

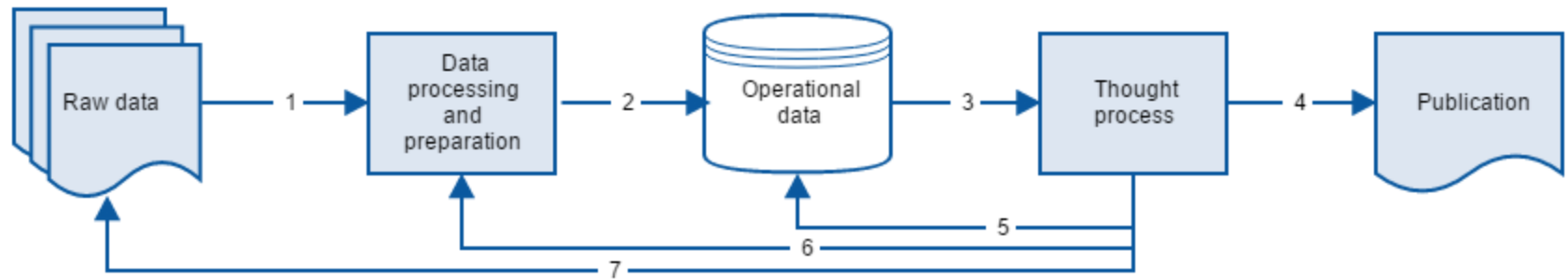
# Chemical databases – challenges and solutions

*Valery Tkachenko*  
*Royal Society of Chemistry*

GC&E  
June 2016



## Science data publishing workflow



# Chemical database

NamesAndSynonyms  
Ethyl DL-mandelate; ethyl 2-hydroxy-2-phenylacetate

Structure

Cid  
1

CaloguesNumber  
MP-00015

CAS

Formula  
C10H12O3

MolWeight  
180.20

Notes

SellUnit  
5.00

Measure  
g

Currency  
USD

Purity  
99.00

IsAvailable  
Available

TotalAvailable  
3 kg

SaltData  
hydrochloride

Solubility

MP  
35 deg C

BP  
115 deg C

Density  
0.86

logP

logD

a-pogP

c-logP

logS

logSW

TPSA

Catalogue-For-MolPort: 1 out of 1 rows.

Chemical Database		
Chemical Name	Representation	Molar Mass
Benzene	c1ccccc1	78.1118
Ethanol	CCO	46.0684
Freon	ClC(Br)CFFF	197.382
Formaldehyde	cO	30.026
Methane	C	16.0425
Methanol	CO	32.0419
Propanol	CCOC	60.1
Toluene	Cc1ccccc1	92.1384
Indole	c1ccc2cc[nH]c2c1	117.148
Ammonia	N	17.0305

# PubChem

Databases > Upload Services > Help more > Today's Statistics >

Clear

## PubChem



BioAssay



Compound



Substance

aspirin

Go

Limits

Advanced



BioAssay Tools

Structure Search

3D Conformer Tools

Structure Clustering

Try the new PubChem Search

**New** PubChem presents at the 251st American Chemical Society Meeting in San Diego (March 13-17, 2016). Read more at <http://1.usa.gov/1QBp0a>

**New** A new article about the PubChem Compound and Substance Search. [Read more...](#)

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

NCBI Resources How To

Sign in to NCBI

PubChem  
Compound

PubChem Compound aspirin

Create alert Limits Advanced

Search

Help

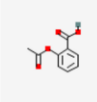
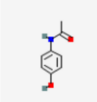
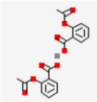
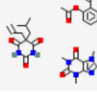

Summary 20 per page Sort by Default order

Send to: Filters: [Manage Filters](#)

### Search results

Items: 1 to 20 of 102

<< First < Prev Page 1 of 6 Next >> Last >>

- ☐   
**aspirin: ACETYLSALICYLIC ACID; 2-Acetoxybenzoic acid ...**  
MW: 180.157420 g/mol MF: C<sub>9</sub>H<sub>8</sub>O<sub>4</sub>  
IUPAC name: 2-acetoxybenzoic acid  
Create Date: 2004-09-16  
CID: 2244  
[Summary](#) [Similar Compounds](#) [Same Parent Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)  
[Active in 132 of 3660 BioAssays](#)
- ☐   
**acetaminophen: 4-Acetamidophenol; Paracetamol ...**  
MW: 151.162560 g/mol MF: C<sub>9</sub>H<sub>9</sub>NO<sub>2</sub>  
IUPAC name: N-(4-hydroxyphenyl)acetamide  
Create Date: 2004-09-16  
CID: 1983  
[Summary](#) [Similar Compounds](#) [Same Parent Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)  
[Active in 62 of 2956 BioAssays](#)
- ☐   
**Calcascorbin: Ascal; Calcium aspirin ...**  
MW: 398.376960 g/mol MF: C<sub>18</sub>H<sub>14</sub>CaO<sub>9</sub>  
IUPAC name: calcium;2-acetoxybenzoate  
Create Date: 2005-08-08  
CID: 6247  
[Summary](#) [Similar Compounds](#) [Same Parent Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)
- ☐   
**Axotal: BUTALBITAL ASPIRIN AND CAFFEINE; BUTAL COMPOUND ...**  
MW: 598.604360 g/mol MF: C<sub>28</sub>H<sub>34</sub>N<sub>6</sub>O<sub>9</sub>  
IUPAC name: 2-acetoxybenzoic acid;5-(2-methylpropyl)-5-prop-2-enyl-1,3...  
Create Date: 2008-07-11  
CID: 24847961  
[Summary](#) [Similar Compounds](#) [Same Parent Connectivity](#) [Mixture/Component Compounds](#) [PubMed \(MeSH Keyword\)](#)
- ☐   
**nitroaspirin: NO-Aspirin 1; Ncx 4016 ...**

### Actions on your results

**BioActivity Analysis**  
Analyze the BioActivities of the compounds

**Structure Clustering**  
Cluster structures based on structural similarity

**Structure Download**  
Download the structures in various formats

**Pathways**  
Analyze pathways containing the compounds

### Refine your results

**Chemical Properties**  
Rule of 5 (29)


**BioActivity Experiments**  
BioAssays, Active (12)  
BioAssays, Tested (19)  
Protein 3D Structures (4)  
Crystal Structure Of Asteropsin A From Marine Sponge Asteropus Sp (1)  
Structure Of Isopropylmalate Dehydrogenase From Thermus Thermophilus In Complex With Ipm, Mn And Nadh (1)  
The Structure Of And Photolytic Induced Changes Of Carbon Monoxide Binding To The Cytochrome Ba3-Oxidase From Thermus Thermophilus (1)  
... All 98 Structures

**BioMedical Annotation**  
Pharmacological Actions (20)  
Anti-Inflammatory Agents, Non-Steroidal (15)  
Cyclooxygenase Inhibitors (8)  
Platelet Aggregation Inhibitors (7)



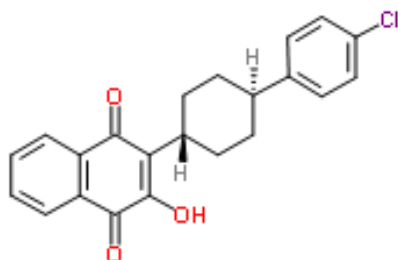
# ChemSpider

Search and share chemistry

- 52 million chemicals and growing
  - Data sourced from >500 different sources
  - Crowdsourced curation and annotation
  - Ongoing deposition of data from our journals and our collaborators
  - A structure centric hub for web-searching
- 

# ChemSpider

Search term: **atovaquone** (Found by approved synonym) [?](#)



[?](#) 2D 3D Save Edit Zoom

 - 2 of 2 defined stereocentres

## Atovaquone

ChemSpider ID: **10482034**

Molecular Formula:  $C_{22}H_{19}ClO_3$

Average mass: 366.837494 Da

Monoisotopic mass: 366.102264 Da

### ▼ Systematic name

2-[trans-4-(4-Chlorophenyl)cyclohexyl]-3-hydroxy-1,4-naphthoquinone

► SMILES and InChIs

► Cite this record

Wikibox

Embed

Deprecate

Watch this record

Manage data slice

# ChemSpider

## ▼ Names and Identifiers

Names and Synonyms Database ID(s)

Validated by Experts, Validated by Users, Non-Validated, Removed by Users, Redirected by Users, Redirect Approved by Experts

(-)-Cholesterol

(3b)-cholest-5-en-3-ol

(3S,8S,9S,10R,13R,14S,17R)-10,13-Dimethyl-17-[(2R)-6-methyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopent a[a]phenanthren-3-ol

(3S,8S,9S,10R,13R,14S,17R)-10,13-Diméthyl-17-[(2R)-6-méthyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tétradécahydro-1H-cyclopent a[a]phénanthrén-3-ol [French]

(3β)-cholest-5-en-3-ol [ACD/IUPAC Name]

(3β)-Cholest-5-en-3-ol [German] [A]

(3β)-Cholest-5-én-3-ol [French] [A]

## ▼ ChemSpider Searches

## ▼ Properties

Experimental data Predicted - ACD/Labs Predicted - EPI Suite Predicted - ChemAxon

Data supplied by datasources and users.

### • Experimental Physico-Chemical Properties

Experimental Melting Point: ?

149 °C Tokyo Chemical Industry Ltd C0318

147-150 °C Alfa Aesar

148-150 °C Oxford University Chemical Safety Data <http://msds.chem.ox.ac.uk/CH/cholesterol.html>

147-150 °C Alfa Aesar A11470

Experimental Boiling Point: ?

360 °C Alfa Aesar

360 °C Oxford University Chemical Safety Data <http://msds.chem.ox.ac.uk/CH/cholesterol.html>

360 °C Alfa Aesar A11470

Experimental Optical Rotation: ?

-36 Alfa Aesar A11470

Experimental Gravity: ?

1.067 g/mL Alfa Aesar A11470


### • Predicted Physico-Chemical Properties


Predicted Melting Point: ?

149 °C Tokyo Chemical Industry Ltd

149 °C Tokyo Chemical Industry Ltd C0318

## Search external sites for this structure:

 Search Google Scholar (by synonym)

 Search Google for exact structure

 Search Google for structures with same skeleton

ACD/Labs10281015312D

```
9 9 0 0 0 0 0 0 0 0 0 1 v2000
1.0787 0.0000 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 -0.7824 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.4120 -2.0463 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7407 -2.0463 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.1528 -0.7824 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5231 -3.1204 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.9815 -4.3380 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3.8472 -2.9861 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4.6296 -4.0602 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
1 5 1 0 0 0 0
1 2 1 0 0 0 0
2 3 2 0 0 0 0
3 4 1 0 0 0 0
4 5 2 0 0 0 0
4 6 1 0 0 0 0
6 7 2 0 0 0 0
6 8 1 0 0 0 0
8 9 1 0 0 0 0
```

M END

> <Catalog\_Number>

AS1-0101

> <CAS\_Number>

2703-17-5

> <Name>

Methyl 1H-pyrrole-3-carboxylate

\$\$\$\$

ACD/Labs10281015312D

```
8 8 0 0 0 0 0 0 0 0 0 1 v2000
1.0794 0.0000 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 -0.7803 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.4118 -2.0460 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7426 -2.0460 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.1544 -0.7803 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5228 -3.1210 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3.8450 -2.9823 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.9810 -4.3348 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
1 2 1 0 0 0 0
1 5 1 0 0 0 0
2 3 2 0 0 0 0
3 4 1 0 0 0 0
4 5 2 0 0 0 0
4 6 1 0 0 0 0
6 7 2 0 0 0 0
6 8 1 0 0 0 0
```

M END

> <Catalog\_Number>

AS1-0102

> <CAS\_Number>

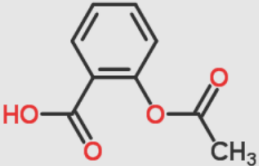
3/3/931



# ChemSpider curation

**Aspirin**

Molecular Formula:  $C_9H_8O_4$   
Average mass: 180.157 Da  
Monoisotopic mass: 180.042252  
ChemSpider ID: 2157



[3D](#)

**More details:**

analgesic anti-inflammatory drug antipyretic antirheumatic drug

**Names and identifiers** Properties Searches Spectra Vendors Aliases

Names and Synonyms Database ID(s)

Validated by Experts, Validated by Users, Non-Validated, Removed by Users

2-(Acetyloxy)benzoic acid  
200-064-1 [\[EINECS\]](#)  
2-Acetoxybenzenecarboxylic acid  
2-Acetoxybenzoesäure [German] [ACD/IUPAC Name]  
2-Acetoxybenzoic acid [ACD/IUPAC Name]  
2-Acetyloxybenzoic acid  
2-Carboxyphenyl acetate  
50-78-2 [\[RN\]](#)  
A.S.A.  
Acesan [Trade name]  
[More...](#)

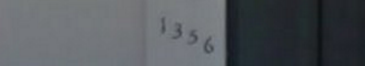
**Add Tag**

Tag Label - Add new or select using autocomplete (100 chars max)

Definition - A textual description of the tag (1000 chars max)

External Url - A link to an external resource which describes this tag

Comment - Optional



Type the text

[Privacy & Terms](#)

**COMMENT ON THIS RECORD**

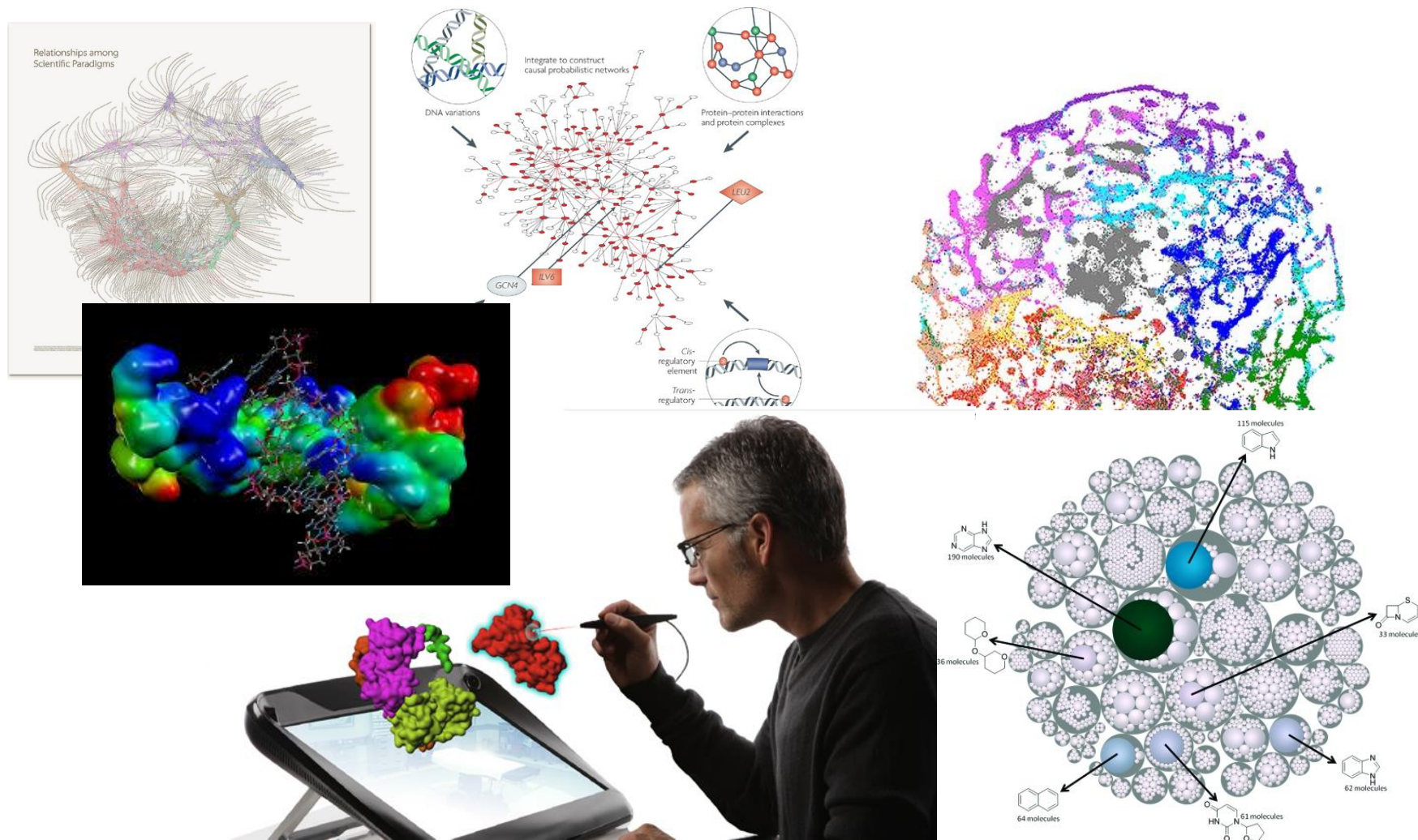
Featured data source

The Merck Index Online has more data on this compound

**EDIT**

**ADD TAG** **CANCEL**

# Dimensions and complexity of science



# ChemSpider Synthetic Pages

Compounds


Reaction

Analytical Data

Text and References

**Tandem devinylation-diformylation** **2,4-Diformyl-1,3,5-trimethoxybenzene**

SyntheticPage 714  
DOI: 10.1039/B7F714  
Submitted Jan 29, 2014, published Feb 02, 2014  
Sandip Bharate (sbharate@slim.ac.in)



**Chemicals Used**

1,3,5-Trimethoxy-2-((E)-3-methylbut-1-en-1-yl)benzene  
Trifluoroacetic acid (BD Fine Chemicals Ltd)  
Hexamethylenetetramine (HMTA) (Sigma Aldrich), ACS reagent, ≥99.0%, 398160  
Sodium bicarbonate (BD Fine Chemicals Ltd)  
Ethyl acetate (BD Fine Chemicals Ltd), directly used without drying

**Procedure**

To a solution of 1,3,5-trimethoxy-2-((E)-3-methylbut-1-en-1-yl)benzene (100 mg, 1 mmol) in TFA (5 ml) was added hexamethylenetetramine (237 mg, 4 mmol) and resulting mixture was heated to reflux at 120 °C for 6 h. Completion of the reaction was monitored by TLC (by observing disappearance of starting material on TLC). The reaction mixture was cooled to room temperature and was neutralized with saturated NaHCO<sub>3</sub> solution and extracted with EtOAc (50 ml x 3). Combined organic layer was dried over anhydrous sodium sulphate and evaporated on rotary evaporator to give a yellow oil. Purification by silica gel (mesh 100-200) column chromatography using 40% EtOAc: hexane as eluent gave 2,4-diformyl-1,3,5-trimethoxybenzene as light yellow solid (80 mg, 84% yield). The product was characterized by comparison of melting point and NMR data with literature values (Kuhnert, N.; Rossignolo, G. M.; Lopez-Periago, A. *Org. Biomol. Chem.* 2003, 1, 1157-1170. <http://dx.doi.org/10.1039/B212102F>)

**Author's Comments**

Variety of aliphatic vinyl groups can be utilized but the reaction is limited to 1,3,5-trimethoxybenzene (For details, see: *Tetrahedron Lett.* 2013, 54, 2913-2915. <http://dx.doi.org/10.1016/j.tetlet.2013.03.05>)

**Data**

2,4-Diformyl 1,3,5-trimethoxy benzene: Light yellow solid; m.p. 70-72 °C;  
1H NMR (CDCl<sub>3</sub>, 400 MHz) δ ppm 10.33 (s, 2H), 6.28 (s, 1H), 4.13 (s, 6H), 3.95 (s, 3H);  
IR (CHCl<sub>3</sub>) 3901, 3735, 3420, 2951, 2928, 2860, 1723, 1679, 1589, 1480, 1453, 1439, 1420, 1382, 1309, 1235, 1221, 1148, 1107, 1072, 1011 cm<sup>-1</sup>  
ESI-MS: m/z 225.07 [M+H]<sup>+</sup>, 247.05 [M+Na]<sup>+</sup>, 263 [M+K]<sup>+</sup>;  
HRMS: m/z 225.0761 calcd for C<sub>11</sub>H<sub>10</sub>O<sub>5</sub> + H<sup>+</sup> (225.0757).

**Lead Reference**

Bharate, S.B.; Mudududda, R.; Sharma, R.; Vishwakarma, R.A. The first method for C-vinylation of aromatic systems. *Tetrahedron Lett.* 2013, 54, 2913-291. <http://dx.doi.org/10.1016/j.tetlet.2013.03.05>

**Other References**

Duff, J. C.; Billis, E. J. *J. Chem. Soc.* 1932, 1997-1998. <http://dx.doi.org/10.1039/UR9320001997>  
Duff, J. C.; Billis, E. J. *J. Chem. Soc.* 1934, 1305-1308. <http://dx.doi.org/10.1039/UR9340001305>

**Supplementary Information**

1H NMR spectra (1H NMR spectra.doc)  
This page has been viewed approximately 365 times since records began.  
Get structure file (.cdx, .sk2, .mol)

**Keywords:** carboxylic compounds, devinylation, diformylation, Duff reaction, elimination, TFA, vinylbenzenes

# Our World is hyperconnected







# Data quality issues

Robochemistry

Proliferation of errors in public and private databases

Automated quality control system



# Standards?

## Blue Book [\[ edit \]](#)

**Nomenclature of Organic Chemistry**, commonly referred to by chemists as the **Blue Book**, is a collection of recommendations on [organic chemical nomenclature](#) published at irregular intervals by the [International Union of Pure and Applied Chemistry](#) (IUPAC). A full edition was published in 1979,<sup>[1]</sup> an abridged and updated version of which was published in 1993 as **A Guide to IUPAC Nomenclature of Organic Compounds**.<sup>[2]</sup> Both of these are now [out-of-print](#) in their paper versions, but are available free of charge in electronic versions. After the release of a draft version for public comment in 2004<sup>[3]</sup> and the publication of several revised sections in the journal *Pure and Applied Chemistry*, a fully revised version was published in print in 2013.<sup>[4]</sup>

## Gold Book [\[ edit \]](#)

The **Compendium of Chemical Terminology** is a book published by the [International Union of Pure and Applied Chemistry](#) (IUPAC) containing internationally accepted definitions for terms in [chemistry](#). Work on the first edition was initiated by [Victor Gold](#), hence its informal name, the **Gold Book**.

The first edition was published in 1987 (ISBN 0-63201-765-1) and the second edition (ISBN 0-86542-684-8), edited by A. D. McNaught and A. Wilkinson, was published in 1997. A slightly expanded version of the *Gold Book* is also freely searchable online. Translations have also been published in French, Spanish and Polish.

## Green Book [\[ edit \]](#)

**Quantities, Units and Symbols in Physical Chemistry**, commonly known as the **Green Book**, is a compilation of terms and symbols widely used in the field of physical chemistry. It also includes a table of physical constants, tables listing the properties of elementary particles, chemical elements, and nuclides, and information about conversion factors that are commonly used in physical chemistry. The most recent edition is the third edition (ISBN 978-0-85404-433-7), originally published by IUPAC in 2007. A second printing of the third edition was released in 2008; this printing made several minor revisions to the 2007 text. A third printing of the third edition was released in 2011. The text of the third printing is identical to that of the second printing.

## Orange Book [\[ edit \]](#)

The **Compendium of Analytical Nomenclature** is a book published by the [International Union of Pure and Applied Chemistry](#) (IUPAC) containing internationally accepted definitions for terms in [analytical chemistry](#). It has traditionally been published in an orange cover, hence its informal name, the **Orange Book**.

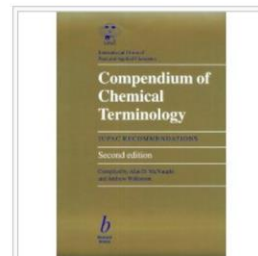
Although the book is described as the "Definitive Rules", there have been three editions published; the first in 1978 (ISBN 0-08022-008-8), the second in 1987 (ISBN 0-63201-907-7) and the third in 1998 (ISBN 0-86542-615-5). The third edition is also available online. A Catalan translation has also been published (1987, ISBN 84-7283-121-3).

## Purple Book [\[ edit \]](#)

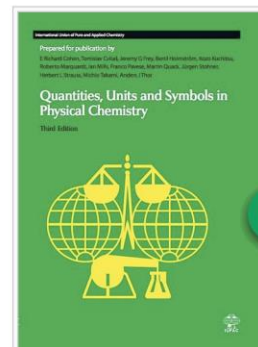
The first edition of the **Compendium of Macromolecular Terminology and Nomenclature**, known as the **Purple Book**, was published in 1991 and is now out of print.

## Red Book [\[ edit \]](#)

**Nomenclature of Inorganic Chemistry**, by chemists commonly referred to as the **Red Book**, is a collection of recommendations on [inorganic chemical nomenclature](#). It is published



The front cover of the second edition of the *Compendium of Chemical Terminology*. 63



# Standards?

Online Browsing Platform (OBP)



Search

ISO 11238:2012(en) x

**ISO 11238:2012(en)** Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on substances

## Table of contents

Available in: en fr

Foreword

Introduction

1 Scope

2 Terms, definitions, symbols and abbreviations

2.1 Terms and definitions

2.2 Symbols and abbreviated terms

3 Requirements

3.1 General

3.2 Concepts required for the unique identification

3.3 Concepts required for the description

3.4 Naming of substances

3.5 Requirements for unique identifiers

3.6 Types of substances

3.7 Defining specified substances

Annex A Existing identifiers and molecular structure representation

A.1 Identifiers

A.2 Molecular structure representation

Bibliography

## Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 11238 was prepared by Technical Committee ISO/TC 215, *Health informatics*.

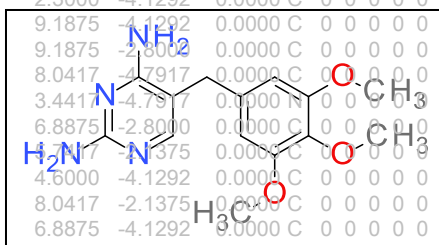
## Introduction

# FDA Substance Registry System SPL Substance Index Files

-FDASRS-04291423352D

21 22 0 0 0 0 0 0 0 0999 V2000

2.3000	-2.8000	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.4417	-2.1375	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.6000	-2.8000	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.3000	-4.1292	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.1875	-4.1292	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.1875	-2.8000	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.0417	-4.7917	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.4417	-4.7917	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6.8875	-2.8000	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.6000	-4.1292	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.0417	-2.1375	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6.8875	-4.1292	0.0000	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.4417	-0.8125	0.0000	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



## ■ Chemical substance

- UNII: AN164J8Y0X
- Chemical structure (MOLFILE)

- InChI=1S/C14H18N4O3/c1-19-10-5-8(6-11(20-2)12(10)21-3)4-9-7-17-14(16)18-13(9)15/h5-7H,4H2,1-3H3,(H4,15,16,17,18)
- IEDVJHCEMCRBQM-UHFFFAOYSA-N

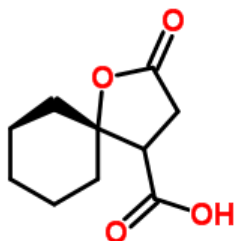
## ■ Biological substance (plant)

- UNII: 1KE45XD28S
- Bibliographic reference: Cichorium intybus L.



# ChemSpider issues

Search term: **85940** (Found by CSID) ?



? 2D 3D Save Zoom

## 2-oxo-1-oxaspiro[4.5]decane-4-carboxylic acid

ChemSpider ID: **85940**

Molecular Formula:  $C_{10}H_{14}O_4$

Average mass: 198.215805 Da

Monoisotopic mass: 198.089203 Da

▼ Systematic name

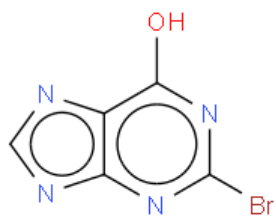
2-Oxo-1-oxaspiro[4.5]decane-4-carboxylic acid

► SMILES and InChIs

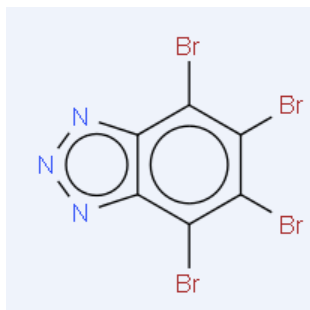
► Cite this record

# DrugBank dataset (6516 records)

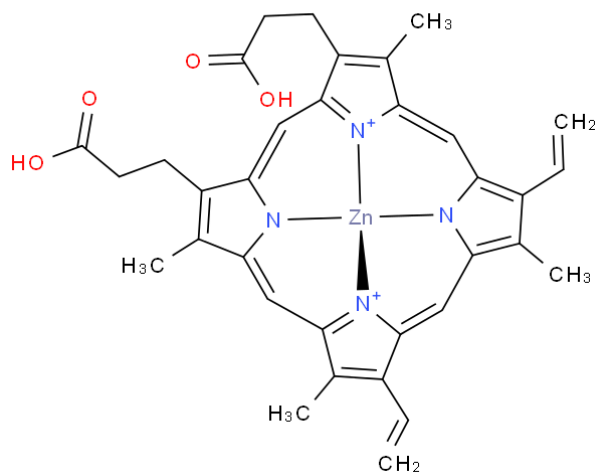
~60 records that can't be dearomatized unambiguously



DB04283

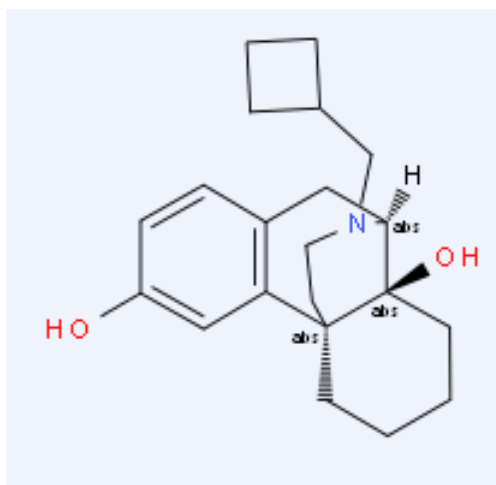


DB04462

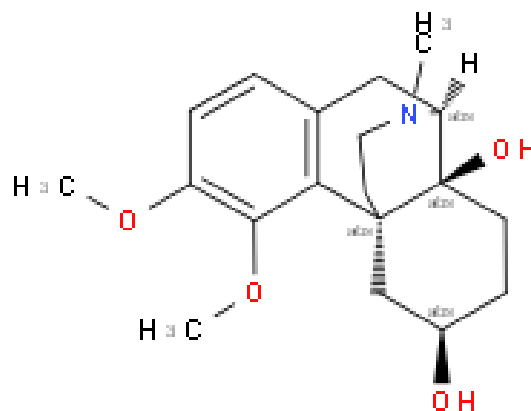


DDB04009

2 records where Smiles, InChI, and name did not match the structure



DB00611

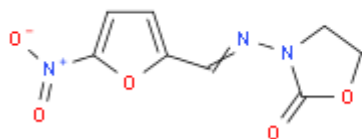
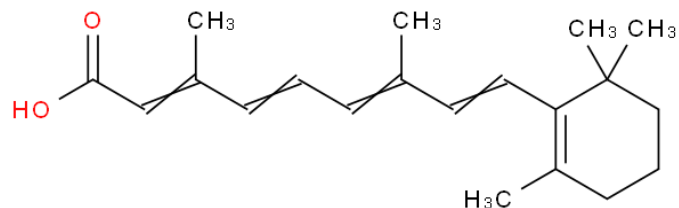


DB01547

## ~40 records where InChIs did not match the structure

DrugBank ID: DB00755

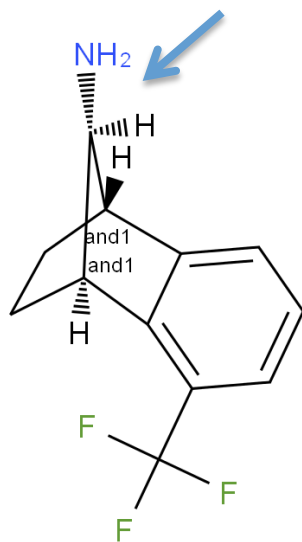
InChI=1S/C20H28O2/c1-15(8-6-9-16(2)14-19(21)22)11-12-18-17(3)10-7-13-20(18,4)5/h6,8-9,11-12,14H,7,10,13H2 1-5H3 (H 21 22)/b9-6+ 12-11+ 15-8+ 16-14+



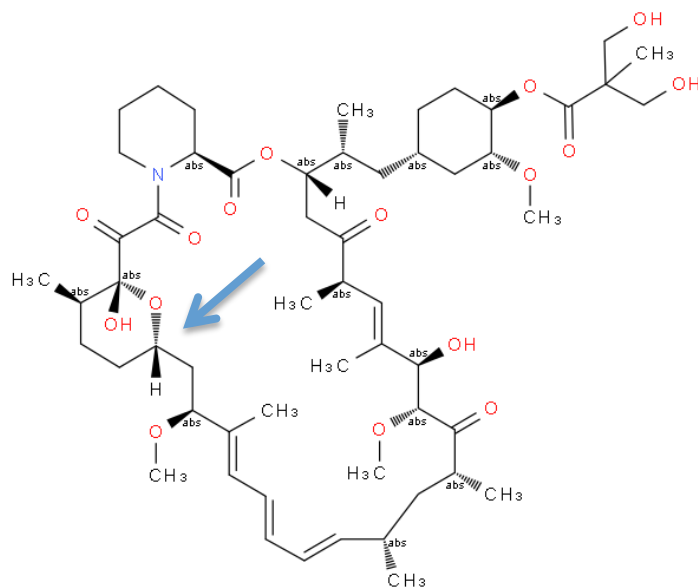
Warn	contains unknown stereo bond
Warn	depositor-specified name(s) do not match the structure : furazolidone
Warn	depositor-specified InChIs do not match the structure : InChI=1S/C8H7N3O5/c12-8-10(3-4-15-8)9-5-6-1-2-7(16-6)11(13)14/h1-2,5H,3-4H2/b9-5+

DruGBank ID: DB00614

7 records with 2 stereo bonds at chiral atoms



DB08128



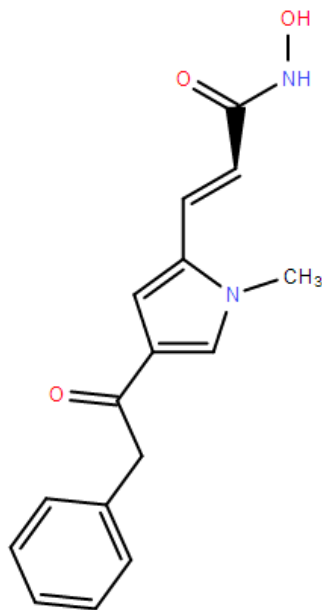
DB06287

J. Brechner, IUPAC  
Graphical Representation of  
stereochem. configurations  
Section: ST-1.1.10



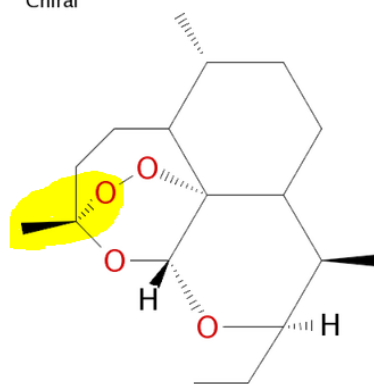
*Not acceptable*

“Direction of bond makes no sense”



# “Stereo types of the opposite bonds mismatch”

Chiral



CHEMBL12270

1916

J. BRECHER

## ST-1.1.10

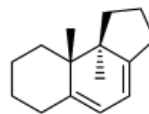
Two plain bonds, one solid wedged bond, and one hashed wedged bond, with the two plain bonds separated by other than 180° and not depicted as adjacent.



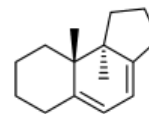
Not acceptable



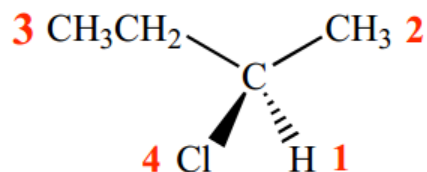
Not acceptable



Not acceptable



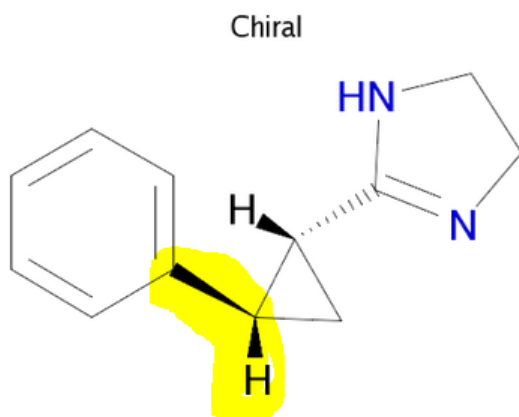
Preferred



<http://www.iupac.org/publications/pac/2006/pdf/7810x1897.pdf>



“Stereo types of non-opposite bonds match”



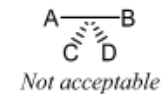
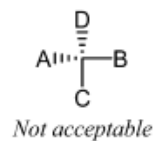
CHEMBL93192

*Graphical representation of stereochemical configuration*

1917

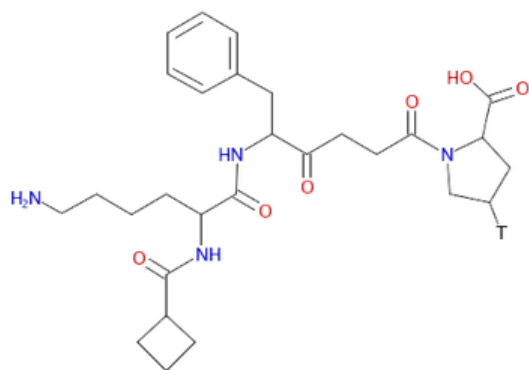
**ST-1.1.11**

Two wedged bonds of the same type (either solid wedged or hashed wedged) and two plain bonds, with each adjacent pair of bonds separated by 180° or less, and with similar bond types depicted as adjacent.



These depictions are formally ambiguous and cannot be interpreted with certainty. They should never be used.

“atom not recognized”



CHEMBL10002

In molfile:

2.6833 -3.5125 0.0000 T 0 0 0 0 0 0 0

Should be atom from periodic table

No mass difference in atom line

No “M ISO” in connection table

# Open PHACTS



**GlaxoSmithKline – Coordinator**  
**Universität Wien – Managing entity**  
Technical University of Denmark  
University of Hamburg, Center for Bioinformatics  
BioSolveIT GmbH  
Consorti Mar Parc de Salut de Barcelona  
Leiden University Medical Centre  
Royal Society of Chemistry  
Vrije Universiteit Amsterdam  
Novartis  
Merck Serono  
H. Lundbeck A/S  
Eli Lilly  
Netherlands Bioinformatics Centre  
Swiss Institute of Bioinformatics  
ConnectedDiscovery  
EMBL-European Bioinformatics Institute  
Janssen Esteve Almirall  
OpenLink Scibite  
The Open PHACTS Foundation  
Spanish National Cancer Research Centre  
University of Manchester  
Maastricht University  
Aqnowledge  
University of Santiago de Compostela  
Rheinische Friedrich-Wilhelms-Universität Bonn  
AstraZeneca  
Pfizer



[info@openphactsfoundation.org](mailto:info@openphactsfoundation.org)



@Open\_PHACTS

# Why is it so hard to....

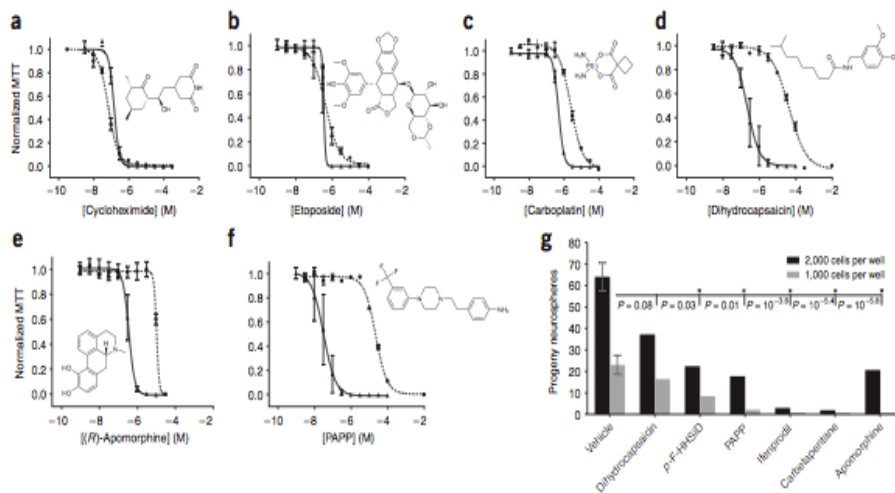
LETTERS

NATURE CHEMICAL BIOLOGY VOLUME 3 NUMBER 5 MAY 2007

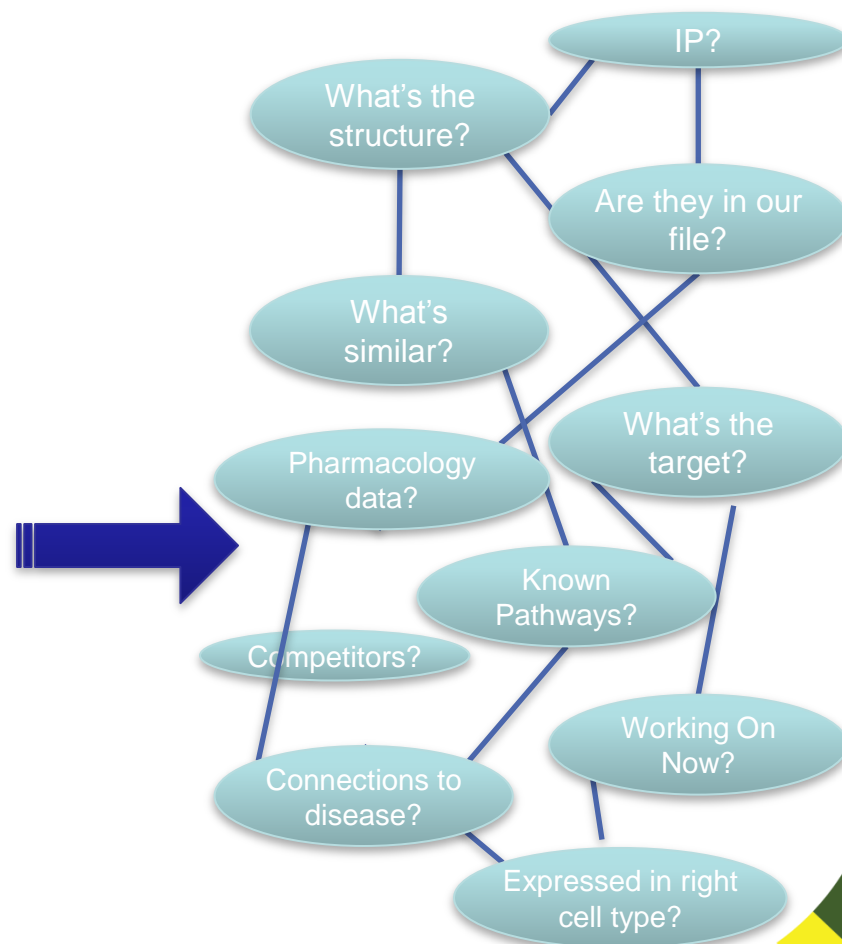
nature  
chemical biology

## Chemical genetics reveals a complex functional ground state of neural stem cells

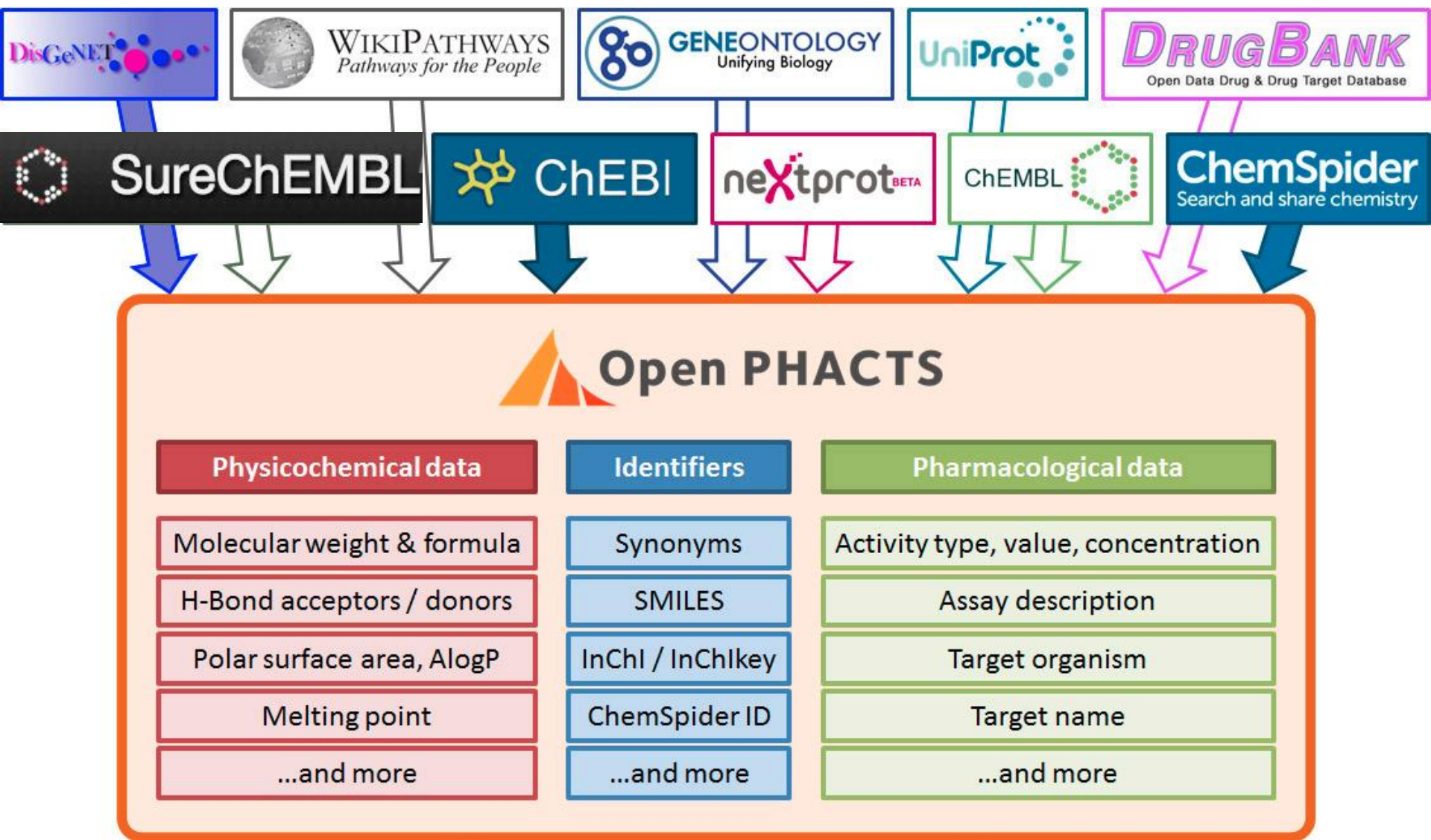
Phedias Diamandis<sup>1-4</sup>, Jan Wildenhain<sup>4</sup>, Ian D Clarke<sup>1,2</sup>, Adrian G Sacher<sup>1,2</sup>, Jeremy Graham<sup>1,2</sup>, David S Bellows<sup>3</sup>, Erick K M Ling<sup>1,2,5</sup>, Ryan J Ward<sup>1,2,5</sup>, Leanne G Jamieson<sup>1,2,5</sup>, Mike Tyers<sup>3,4</sup> & Peter B Dirks<sup>1,2,5,6</sup>



**Figure 2** Identification of potent NPC-specific compounds. (a-f) Dose-response curves and chemical structures of controls: cycloheximide (a), etoposide (b) and carboplatin (c), and of selected newly identified compounds: dihydrocapsaicin (d), apomorphine (e) and PAPP (f). Each plot shows the fitted sigmoidal logistic curve to MTT proliferation assay readings of both astrocytes (-●-) and neurosphere cultures (-▲-). Values represent the mean and

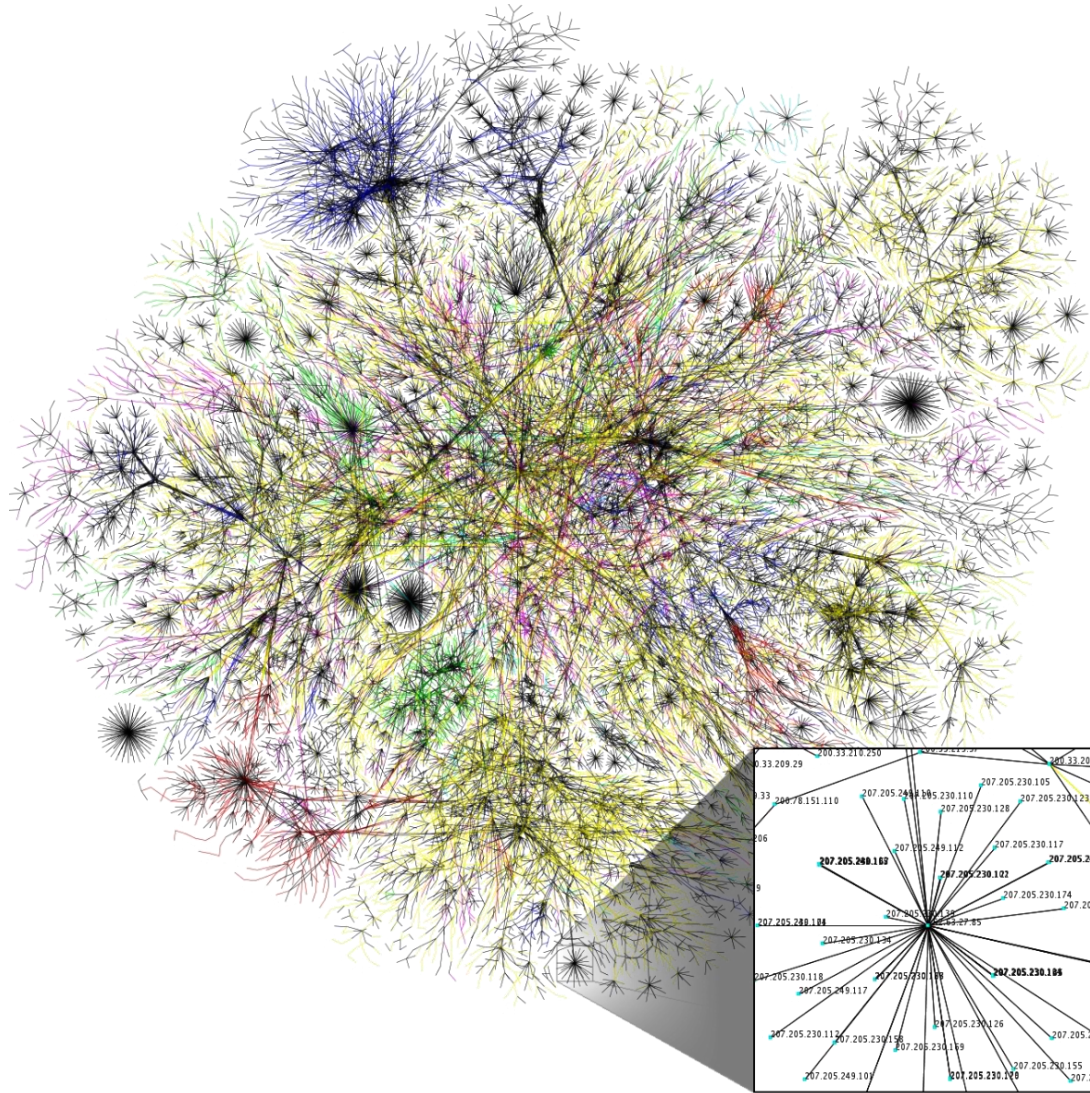


# Knowledge is federated



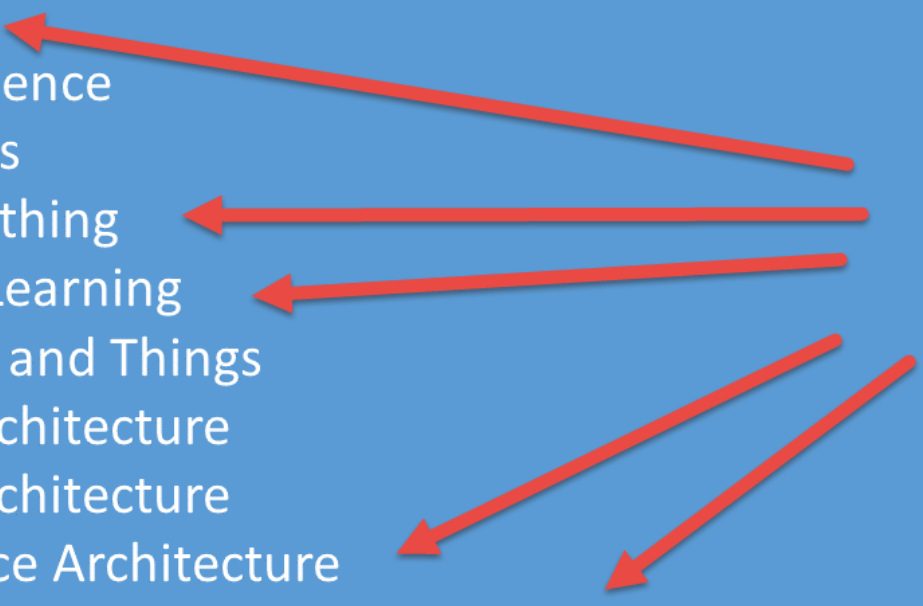


# The World we live in



# The World we are heading into

## Gartner's Top 10 Strategic Technology Trends for 2016

1. The Device Mesh
  2. Ambient User Experience
  3. 3D-Printing Materials
  4. Information of Everything
  5. Advanced Machine Learning
  6. Autonomous Agents and Things
  7. Adaptive Security Architecture
  8. Advanced System Architecture
  9. Mesh App and Service Architecture
  10. Internet of Things Architecture and Platforms
- 



<http://www.wired.com/2014/04/google-project-ara/>

<http://www.wsj.com/articles/googles-modular-phones-to-go-on-sale-next-year-1463783371>



# CVSP

## Deposition Gateway

[Home](#)[Submit](#)[Depositions](#)[Profile](#)

Welcome, Alexey Pshenichnov!

[Sign out](#)

### Submit new deposition

Datasource

Select... 

File

Validate



Standardize



Calculate Properties



Parents Generation



Public



## Data Repository

### Deposition Gateway

[Home](#)[Submit](#)[Depositions](#)[Profile](#)

Welcome, Alexey Pshenichnov!

[Sign out](#)Depositions **15**

ID	Date	Datasource	Depositor	File Name	Status
cada2290-d1c8-4923-945c-1f4ac495c0f4	10/7/2015	WikiPathways	Aleksey Pshenichnov	WikiPathways.sdf	Processed
39df661f-c9f8-4490-9eae-f96b0261ad34	10/7/2015	Thomson Pharma	Aleksey Pshenichnov	ThomsonReuters.sdf.gz	Processed
7f8eaadd-7f01-42d5-b44a-96a3f8a11737	10/7/2015	MeSH	Aleksey Pshenichnov	mesh20151407final.sdf.zip	Deposited2GCN
600ca202-638c-4b18-9b2c-2cfe0cf87211	10/8/2015	ChEBI	Aleksey Pshenichnov	ChEBI_lite.sdf.gz	Deposited2GCN
859cea6d-33e0-475c-994b-c3c52dbcf71	10/8/2015	DrugBank	Aleksey Pshenichnov	drugbank.zip	Deposited2GCN
bf4b356e-381e-4b14-b54f-2f76b6528085	10/8/2015	Human Metabolome Database	Aleksey Pshenichnov	hmdb.zip	Deposited2GCN
b7b3eda8-35b0-44ba-8775-e76452460b87	10/9/2015	ChEMBL	Aleksey Pshenichnov	chembl_20.sdf	Deposited2GCN
01931b57-5498-4416-91bd-7dd83139f825	10/20/2015	MeSH	Aleksey Pshenichnov	mesh_20151016.zip	Processed
a9ca97f2-61e9-4df3-a3f6-2d18e3df1102	10/21/2015	PDB	Aleksey Pshenichnov	pdb3d_2.zip	Deposited2GCN

# CVSP – submission details

## Data Repository Deposition Gateway



[Home](#) [Submit](#) [Depositions](#) [Profile](#)

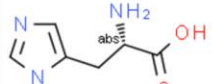
Welcome, Alexey Pshenichnov! [Sign out](#)

Deposition

[Annotations](#) [Jobs](#) [Chunks](#) [Delete](#) [Delete From GCN](#)

**Guid** 859cea6d-33e0-475c-994b-c3c52dbcfd71  
**Public** No  
**Datasource** DrugBank  
**Status** Deposited to GCN  
**Files** drugbank.zipall.sdf  
**Submitted** 2015/10/08  
**Records** **6837** [Errors - 41](#) [Warnings - 755](#) [Information - 2668](#)  
**Processing Parameters**  
**Validate** true  
**Standardize** true  
**PropertiesCalculation** true  
**ParentsGeneration** true

Records **6837**

Ordinal	Original	Issues
9		

Filter

By REGIDs

Example: regid1, regid2

Severities

- ☒ Error (41)
- ☐ Warning (755)
- ☐ Information (2668)

By Ordinals

Example: 1, 9, 23

Issue Types

- ☒ 100.26 - Smiles generation failed (1)
- ☒ 100.60 - Not able to properly dearomatize (14)
- ☒ 400.12 - Standardization failed (25)
- ☒ 200.20 - Ambiguous H (indigo) (2)
- ☒ 100.27 - Canonical smiles generation failed (7)
- ☒ 100.29 - InChI generation failed (4)
- ☒ 100.5 - Contains non aromatic query bond(s) (2)
- ☒ 500.1 - Processing operation failed (38)

Close

Reset

Apply

# CVSP – issues review

Records <b>41</b> - filtered by Severity (Error) and Issue Types (100.26, 100.60, 400.12, ...)				<div>Filter</div> <div></div>	
Ordinal	Original	Issues	Standardized		
7	<div></div>	<div><div> Contains non aromatic query bond(s)</div><div> Standardization failed</div><div> Processing operation failed</div><div> Processing operation failed</div><div> Processing operation failed</div><div>more...</div></div>			
571	<div></div>	<div><div> InChI generation failed</div><div> Radical count failed</div><div> Standardization failed</div><div> Processing operation failed</div><div> Processing operation failed</div><div>more...</div></div>	<div></div>		
813	<div><div></div><div>NH<sub>3</sub></div><div>NH<sub>3</sub></div></div>	<div><div> Custom user validation information</div><div> Custom user validation information</div><div> More than one instance of a same molecule</div><div> Standardization failed</div></div>	<div><div></div><div>NH<sub>3</sub> NH<sub>3</sub></div></div>		

# CVSP - mapping

**Data Repository**  
Deposition Gateway

Home Submit Depositions

ROYAL SOCIETY OF CHEMISTRY

Depositions by Pshenichnov! Sign out

Add Annotation

External ID DATABASE\_ID

Cancel Add

Deposition's Annotations

Field	Annotation
DATABASE_ID	External ID

+ Add Annotation

Back

Add Annotation

External ID

External ID

InChI

InChI Key

SMILES

Synonym

Xref

Comment

DATABASE\_ID

Add Annotation

External ID

DATABASE\_ID

DATABASE\_ID

DATABASE\_NAME

SMILES

INCHI\_IDENTIFIER

INCHI\_KEY

FORMULA

MOLECULAR\_WEIGHT

EXACT\_MASS

JCHEM\_ACCEPTOR\_COUNT

JCHEM\_AVERAGE\_POLARIZABILITY

JCHEM\_BIOAVAILABILITY

JCHEM\_DONOR\_COUNT

JCHEM\_FORMAL\_CHARGE

JCHEM\_GHOSE\_FILTER

JCHEM\_IUPAC

ALOGPS\_LOGP

JCHEM\_LOGP

ALOGPS\_LOGS

JCHEM\_MDDR\_LIKE\_RULE

JCHEM\_NUMBER\_OF\_RINGS

# CVSP – rules

[Feedback](#) [Help](#)

[Home](#) [Submit](#) [Depositions](#) [Profile](#)

☐ Email me when my submissions are processed

[My Rules](#) [Default CVSP Rules](#) [Community Rules](#) [Private User Rules](#)

ID	Title
1	Acid-Base rule set (default)
2	Validation rule set (default)
3	Standardization rule set (default)

Content Type	Validation rule set
Owner	CVSP
Date created	26/11/2014 20:33:29
Date revised	26/11/2014 20:33:29
Passed XML Validation:	True
Title	<input type="text" value="Validation rule set (default)"/>
Description	<input type="text" value="Validation rule set (default)"/>
XML Content	<pre>&lt;?xml version="1.0" encoding="utf-8" ?&gt; &lt;rules&gt;   &lt;moleculerules&gt;     &lt;!-- The SMARTS tests below are complimentary to set of validations that CVSP does by default --&gt;      &lt;Warning message="Contains cyclobutane" description="[CX4;H2;r4]1[CX4;H2;r4][CX4;H2;r4][CX4;H2;r4]1"&gt;       &lt;test name="SMARTSTest" param="[CX4;H2;r4]1[CX4;H2;r4][CX4;H2;r4][CX4;H2;r4]1"&gt;     &lt;/Warning&gt;     &lt;Warning message="Contains ethane" description="[CX4;H3][CX4;H3]"&gt;       &lt;test name="SMARTSTest" param="[CX4;H3][CX4;H3]"&gt;     &lt;/Warning&gt;     &lt;Warning message="Contains S with no explicit bonds" description="[S,D0]"&gt;       &lt;test name="SMARTSTest" param="[S,D0]"&gt;     &lt;/Warning&gt;     &lt;Warning message="Contains B with no explicit bonds" description="[B,D0]"&gt;       &lt;test name="SMARTSTest" param="[B,D0]"&gt;     &lt;/Warning&gt;      &lt;Warning message="Contains methane" description="[CX4;H4]"&gt;       &lt;test name="SMARTSTest" param="[CX4;H4]"&gt;     &lt;/Warning&gt;</pre>

[Revise](#) [Clone](#)

# CVSP – custom rules

[My Rules](#) [Default CVSP Rules](#) [Community Rules](#) [Private User Rules](#)

My Acid-Base Rules

[New Acid-Base Rules](#)

My Validation Rules

[New Validation Rules \(Smarts\)](#)

My Standardization Rules

[New Standardization Rules \(modules, Smir\)](#)

[Home](#) [Submit](#) [Depositions](#) [Profile](#) [Feedback](#) [Help](#)

☐ Email me when my submissions are processed

[My Rules](#) [Default CVSP Rules](#) [Community Rules](#) [Private User Rules](#)

Only Admins can see the content below. Only 3 default rules have to be set by admins: one validation, one acid-base, and one standardization. Default rules are visible to everybody, but only Admins can revise them.

User can make their own rule public (via Revise page) but it has to follow by Admin's approval to be visible to other users.

ID	User Name	Title	Platform Default	Shared / Approved
1	Karen Karapetyan	Acid-Base rule set (default)	True	False / False
2	Karen Karapetyan	Validation rule set (default)	True	False / False
3	Karen Karapetyan	Standardization rule set (default)	True	False / False
4	Alex Tester	Cloned content: Acid-Base rule set (default)	False	False / False

**METHODOLOGY**

**Open Access**



# The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets

Karen Karapetyan<sup>1\*</sup>, Colin Batchelor<sup>2</sup>, David Sharpe<sup>2</sup>, Valery Tkachenko<sup>1</sup> and Antony J Williams<sup>1,3</sup>

## Abstract

**Background:** There are presently hundreds of online databases hosting millions of chemical compounds and associated data. As a result of the number of cheminformatics software tools that can be used to produce the data, subtle differences between the various cheminformatics platforms, as well as the naivety of the software users, there are a myriad of issues that can exist with chemical structure representations online. In order to help facilitate validation and standardization of chemical structure datasets from various sources we have delivered a freely available internet-based platform to the community for the processing of chemical compound datasets.

**Results:** The chemical validation and standardization platform (CVSP) both validates and standardizes chemical structure representations according to sets of systematic rules. The chemical validation algorithms detect issues with submitted molecular representations using pre-defined or user-defined dictionary-based molecular patterns that are chemically suspicious or potentially requiring manual review. Each identified issue is assigned one of three levels of severity - Information, Warning, and Error - in order to conveniently inform the user of the need to browse and review subsets of their data. The validation process includes validation of atoms and bonds (e.g., making aware of query atoms and bonds), valences, and stereo. The standard form of submission of collections of data, the SDF file, allows the user to map the data fields to predefined CVSP fields for the purpose of cross-validating associated SMILES and InChIs with the connection tables contained within the SDF file. This platform has been applied to the analysis of a large number of data sets prepared for deposition to our ChemSpider database and in preparation of data for the Open PHACTS project. In this work we review the results of the automated validation of the DrugBank dataset, a popular drug and drug target database utilized by the community, and ChEMBL 17 data set. CVSP web site is located at <http://cvsp.chemspider.com/>.

**Conclusion:** A platform for the validation and standardization of chemical structure representations of various formats has been developed and made available to the community to assist and encourage the processing of chemical structure files to produce more homogeneous compound representations for exchange and interchange between online





[Personal](#)

[Open source](#)

[Business](#)

[Explore](#)

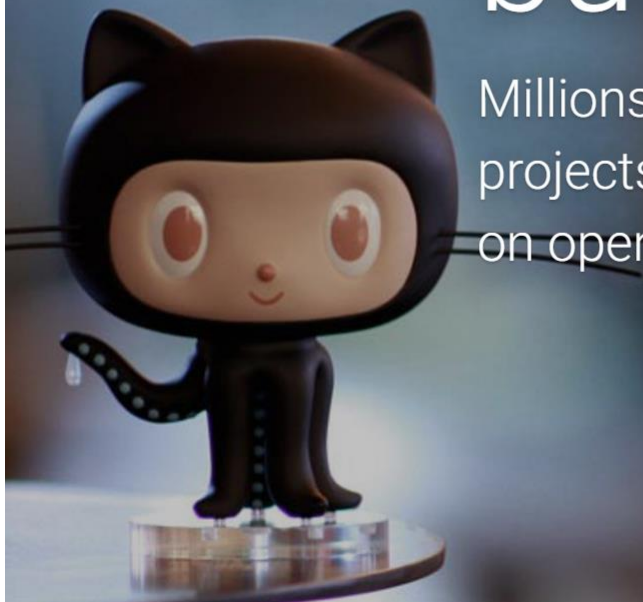
[Pricing](#)

[Blog](#)

[Support](#)

# How people build software

Millions of developers use GitHub to build personal projects, support their businesses, and work together on open source technologies.





&lt;&gt; Code

🔗 Pull requests 0

📈 Pulse

📊 Graphs

⚙️ Settings

## Chemistry Validation and Standardization Platform — Edit

🕒 2 commits

🌿 1 branch

📦 0 releases

👤 1 contributor

Branch: master ▾

New pull request

Create new file

Upload files

Find file

Clone or download ▾

🏠 valt Latest update

Latest commit dd75bef 4 minutes ago

📁 Source

Latest update

4 minutes ago

📄 .gitignore

Latest update

4 minutes ago

📄 README.md

Initial commit

12 days ago

📄 README.md

# cvsp

Chemistry Validation and Standardization Platform



# Thank you

**Email:** [tkachenkov@rsc.org](mailto:tkachenkov@rsc.org)

**Slides:**

<http://www.slideshare.net/valerytkachenko16>

