PubChem: A resource for chemical health and safety

Evan Bolton, Ph.D.
Disclaimer

• NLM does not endorse or recommend any commercial products, processes, or services. The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

• It is not the intention of NLM to provide specific medical advice, but rather to provide information to better understand health and disease. Specific medical advice will not be provided, and NLM urges you to consult with a qualified health professional for diagnosis and for answers to your personal medical questions.
PubChem is a data repository

• World's largest collection of freely accessible chemical information.

• Helps researchers make sense of the biological roles and health effects of chemicals on human health and the environment.

Many page types

- Compound
- Gene
- Protein
- BioAssay
- Substance
- Patent
- Pathway, Taxonomy, Cell-line, and more

**PubChem** Dopamine (Compound)

**PubChem** ATP13A2 - ATPase cation transporting 13A2 (human) (Gene)

**PubChem** PDZ domain-containing protein 11 (Protein)

**PubChem** siRNA Circadian Assay (BioAssay)

**PubChem** GNF169433 (Substance)

**PubChem** COMPOSITIONS AND METHODS FOR TREATMENT OF NEU...

**PubChem** Ion channel transport (Pathway)
A lot of data enabling many use cases

<table>
<thead>
<tr>
<th>Collection</th>
<th>Live Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Periodic Table of Elements</td>
<td>118</td>
<td>Interactive periodic table with up-to-date element properties</td>
</tr>
<tr>
<td>Compounds</td>
<td>115,668,812</td>
<td>Unique chemical structures extracted from contributed PubChem</td>
</tr>
<tr>
<td>Substances</td>
<td>307,633,237</td>
<td>Information about chemical entities provided by PubChem</td>
</tr>
<tr>
<td>BioAssays</td>
<td>1,626,630</td>
<td>Biological experiments provided by PubChem contributors</td>
</tr>
<tr>
<td>Bioactivities</td>
<td>292,123,746</td>
<td>Biological activity data points reported in PubChem BioAssays</td>
</tr>
<tr>
<td>Genes</td>
<td>112,728</td>
<td>Gene targets tested in PubChem BioAssays and those involved</td>
</tr>
<tr>
<td>Proteins</td>
<td>186,035</td>
<td>Protein targets tested in PubChem BioAssays and those involved</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>113,693</td>
<td>Organisms of targets tested in PubChem BioAssays and those involved</td>
</tr>
<tr>
<td>Pathways</td>
<td>240,671</td>
<td>Interactions between chemicals, genes, and proteins</td>
</tr>
<tr>
<td>Cell Lines</td>
<td>1,986</td>
<td>Information about cell lines</td>
</tr>
<tr>
<td>Literature</td>
<td>38,884,316</td>
<td>Scientific publications with links in PubChem</td>
</tr>
<tr>
<td>Patents</td>
<td>42,020,426</td>
<td>Patents with links in PubChem</td>
</tr>
<tr>
<td>Data Classifications</td>
<td>71</td>
<td>Browse the distribution of PubChem data among nodes in</td>
</tr>
<tr>
<td>Data Sources</td>
<td>932</td>
<td>Organizations contributing data to PubChem</td>
</tr>
</tbody>
</table>
Two primary archival databases

PubChem Substance and Compound databases
DOI: 10.1093/nar/gkv951
PMID: 26400175
PMCID: PMC4702940
Compound Summary

- Top-level summary
- Nested navigation menu
- >500 different data fields
  - Experimental properties, Spectra, Literature, Toxicity, Bioactivities, Chemical vendors, Pharmacology, Patents, Pathways, Health & Safety, Classifications, …
- Clear provenance
- Annotation from authoritative and curated sources

Toluene

See also: Benzene, toluene, ethylbenzene and xylene (component of); Benzene, toluene and xylene (component of); Laboratory-grade xylene (impurity of).

PubChem CID: 1140

Structure

Chemical Safety

- Flammable
- Irritant
- Health Hazard

Laboratory Chemical Safety Summary (LCSS) Datasheet

Molecular Formula: C₇H₈
C₇H₈CH₃

Synonyms
- toluene
- methylbenzene
- tolul
- 108-88-3
- Phenylmethane

Molecular Weight: 92.14 g/mol
Computed by PubChem 2.2 (PubChem release 2021.10.14)

Dates
- Create: 2004-09-16
- Modify: 2023-08-12

Description
Toluene is a clear, colorless liquid with a distinctive smell. Toluene occurs naturally in crude oil and in the tolu tree. It is also produced in the process of making gasoline and other fuels from crude oil and making coke from coal. Toluene is used in making paints, paint thinners, fingernail polish, lacquers, adhesives, and rubber and in some printing and leather tanning processes.

Agency for Toxic Substances and Disease Registry (ATSDR)
Hundreds of compound-specific annotations are available

The PubChem Compound TOC (Table of Contents) classification contains a browse-able tree of all available sections within PubChem Compound records.
Find records with a specific annotation
Hundreds of data sources

The PubChem Data Sources interface contains a list of all available data sources within PubChem that one can search, filter, and download.
Find data sources and what they provide

<table>
<thead>
<tr>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Find out who contributed what to PubChem.</strong> Interested in becoming a PubChem contributor? Learn how to get started with a PubChem submission.</td>
</tr>
<tr>
<td>80 sources</td>
</tr>
</tbody>
</table>

**FILTER BY** |
- **Data Type**
  - Annotations (85)
  - Citations (2)
  - ClinicalTrials.gov (2)
  - Classification (15)
  - Clinical (15)
  - ClinicalProtocols (2)
  - On-Hold Biotechs (2)
  - Pathways (1)
  - On-Hold Substances (1)

**Source** |
- ClinicalTrials.gov
  - Governmental Organizations
    - United States
    - ClinicalTrials.gov
    - United States
    - 17,732 Annotations
    - 2023/08/12
- NLM ReNorm Terminology
  - Governmental Organizations
    - United States
    - 6,896 Annotations
    - 2023/08/11
- NIST Physical Measurement Laboratory
  - Governmental Organizations
    - United States
    - 118 Annotations
    - 2023/08/11

**DATA SOURCES** |
- NIOSH Manual of Analytical Methods
  - Governmental Organizations
    - United States
    - 596 Annotations
    - 2023/08/11

**NIOSH Manual of Analytical Methods**

NIOSH Manual of Analytical Methods (NMAM) is a collection of methods for sampling and analysis of contaminants in workplace air, and in the blood and urine of workers who are occupationally exposed. These methods have been developed or adapted by NIOSH or its partners and have been evaluated according to established experimental protocols and performance criteria.

**Organization** |
- CDC
- Governmental Organizations

**Category** |
- Governmental Organizations

**URL** |

**License Note** |
- The information provided using CDC Web site is only intended to be general summary information to the public. It is not intended to take the place of either the written law or regulations.

**License URL** |
- https://www.cdc.gov/Other/disclainer.html

**Contact Name** |
- NIOSH CDC

**Address** |
- 1600 Clifton Road, Atlanta, GA, United States, 30329-4027

**Data Source ID** |
- 3712

**Data in PubChem** |
- 596 Annotations

**Last Updated** |
- 2023/08/11

**Annotations from NIOSH Manual of Analytical Methods**

1 annotation topic

596 total annotation data items

NIOSH Analytical Methods (Compound) 

Download
Health and Safety data in PubChem

• More than 100 data sources provide health and safety, toxicity, and property data

Are we missing data sources that you use? What should be added?
Describing the indescribable (UVCBs)

Asbestos

Description: Asbestos is the name given to a group of six different fibrous minerals (amosite, chrysotile, crocidolite, and the fibrous varieties of tremolite, actinolite, and anthophyllite) that occur naturally in the environment. Asbestos minerals have separable long fibers that are strong and flexible enough to be spun and woven and are heat resistant. Because of these characteristics, asbestos has been used for a wide range of manufactured goods, mostly in building materials (roofing shingles, ceiling and floor tiles, paper products, and asbestos cement products), friction products (automobile clutch, brake, and transmission parts), heat-resistant fabrics, packaging, gaskets, and coatings. Some vermiculite or talc products may contain asbestos.

PubChem CID: Not available because this is not a discrete structure.

Molecular Formula: [Mg₆Si₄O₁₀(OH)₈]

Signal: Danger

GHS Hazard Statements:
- H350: May cause cancer [Danger: Carcinogenicity]
- H372: *F: Causes damage to organs through prolonged or repeated exposure [Danger: Specific target organ toxicity, repeated exposure]

Precautionary Statement Codes:
- P203, P260, P264, P270, P280, P318, P319, P405, and P501
  (The corresponding statement to each P-code can be found at the GHS Classification page.)

10.1.2 Hazard Classes and Categories:
- Carc. 1A
- STOT RE 1
- EU REGULATION (EC) No 1272/2008
- Carc. 1A (100%)
- European Chemicals Agency (ECHA)
Describing the indescribable (UVCBs)
Describing the indescribable (UVCBs)

Recently introduced, the PubChem Reference Collection includes a set of records that include undefined, variable, complex mixtures, and biologics. It also includes reference information from curated and authoritative data sources such as chemical structure and chemical names.

Internally, we call these substance records “concepts” and they help PubChem improve the quality of information.

### DATA SOURCES

**PubChem Reference Collection**

The PubChem Reference Collection contains a reference set of chemical substances derived from authoritative and curated data sources. These substances contain the chemical names and (when applicable) a representative chemical structure used by PubChem to link to chemical substance annotations. This collection includes structures that lack full atomic-level detail: polymers, complex mixtures, unknown or variable composition, biologics, reaction products, and more. The quality of this data set may be annotated as a comment relative to the availability of authoritative and curated data sources.

<table>
<thead>
<tr>
<th>Organization</th>
<th>PubChem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Governmental Organizations</td>
</tr>
<tr>
<td>Contact Name</td>
<td>PubChem Staff</td>
</tr>
<tr>
<td>Address</td>
<td>8600 Rockville Pike, Bethesda, MD, United States, 20894</td>
</tr>
<tr>
<td>Data Source ID</td>
<td>26136</td>
</tr>
<tr>
<td>Data in PubChem</td>
<td>53,547 Live Substances</td>
</tr>
<tr>
<td>Last Updated</td>
<td>2023/07/25</td>
</tr>
</tbody>
</table>
Engaging the community

Multi-year engagements with many parties involved

.. here are some examples
Engaging the community

ACS Professional Divisions have been pivotal
• CINF (Leah McEwen) and CHAS (Ralph Stuart)
• Answer key questions necessary for progress
  • What is needed and how to present it?
  • Provided a trajectory towards improvements
    • Use cases, content, and interfaces
• PubChem LCSS and other engagements

Laboratory Chemical Safety Summary (LCSS) views now available in PubChem

Posted on August 17, 2015

The PubChem Laboratory Chemical Safety Summary (LCSS) provides pertinent chemical health and safety data for a given PubChem Compound record. The PubChem LCSS is a community effort involving professionals in health and safety, chemistry librarianship, informatics, and other specialties.
Engaging the community

Univ. California, CHAS, IUPAC, ...

- Series of meetings, workshops, and interactions
- Many practical issues to overcome
- Mixtures are very important
- MInChI, IUPAC InChI QR Codes, ...
- Big influence on PubChem efforts

Specification of International Chemical Identifier (InChI) QR codes for linking labels on containers of chemical samples to digital resources (IUPAC Recommendations 2021)

Jeremy G. Frey, Richard M. Hartshorn and Leah R. McEwen

From the Journal Pure and Applied Chemistry
https://doi.org/10.1515/pac-2021-0604

Abstract

This article discusses the ways of linking physical objects to digital information relevant to chemical entities, specifically those that can be described by the use of the IUPAC International Chemical Identifier (InChI). It makes recommendations on the form of the computer readable components of labels provided for chemicals and materials that are used on product/sample containers and on the associated documentation that is used when transporting these containers (either internally or during export/import). The focus is on specification of the content of the 2D Quick Response bar codes required to describe the molecular content of the containers and link to digital resources to supplement that provided on a physical label. The necessary technical and (possible) business infrastructure necessary to support the use of the InChI and InChIKey for rapid recall of relevant information is considered here and suggestions are made.

Keywords: cheminformatics; InChI; InChIKey; International Chemical Identifier; IUPAC; QR code
Engaging the community

Pistoia Alliance Chemical Safety Library, ...
- Chemical reaction safety importance
- Need for community-based resources for (unexpected) reactivity hazards
- *PubChem CAMEO Chemical Reactivity Classification integration*
- *Pistoia Alliance CSL data source*
Engaging the community

European Chemicals Agency (ECHA), CHAS, ...

- Difficulties disseminating REACH data
- Programmatic access tools (APIs)
- Primary source of GHS information
- How to summarize many GHS reports?
- Improved PubChem GHS displays, classification
Engaging the community

Environmental Protection Agency (EPA), Norman SLE, ...

- Chemical information is highly nuanced
- Regulatory needs can be different
- Many new data sources and improved integration
- Describing the “indescribable” (UVCBs)
- PubChem “concept” infrastructure introduced
Engaging the community

Chemical Abstract Services, EPA, IUPAC, ...

- Improving the quality and trust of information
- Clarifying CAS Registry Number and chemical structure associations
- Scoping of “common” chemicals
- **Revamp of the CAS Common Chemistry website**

https://commonchemistry.cas.org/
Summing it all up

Community engagements are key
Heavily influence changes made to PubChem

It is all about the use cases
What does the community need? Why does the community need it?

PubChem is your safety resource
How can we help?
PubChem Crew ...

Evan Bolton
Jie Chen
Tiejun Cheng
Asta Gindulyte
Jane He
Siqian He
Sunghwan Kim

Qingliang Li
Ben Shoemaker
Paul Thiessen
Bo Yu
Leonid Zaslavsky
Jian Zhang

Special thanks to the NCBI Help Desk, the NCBI Systems teams, and past PubChem group members
Special thanks

- All PubChem contributors and collaborators
- Health and Safety collaborators (especially Ralph Stuart, Leah McEwen)
- Exposomics collaborators (especially Emma Schymanski, NORMAN SLE)
- IUPAC collaborators (especially InChI contributors, Leah McEwen)
- Software collaborators
  - NextMove Software (Roger Sayle, John May) and Daniel Lowe / Noel O’Boyle
  - Xemistry GmbH (Wolf D. Ihlenfeldt)
  - OpenEye Scientific Software

- This research was supported by the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM), National Institutes of Health (NIH)
Bring your data into the light

Open science is grand ...

... so others can build off it and advance the (data) science

Help identify and close knowledge gaps

Support interpretation of complex data

Increase the utility of your research

many hands make light work

Information in the public domain helps everybody

more data is more data

making data FAIR

Science is data driven

many data gaps

U.S. National Library of Medicine
National Center for Biotechnology Information
NCBI