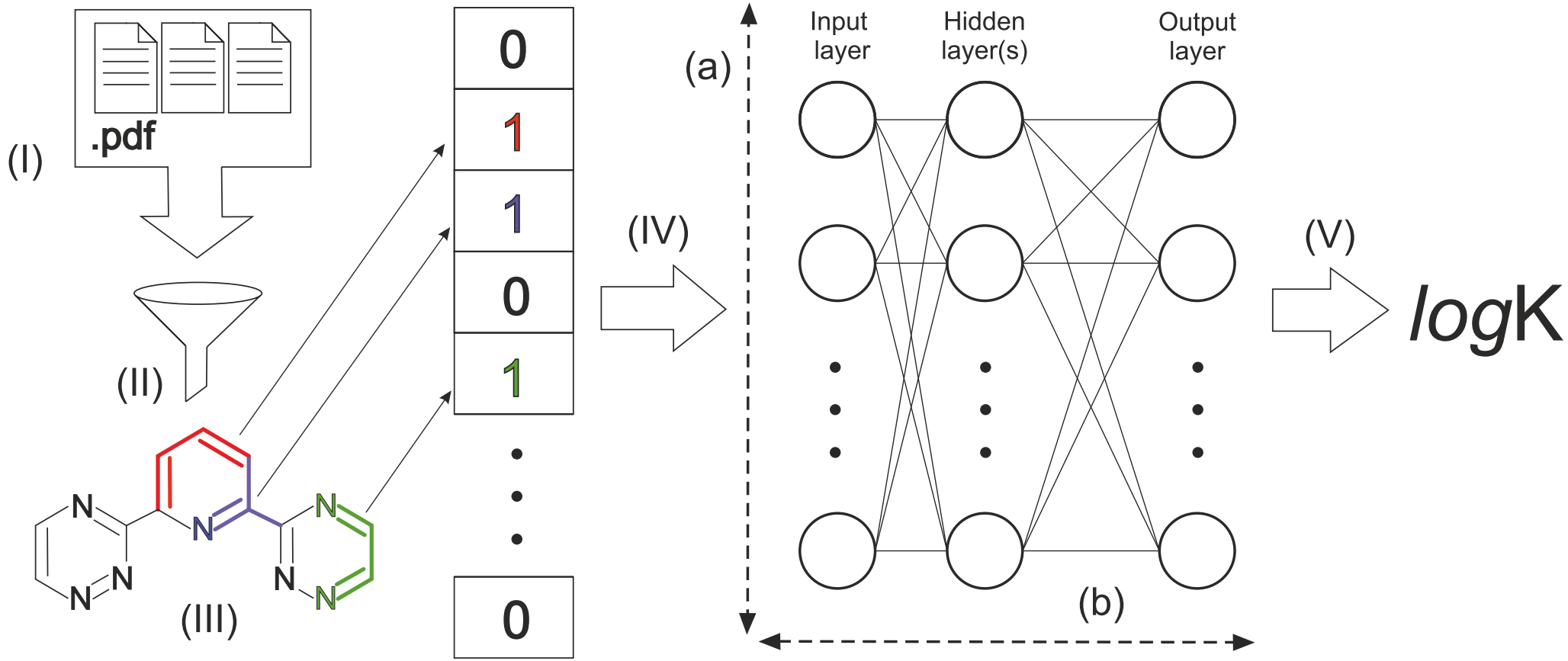
Химические данные

Data

Data as an input



10.3390/molecules26113237

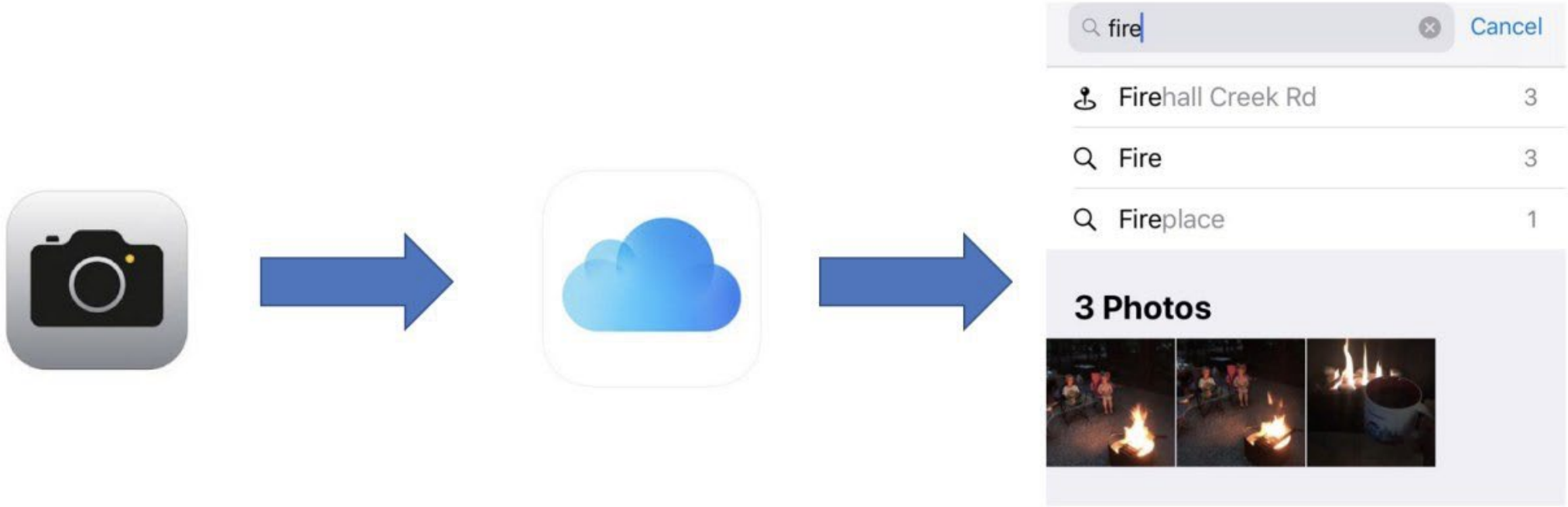
Peculiarities of chemical data

- Main entity is a chemical structure (rarely: spectra, electron structure)
- Various sizes
- Not machine-readable ‘as is’. We need to convert it
- Non-standard formats

What do we have

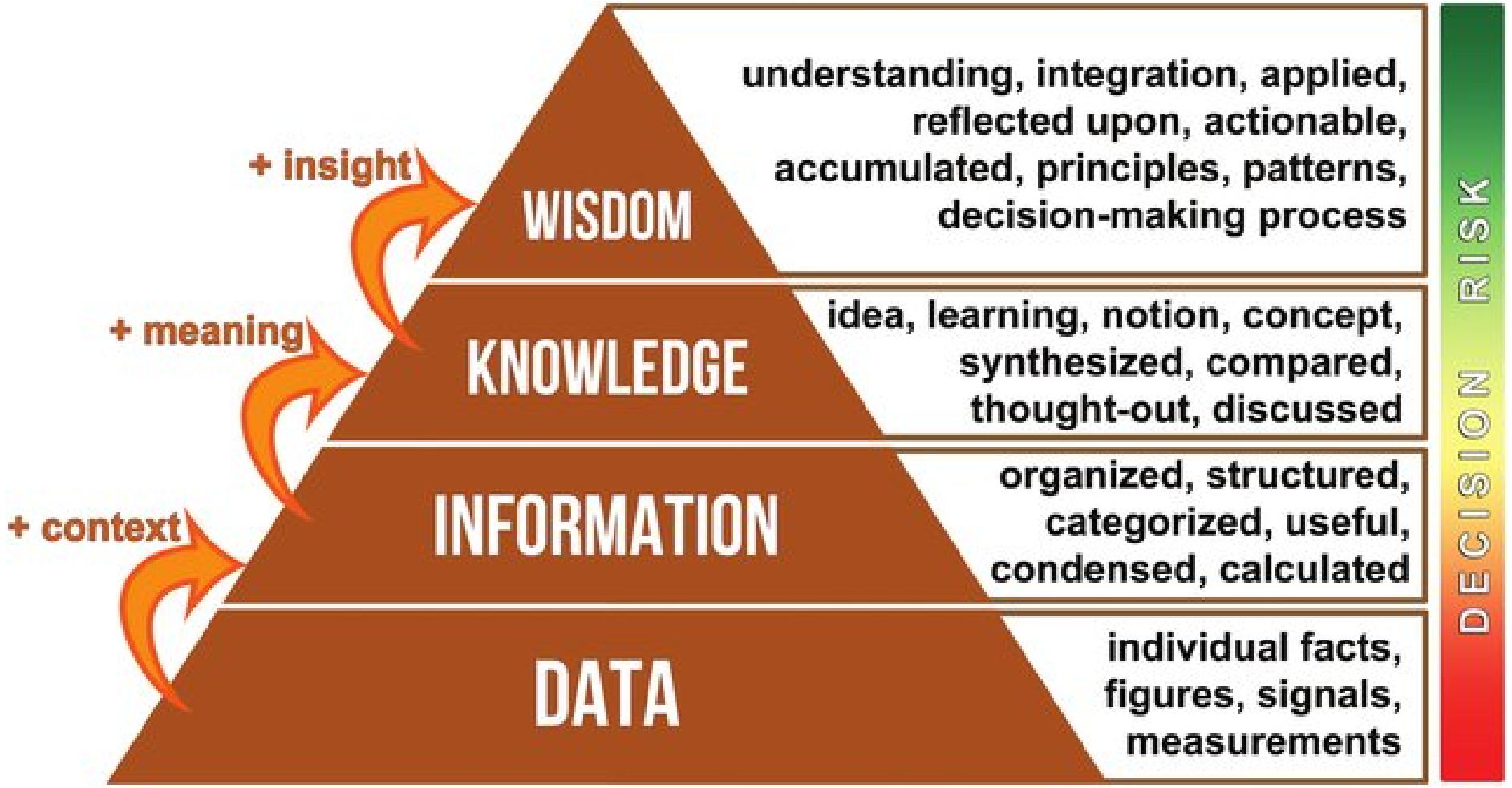
- Laboratory notebooks (paper?)
- Instrument data (Vendor format? Unannotated? Uncurated?)
- What we publish (successes)
- What we don’t publish (fails)

Data Lake for dummies



https://medium.com/rock-your-data/getting-started-with-data-lake-4bb13643f9

Big data



DOI: 10.1093/af/vfz005

Curated data

- Очистка
- Оценка качества
- Аннотация
- ….

Labelled data

Fair data

Findable	Data and materials enriched with metadata assigned with a unique identifier
Accessible	Data and metadata stored in a trusted repository with an open and free protocol. Accessible by machines and humans
Interoperable	Using vocabularies and public domain ontologies the metadata can be referenced and linked
Reusable	Additional documentation and protocols describing the acquisition of the data, licensed with a detailed provenance

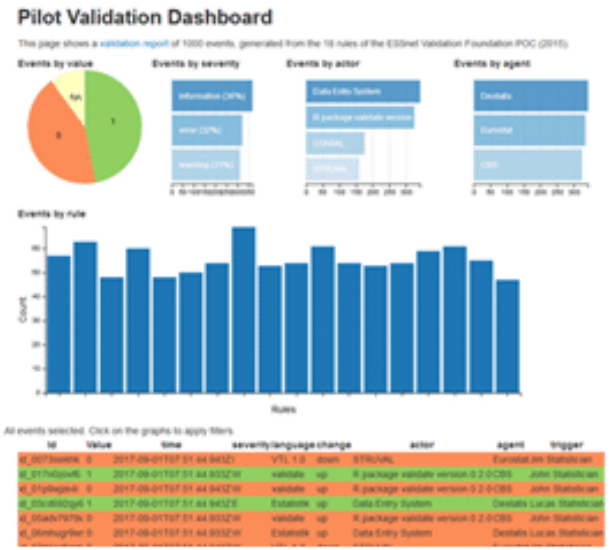
DOI:10.1016/j.ejmp.2021.01.083

Human readable

Machine readable


```
[
  {
    "rule": {
      "expression": "check(DS.hours_worked between 1 and 80)",
      "severity": "warning"
    },
    "event": {
      "time": "2017-09-01T07:51:44.933Z",
      "actor": "Eurostat"
    },
    "data": {},
    "value": "0" // failed
  },
  {
    "rule": {
      "expression": "cost + profit == turnover",
      "severity": "error"
    },
    "event": {
      "time": "2017-09-01T07:51:46.933Z",
      "actor": "Eurostat"
    },
    "data": {},
    "value": "1" // passed
  },
  ...
  ...
  ...
]
```


Human readable



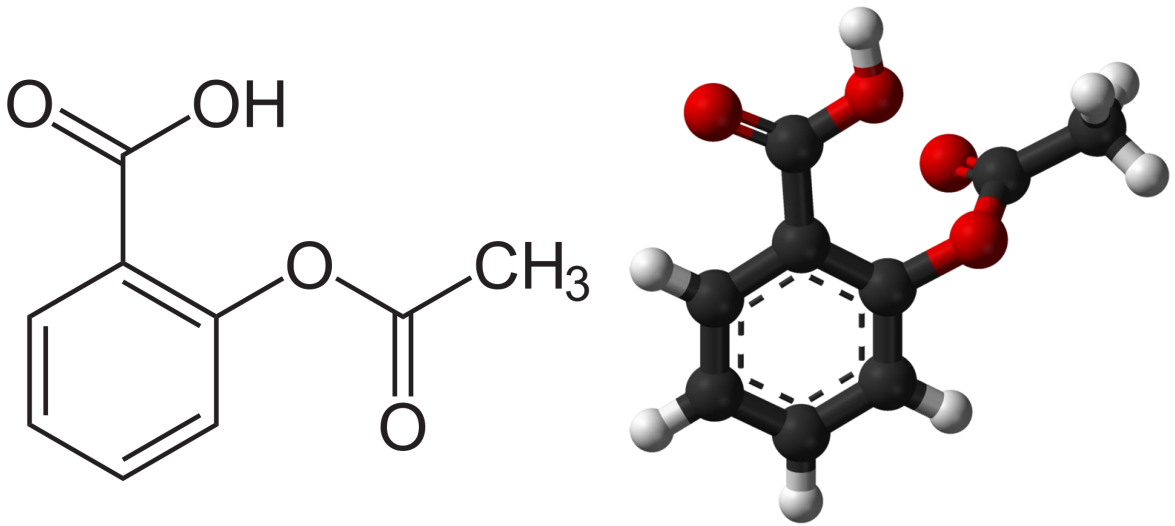
markdown

html pdf word

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T5_Netherlands_BOSCH_Paper.pdf

Formats

Базовое представление, молекулярный граф



2-ацетилоксибензойная кислота,
ацетилсалициловая кислота, C9H8O4

	C	C	O	O	C	C	C	C	C	C	C	O	O
C	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	1.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
O	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
O	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	1.0	0.0	1.5	0.0	0.0	0.0	1.5	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.5	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	1.5	0.0	1.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0	1.0
O	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
O	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

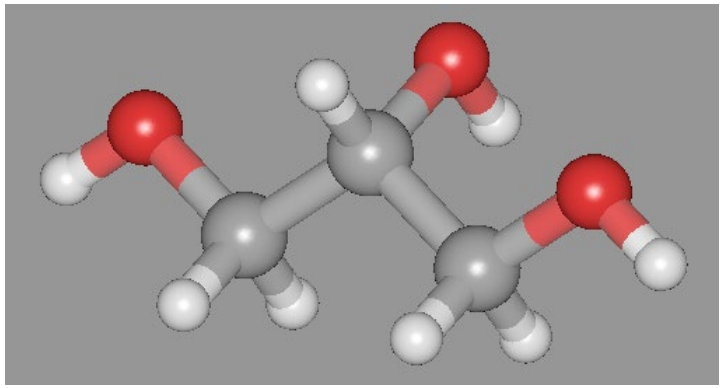
.xyz

Cartesian

(Формат .xyz)

1) Простой

Глицерин



Internal

(Z-matrix)

1) Химический смысл
2) Сложнее в построении
3) Много способов задать

```
14
Comments line
C          -1.2540710000          0.3888570000         -0.1001430000
O          -1.7850710000          1.7148570000         -0.0311420000
C           0.2339290000          0.4098570000          0.2938570000
O           0.9539290000          1.2578570000         -0.6041420000
C           0.8059290000         -1.0181430000          0.2188580000
O           2.1879280000         -0.9981430000          0.5848580000
H          -1.3540710000          0.0168570000         -1.0991420000
H          -1.7930710000         -0.2461430000          0.5718580000
H          -2.7130710000          1.7018570000         -0.2771420000
H           0.3339290000          0.7818570000          1.2918580000
H           0.8649290000          0.9248570000         -1.5001420000
H           0.2679290000         -1.6541430000          0.8908580000
H           0.7069290000         -1.3911430000         -0.7791430000
H           2.5439290000         -1.8891430000          0.5388580000
```

N	Symbol	R	Angle	Dihedral	NR	NA	ND
1	C						
2	O	RO2			1		
3	C	RC3	AC3		2	1	
4	O	RO4	AO4	DO4	3	2	1
5	C	RC5	AC5	DC5	4	3	2
6	O	RO6	AO6	DO6	5	4	3
7	H	RH7	AH7	DH7	6	5	4
8	H	RH8	AH8	DH8	7	6	5
9	H	RH9	AH9	DH9	8	7	6
10	H	RH10	AH10	DH10	9	8	7
11	H	RH11	AH11	DH11	10	9	8
12	H	RH12	AH12	DH12	11	10	9
13	H	RH13	AH13	DH13	12	11	10
14	H	RH14	AH14	DH14	13	12	11

.mol

```
Benzene, ID: C001
  SJC    20160623 1   1.00000     0.00000
Example Benzene mol file.
  6  6  0     0  0               1 V2000
   -1.3961    0.0013   -0.0504 C   0  0  0  0  0  0           0  0  0
   -0.7402   -0.3516    1.1313 C   0  0  0  0  0  0           0  0  0
    0.6556    -0.344    1.1844 C   0  0  0  0  0  0           0  0  0
    1.3956    0.0123    0.0546 C   0  0  0  0  0  0           0  0  0
    0.7398    0.3611   -1.1284 C   0  0  0  0  0  0           0  0  0
    -0.656    0.3577   -1.1803 C   0  0  0  0  0  0           0  0  0
  2  1  2  0     0  0
  1  3  1  0     0  0
  4  2  1  0     0  0
  3  6  2  0     0  0
  5  4  2  0     0  0
  6  5  1  0     0  0
M  END
```

DOI:10.1186/s13321-016-0168-9

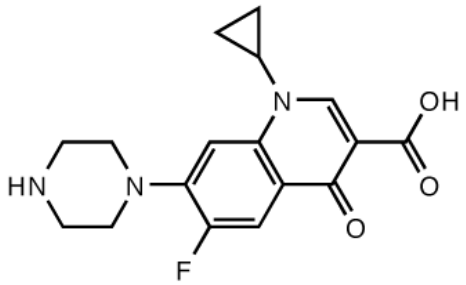
.sdf

```
NCGC00015959-03
  Marvin  07111412562D

 25 30  0  0  0  0            999 V2000
    3.4098   -1.3130    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
    4.8329   -1.3130    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4098   -2.1380    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.1248   -2.5436    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.6948   -2.5436    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.8329   -2.1380    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.1248   -0.8937    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.5547   -0.8937    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0

  1  3  1  0  0  0  0
  1  7  2  0  0  0  0
  1 25  1  0  0  0  0
  2  7  1  0  0  0  0
  2  6  2  0  0  0  0
  2  8  1  0  0  0  0
  3  4  2  0  0  0  0
  3  5  1  0  0  0  0
  4 13  1  0  0  0  0
  4  6  1  0  0  0  0
  5  9  1  0  0  0  0

M  CHG  1   1   1
M  END
>  <Formula>
C20H14NO4

>  <FW>
332.3289

>  <DSSTox_CID>
25204

>  <Active>
1
```

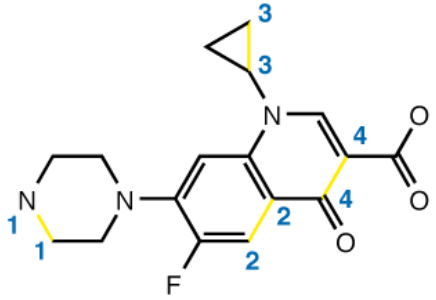
DOI:10.3906/elk-1809-9

SMILES

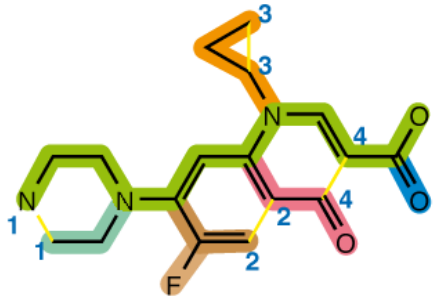
A



B



C



D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

- Игнорируем водороды (если это не критично)
- Выбираем самую длинную цепочку
- Разрываем циклические фрагменты
- Записываем (-, =, #, : или не записываем) связи
- Записываем боковые цепочки в скобках

Но:

- С какого атома начать?
- Ароматические системы?

https://ru.wikipedia.org/wiki/SMILES

Other ASCII representations

- InChi
- SMARTS
- SYBYL
- SELFIES
- …

.cif

```
data_1000041
loop_
_publ_author_name
'Abrahams, S C'
'Bernstein, J L'
_publ_section_title                 Accuracy of an automatic diffractometer. ...
_journal_coden_ASTM                 ACCRA9
_journal_name_full                  'Acta Crystallographica (1,1948-23,1967)'
_journal_page_first                 926
_journal_page_last                  932
_journal_paper_doi                  10.1107/S0365110X65002244
_journal_volume                     18
_journal_year                       1965
_chemical_formula_structural        'Na Cl'
_chemical_formula_sum               'Cl Na'
_chemical_name_systematic           'Sodium chloride'
_space_group_IT_number              225
_symmetry_cell_setting              cubic
_symmetry_Int_Tables_number         225
_symmetry_space_group_name_Hall     '-F 4 2 3'
_symmetry_space_group_name_H-M      'F m -3 m'
_cell_angle_alpha                   90
_cell_angle_beta                    90
_cell_angle_gamma                   90
_cell_formula_units_Z               4
_cell_length_a                      5.62
_cell_length_b                      5.62
_cell_length_c                      5.62
_cell_volume                        177.5
_refine_ls_R_factor_all             0.022
_cod_database_code                  1000041
loop_
_symmetry_equiv_pos_as_xyz
x,y,z
y,z,x
...
1/2+z,y,1/2-x
1/2+z,1/2+y,-x
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_symmetry_multiplicity
_atom_site_Wyckoff_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
_atom_site_attached_hydrogens
_atom_site_calc_flag
Na1 Na1+ 4 a 0. 0. 0. 1. 0 d
Cl1 Cl1- 4 b 0.5 0.5 0.5 1. 0 d
loop_
_atom_type_symbol
_atom_type_oxidation_number
Na1+ 1.000
Cl1- -1.000
```

DOI:10.1186/s13321-016-0168-9

.pdb

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

DOI:10.1007/978-3-642-54816-1_4

DataBases

Pubchem

National Library of Medicine
National Center for Biotechnology Information

PubChem About Blog Submit Contact Search PubChem

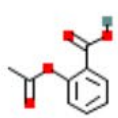
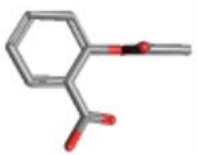
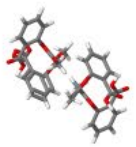
COMPOUND SUMMARY

Aspirin

Cite Download

CONTENTS

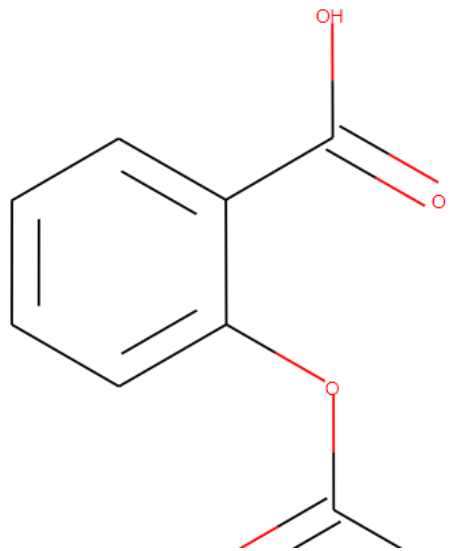
- Title and Summary
- 1 Structures
- 2 Names and Identifiers
- 3 Chemical and Physical Properties
- 4 Spectral Information
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication

PubChem CID	2244
Structure	 2D  3D  Crystal

Find Similar Structures

Compound Report Card

Name And Classification



ID:	CHEMBL25
Name:	ASPIRIN
Max Phase:	4 Approved
Molecular Formula:	C9H8O4
Molecular Weight:	180.16
ChEMBL Synonyms:	Acetylsalicylic Acid, ACETYLSALICYLIC ACID, Aspirin, ASPIRIN, BAY1019036, NSC-27223, NSC-406186
Synonyms From Alternate Forms:	ACETYLSALICYLATE LYSINE, ASPIRIN DL-LYSINE, Lysine Acetylsalicylate, LYSINE ACETYLSALICYLATE
	8-HOUR BAYER, Acetosalic Acid, ACETYLSALIC ACID, Acetylsalicylic Acid, ALKA RAPID, ANADIN ALL NIGHT, ANGETTES 75, ASPIRIN, ASPRO CLR

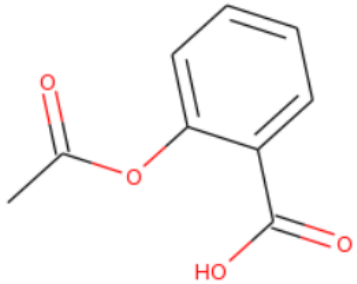
ZINC53 (Aspirin)

In: anodyne bb fda for-sale in-stock natural-products

Google Wikipedia PubMed

Added	Availability	Since	Mwt	logP	Download
2005-09-27	In-Stock	2015-08-07	180.159	1.31	

Mol Formula	Rings	Heavy Atoms	Hetero Atoms	Fraction sp^3	Tranche
C9H8O4	1	13	4	0.11	ADAA



Draw

SMILES	CC(=O)Oc1ccccc1C(=O)O
InChI	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
InChI Key	BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Available 3D Representations

PDB

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | Documentation | Careers | MyPDB

RCSB PDB PROTEIN DATA BANK

185158 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search terms or PDB ID(s).

Advanced Search | Browse Annotations

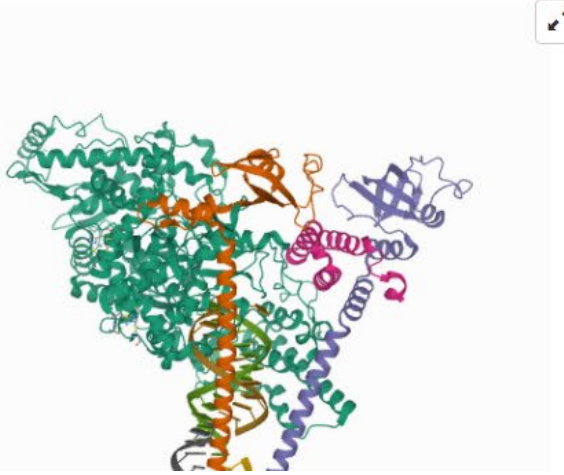
Help

PDB-101 | Worldwide PDB | EMDataResource | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Celebrating 50 YEARS OF Protein Data Bank

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Versions

Biological Assembly 1



Display Files | Download Files

6YYT

Structure of replicating SARS-CoV-2 polymerase

DOI: 10.2210/pdb6YYT/pdb **EMDataResource:** EMD-11007

Classification: VIRAL PROTEIN
Organism(s): Severe acute respiratory syndrome coronavirus 2, synthetic construct
Expression System: Spodoptera aff. frugiperda 1 BOLD-2017, Escherichia coli BL21(DE3)
Mutation(s): No

Deposited: 2020-05-06 **Released:** 2020-05-13
Deposition Author(s): Hillen, H.S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D., Cramer, P.
Funding Organization(s): German Research Foundation (DFG), European Research Council (ERC), Volkswagen

ICSD

ICSD

Welcome to ICSD Web. Logged in: Ruehl, Stephan

FIZ Karlsruhe | Contact
Logout

Content Selection

- [x] Experim. inorganic structures
- [x] Experim. metal-organic str.
- [x] Theoretical structures

Navigation

- Basic search & retrieve

Advanced search & retrieve

- Bibliography
- Cell
- Chemistry
- Symmetry
- Crystal Chemistry
- Structure Type
- Experimental Information
- DB Info

Query Management

- Manage Queries
- List Combined Queries
- Create Combined Query

ICSD links

- ICSD News

Basic Search & Retrieve

Bibliography

Authors

Year of Publication

Title of Journal

Title of Article

Chemistry

Composition — Periodic Table

Number of Elements

Cell

Cell Parameters

Cell Volume

Tolerance +/- %

Symmetry

Space Group Symbol

Space Group Number

Crystal System

Centering

Exp. Info. & Ref. Data

New Data Only

PDF Number

Temperature — K

ICSD Collection Code

Pressure — MPa

Clear Basic Search

Count Basic Search

Search Action

Run Query | Clear Query

Search Summary

Basic Search: -

Query History

Number of queries: 23

Clear Query History

2019-06-11T09:11	1
2019-06-04T14:28	7432
2019-06-04T14:26	9418
2019-06-04T14:24	9585
2019-06-04T14:22	9633
2019-06-04T14:19	9780
2019-06-04T14:17	9514
2019-06-04T14:14	9374
2019-06-04T09:12	9522
2019-06-04T09:11	9717
2019-06-04T09:09	9722

CSD

Organic
43%

Metal-Organic
57%

At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po

Not Polymeric
89%

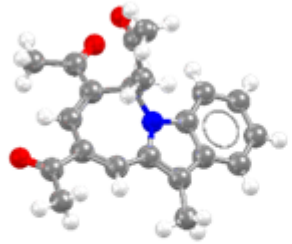
Polymeric: 11%

Single Component
56%

Multi Component
44%

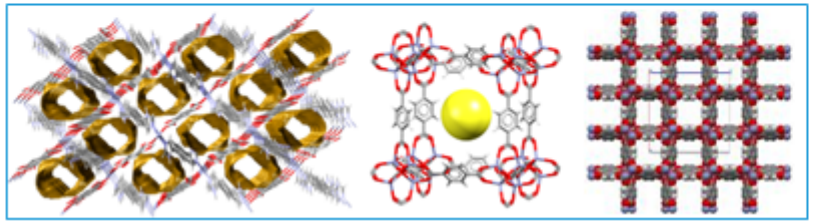
Organic

- Drugs
- Agrochemicals
- Pigments
- Explosives
- Protein ligands



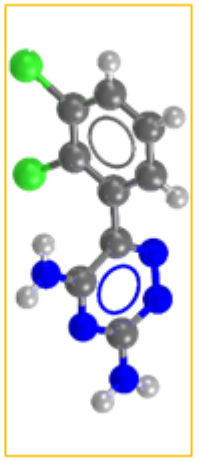
Metal-Organic

- Metal Organic Frameworks
- Models for new catalysts
- Porous frameworks for gas storage
- Fundamental chemical bonding



Additional data

- 10,860 polymorph families
- 169,218 melting points
- 840,667 crystal colours
- 700,002 crystal shapes
- 23,622 bioactivity details
- 9,740 natural source data
- > 250,000 oxidation states



Links/subsets

- Drugbank
- Druglike
- MOFs
- PDB ligands
- PubChem
- ChemSpider
- Pesticides

https://www.ccdc.cam.ac.uk/solutions/csd-core/components/csd/

Crystallography Open Database

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

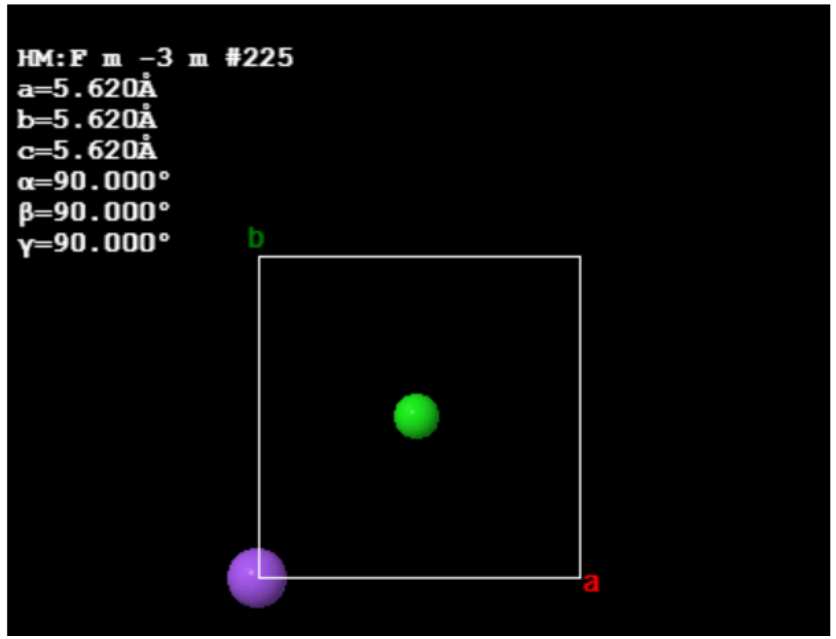
Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

Documentation
COD Wiki
Obtaining COD
License
Querying COD
Citing COD
COD Mirrors

Information card for entry 1000041

1000040 << **1000041** >> 1000042

Preview



‘Good’ database

- Big
- Curated
- Conventional formats, exports (downloads), API
- Computational-ready
- (Multi)labelled
- Free


More DataBases?

Other: CORE-MOF

Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019

Yongchul G. Chung*, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann*, David S. Sholl*, and Randall Q. Snurr*

Cite this: *J. Chem. Eng. Data* 2019, 64, 12, 5985–5998
Publication Date: November 4, 2019
https://doi.org/10.1021/acs.jced.9b00835

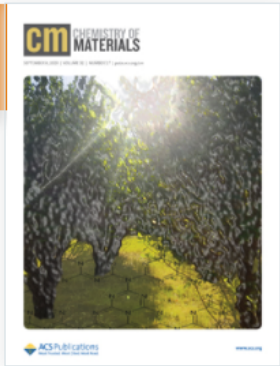
Journal of Chemical & Engineering Data

Other: 2D-hybrid perovskites

Database of Two-Dimensional Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning

Ekaterina I. Marchenko, Sergey A. Fateev, Andrey A. Petrov, Vadim V. Korolev, Artem Mitrofanov, Andrey V. Petrov, Eugene A. Goodilin, and Alexey B. Tarasov*

Cite this: *Chem. Mater.* 2020, 32, 17, 7383–7388
Publication Date: August 11, 2020
https://doi.org/10.1021/acs.chemmater.0c02290



Chemistry of Materials

Read Online | PDF (3 MB) | SI Supporting Info (1) »

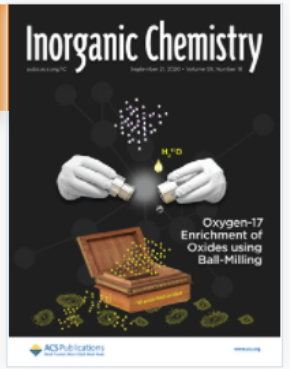
SUBJECTS: Layers,

Other: HERDB

Heavy-Element Reactions Database (HERDB): Relativistic ab Initio Geometries and Energies for Actinide Compounds

Nikolai Andreadi*, Artem Mitrofanov, Petr Matveev, Anna Volkova, and Stepan Kalmykov

Cite this: *Inorg. Chem.* 2020, 59, 18, 13383–13389
Publication Date: September 2, 2020
https://doi.org/10.1021/acs.inorgchem.0c01746



Inorganic Chemistry

Abstract

Actinide chemistry appears to be a challenge for both experimentalists and theoreticians. Radioactivity and computational obstacles lead to a lack of