lipidomics reveal perturbation of sphingolipid metabolism by a novel anti-trypanosomal 3-(oxazolo[4,5-b]pyridine-2-yl)anilide

Darren J Creek

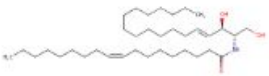
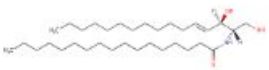
Trypanosoma brucei is the causative agent of human A
therapies are limited and there is an urgent need to dev
(OXPA), was initially identified in a phenotypic screen a
shown to be non-toxic and to be active against a numb

Trypanosoma brucei | blank

PUBLICATIONS

Metabolomics and lipidomics reveal perturbation of sphingolipid metabolism by a novel anti...

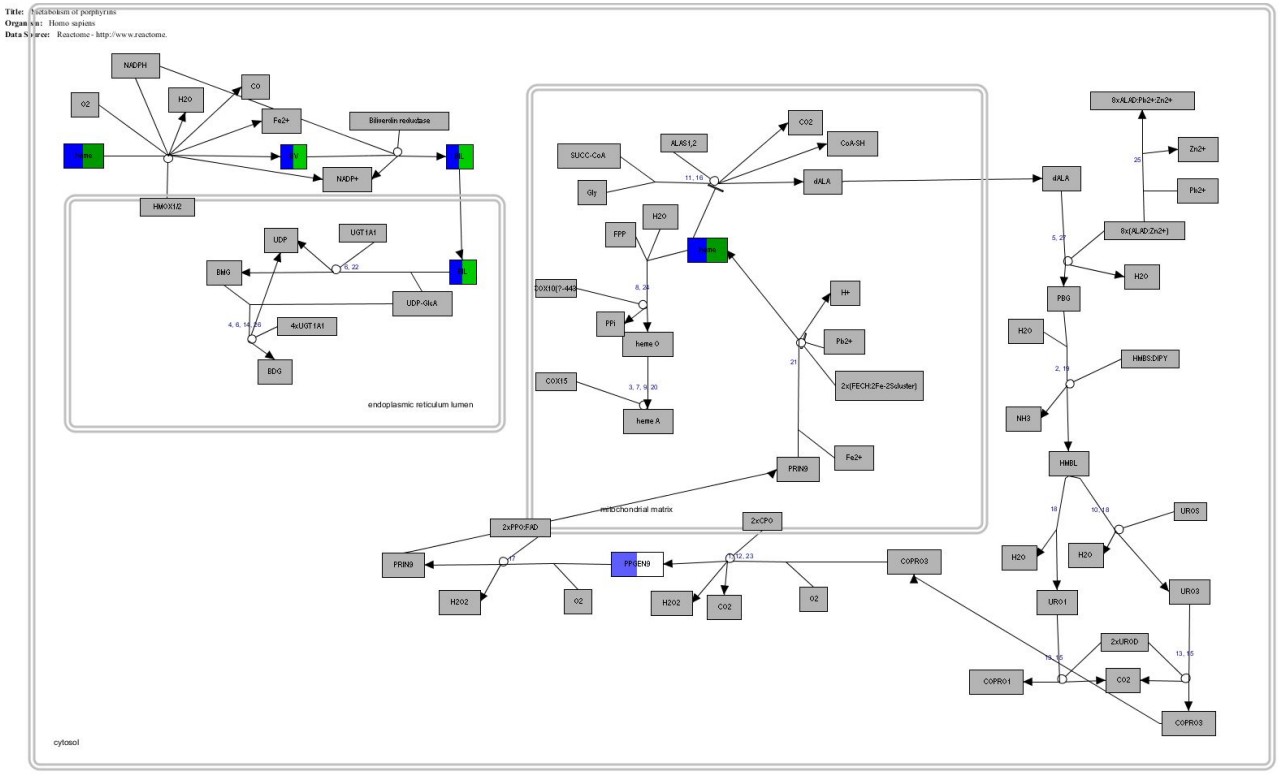
Daniel Stoessel, Cameron J. Nowell, Amy J. Jones, Lori Ferri...

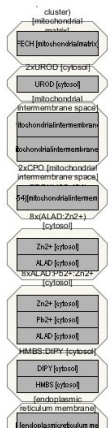
Structure	Database identifier	Chemical formula	SMILES
	CHEBI:77996	C36H69NO3	CCCCCCCCCCCCCC=C[C@@H](O)[C@H](CO)NC(=O)CCCCCCCC=C/CCCCCCCC
	CHEBI:138507	C33H65NO3	[C@]([C@@](/C=C/CCCCCCCCCCC)(O)[H])(NC(=O)CCCCCCCCCCCCCCCC)([H])CO
	CHEBI:67035	C36H73NO4	

NUTRIM

Maastricht University | Maastricht UMC+

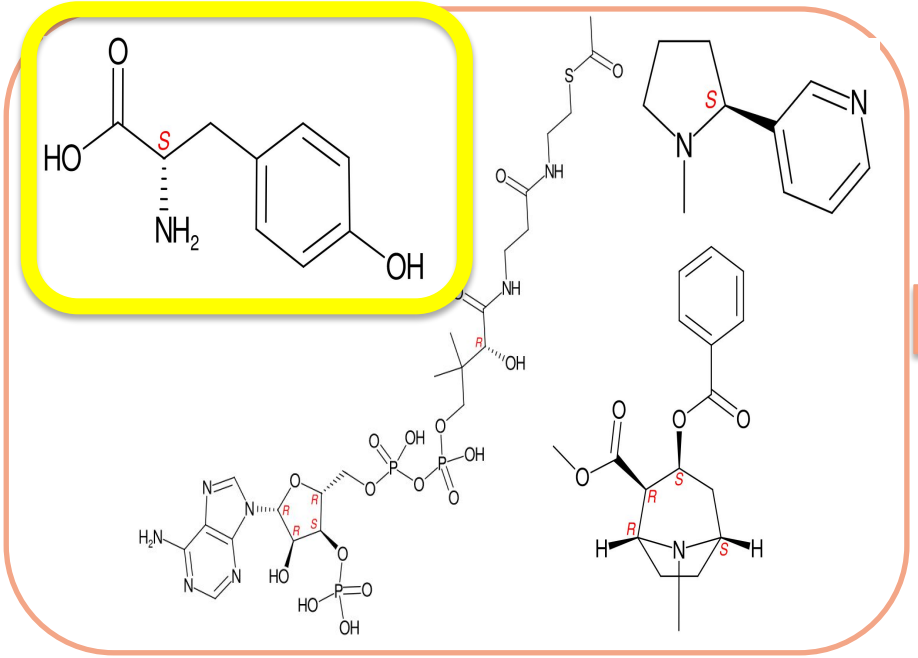
The effect of troglitazone on heme biosynthesis

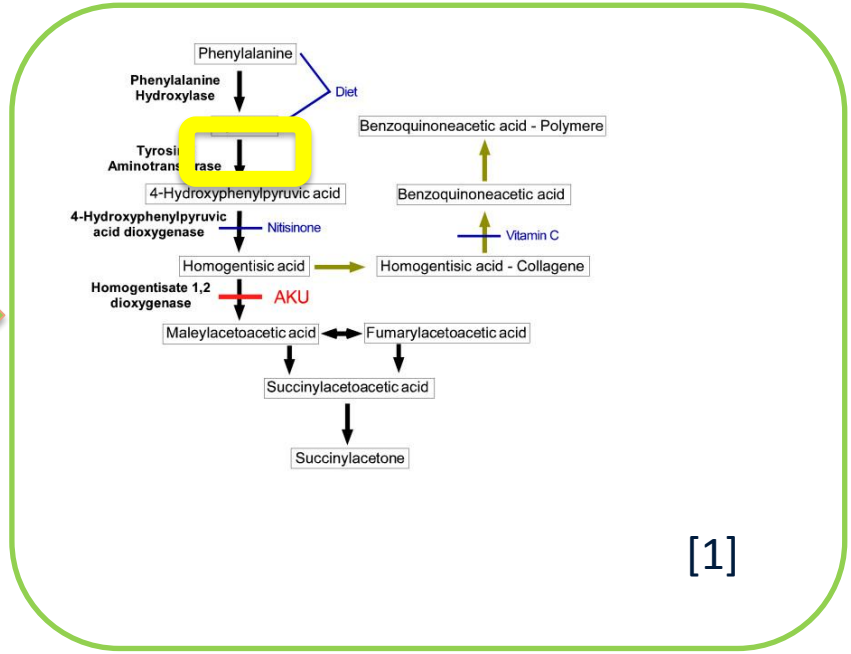




Metabolism of porphyrins (Homo sapiens), 2018, Reactome Team, Anwesha Bohler, Egon Willighagen, Martijn van Iersel, http://identifiers.org/wikipathways/WP1852

Linking metabolomics data to pathways…





[1]

Sparseness of Data

Amount of data

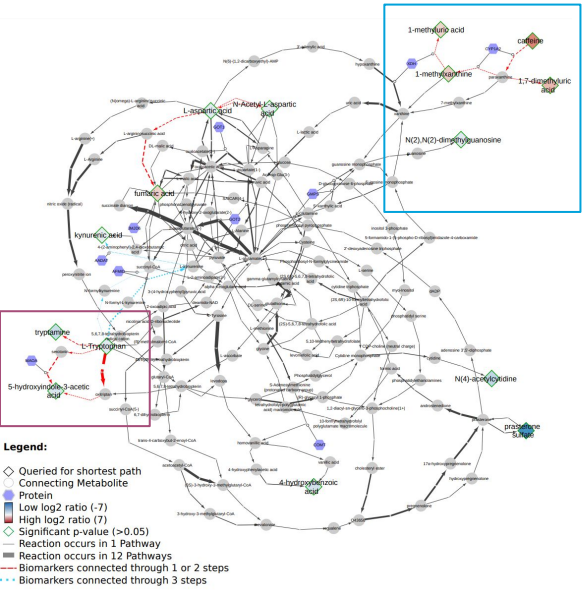
Identifier mapping

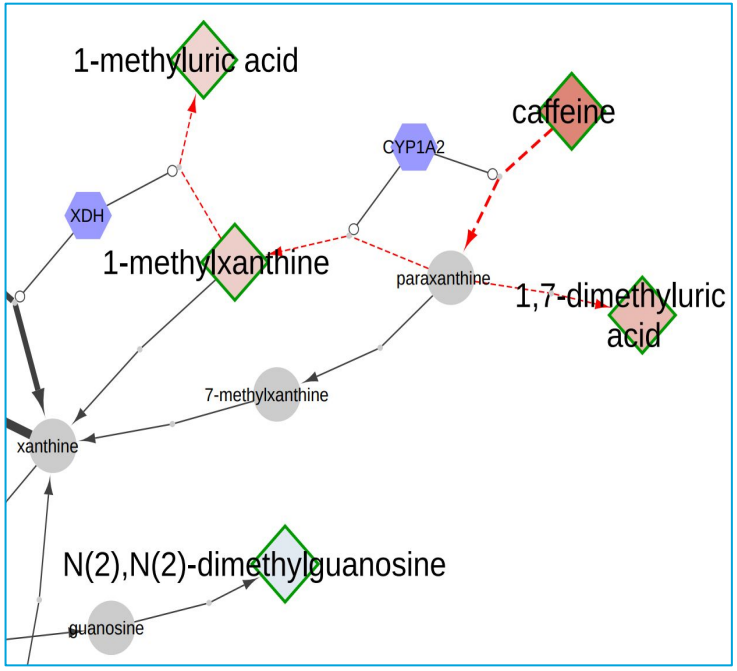
NUTRIM

[1] Lindner, et al. BMC Ophthalmology (2014)

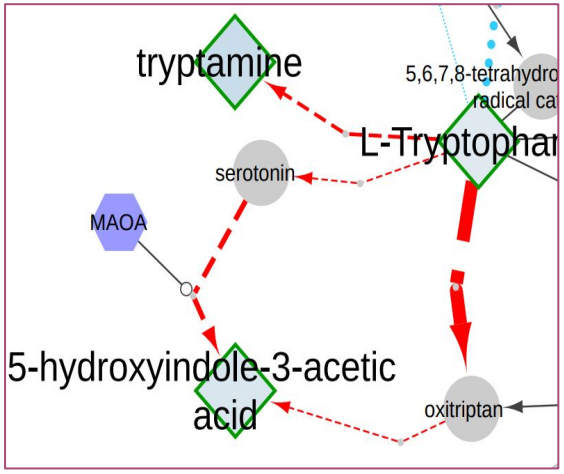
Maastricht University | Maastricht UMC+

Name	Technique	Matrix	Metabolites of interest detected with:	Male/ Female	Age range (y)	# Identified metabolites	# age related	shortest path linked biomarkers
MTBLS404 [1]	LC-HRMS	Urine	Spearman rank correlation test, Orthogonal partial least-squares (OPLS)	100/83	40.9 ± 10.3	120	30	14





"These metabolites cannot be regarded as individual variables in modeling approaches"



"Oxitriptan is an immediate precursor to serotonin, known to decline with age [2]"

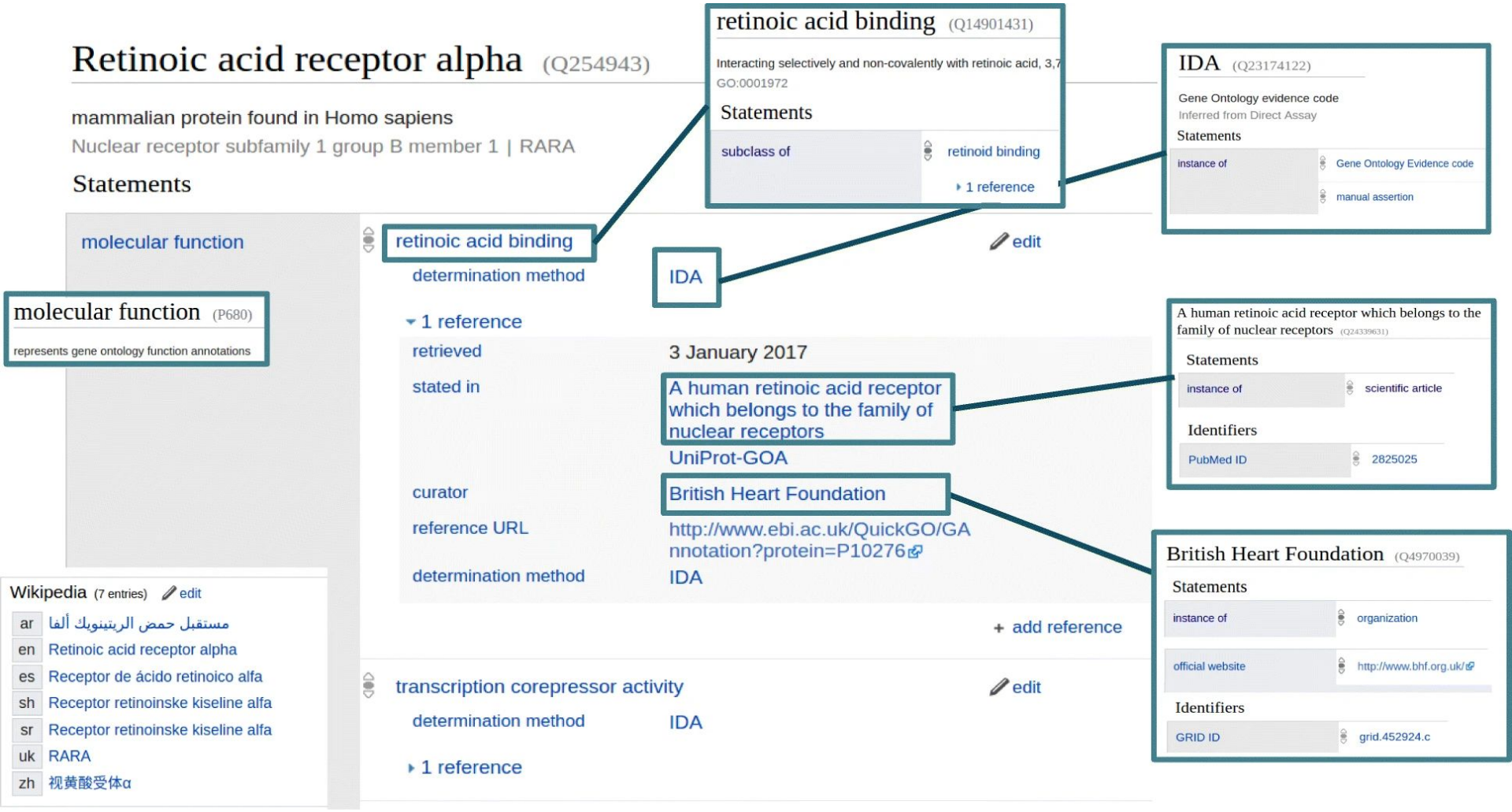
Maastricht University

Slenter D, et al. Discovering life's directed metabolic (sub)paths to interpret human biochemical markers using the DSMN tool. Digital Discovery. 2023; DOIL 10.1039.D3DD00069A.

[1] Thévenot, Etienne A., *et al*., J. Prot. Res. (2015). DOI: 10.1021/acs.jproteome.5b00354
[2] R. Peters, Postgrad. Med. J. (2006). DOI: 10.1136/pgmj.2005.036665

Wikidata and Scholia

Open infrastructure



NUTRIM

Waagmeester, A. *et al. BMC Biol* 19, 12 (2021). https://doi.org/10.1186/s12915-020-00940-y

Maastricht University | Maastricht UMC+

Research Ideas and Outcomes 1: e7573
doi: 10.3897/rio.1.e7573
OPEN ACCESS
H2020 Grant Proposal

Enabling Open Science: Wikidata for Research (Wiki4R)

Daniel Mietchen[‡], Gregor Hagedorn[‡], Egon Willighagen[§], Mariano Rico[|], Asunción Gómez-Pérez[|], Eduard Aibar[¶], Karima Rafes[#], Cécile Germain[¤], Alastair Dunning[«], Lydia Pintscher[»], Daniel Kinzler[»]

‡ Museum für Naturkunde Berlin, Berlin, Germany
§ Maastricht University, Maastricht, Netherlands
| Universidad Politécnica de Madrid (UPM), Madrid, Spain
¶ Universitat Oberta de Catalunya, Barcelona, Spain
# Inria, Paris, France
¤ Université Paris-Sud, Paris, France
« The European Library, Den Haag, Netherlands
» Wikimedia Deutschland e.V., Berlin, Germany

Corresponding author: Daniel Mietchen (daniel.mietchen@mfn-berlin.de)

Reviewable | v1

Received: 21 Dec 2015 | Published: 22 Dec 2015

Citation: Mietchen D, Hagedorn G, Willighagen E, Rico M, Gómez-Pérez A, Aibar E, Rafes K, Germain C, Dunning A, Pintscher L, Kinzler D (2015) Enabling Open Science: Wikidata for Research (Wiki4R). Research Ideas and Outcomes 1: e7573. doi: 10.3897/rio.1.e7573

Abstract

Wiki4R will create an innovative virtual research environment (VRE) for Open Science at scale, engaging both professional researchers and citizen data scientists in new and potentially transformative forms of collaboration. It is based on the realizations that (1) the structured parts of the Web itself can be regarded as a VRE, (2) such environments depend on communities, (3) closed environments are limited in their capacity to nurture thriving communities. Wiki4R will therefore integrate Wikidata, the multilingual semantic backbone behind Wikipedia, into existing research processes to enable transdisciplinary research and reduce fragmentation of research in and outside Europe. By establishing a advance efforts by the open-data community to build a rich and heterogeneous network of scientific knowledge. That knowledge network could, in turn, be the foundation for many computational tools, applications and analyses.

Most data- and knowledge-integration initiatives fall on either end of a spectrum. At one end, centralized efforts seek to bring multiple knowledge sources into a single database (see, for example, Mungall et al., 2017): this approach has the advantage of data alignment according to a common data model and of enabling high performance queries. However, centralized resources are difficult and expensive to maintain and expand (Chandras et al., 2009; Gabella et al., 2018), at least in part because of

European Semantic Web Conference
↳ ESWC 2017: **The Semantic Web: ESWC 2017 Satellite Events** pp 237–259 | Cite as

Scholia, Scientometrics and Wikidata

Finn Årup Nielsen ✉, Daniel Mietchen & Egon Willighagen

Conference paper | Open Access | First Online: 08 November 2017

8313 Accesses | 36 Citations | 30 Altmetric

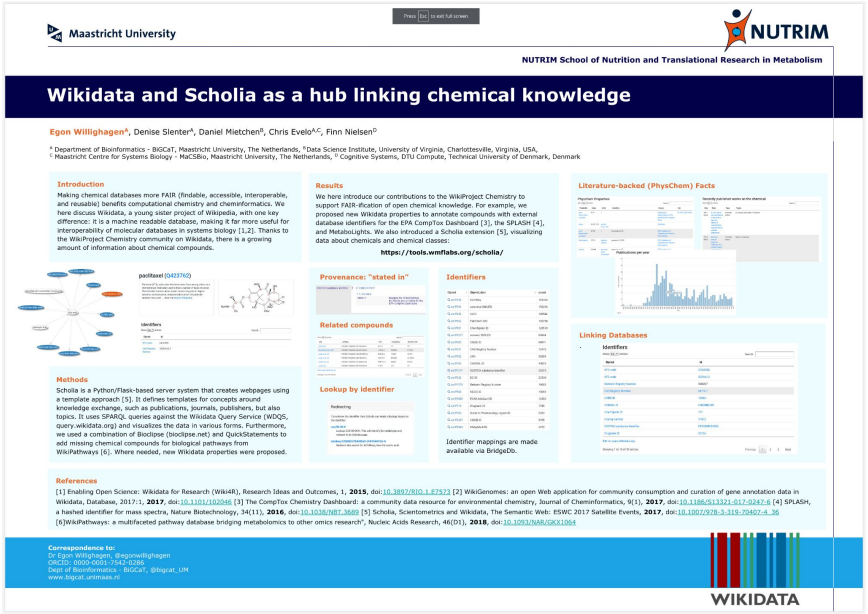
Part of the Lecture Notes in Computer Science book series (LNISA,volume 10577)

Abstract

Scholia is a tool to handle scientific bibliographic information through Wikidata. The Scholia Web service creates on-the-fly scholarly profiles for researchers, organizations, journals, publishers, individual scholarly works, and for research topics. To collect the data, it queries the SPARQL-based Wikidata Query Service. Among several display formats available in Scholia are lists of publications for individual researchers and organizations, plots of publications per year, employment timelines, as well as co-author and topic networks and citation graphs. The Python package implementing the Web service is also able to format Wikidata bibliographic entries for use in LaTeX/BIBTeX. Apart from detailing Scholia, we describe how Wikidata has been used for bibliographic information and we also provide some scientometric statistics on this information.

Keywords

Wikidata | Wikipedias | Functional Requirements For Bibliographic Records (FRBR) | Public Library Of Science (PLOS)



SCIENCE FORUM

Wikidata as a knowledge graph for the life sciences

Abstract Wikidata is a community-maintained knowledge base that has been assembled from repositories in the fields of genomics, proteomics, genetic variants, pathways, chemical compounds, and diseases, and that adheres to the FAIR principles of findability, accessibility, interoperability and reusability. Here we describe the breadth and depth of the biomedical knowledge contained within Wikidata, and discuss the open-source tools we have built to add information to Wikidata and to synchronize it with source databases. We also demonstrate several use cases for Wikidata, including the crowdsourced curation of biomedical ontologies, phenotype-based diagnosis of disease, and drug repurposing.

ANDRA WAAGMEESTER[†], GREGORY STUPP[†], SEBASTIAN BURGSTALLER-MUEHLBACHER, BENJAMIN M GOOD, MALACHI GRIFFITH, OBI L GRIFFITH, KRISTINA HANSPERS, HENNING HERMJAKOB, TOBY S HUDSON, KEVIN HYBISKE, SARAH M KEATING, MAGNUS MANSKE, MICHAEL MAYERS, DANIEL MIETCHEN, ELVIRA MITRAKA, ALEXANDER R PICO, TIMOTHY PUTMAN, ANDERS RIUTTA, NURIA QUERALT-ROSINACH, LYNN M SCHRIML, THOMAS SHAFEE, DENISE SLENTER, RALF STEPHAN, KATHERINE THORNTON, GINGER TSUENG, ROGER TU, SABAH UL-HASAN, EGON WILLIGHAGEN, CHUNLEI WU AND ANDREW I SU*

Introduction

Integrating data and knowledge is a formidable challenge in biomedical research. Although new scientific findings are being discovered at a rapid pace, a large proportion of that knowledge is either locked in data silos (where integration is hindered by differing nomenclature, data models, and licensing terms; *Wilkinson et al., 2016*) or locked away in free-text. The lack of an integrated and structured version of biomedical knowledge hinders efficient querying or mining of that information, thus preventing the full utilization of our accumulated scientific knowledge.

Recently, there has been a growing emphasis within the scientific community to ensure all sci-

*For correspondence: asu@scripps.edu
†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.
Funding: See page 12
Reviewing editor: Peter Rodgers, eLife, United Kingdom

Waagmeester et al. BMC Biology (2021) 19:12
https://doi.org/10.1186/s12915-020-00940-y

BMC Biology

METHODOLOGY ARTICLE Open Access

A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses

Andra Waagmeester[1], Egon L. Willighagen[2], Andrew I. Su[3], Martina Kutmon[2,4], Jose Emilio Labra Gayo[5], Daniel Fernández-Álvarez[5], Quentin Groom[6], Peter J. Schaap[7], Lisa M. Verhagen[8] and Jasper J. Koehorst[7*]

Abstract

Background: Pandemics, even more than other medical problems, require swift integration of knowledge. When caused by a new virus, understanding the underlying biology may help finding solutions. In a setting where there are a large number of loosely related projects and initiatives, we need common ground, also known as a "commons." Wikidata, a public knowledge graph aligned with Wikipedia, is such a commons and uses unique identifiers to link knowledge in other knowledge bases. However, Wikidata may not always have the right schema for the urgent questions. In this paper, we address this problem by showing how a data schema required for the integration can be modeled with entity schemas represented by Shape Expressions.

Results: As a telling example, we describe the process of aligning resources on the genomes and proteomes of the SARS-CoV-2 virus and related viruses as well as how Shape Expressions can be defined for Wikidata to model the knowledge, helping others studying the SARS-CoV-2 pandemic. How this model can be used to make data between various resources interoperable is demonstrated by integrating data from NCBI (National Center for Biotechnology Information) Taxonomy, NCBI Genes, UniProt, and WikiPathways. Based on that model, a set of automated applications or bots were written for regular updates of these sources in Wikidata and added to a platform for automatically running these updates.

Conclusions: Although this workflow is developed and applied in the context of the COVID-19 pandemic, to demonstrate its broader applicability it was also applied to other human coronaviruses (MERS, SARS, human coronavirus NL63, human coronavirus 229E, human coronavirus HKU1, human coronavirus OC4).

Keywords: COVID-19, Wikidata, Linked data, SHEx, Open Science

Background

The coronavirus disease 2019 (COVID-19) pandemic, [1]. In response to the pandemic, many research groups have started projects to understand the SARS-CoV-2

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community

Andrea Jacobs,* Dustin Williams, Katherine Hickey, Nathan Patrick, Antony J. Williams, Stuart Chalk, Leah McEwen, Egon Willighagen, Martin Walker, Evan Bolton, Gabriel Sinclair, and Adam Sanford

Cite This: J. Chem. Inf. Model. 2022, 62, 2737−2743

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: CAS Common Chemistry (https://commonchemistry.cas.org/) is an open web resource that provides access to reliable chemical substance information for the scientific community. Having served millions of visitors since its creation in 2009, the resource was extensively updated in 2021 with significant enhancements. The underlying dataset was expanded from 8000 to 500,000 chemical substances and includes additional associated information, such as basic properties and computer-readable chemical structure information. New use cases are supported with enhanced search capabilities and an integrated application programming interface. Reusable licensing of the content is provided through a Creative Commons Attribution-Non-Commercial (CC-BY-NC 4.0) license allowing other public resources to integrate the data into their systems. This paper provides an overview of the enhancements to data and functionality, discusses the benefits of the contribution to the chemistry community, and summarizes recent progress in leveraging this resource to strengthen other information sources.

INTRODUCTION

CAS, a division of the American Chemical Society, has collected, curated, and analyzed the world's published science as part of its vision to improve people's lives through the transforming power of chemistry since 1907.[1,2] Scientists, manufacturers, regulators, and data scientists around the world rely on CAS for accurate information on chemical substances. CAS Common Chemistry, an open resource based on a subset of chemical substance content from CAS REGISTRY, was first launched in 2009 by CAS to strengthen the accuracy of publicly available scientific information.[3]

CAS Common Chemistry was established to provide a reliable source of chemical identifiers and associated information to the general public as part of the mission of the ACS. It enables millions of visitors to obtain reliable scientific information on nearly 500,000 substances through search or application programming interface (API) functionality. Users leverage this information in a variety of ways, including in teaching and learning, to promote safe practices, and to support research. Additionally, Wikipedia has leveraged the resource since its inception in 2009 to provide accurate CAS Registry Numbers for the most ubiquitous chemical substances.[4] For each included substance, a substance detail page provides key attributes as well as a citation to support referencing in academic studies. For examples of search results, see substance detail pages for caffeine, aspirin, and benzene, as well as Figures 1 and 2.

tion, validated CAS Registry Numbers, and additional chemical representations.

Expanded Dataset Sourced from CAS REGISTRY. CAS Common Chemistry contains substances and related data from CAS REGISTRY, the largest scientist-curated chemical substance database in the world, crossing 250 million chemical substances in April 2021 and growing daily.[5−7] In response to requests from the scientific community, the number of substances openly released in CAS Common Chemistry was recently expanded to nearly 500,000 substances. This is a dramatic increase from the 8000 substances initially available. The collection represents substances and related data for chemicals of concern, consumer product ingredients, commonly regulated chemicals, and chemicals frequently used in undergraduate chemistry curricula.

Enhanced Associated Chemical Information. Each substance in the resource includes its chemical name, chemical structure image, molecular formula, and molecular weight. The resource is further enhanced by chemical synonyms, which may include systematic chemical names, common names, and trade names for each chemical substance. Basic substance properties—boiling point, melting point, and density—are also included where available. This is demonstrated in Figure 3a,b.

For users of CAS SciFinder, a quick link is also included to launch the product and quickly discover more information about

Received: March 4, 2022

D578–D586 Nucleic Acids Research, 2022, Vol. 50, Database issue *Published online 29 October 2021*
https://doi.org/10.1093/nar/gkab991

Complex Portal 2022: new curation frontiers

Birgit H. M. Meldal[1], Livia Perfetto[1,2], Colin Combe[3], Tiago Lubiana[4], João Vitor Ferreira Cavalcante[5], Hema Bye-A-Jee[1], Andra Waagmeester[6], Noemi del-Toro[1], Anjali Shrivastava[1], Elisabeth Barrera[1], Edith Wong[7], Bernhard Mlecnik[8,9,10,11], Gabriela Bindea[8,9,10], Kalpana Panneerselvam[1], Egon Willighagen[12], Juri Rappsilber[3,13], Pablo Porras[1], Henning Hermjakob[1,*] and Sandra Orchard[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Fondazione Human Technopole, 20157 Milan, Italy, [3]Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK, [4]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, Av. Professor Lineu Prestes 580, CEP 05508-000 São Paulo SP, Brasil, [5]Bioinformatics Multidisciplinary Environment (BioME), Digital Metropolis Institute, Federal University of Rio Grande do Norte, Av. Odilon Gomes de Lima 1722, Capim Macio, 59078-400 Natal/RN, Brasil, [6]Micelio, Veltvijckxlaan 305, 2180 Ekeren, Belgium, [7]Department of Genetics, School of Medicine, Stanford University, Palo Alto, CA, USA, [8]Laboratory of Integrative Cancer Immunology, INSERM, 75006 Paris, France, [9]Equipe Labellisée Ligue Contre le Cancer, 75006 Paris, France, [10]Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris, 75006 Paris, France, [11]Inovarion, 75005 Paris, France, [12]Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands and [13]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

Received September 15, 2021; Revised October 07, 2021; Editorial Decision October 08, 2021; Accepted October 10, 2021

ABSTRACT

The Complex Portal (www.ebi.ac.uk/complexportal) is a manually curated, encyclopaedic database of macromolecular complexes with known function from a range of model organisms. It summarizes complex composition, topology and function along with links to a large range of domain-specific resources (i.e. wwPDB, EMDB and Reactome). Since the last update in 2019, we have produced a first draft complexome for *Escherichia coli*, maintained and updated that of *Saccharomyces cerevisiae*, added over 40 coronavirus complexes and increased the human complexome to over 1100 complexes that include approximately 200 complexes that act as targets for viral proteins or are part of the immune system. The display of protein features in ComplexViewer has been improved and the participant ta-

now CC0 to encourage data reuse. Users are encouraged to get in touch, provide us with feedback and send curation requests through the 'Support' link.

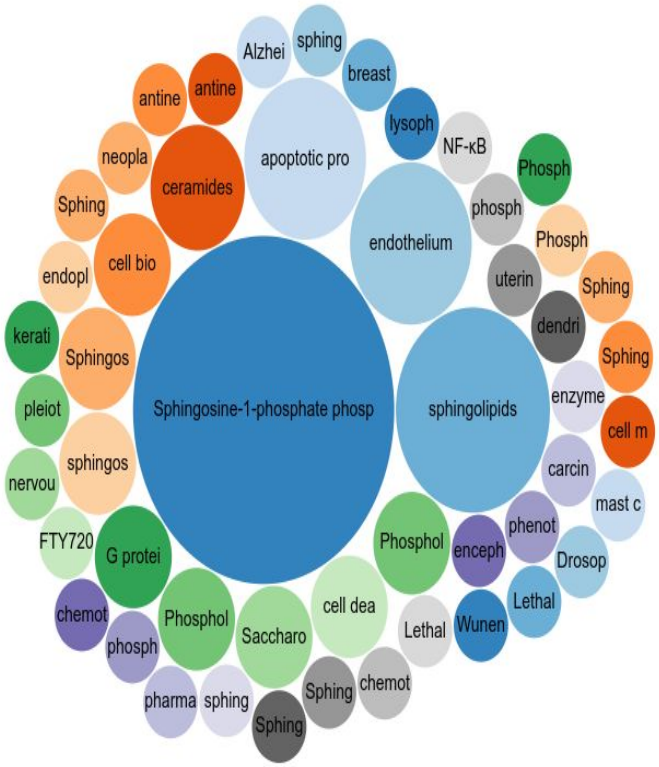
INTRODUCTION

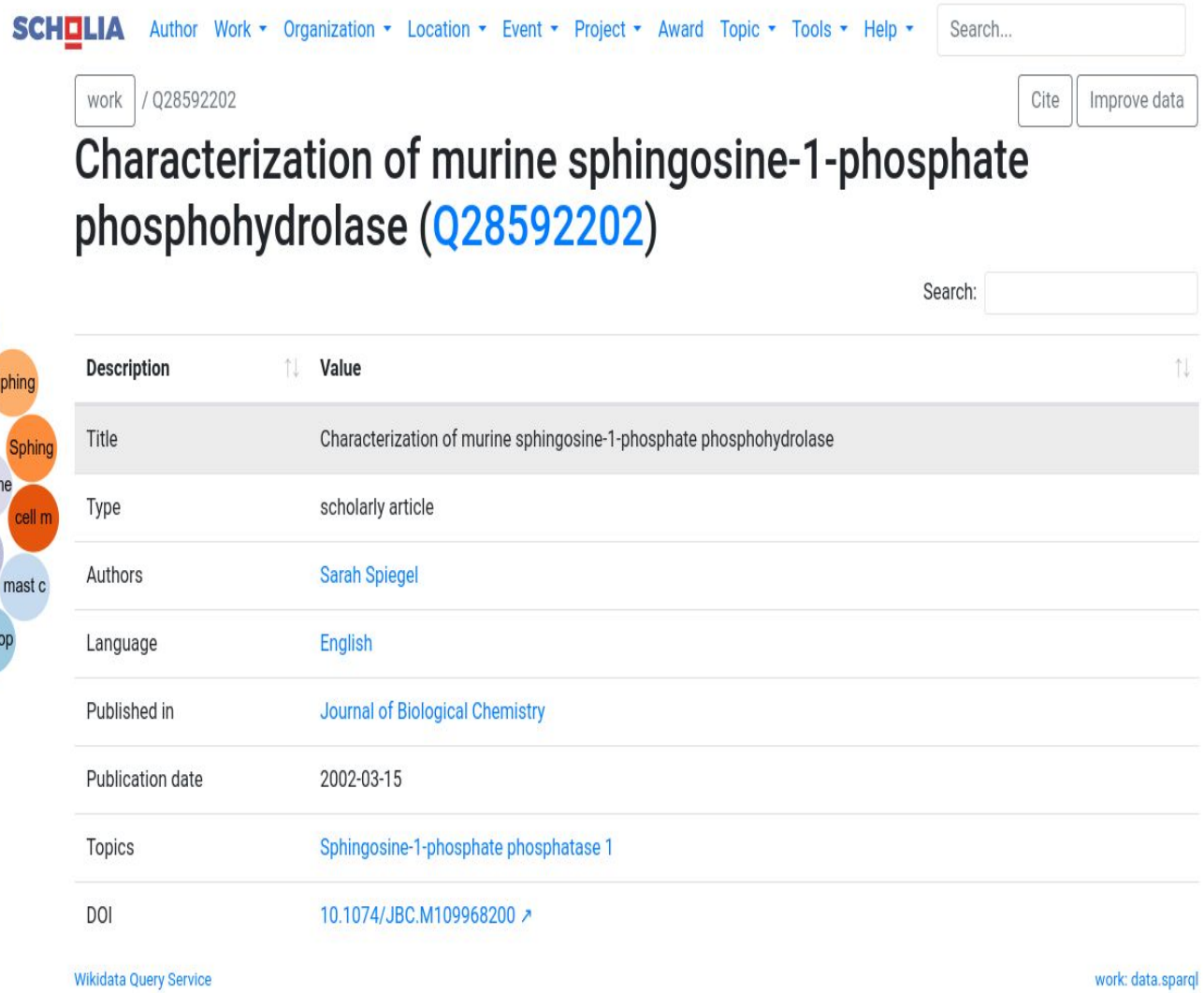
Protein complexes, stable functional assemblies consisting of two or more associated polypeptide chains, are responsible for driving and regulating many cellular processes. Multi-chain assemblies perform many functions, including (a) positioning molecules involved in the same process in close proximity (b) bringing structure to disordered regions of proteins and (c) creating novel substrate binding sites at subunit interfaces. These assemblies can contain additional molecules, such as nucleic acids and small molecules. In budding yeast (1), around one in three proteins have a function in stable heteromeric complexes and in bacteria around one in five (see below).

NUTRIM

Maastricht University | Maastricht UMC+

Scholia and Wikidata





NUTRIM

https://doi.org/10.1007/978-3-319-70407-4_36

Scholia

Related works

Related works from co-citation analysis

Reload

Show 10 entries

Search:

Count	Work
30	Sphingosine-1-phosphate: an enigmatic signalling lipid
29	Molecular cloning and characterization of a lipid phosphohydrolase that degrades sphingosine-1- phosphate and induces cell death
28	Identification and characterization of a novel human sphingosine-1-phosphate phosphohydrolase, hSPP2
27	Sphingosine-1-phosphate phosphohydrolase in regulation of sphingolipid metabolism and apoptosis
24	Sphingosine-1-phosphate as second messenger in cell proliferation induced by PDGF and FCS mitogens
22	Sphingoid base 1-phosphate phosphatase: a key regulator of sphingolipid metabolism and stress response.
21	Identification and characterization of Saccharomyces cerevisiae dihydrosphingosine-1-phosphate phosphatase
19	Sphingosine-1-phosphate as a ligand for the G protein-coupled receptor EDG-1
19	Suppression of ceramide-mediated programmed cell death by sphingosine-1-phosphate
18	Dual actions of sphingosine-1-phosphate: extracellular through the Gi-coupled receptor Edg-1 and intracellular to regulate proliferation and survival

Wikidata Query Service

work: related-works.sparql

Showing 1 to 10 of 500 entries

Previous 1 2 3 4 5 ... 50 Next

NUTRIM

Maastricht University | Maastricht UMC+

Scholia

Citations to the work

Reload

Recent citations to the work

Show 10 entries

Search:

Citations	Publication date	Citing work
848	2003-05-01	Sphingosine-1-phosphate: an enigmatic signalling lipid
401	2013-12-01	The Concise Guide to PHARMACOLOGY 2013/14: enzymes.
304	2008-06-13	"Inside-out" signaling of sphingosine-1-phosphate: therapeutic targets
168	2003-09-03	Phosphorylation and action of the immunomodulator FTY720 inhibits vascular endothelial cell growth factor-induced vascular permeability.
83	2002-09-16	Sphingosine-1-phosphate phosphohydrolase in regulation of sphingolipid metabolism and apoptosis
81	2005-04-01	Integral membrane lipid phosphatases/phosphotransferases: common structure and diverse functions
74	2007-03-22	Intracellular generation of sphingosine 1-phosphate in human lung endothelial cells: role of lipid phosphate phosphatase-1 and sphingosine kinase 1.
71	2003-01-10	Identification and characterization of a novel human sphingosine-1-phosphate phosphohydrolase, hSPP2
64	2009-02-27	FTY720 inhibits ceramide synthases and up-regulates dihydrosphingosine 1-phosphate formation in human lung endothelial cells
39	2012-06-23	Shaping the landscape: metabolic regulation of S1P gradients

Wikidata Query Service

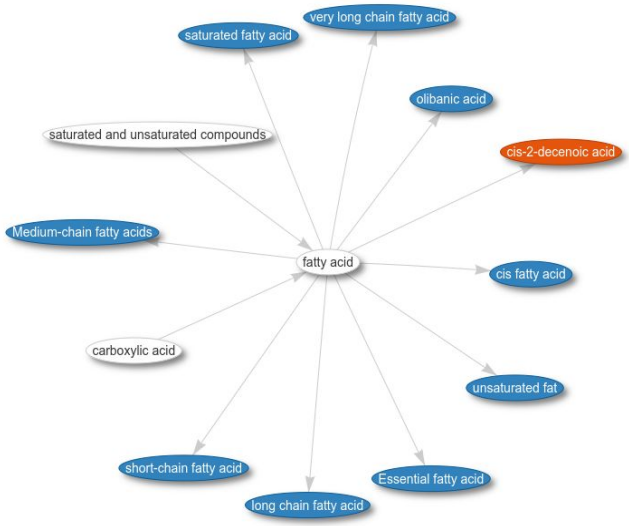
work: citations.sparql

Showing 1 to 10 of 41 entries

NUTRIM

Maastricht University | Maastricht UMC+

Wikidata / Scholia



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.

Show 10 entries Search:

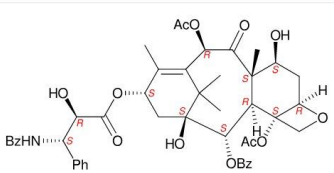
Mol	InChIKey	CAS	ChemSpider	PubChem CID
acetic acid	QTBSBXVTEAMEQO-UHFFFAOYSA-N	64-19-7	171	176
deuterated acetic acid	QTBSBXVTEAMEQO-GUEYOVJQSA-N	1186-52-3	2006083	2723903
acetic acid c-14	QTBSBXVTEAMEQO-HQMMCQRPSA-N	2845-03-6	144444	164769
acetic acid c-13	QTBSBXVTEAMEQO-VQEHIDDOSA-N	1563-79-7	8329490	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 Previous 1 Next

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 10 entries Search:

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

2019: 10.3897/rio.5.e35820
2017: 10.6084/m9.figshare.6356027.v1

NUTRIM

Maastricht University | Maastricht UM SCHOLIA

What I want you to remember

What I want you to remember

- Metabolomics needs more cheminformatics and knowledge graphs
- We can be really precise in our uncertainties
  - but we tend not to
- Need to bridge different information sources
  - thus, models (of reality) and language we use
- Too much knowledge is published unFAIR
- Open Science continues to build up
  - the open license makes reuse practically feasible

Giving metabolites (and lipids) a chemical and biological context with open science

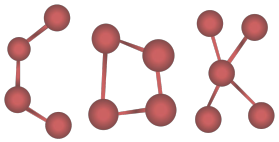
Egon Willighagen, BiGCaT team and collaborations
@egonw@social.edu.nl
0000-0001-7542-0286

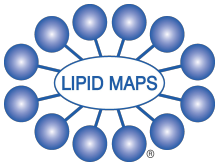
NMetC, Turku 2024-08-26
doi:10.5281/zenodo.13373746

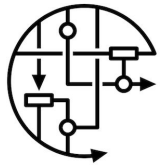
NUTRIM
School of Nutrition and Translational Research in Metabolism

Maastricht University | Maastricht UMC+

Funded projects







Horizon2020
European Union Funding
for Research & Innovation

Acknowledgements

Sean Ekins, Denny Vrandečić, Rudi Balling, John Wilbanks, Roderic D. M. Page, Peter Murray-Rust, Andrew S.I.D. Lang, Antony John Williams, Vinod Scaria, Alfonso Valencia, Shareen Doak, Pieter C. Dorrestein, Lane Rasberry, Carole Goble, Barend Mons, Gerhard Ecker, Henry S. Rzepa, Magnus Manske, Antonio Rosato, Ferran Sanz, Jose A. Seoane, Katherine Thornton, David J. Wild, Lutgarde Buydens, Ron Wehrens, Rajarshi Guha, Janna Hastings, **Daniel Mietchen, Finn Årup Nielsen**, Barbara Zdrazil, Stefan Senger, Binne Zwanenburg, Roeland J. M. Nolte, Alejandra González-Beltrán, Lars Eijssen, Marta Cascante Serratosa, Philippe Rocca-Serra, Chris T. Evelo, Christine Chichester, Emma L. Schymanski, Gregory Stupp, Rianne Fijten, Linda Rieswijk, Elvira Mitraka, Sebastian Burgstaller-Muehlbacher, Gaël Varoquaux, Niklas Blomberg, Christoph Steinbeck, Robert David Stevens, Steve Pettifer, Jean-Claude Bradley, Daniela Digles, Anna Gaulton, Arvid Berg, Marcus D. Hanwell, Johannes Wagener, Daniel M Lowe, Stefan Kuhn, Sajjan S Mehta, Ola Engkvist, Jakub Galgonek, Harold Solbrig, Vladimir Lobaskin, Igor Linkov, Pierre Lindenbaum, Peter Hoet, Lucian Farcal, Stacey L Harper, George Komatsoulis, Christina K. Pikas, Mark A. Musen, Fernando Pérez, Timothy Elliott Putman, Samuel Kaski, David S Wishart, Jasper Koehorst, Lee Harland, Alasdair J. G. Gray, **Martina Summer-Kutmon**, Roland Grafström, Bengt Fadeel, Friederike Ehrhart, Evan E. Bolton, Kenneth Haug, Oliver Kohlbacher, Jean-Loup Faulon, John P. Overington, Laura I. Furlong, Mark R. Viant, Paul Jennings, Claire O'Donovan, Mathieu Vinken, Alice Limonciel, Venkata P Satagopam, Thomas Shafee, Rob Hooft, Jeremy G. Frey, Eduard Aibar, Christoph Helma, Philip Doganis, Robert M Hanson, Bart Smeets, Lyn Healy, Edgar Luttmann, René de Gelder, Achim Zielesny, Mariano Rico, Gillian Townend, David van Enckevort, Alastair Dunning, Henk J van Kranen, Annika Jacobsen, Jan Aerts, Fernando Torres Andón, Peter Woollard, Marco Roos, Nikolay Kochev, Karima Rafes, Markus Hegi, Emilio Benfenati, Cécile Germain, Glenn Myatt, David Vorgrimmler, Bin Chen, Jörg K Wegner, Daniel Lemire, Jaap Heringa, C. Titus Brown, Penny Nymark, Oliver Fiehn, Steffen Neumann, Henning Hermjakob, Cameron Neylon, Lydia Pintscher, Tomasz Puzyn, Peter J. Schaap, Dario Greco, Alexander Mazein, **Matthew J Conroy**, Venkata Chandrasekhar Nainala, Benjamin M. Good, Andrew I. Su, Maria Rigau, Jonathan Bisson, Jorrit Poelen, Alexander Malic, Lauren J Dupuis, Natasha M Sanabria, Joris T.K. Quik, Panagiotis Isigonis, Eleonora Longhin, Jiří Vondrášek, Ineke Malsch, Francesco Messina, Anna Rybinska-Fryca, Serena Bonaretti, Willem Melssen, Kaido Tämm, James G Graham, Maria Sorokina, Annsofie Andersson, Mark D. Hoover, Mingxun Wang, Brian Halligan, Anna Niarakis, Ammar Ammar, Thomas E Exner, Charles Auffray, Aurelio Orta-Resendiz, Tobias Kuhn, Marcel Leist, Marvin Martens, Rafael C. Jimenez, Anna Forsby, Warwick Dunn, Ahmed Ghallab, Stuart J. Chalk, Christoph van Thriel, Carlos Fernandez-Lozano, Tomáš Pluskal, Susan Coort, Paul Groth, Nina Jeliazkova, Ola Spjuth, Andra Waagmeester, Barry Hardy, Vedrin Jeliazkov, Ryan Miller, Jérôme Pansanel, Carl Mäsak, Gilleain Torrance, Anja Jentzsch, Kristina M. Hettne, Elisa Cirillo, Anders Riutta, **Denise Slenter**, Shamith A Samarajiwa, Nathan Salomonis, Thomas Hankemeier, John W Mayfield, Andreas Bender, Martin Paul Brändle, José Emilio Labra Gayo, Najko Jahn, Arnaud Gaudry, Zsuzsanna Hollander, Marc Gillespie, Jean-Luc Wolfender, Pierre-Marie Allard, Nick Lynch, Dan Brickley, Anna Beronius, Emmanuel Barillot, Andreas Dräger, Albert Meroño-Peñuela, Angela Serra, Michel Dumontier, **Alexander R. Pico**, Jarl Wikberg, Noel O'Boyle, Haralambos Sarimveis, Pekka Kohonen, Cristian R. Munteanu, Anwesha Bohler, Martin Eklund, Georgia Tsiliki, Thomas Kelder, Kristina Hanspers, John Rumble, Christine Ogilvie Hendren, Harry Vrieling, Toby Hudson, Enrico Mombelli, Rosemarie Marchan, Jan G Hengstler, Matteo Floris, Ewy A Mathé, Tim Ebbels, Daniel Schober, Gareth Owen, Johannes Rainer, Miquel Rojas-Chertó, Fabien Jourdan, Christian Hoppe, Hanna L. Karlsson, Iseult Lynch, Quentin Groom, Masanori Arita, Salvador Capella-Gutierrez, Joaquín Dopazo, Akira Funahashi, Martin Himly, Guido F Pauli, Antreas Afantitis, Jonathan P. Tennant, Rebecca Ceder, Gregor Hagedorn, Susanna-Assunta Sansone, Lynn Schriml, Lars Carlsson, Michiel E. Adriaens, Tobias Kind, Gert Wohlgemuth, David M Jessop, Lezan Hawizy, Reza M. Salek, Asunción Gómez Pérez, Samuel Lampa, Maris Lapins, Obi Griffith, Krister Wennerberg, Tobias Schulze, Juuso A Parkkinen, Kairi Koort, Merlijn van Rijswijk, Charlie Beirnaert, Namrata Kale, Etienne A Thévenot, Ginger Tsueng, Sandra Karcher, Maria Dusinska, Elise Rundén-Pran, Irene Liampa, Laurent Winckers, Antonio Marcomini, Reinhard Schneider, Daniel Fernández-Álvarez, Lisa M. Verhagen, **Adriano Rutz**, Benjamin D Trump, Mark R Wiesner, J.P.M. van Rijn, Nhung Pham, Nathan Bossa, Qasim Chaudhry, Ali Beitollahi, Peter Buijnsters, Martijn P van Iersel, Jonathan Alvarsson, Matthias Samwald, Eric G. Prud'hommeaux, Colin Batchelor, Nuno Nunes, Valery Tkachenko, Sam E. Adams, Geoffrey R. Hutchison, Susie Stephens, Christian Brenninkmeijer, Antonis Loizou, Arfon Smith, Malachi Griffith, Mark D. Wilkinson, Núria Queralt Rosinach, Daniel Kinzler, Simon J. Coles, Leopold M G Curfs, Jörg Kurt Wegner, Patrik Rydberg, Karol M Langner, Kalai Vanii Jayaseelan, Sravanthi R Sinha, Jonathan Mélius, Charalampos Chomenidis, Pantelis Sopasakis, Janos Hajagos, Nico Adams, Neil Saunders, Valentin Georgiev, Jan Stanstrup, Sarah M Keating, Jose L Izquierdo-Garcia, Hendrik Treutler, Kristian Peters, Kevin Hybiske, Erik Schultes, Michael Witting, Agustin Gonzalez-Vicente, Erik Danen, Chunlei Wu, Gianluigi Zanetti, Nico Sommerdijk, Martin Feiters, Annemarie H Meijer, Rick Helmus, Susanne Hougaard Bennekou, Nils Hoffmann, Cristina Cadenas, Marek Ostaszewski, Melanie Imming, Ralf Stephan, Corey D. Broeckling, Axel P. Mustad, Bjørn E V Koch, Michael A Stravs, Roger Tu, Michael Mayers, Sabah Ul-Hasan, Thomas Naake, Garima Thakur, Andrea Terron, Natalie Meyers, Jeaphianne van Rijn, Ozan Çınar, Jente Houweling

NUTRIM

https://scholia.toolforge.org/author/Q20895241

Maastricht University | Maastricht UMC+