

# Open chemistry registry and mapping platform based on open source cheminformatics toolkits

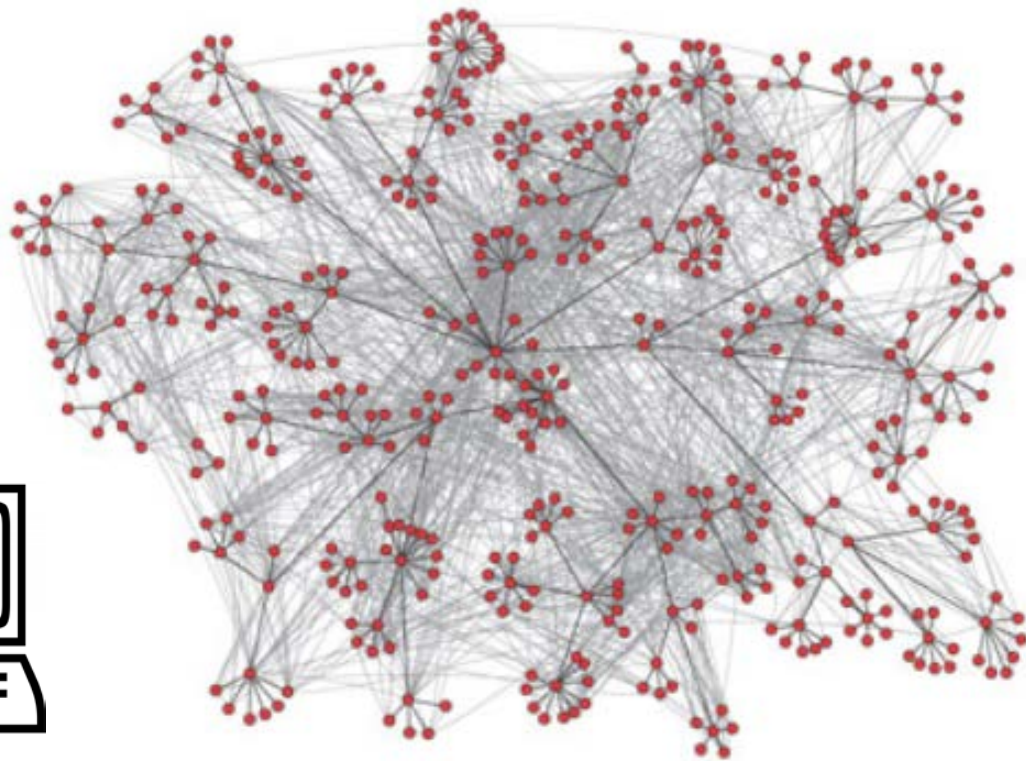
Fall ACS 2017, Washington, DC

*Valery Tkachenko, Denise Slenter, Nina Jeliaskova, Anna Gaulton,  
Antony Williams, Christoph Steinbeck, Chris Evelo, Egon Willighagen*





# We live in a hyperconnected World





# Data quality issues

Robochemistry

Proliferation of errors in public and private databases

Automated quality control system



# Standards and authorities

## *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**.

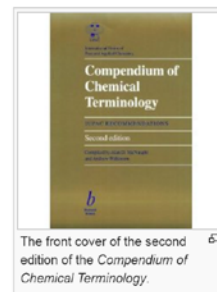
Although the book is described as the "Definitive Rules", there have been three editions published; the first in 1978 (ISBN 0-08022-008-8), the second in 1987 (ISBN 0-63201-907-7) and the third in 1998 (ISBN 0-86542-615-5). The third edition is also available online. A Catalan translation has also been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

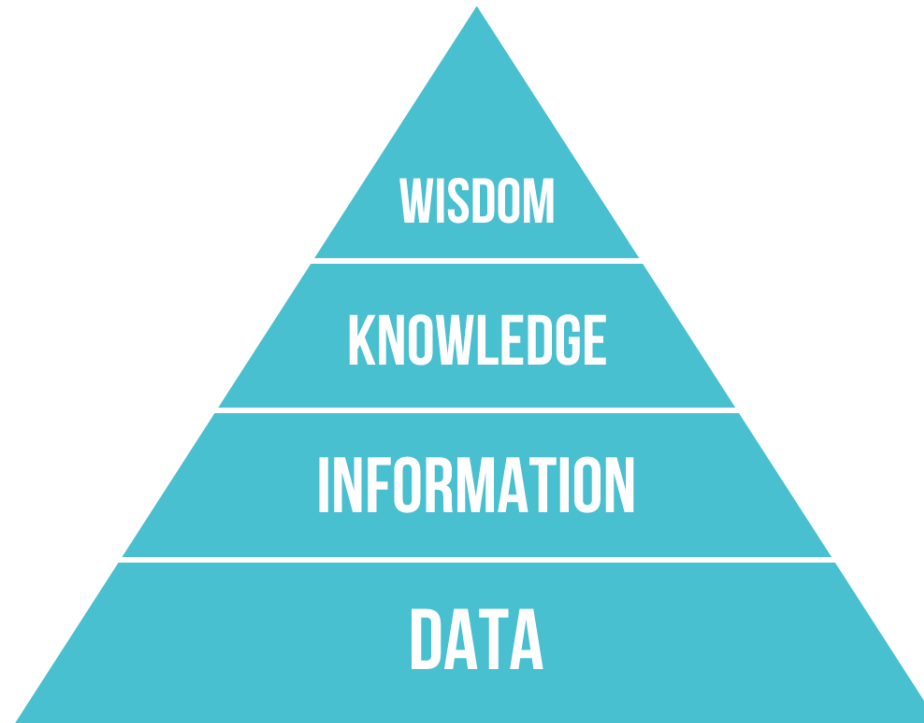
## *Red Book* [edit]

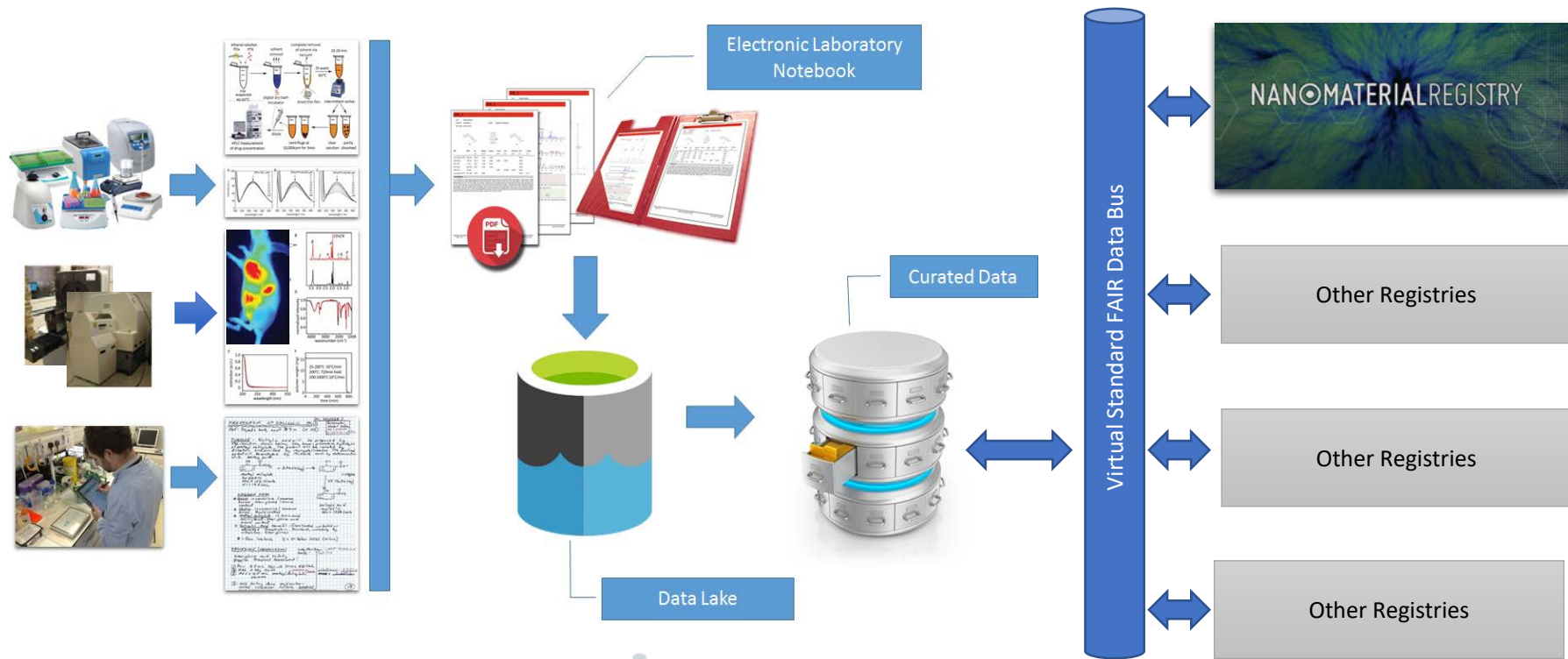
**Nomenclature of Inorganic Chemistry**, by chemists commonly referred to as the **Red Book**, is a collection of recommendations on [inorganic chemical nomenclature](#). It is published

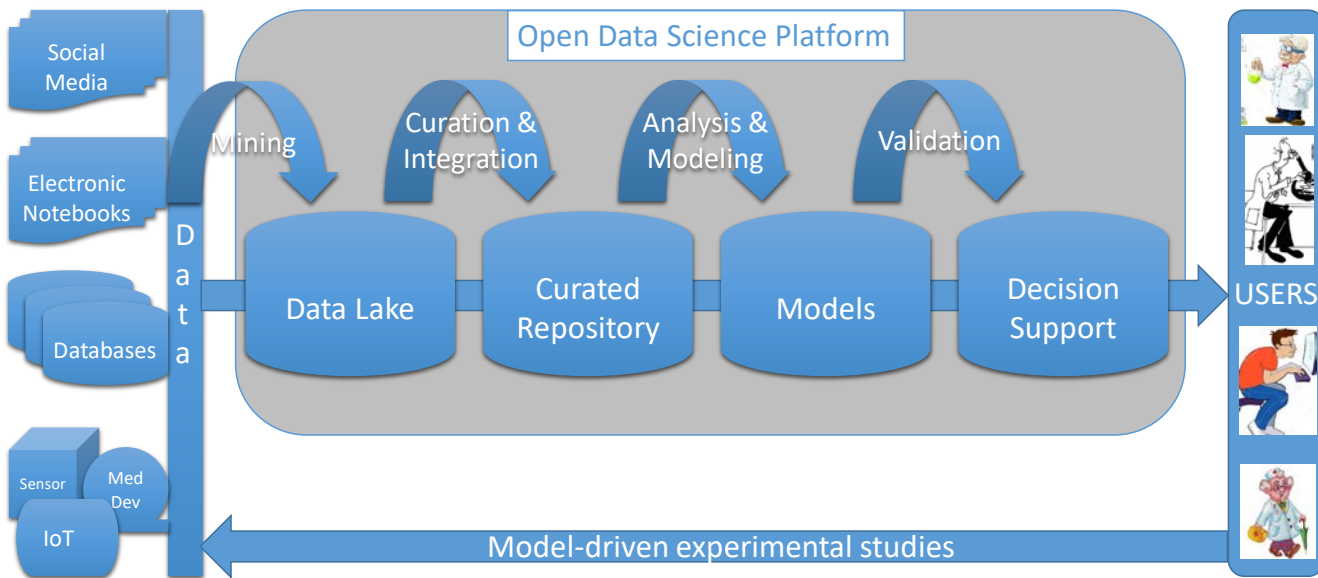


The front cover of the second edition of the *Compendium of Chemical Terminology*.










# OpenPHACTS



**GlaxoSmithKline – Coordinator**  
**Universität Wien – Managing entity**  
Technical University of Denmark  
University of Hamburg, Center for  
Bioinformatics  
BioSolveIT GmbH  
Consorci 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....

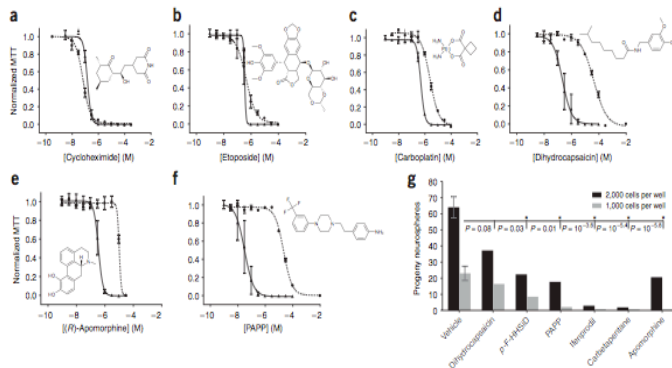
LETTERS

NATURE CHEMICAL BIOLOGY VOLUME 3 NUMBER 5 MAY 2007

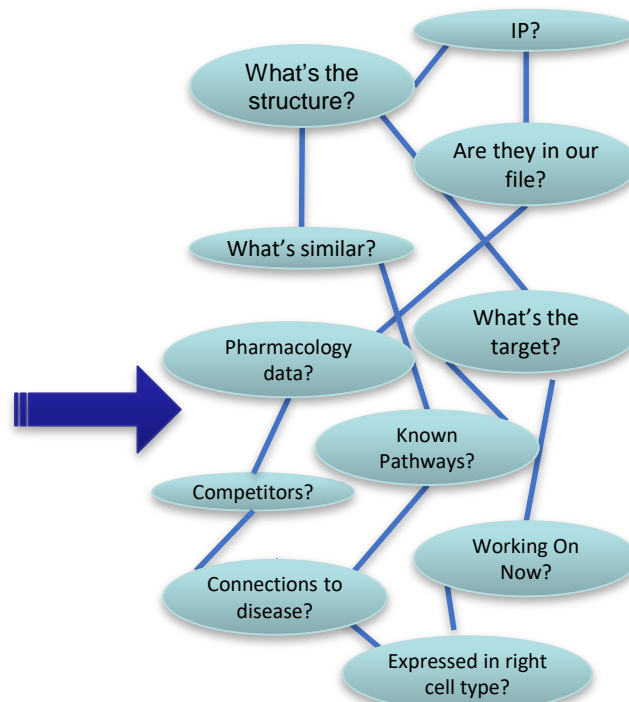
nature  
chemical biology

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

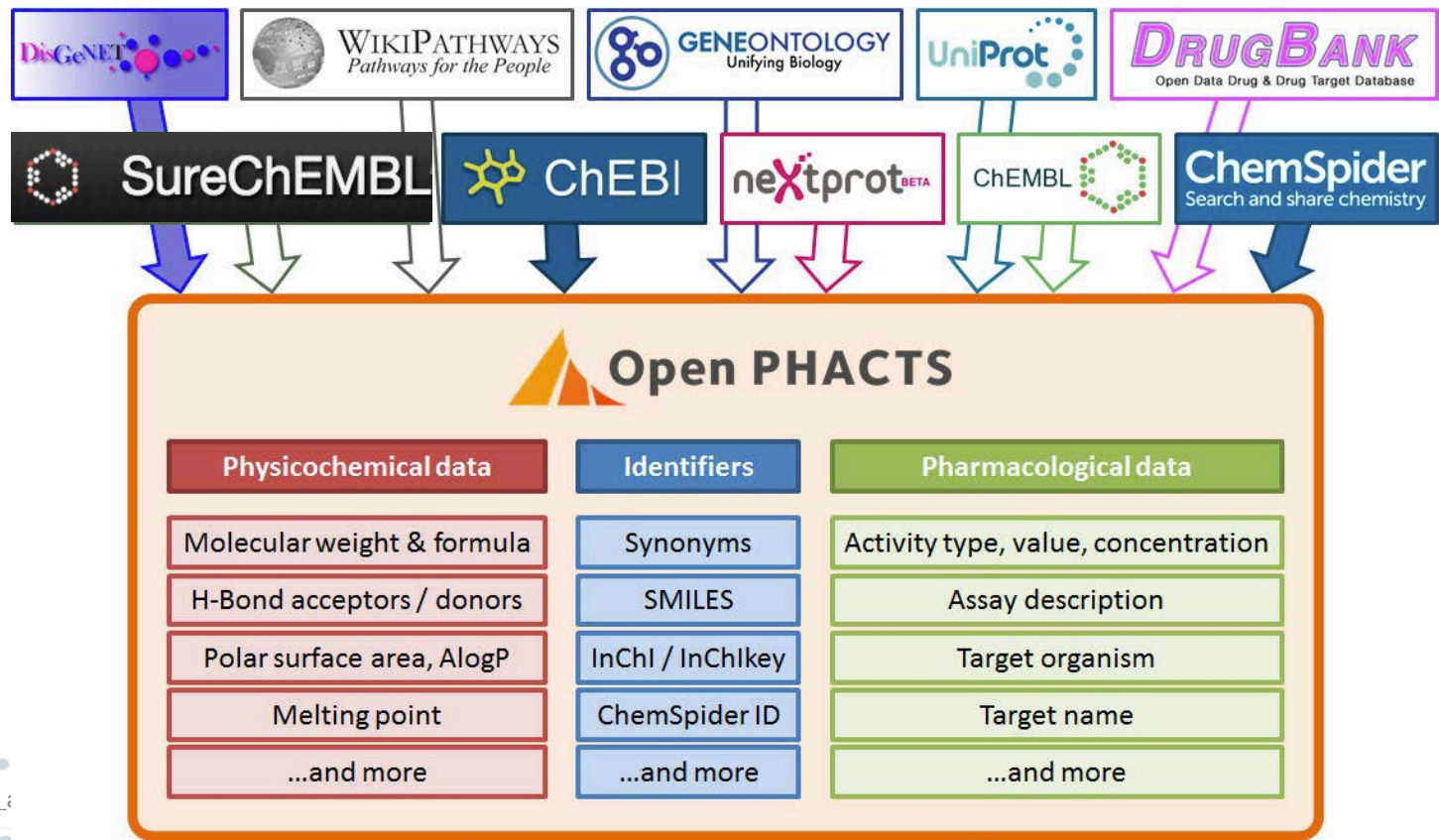
Phedias Diamandis<sup>1,4</sup>, Jan Wildenhain<sup>4</sup>, Ian D Clarke<sup>1,2</sup>, Adrian G Sacher<sup>1,2</sup>, Jeremy Graham<sup>1,2</sup>, David S Bellows<sup>3</sup>, Erick K M Ling<sup>1,2,5</sup>, Ryan J Ward<sup>1,2,5</sup>, Leanne G Jamieson<sup>1,2,5</sup>, Mike Tyers<sup>3,4</sup> &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



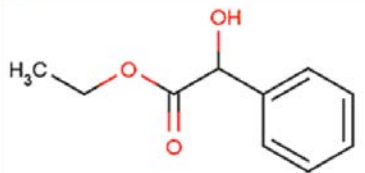
# Knowledge is federated



# Chemical structures representation

NamesAndSynonyms  
Ethyl 2-hydroxy-2-phenylacetate

Structure



Cid  
1

CatalogNumber  
MP-00015

CAS

Formula  
C10H12O3

MolWeight  
180.20

Notes

SellUnit  
5.00

Measure  
g

Currency  
USD

Purity  
98.00

IsAvailable  
Available

TotalAvailable  
31g

SubData  
Hydrohalide

Solubility

MP  
35 deg C

BP  
115 deg C

Density  
0.86

logP

logD

CatalogForMolPort: 1 out of 1 rows.

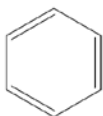
Chemical Database		
Chemical Name	Representation	Molar Mass
Benzene	c1ccccc1	78.1118
Ethanol	CCO	46.0684
Freon	ClC(Br)CFFF	197.382
Formaldehyde	cO	30.026
Methane	C	16.0425
Methanol	CO	32.0419
Propanol	CCOC	60.1
Toluene	Cc1ccccc1	92.1384
Indole	c1ccc2cc[nH]c2c1	117.148
Ammonia	N	17.0305

```
ACD/Labs10281015312D
9 9 0 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
0.0000 -0.7824 0.0000 C 0 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 0
1.7407 -2.0463 0.0000 C 0 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 0
2.5231 -3.1204 0.0000 C 0 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 0
3.8472 -2.9861 0.0000 O 0 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 0
1 5 1 0 0 0 0 0
1 2 1 0 0 0 0 0
2 3 2 0 0 0 0 0
3 4 1 0 0 0 0 0
4 5 2 0 0 0 0 0
4 6 1 0 0 0 0 0
6 7 2 0 0 0 0 0
6 8 1 0 0 0 0 0
8 9 1 0 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 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
0.0000 -0.7803 0.0000 C 0 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 0
1.7426 -2.0460 0.0000 C 0 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 0
2.5228 -3.1210 0.0000 C 0 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 0
1.9810 -4.3348 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0 0 0 0 0
1 5 1 0 0 0 0 0
2 3 2 0 0 0 0 0
3 4 1 0 0 0 0 0
4 5 2 0 0 0 0 0
4 6 1 0 0 0 0 0
6 7 2 0 0 0 0 0
6 8 1 0 0 0 0 0
M END
> <Catalog_Number>
AS1-0102
> <CAS_Number>
3/931
```

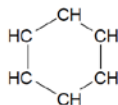


# InChI (<http://www.inchi-trust.org/>)

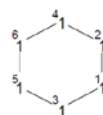
Input Structure



Normalized Structure



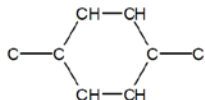
Canonical Numbering



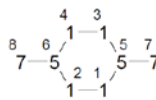
Input Structure



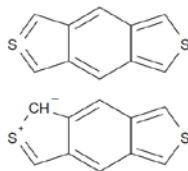
Normalized Structure



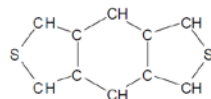
Canonical Numbering



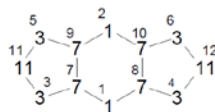
Input Structures



Normalized Structure



Canonical Numbering



```
{InChI version}
1. Main Layer (M):
/{formula}
/c{connections}
/h{H_atoms}
2. Charge Layer
/q{charge}
/p{protons}
3. Stereo Layer
/b{stereo:dbond}
/t{stereo:sp3}
/m{stereo:sp3:inverted}
/s{stereo:type (1=abs, 2=rel, 3=rac)}
4. Isotopic Layer (M):
/i{isotopic:atoms}*
/h{isotopic:exchangeable H}
/b{isotopic:stereo:dbond}
/t{isotopic:stereo:sp3}
/m{isotopic:stereo:sp3:inverted}
/s{isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
5. Fixed H Layer (F):
/f{fixed_H:formula}*
/h{fixed_H:H fixed}
/q{fixed_H:charge}
/b{fixed_H:stereo:dbond}
/t{fixed_H:stereo:sp3}
/m{fixed_H:stereo:sp3:inverted}
/s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}
(6.) Fixed/Isotopic Combination (FI)
/i{fixed_H:isotopic:atoms}*
/b{fixed_H:isotopic:stereo:dbond}
/t{fixed_H:isotopic:stereo:sp3}
/m{fixed_H:isotopic:stereo:sp3:inverted}
/s{fixed_H:isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
/o{transposition}
```

# Chemistry Validation and Standardization Platform

**Deposition Gateway**

Home **Submit** Depositions Profile

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

Submit new deposition

Datasource

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
<a href="#">cada2290-d1c8-4923-945c-1f4ac495c0f4</a>	10/7/2015	WikiPathways	Aleksey Pshenichnov	WikiPathways.sdf	Processed
<a href="#">39df661f-c9f8-4490-9eae-f96b0261ad34</a>	10/7/2015	Thomson Pharma	Aleksey Pshenichnov	ThomsonReuters.sdf.gz	Processed
<a href="#">7f8eaadd-7f01-42d5-b44a-96a3f8a11737</a>	10/7/2015	MeSH	Aleksey Pshenichnov	mesh20151407final.sdf.zip	Deposited2GCN
<a href="#">600ca202-638c-4b18-9b2c-2cfe0cf87211</a>	10/8/2015	ChEBI	Aleksey Pshenichnov	ChEBI_lite.sdf.gz	Deposited2GCN
<a href="#">859cea6d-33e0-475c-994b-c3c52dbcd71</a>	10/8/2015	DrugBank	Aleksey Pshenichnov	drugbank.zip	Deposited2GCN
<a href="#">bf4b356e-381e-4b14-b54f-27f6b6528085</a>	10/8/2015	Human Metabolome Database	Aleksey Pshenichnov	hmdb.zip	Deposited2GCN
<a href="#">b7b3eda8-35b0-44ba-8775-e76452460b87</a>	10/9/2015	ChEMBL	Aleksey Pshenichnov	chembl_20.sdf	Deposited2GCN
<a href="#">01931b57-5498-4416-91bd-7dd83139f825</a>	10/20/2015	MeSH	Aleksey Pshenichnov	mesh_20151016.zip	Processed
<a href="#">a9ca97f2-61e9-4df3-a3f6-2d18e3df1102</a>	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

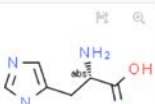
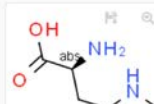
Annotations Jobs Chunks Delete Delete From GCN

**Guid** 859cea6d-33e0-475c-994b-c3c52dbcf71  
**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**

<b>Validate</b>	true
<b>Standardize</b>	true
<b>PropertiesCalculation</b>	true
<b>ParentsGeneration</b>	true

Records 6837

Ordinal	Original	Issues	Standardized
9			

Filter

By REGIDs  
Example: regid1, regid2

By Ordinals  
Example: 1, 9, 23

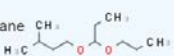
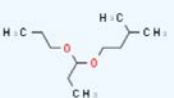
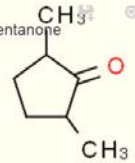
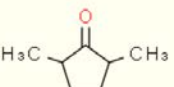
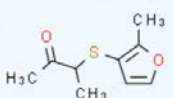
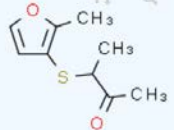

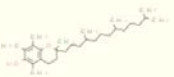
Severities


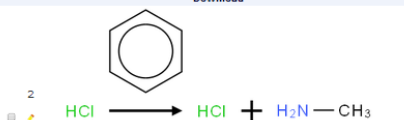
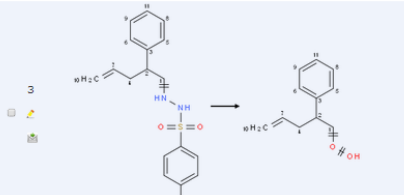
- ☒ Error (41)
- ☐ Warning (755)
- ☐ Information (2668)

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)

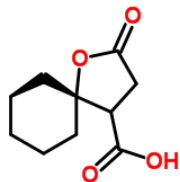
Close Reset Apply

Ordinal	REGID	Original	Issues	Standardized
4	1-isopentyloxy-1-propoxypropane		Contains completely undefined stereo - enantiomers	
15	2,5-dimethylcyclopentanone		Contains completely undefined stereo - mixtures	
20	3-((2-methyl-3-furyl)thio)-2-butanone		Contains completely undefined stereo - enantiomers	
22	3,4-dihydro-2,5,7,8-tetramethyl-2-(4,8,12-trimethyltridecyl)-1-benzopyran-6-ol		Contains completely undefined stereo - mixtures	

Original	% of Mapped Atoms	Issues
 Download	63	<p>Warn more than one instance of same molecule found in reactants : [InChI=1S/C5H6N2O2/c1-3-4(5(8)9)7-2-6-3/h2H,1H3,(H,6,7)(H,8,9)]</p> <p>Warn more than one instance of same molecule found in reactants : [InChI=1S/K/g+1]</p> <p>Warn percentage of mapped atoms: 60-80%</p> <p>Warn 0 bond made or broken, 0 bond order changed</p>
 Download	0	<p>Warn same molecule found in reactants and products : [InChI=1S/ClH/h1H]</p> <p>Warn percentage of mapped atoms: &lt;20%</p> <p>Warn 0 bond made or broken, 0 bond order changed</p>
 Download	61	<p>Warn percentage of mapped atoms: 60-80%</p> <p>Info 1 bond made or broken</p>



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

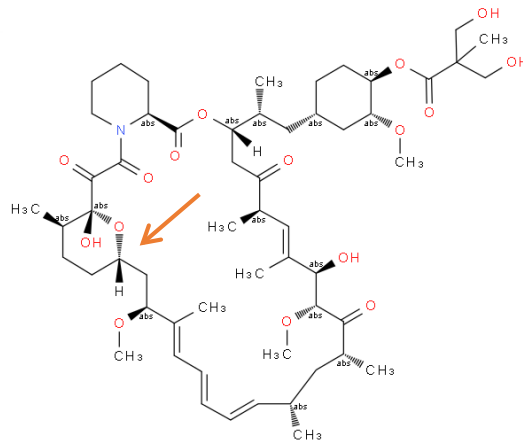
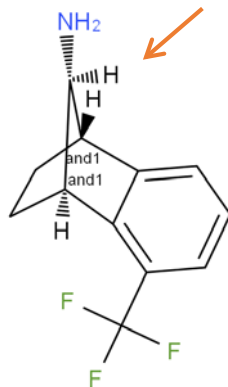
Monoisotopic mass: 198.089203 Da

▼ Systematic name

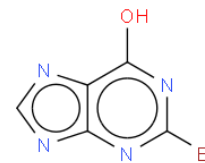
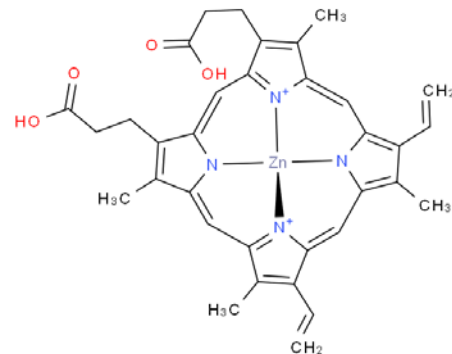
2-Oxo-1-oxaspiro[4.5]decane-4-carboxylic acid

► SMILES and InChIs

► Cite this record



DB06287



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

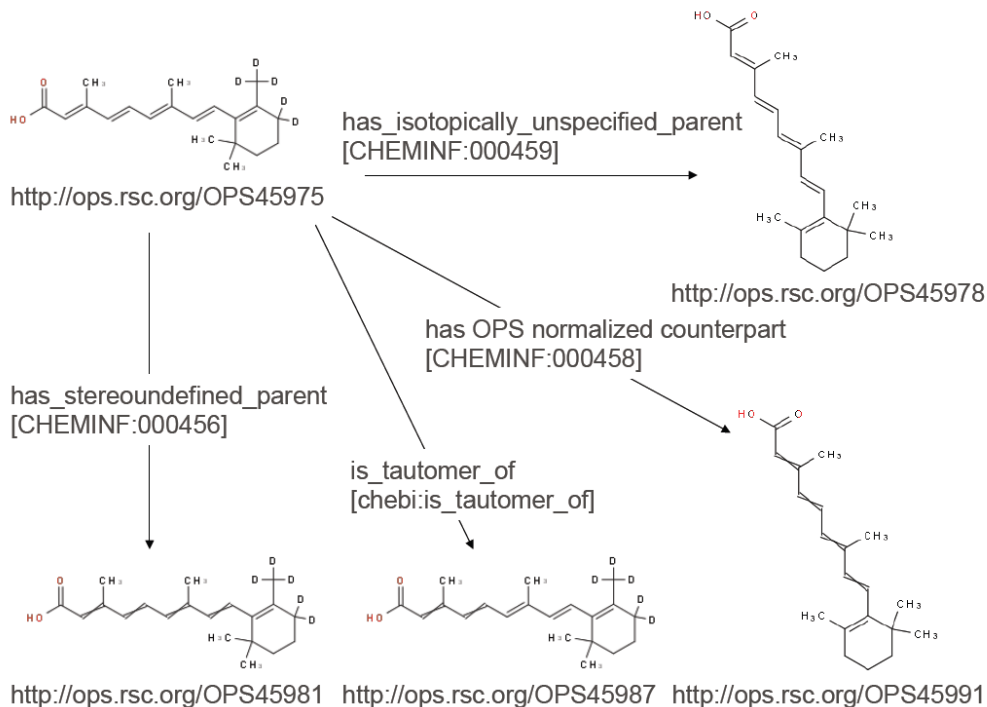


*Not acceptable*





# Chemical Lenses





# CVSP - mapping

**Add Annotation**

External ID  DATABASE\_ID

Cancel Add

Field	Annotation
DATABASE_ID	External ID

+ Add Annotation

Back

**Add Annotation**

External ID  DATABASE\_ID

Cancel Add

- External ID
- InChI
- InChI Key
- SMILES
- Synonym
- Xref
- Comment

**Add Annotation**

External ID  DATABASE\_ID

Cancel Add

- 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

Home

Submit

Depositions

Profile

Feedback

☐ 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)

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.Smirt)

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

Validation rule set (default)

Description

Validation rule set (default)

XML Content

```
<?xml version="1.0" encoding="utf-8" ?>
<rules>
  <moleculerules>
    <!-- The SMARTS tests below are complimentary to set of validations that CVSP does by default -->
    <Warning message="Contains cyclobutane" description="[CX4;H2;4]1[CX4;H2;4][CX4;H2;4][CX4;H2;4]1">
      <test name="SMARTSTest" param="[CX4;H2;4]1[CX4;H2;4][CX4;H2;4][CX4;H2;4]1">
    </Warning>
    <Warning message="Contains ethane" description="[CX4;H3][CX4;H3]">
      <test name="SMARTSTest" param="[CX4;H3][CX4;H3]">
    </Warning>
    <Warning message="Contains S with no explicit bonds" description="[S,D0]">
      <test name="SMARTSTest" param="[S,D0]">
    </Warning>
    <Warning message="Contains B with no explicit bonds" description="[B,D0]">
      <test name="SMARTSTest" param="[B,D0]">
    </Warning>
    <Warning message="Contains methane" description="[CX4;H4]">
      <test name="SMARTSTest" param="[CX4;H4]">
    </Warning>
  </moleculerules>
</rules>
```

Revise

Clone



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



# What exactly CRS provides?

## 1. Chemistry processing

- Validation
- Standardization
- Properties generation
- Properties retrieval

## 2. Export

- RDF
- SDF

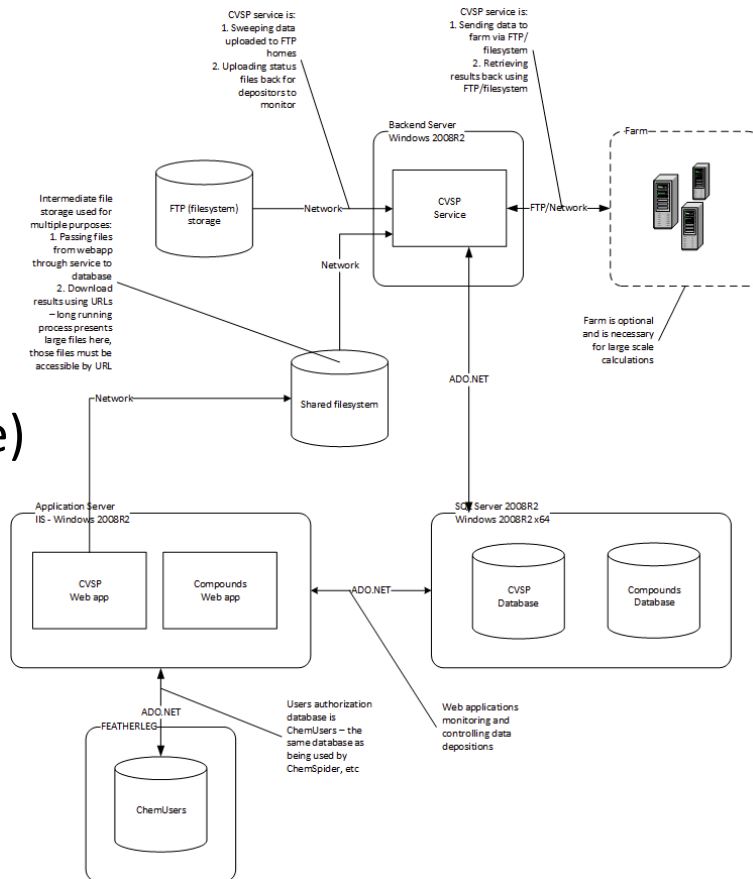
## 3. API

- Domain-specific searches
- Chemical visualization
- Properties
- Conversions



# Subsystems

- CVSP (frontend, backend, database)
- Compounds (frontend, database)
- OpenPHACTS API (frontend, database)
- Datasources registry (frontend, database)
- Processing farm (optional)





# CVSP v1.0 vs v2.0

- Indigo → Indigo, CDK, RDKit
- Windows → Platform independent (.NET Core, Docker)
- Web App → Libraries (.NET, Python)
- 3-tier → Microservices





# CVSP on Jupyter - validation

```
In [1]: from CVSP_alpha_indigo import get_infolist, get_ontology_rules, standardize, add_ontology, validate
```

```
validate('validation.xml', 'val_test_1.mol', format='mol')
```

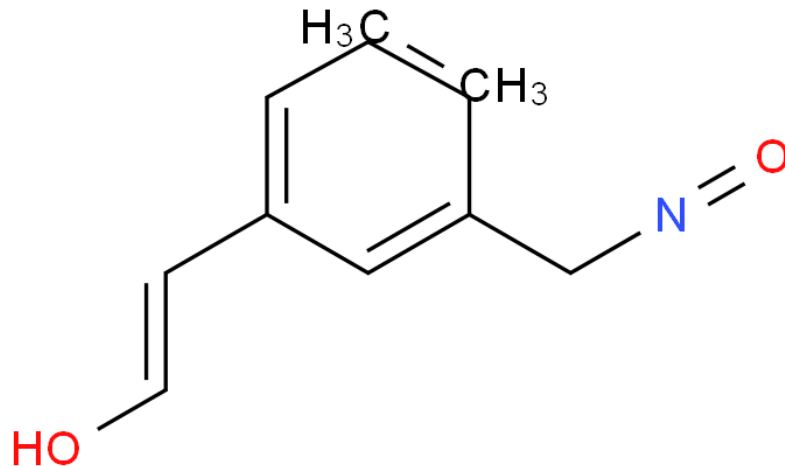
Warnings:

Contains ethane

Information:

Contains enol function

Contains nitroso form of oxime

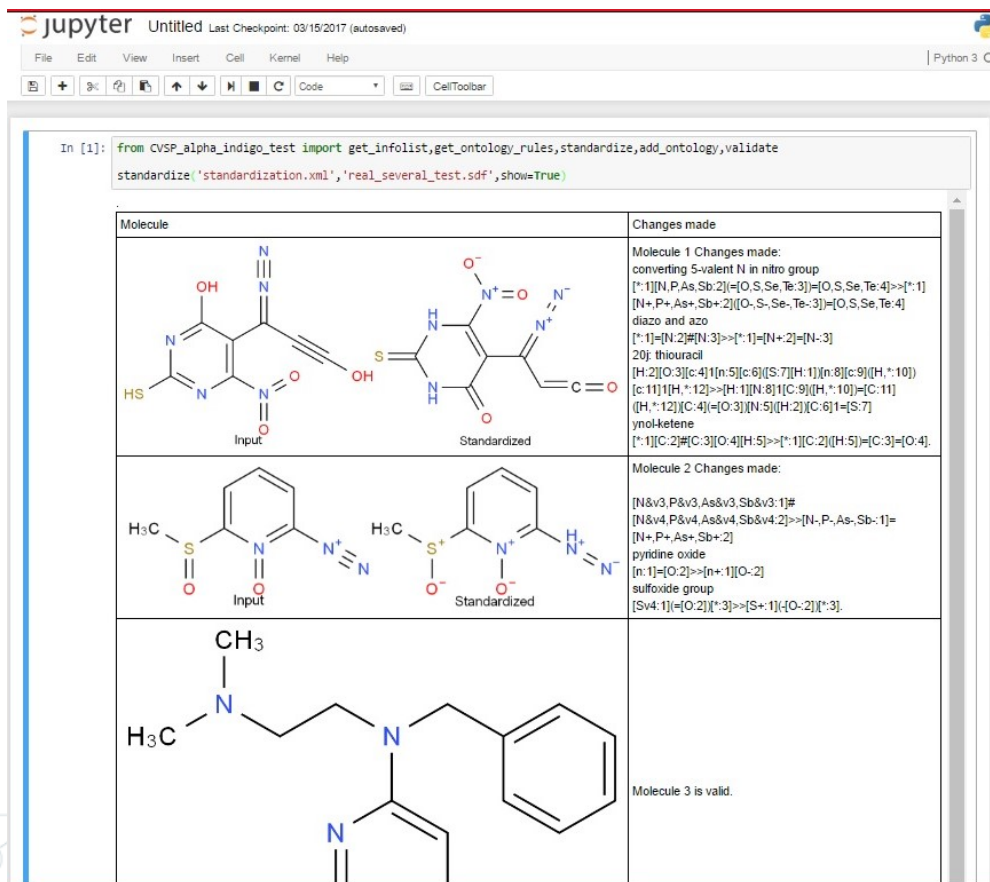
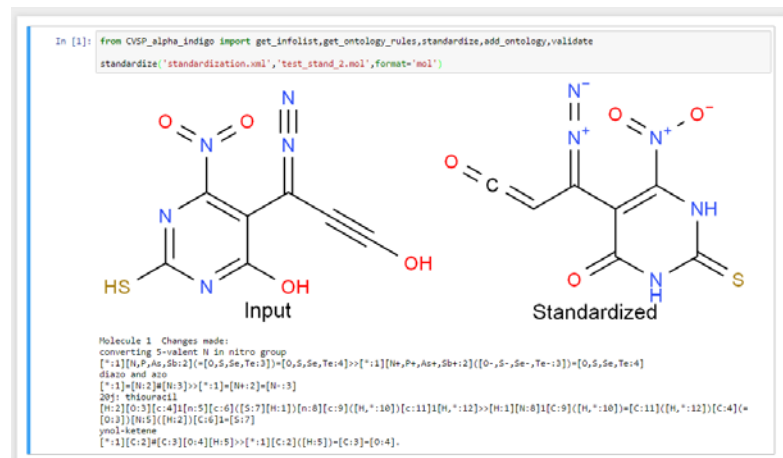


Warnings:

Contains ethane



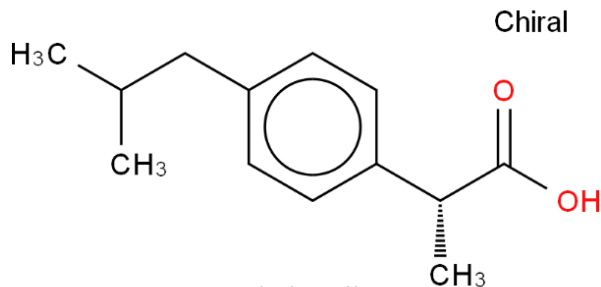
# CVSP on Jupyter - standardization



# CVSP on Jupyter – ontologies, chemotypes, etc

```
In [1]: from CVSP_alpha_indigo import get_infolist, get_ontology_rules, standardize, add_ontology, validate
        add_ontology('ontology.xml', 'val_test_1.mol', Format='mol')
```

CHEBI:26004  
CHEBI:2571  
CHEBI:33854  
CHEBI:33655  
CHEBI:33659  
CHEBI:36537  
CHEBI:22712  
CHEBI:36586  
CHEBI:33575



phenylpropanoid  
aliphatic alcohol  
aromatic compound  
organic aromatic compound  
polyatomic entity  
benzene  
carbonyl compound  
carboxylic acid

```
1 val_test_1.mol
2 -INDIGO-03191712312D
3
4 15 15 0 0 1 0 0 0 0 0 0999 V2000
5 -2.8579 1.2375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
6 -2.1434 0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
7 -2.1434 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
8 -1.4289 1.2375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
9 -0.7145 0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
10 -0.7145 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
11 0.0000 -0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
12 0.7145 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
13 0.7145 0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
14 0.0000 1.2375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
15 1.4289 -0.4125 0.0000 C 0 1 0 0 0 0 0 0 0 0 0 0
16 1.4289 -1.2375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
17 2.1434 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
18 2.1434 0.8250 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
19 2.8579 -0.4125 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
20 1 2 1 0 0 0 0
21 2 3 1 0 0 0 0
22 2 4 1 0 0 0 0
23 4 5 1 0 0 0 0
24 5 6 4 0 0 0 0
25 6 7 4 0 0 0 0
26 7 8 4 0 0 0 0
27 8 9 4 0 0 0 0
28 9 10 4 0 0 0 0
29 5 10 4 0 0 0 0
30 8 11 1 0 0 0 0
31 11 12 1 6 0 0 0
32 11 13 1 0 0 0 0
33 13 14 2 0 0 0 0
34 13 15 1 0 0 0 0
35 M END
36 > <CHEBI>
37 CHEBI:26004
38 CHEBI:2571
39 CHEBI:33854
40 CHEBI:33655
41 CHEBI:33659
42 CHEBI:36537
43 CHEBI:22712
44 CHEBI:36586
45 CHEBI:33575
46
47 $$$$
48
```



## ONE PLACE TO STORE YOUR DATA



### UPLOAD AND ORGANIZE

You can upload your files, transfer them from Google Drive, Box.com, DropBox.com etc. And - manage your data your own way!

### MACHINE LEARNING

Data models, algorithms and pipelines for cheminformatics and drug discovery.

### SHARE & ANNOTATE

We support widely used vocabularies. Plus, you can add your own. Share your work with others.





# Open Science Data Repository (OSDR)

The screenshot displays the OSDR web application interface. At the top, a dark navigation bar contains the OSDR logo, the text "OPEN SCIENCE DATA REPOSITORY", and the subtext "Powered by Dataledger™". On the right side of the navigation bar are links for "Home" and "Organize", and a user profile icon labeled "valt".

Below the navigation bar, the interface is divided into two main sections. On the left is a sidebar with a list of categories, each with an icon and a count in a grey pill-shaped box:

- Articles: 1233
- Images: 10000
- Research: 200
- Structures: 30012645
- Crystals: 86659
- Reactions: 3144655
- Spectra: 6094
- Datasets: 2133

The main content area on the right is titled "DRAFTS" and features a search icon, a plus icon, an upload icon, and a view toggle icon. It displays a grid of draft folders, each with a folder icon and a label:

- Biologics
- Chemicals
- Crystals
- Datasets
- Documents
- Materials
- QSAR
- Reactions
- Spectra
- Spectra

# Chemical processing

- Support for chemical formats
- Chemistry validation and standardization
- Automatic processing and visualization

The screenshot displays the Open Science Data Repository interface, which is powered by DataLedger. The header includes the Open Science logo, the text "DATA REPOSITORY", and navigation links for "Home" and "Organize". A user profile icon labeled "valt" is also present. On the left, a sidebar lists various data types with their respective counts: Articles (1233), Images (10000), Research (200), Structures (30012645), Crystals (86659), Reactions (3144655), Spectra (6094), and Datasets (2133). The main content area, titled "DRAFTS / QSAR", features a grid of 15 data cards. Each card contains a chemical structure or a folder icon and a label. The labels include: data\_Hemeoxyg..., data\_Humancox..., data\_lipinski.mo..., data\_Adrenergic..., data\_Alcoholde..., data\_Arachidon..., data\_Cytochro..., data\_Hemeoxyg..., data\_Humancox..., data\_LaCrossevi..., data\_Nitric\_oxid..., data\_Pseudolysi..., data\_Serotonin..., data\_Tryptopha..., and data\_lipinski.sdf. The interface also includes search, upload, and view icons in the top right corner.



# OSDR - documents

- Integrated text-mining

The OSDR interface displays a sidebar with navigation options: Articles, Images, Research, Structures, Crystals, Reactions, Spectra, and Datasets. The main area is titled 'DRAFTS / ARTICLES' and shows a grid of document thumbnails. A document titled 'BRANDMAIER ET AL...' is selected. Below the grid, a 'Properties from file' section shows metadata for a PDF file: 'DRAFTS / MATERIALS / 2H-AGFE02 / PHYSREVB.91.094434.PDF'. The 'User properties' section shows the reference URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.91.094434>. The description states: 'Magnetic and dielectric properties of the hexagonal triangular lattice antiferromagnet 2H-AgFeO<sub>2</sub> have been studied by neutron diffraction, magnetic susceptibility, specific heat, pyroelectric current, and dielectric constant measurements. The ferroelectric polarization,  $P \approx 5 \mu\text{C}/\text{m}^2$ , has been found to appear below 11 K due to a polar nature of the magnetic ground state of the system. In the temperature range of 11 K  $\leq T \leq 18$  K, an incommensurate spin density wave (ICM1) with the nonpolar magnetic point group  $\text{mmn}1'$  and the  $k_1 = (0, a_2^*, 0, a_2^*)$  propagation vector takes place. Below 14 K, a proper screw ordering (ICM2) and  $k_2 = (0, a_2^*, 0, a_2^*) = (0.385, 0.390)$  appears as a mirror phase which coexists with ICM1 and the ground state down to the lowest measured temperature 5.5 K. No ferroelectric polarization associated with the ICM2 phase was observed in agreement with its nonpolar point group  $2221'$ . Finally, a spiral order with cycloid and proper screw components (ICM3), and  $k_3 = (a_1^*, a_2^*, 0, a_2^*) = (0.0467, 0.349)$  emerges below 11 K as the ground state of the system. Based on the deduced magnetic point group  $21'$ , we conclude that the ferroelectric polarization in ICM3 is parallel to the  $c$  axis and is caused by the inverse Dykhovskii-Moriya effect with  $p_i \propto \epsilon_{ij} \times (S_i \times S_j)$ .' The authors listed are Noriki Terada, Dmitry D. Khalyavin, Pascal Manuel, Yoshitomo Tsuchimoto, and Alexei A. Belik.

The PDF document preview shows the title 'Magnetic ordering and ferroelectricity in multiferroic 2H-AgFeO<sub>2</sub>: Comparison between hexagonal and rhombohedral polytypes'. The authors listed are Noriki Terada, Dmitry D. Khalyavin, Pascal Manuel, Yoshitomo Tsuchimoto, and Alexei A. Belik. The document is from PHYSICAL REVIEW B 91, 094434 (2015). The abstract states: 'Magnetic and dielectric properties of the hexagonal triangular lattice antiferromagnet 2H-AgFeO<sub>2</sub> have been studied by neutron diffraction, magnetic susceptibility, specific heat, pyroelectric current, and dielectric constant measurements. The ferroelectric polarization,  $P \approx 5 \mu\text{C}/\text{m}^2$ , has been found to appear below 11 K due to a polar nature of the magnetic ground state of the system. In the temperature range of 11 K  $\leq T \leq 18$  K, an incommensurate spin density wave (ICM1) with the nonpolar magnetic point group  $\text{mmn}1'$  and the  $k_1 = (0, a_2^*, 0, a_2^*) = (0.385, 0.390)$  propagation vector takes place. Below 14 K, a proper screw ordering (ICM2) and  $k_2 = (0, a_2^*, 0, a_2^*) = (0.385, 0.390)$  appears as a mirror phase which coexists with ICM1 and the ground state down to the lowest measured temperature 5.5 K. No ferroelectric polarization associated with the ICM2 phase was observed in agreement with its nonpolar point group  $2221'$ . Finally, a spiral order with cycloid and proper screw components (ICM3), and  $k_3 = (a_1^*, a_2^*, 0, a_2^*) = (0.0467, 0.349)$  emerges below 11 K as the ground state of the system. Based on the deduced magnetic point group  $21'$ , we conclude that the ferroelectric polarization in ICM3 is parallel to the  $c$  axis and is caused by the inverse Dykhovskii-Moriya effect with  $p_i \propto \epsilon_{ij} \times (S_i \times S_j)$ .' The document is published 17 February 2015; published 30 March 2015.

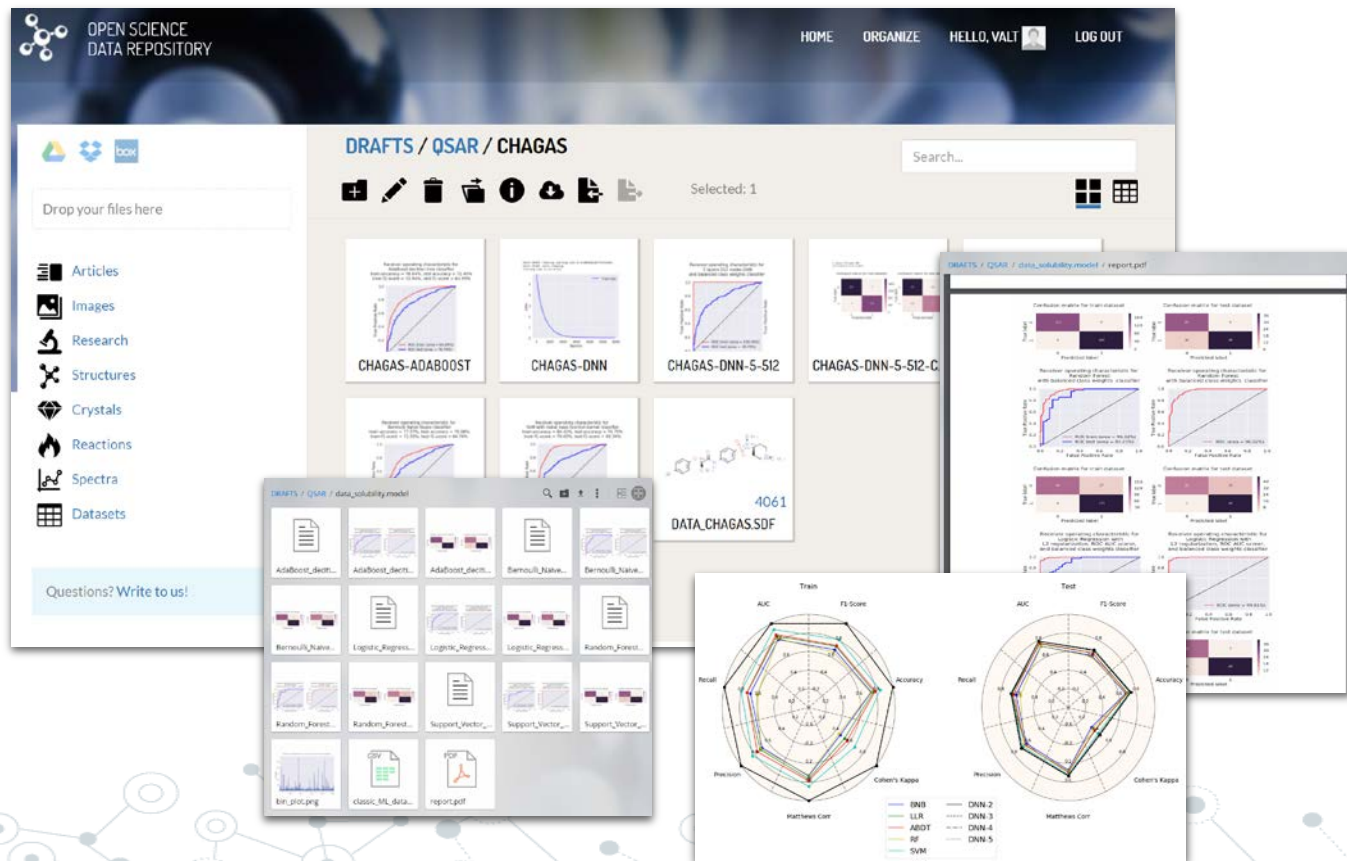
The OSDR interface displays a sidebar with navigation options: Articles, Images, Research, Structures, Crystals, Reactions, Spectra, and Datasets. The main area is titled 'DRAFTS / MATERIALS / LAC003' and shows a grid of document thumbnails. A document titled '13072815.PDF' is selected. Below the grid, a 'Properties from file' section shows metadata for a PDF file: 'DRAFTS / MATERIALS / LAC003 / 13072815.PDF'. The 'User properties' section shows the reference URL: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.91.094434>. The description states: 'Magnetic and dielectric properties of the hexagonal triangular lattice antiferromagnet 2H-AgFeO<sub>2</sub> have been studied by neutron diffraction, magnetic susceptibility, specific heat, pyroelectric current, and dielectric constant measurements. The ferroelectric polarization,  $P \approx 5 \mu\text{C}/\text{m}^2$ , has been found to appear below 11 K due to a polar nature of the magnetic ground state of the system. In the temperature range of 11 K  $\leq T \leq 18$  K, an incommensurate spin density wave (ICM1) with the nonpolar magnetic point group  $\text{mmn}1'$  and the  $k_1 = (0, a_2^*, 0, a_2^*) = (0.385, 0.390)$  propagation vector takes place. Below 14 K, a proper screw ordering (ICM2) and  $k_2 = (0, a_2^*, 0, a_2^*) = (0.385, 0.390)$  appears as a mirror phase which coexists with ICM1 and the ground state down to the lowest measured temperature 5.5 K. No ferroelectric polarization associated with the ICM2 phase was observed in agreement with its nonpolar point group  $2221'$ . Finally, a spiral order with cycloid and proper screw components (ICM3), and  $k_3 = (a_1^*, a_2^*, 0, a_2^*) = (0.0467, 0.349)$  emerges below 11 K as the ground state of the system. Based on the deduced magnetic point group  $21'$ , we conclude that the ferroelectric polarization in ICM3 is parallel to the  $c$  axis and is caused by the inverse Dykhovskii-Moriya effect with  $p_i \propto \epsilon_{ij} \times (S_i \times S_j)$ .' The authors listed are Noriki Terada, Dmitry D. Khalyavin, Pascal Manuel, Yoshitomo Tsuchimoto, and Alexei A. Belik.





# Built-in Machine Learning

- Automated ML pipeline
- Pre-built ML modules
- Comparison between different ML algorithms
- NB, NN, RF, SVM, LR
- DNN





# FAIR Data Principles

## To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

## To be Accessible:

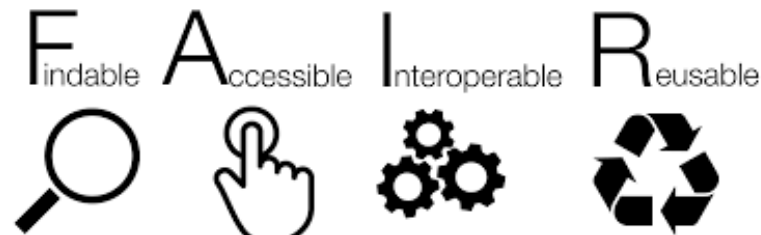
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

## To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

## To be Re-usable:

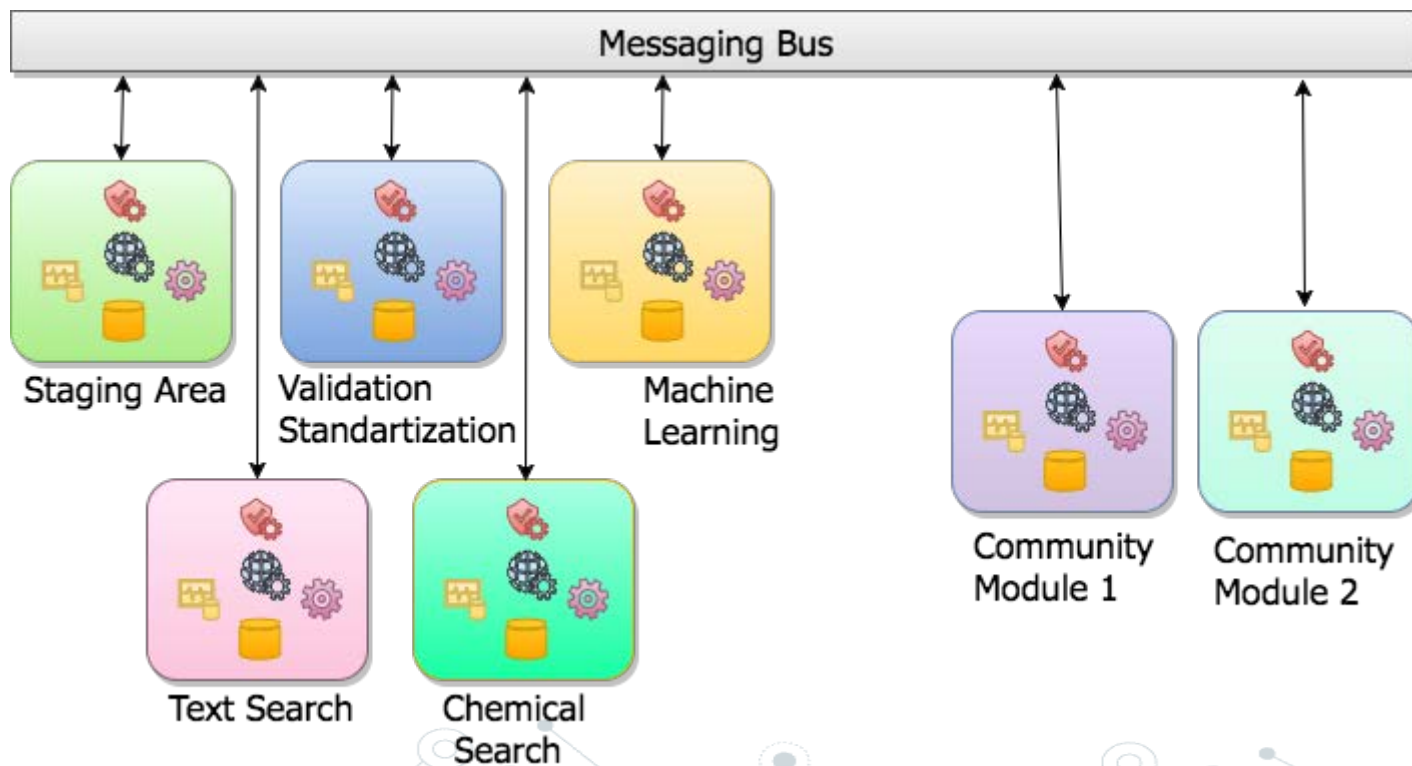
- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.







# Extensible micro-service based architecture





# Summary

- OpenPHACTS Chemistry Registry System (CRS)
- Uses open source toolkits (CDK, RDKit, Indigo)
- Rules are expressed in XML format
- Can handle specific cases via modules
- Supports FAIR data principles
- Evolve and improve continuously





# Thank you!

On Web:

[scidatasoft.com](http://scidatasoft.com)

Contact us:

[info@scidatasoft.com](mailto:info@scidatasoft.com)

Slides:

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

