

PubChem 2017

where are we now and where are we going?

Evan Bolton, Ph.D.

U.S. National Center for Biotechnology Information (NCBI)

Oct. 11, 2017

Global Ingredient Archival System (GInAS) Meeting 2017 at USP



PubChem Resource

<https://pubchem.ncbi.nlm.nih.gov/>

The screenshot shows the main interface of the PubChem website. At the top, there is a navigation bar with links for Databases, Upload, Services, Help, and more. A prominent feature is a box titled "Today's Statistics" containing various compound counts. To the right, there are social media sharing icons for Facebook, Twitter, Google+, and RSS. On the far right, a vertical column lists various tools and services with corresponding icons. Below the main content area, there are two "New" notifications and a link to "more ...". At the bottom, there is a footer with links to Helpdesk, Disclaimer, Privacy Statement, Accessibility, Data Citation Guidelines, and the National Center for Biotechnology Information, NLM, NIH, and HHS.

Today's Statistics ›

Compounds:	92,699,805
Substances:	233,117,297
BioAssays:	1,252,815
Tested Compounds:	2,395,818
Tested Substances:	3,820,358
RNAi BioAssays:	170
BioActivities:	233,516,687
Protein Targets:	10,341
Gene Targets:	22,104

PubChem is an open chemical database resource with the most comprehensive information on the biological activity of chemical substances

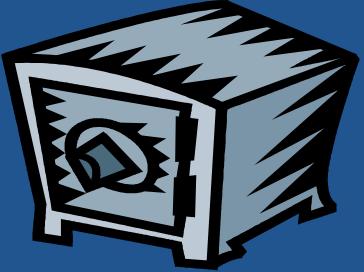
New The PubChem Target Summary page is now launched. [Read more...](#)

New In PubChem, more than 300 thousand chemicals have spectral information, including NMR, IR, Raman, MS and more. [Read more...](#)

more ...

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

U.S. National Library of Medicine

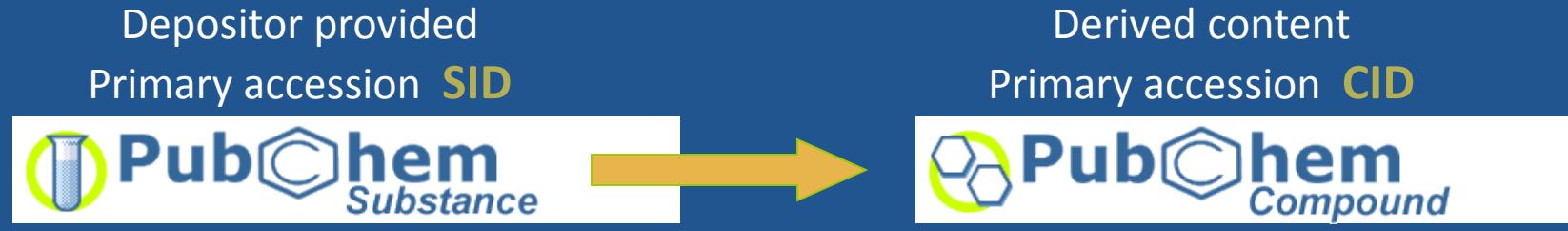


What is PubChem?



- An open archive
 - anyone can contribute
 - biological experiments
 - chemical structures
 - (NCBI) cross-references
 - substance synonyms
 - link-back URLs
 - textual commentary
 - hierarchical annotation
 - records are versioned
 - voluntary data push
 - interconnects chemical biology resources
- A public resource
 - anyone can access
 - search, subset, select, ~~analyze~~, download
 - integrated
 - biomedical literature, sequences, pathways, patents, ontologies, etc.
 - programmatic layers
 - URL-based interfaces
 - NCBI Entrez Utilities
 - PUG [XML, ~~VIEW~~, SOAP, REST, RDF]
 - PubChem Widgets
 - [embed PubChem into your web pages]

PubChem as an archive



Derived content
Primary accession **CID**

**Unique chemical structure
content of PubChem**

Substances not well-defined
are not a “compound”

Compound helps to link
substance records

Substance records keep
provenance clear

Why does a user come to PubChem?

- Google/Yahoo/Baidu suggested it
- I want to buy molecule 'X'
- Publications about molecule 'X'
- Patents/Biological activities
- What is known about the molecule?
 - Physical properties
 - Pharmacology
 - Biological activity
 - Safety information
 - Spectroscopy
 - Toxicity
 - Pathways
 - Etc.
- Deep questions query can decipher

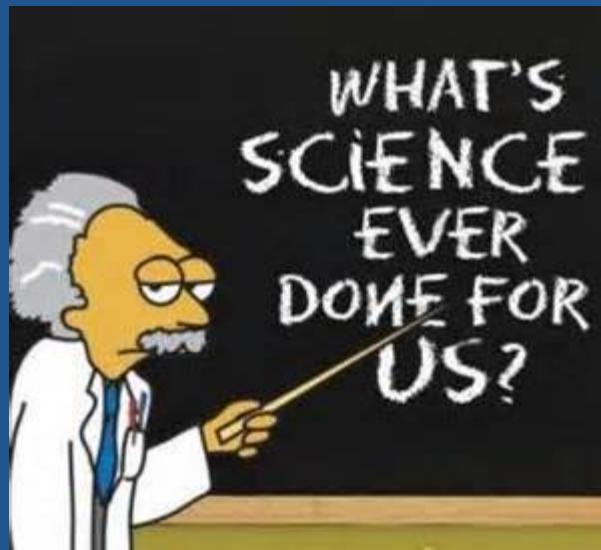


Image credit: <http://blogs.egu.eu/network/palaeoblog/2012/10/31/why-bother-communicating/>



PubChem data complexity

- Many links between large record collections
 - ~230M Substances <-> ~90M Compounds
 - ~90M Compounds <-> ~90M Compounds
 - ~230M Bioactivities <-> ~3M Substances
 - ~230M Bioactivities <-> ~2M Compounds
 - ~230M Bioactivities <-> ~1M BioAssays
 - ~10M PMIDs <-> ~100K Compounds
 - ~3M Patents <-> ~30M Substances
 - ~3M Patents <-> ~15M Compounds
- Sparse and dense data and/or linking
- New types of data, links, and metadata on a regular basis

PubChem linked data overview
<https://pubchem.ncbi.nlm.nih.gov/rdf/>

Scientific information is a tale of two cities ...

- Any (bioactivity) data system is a balancing act between two factions
 - Data creators
 - Data users
- Each has their own needs
 - Creators want to capture all details necessary to reproduce experiment
 - Users want ease of navigation and comparison between experiments
- PubChem tries to please both but caters to the users

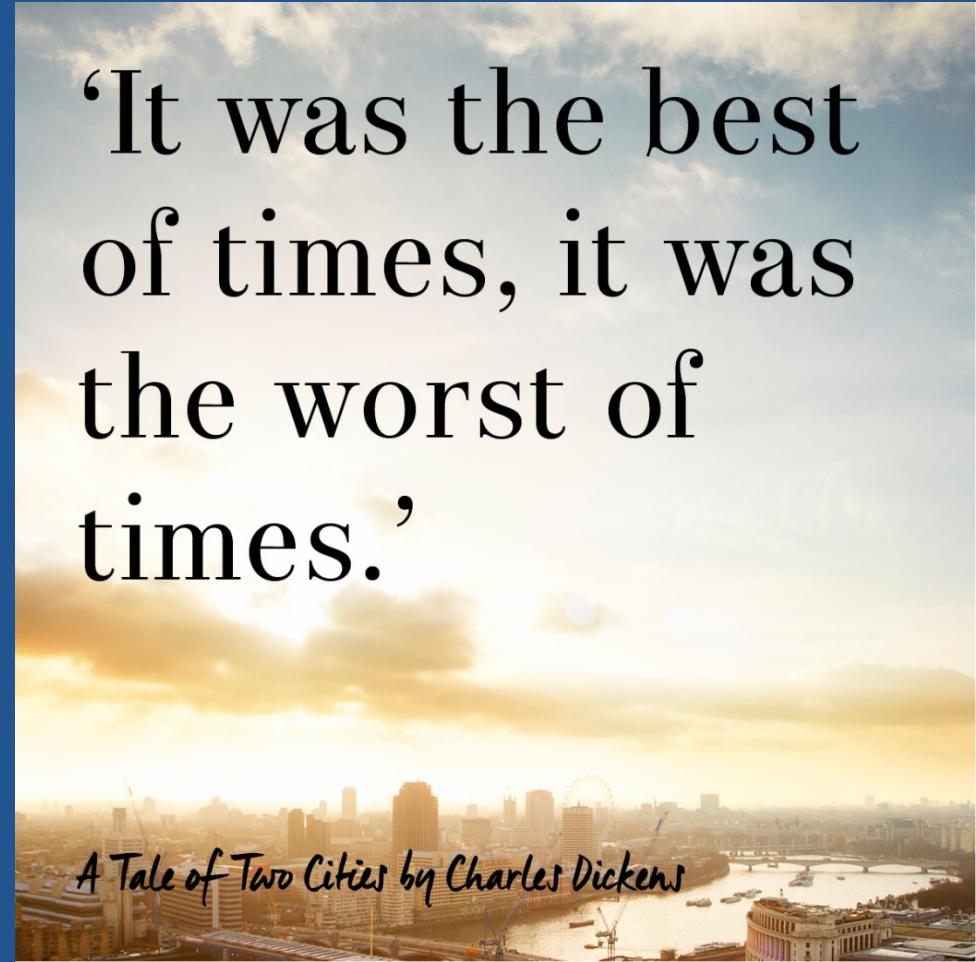
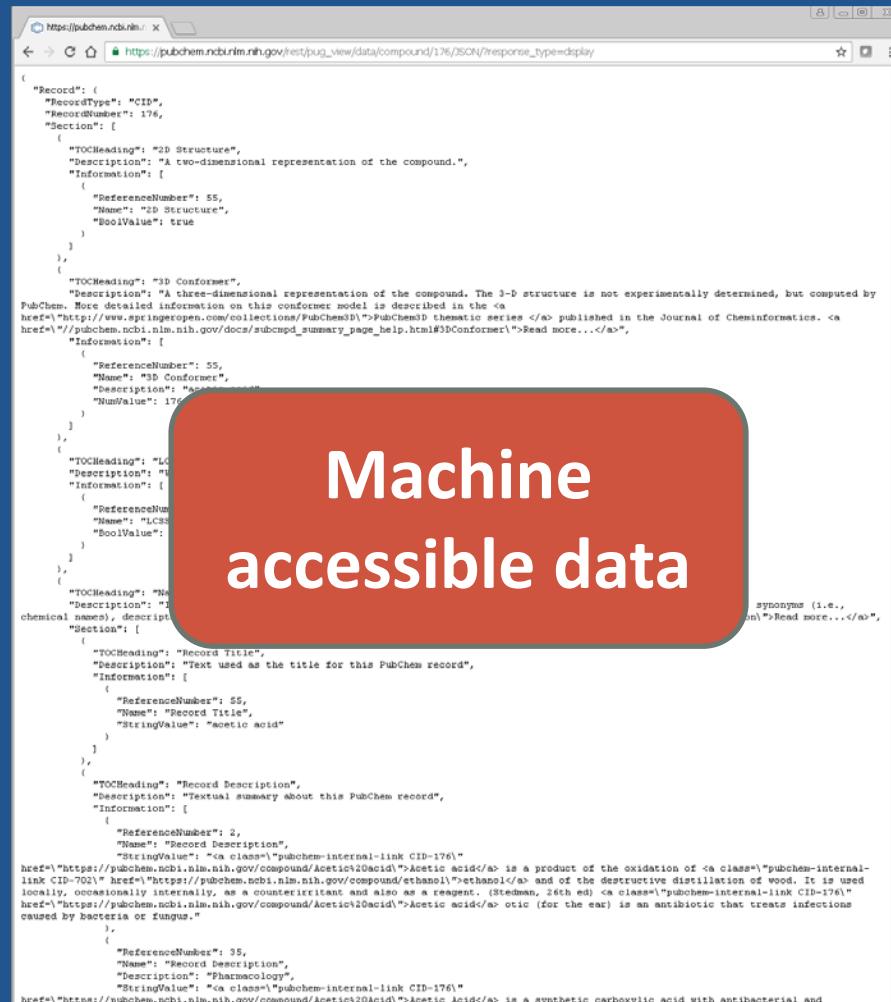


Image credit:

http://redonline.cdnds.net/main/thumbs/18557/a_tale_of_two_cities_best_first_lines_from_books_redonline.co.uk-2.jpg

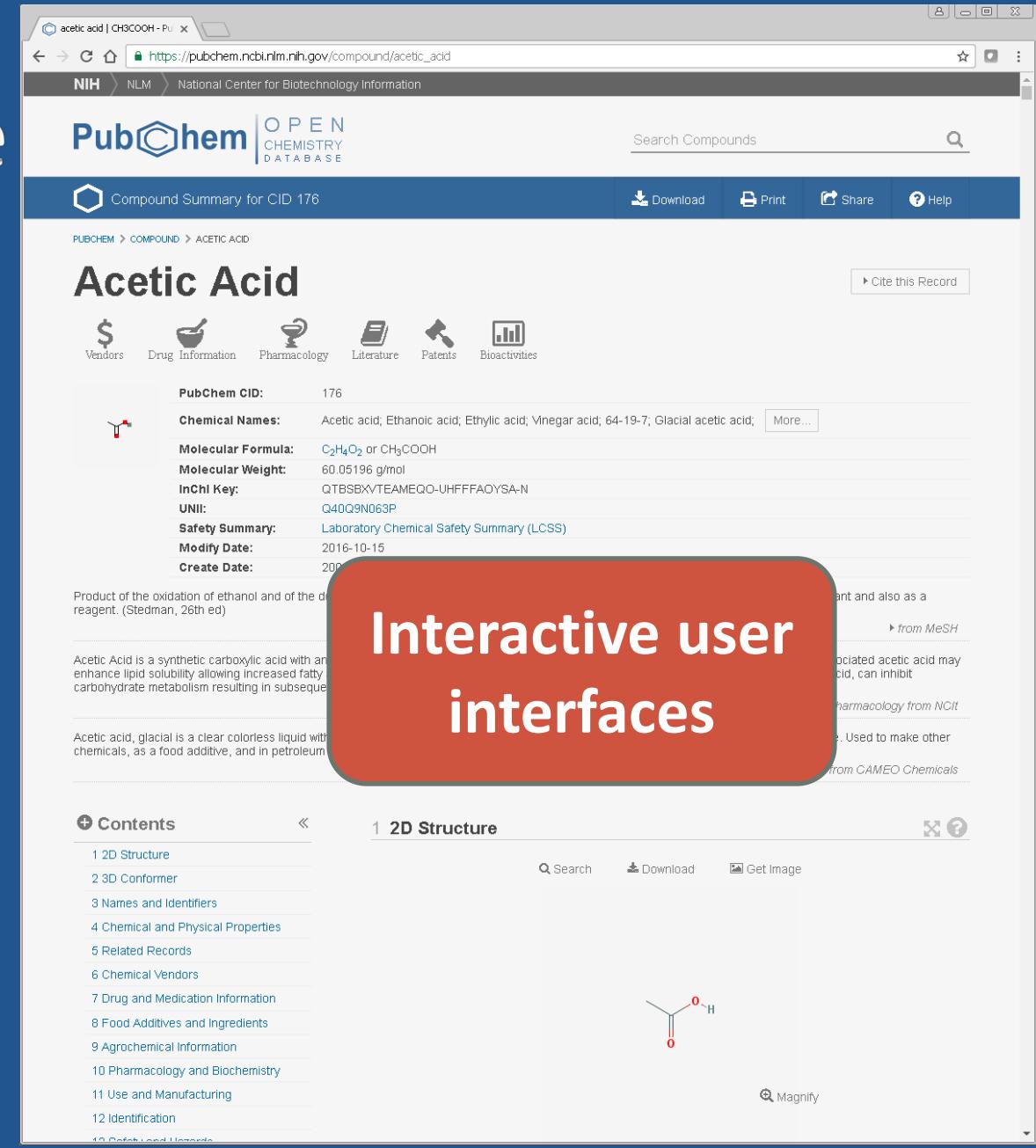
PubChem as a public interface



A screenshot of a web browser displaying a JSON API response for compound CID 176. The JSON structure includes sections for 2D Structure, 3D Conformer, Record Title, Record Description, and various identifiers and descriptions. A red callout box highlights the text "Machine accessible data".

```
{ "Record": { "RecordType": "CID", "RecordNumber": 176, "Section": [ { "TOCHeading": "2D Structure", "Description": "A two-dimensional representation of the compound.", "Information": [ { "ReferenceNumber": 55, "Name": "2D Structure", "BoolValue": true } ] }, { "TOCHeading": "3D Conformer", "Description": "A three-dimensional representation of the compound. The 3-D structure is not experimentally determined, but computed by PubChem. More detailed information on this conformer model is described in the <a href='http://www.springeropen.com/collections/SubChem3D'>PubChem3D thematic series </a> published in the Journal of Cheminformatics. <a href='http://pubchem.ncbi.nlm.nih.gov/docs/subcompd_summary_page_help.html#3DConformer'>Read more...</a>.", "Information": [ { "ReferenceNumber": 55, "Name": "3D Conformer", "Description": "Acetic acid", "NumValue": 176 } ] }, { "TOCHeading": "Record Title", "Description": "The title used for this PubChem record.", "Information": [ { "ReferenceNumber": 55, "Name": "Record Title", "StringValue": "acetic acid" } ] }, { "TOCHeading": "Record Description", "Description": "Textual summary about this PubChem record.", "Information": [ { "ReferenceNumber": 2, "Name": "Record Description", "StringValue": "Acetic acid<br>Acetic acid is a product of the oxidation of ethanol and of the destructive distillation of wood. It is used locally, occasionally internally, as a counterirritant and also as a reagent. (Stedman, 26th ed) <a href='https://pubchem.ncbi.nlm.nih.gov/compound/Acetic%20acid'>Acetic acid</a> is an antibiotic that treats infections caused by bacteria or fungi." } ] }, { "TOCHeading": "Names", "Description": "Synonyms for this compound.", "Information": [ { "ReferenceNumber": 35, "Name": "Record Description", "Description": "Acetic Acid<br>Acetic Acid is a synthetic carboxylic acid with antibacterial and antifungal properties. It can enhance lipid solubility allowing increased fatty acid absorption. Acetic acid is used in the synthesis of carbohydrates and in the metabolism of carbohydrates resulting in subsequent metabolic pathways." } ] } ] }, { "TOCHeading": "Record Title", "Description": "Text used as the title for this PubChem record.", "Information": [ { "ReferenceNumber": 55, "Name": "Record Title", "StringValue": "acetic acid" } ] }, { "TOCHeading": "Record Description", "Description": "Textual summary about this PubChem record.", "Information": [ { "ReferenceNumber": 2, "Name": "Record Description", "StringValue": "Acetic acid<br>Acetic acid is a product of the oxidation of ethanol and of the destructive distillation of wood. It is used locally, occasionally internally, as a counterirritant and also as a reagent. (Stedman, 26th ed) <a href='https://pubchem.ncbi.nlm.nih.gov/compound/Acetic%20acid'>Acetic acid</a> is an antibiotic that treats infections caused by bacteria or fungi." } ] } ] }, { "TOCHeading": "Names", "Description": "Synonyms (i.e., other names) for this chemical name.", "Information": [ { "ReferenceNumber": 35, "Name": "Record Description", "Description": "Acetic Acid<br>Acetic Acid is a synthetic carboxylic acid with antibacterial and antifungal properties. It can enhance lipid solubility allowing increased fatty acid absorption. Acetic acid is used in the synthesis of carbohydrates and in the metabolism of carbohydrates resulting in subsequent metabolic pathways." } ] } ] } ] }
```

Machine
accessible data



A screenshot of the PubChem Compound Summary for Acetic Acid (CID 176). The page includes a header with the PubChem logo and search bar, a navigation menu, and tabs for different data categories like Drug Information, Pharmacology, and Bioactivities. A red callout box highlights the text "Interactive user interfaces".

Acetic Acid

PubChem CID: 176

Chemical Names: Acetic acid; Ethanoic acid; Ethylic acid; Vinegar acid; 64-19-7; Glacial acetic acid; [More...](#)

Molecular Formula: C₂H₄O₂ or CH₃COOH

Molecular Weight: 60.05196 g/mol

InChI Key: QTBSBVTEAMEQO-UHFFFAOYSA-N

UNII: Q40Q9N063P

Safety Summary: Laboratory Chemical Safety Summary (LCSS)

Modify Date: 2016-10-15

Create Date: 2000-01-01

Product of the oxidation of ethanol and of the destructive distillation of wood. It is used locally, occasionally internally, as a counterirritant and also as a reagent. (Stedman, 26th ed)

Acetic Acid is a synthetic carboxylic acid with antibacterial and antifungal properties. It can enhance lipid solubility allowing increased fatty acid absorption. Acetic acid is used in the synthesis of carbohydrates and in the metabolism of carbohydrates resulting in subsequent metabolic pathways.

Acetic acid, glacial is a clear colorless liquid with a sharp, pungent odor. It is miscible with water and organic solvents. Used to make other chemicals, as a food additive, and in petroleum products. Used to make other chemicals, as a food additive, and in petroleum products.

2D Structure

Contents

- 1 2D Structure
- 2 3D Conformer
- 3 Names and Identifiers
- 4 Chemical and Physical Properties
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication Information
- 8 Food Additives and Ingredients
- 9 Agrochemical Information
- 10 Pharmacology and Biochemistry
- 11 Use and Manufacturing
- 12 Identification
- 13 Safety and Hazards

2D Structure

Chemical structure of Acetic Acid: CC(=O)O

Interactive user
interfaces

PubChem Programmatic Services

PUG REST



<http://pubchem.ncbi.nlm.nih.gov/rest/pug>
prolog

[/compound/name/viox](https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/viox)
input

[/property/InChI](https://pubchem.ncbi.nlm.nih.gov/rest/pug/property/InChI)
operation

[/TXT](https://pubchem.ncbi.nlm.nih.gov/rest/pug/property/InChI/output)
output

https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST_Tutorial.html

Scientific information has many CAVEATS!

- Chemical information is a bit of a mess and can be rather nuanced
 - Names, names, and more names (+300M in PubChem)
 - Some standard names are not open and cannot be used/verified without \$\$\$
 - Name/structure associations vary by use case (many overlapping)
 - Acetic acid vs. Acetic acid tri-hydrate
 - Formaldehyde: (gas) vs. Formalin (liquid , 40% formaldehyde w/ water)
 - Sulfuric acid: SO₃ (gas) vs. H₂SO₄ (liquid)
 - Glucose: L/D, ring open/closed (p vs. f), alpha/beta/both vs. monohydrate
 - Large corpus in the ‘wild’ .. data source dependent nuances
- One needs to verify with annotation source(s) as to what is meant
 - i.e., is this the form of the chemical I care about?

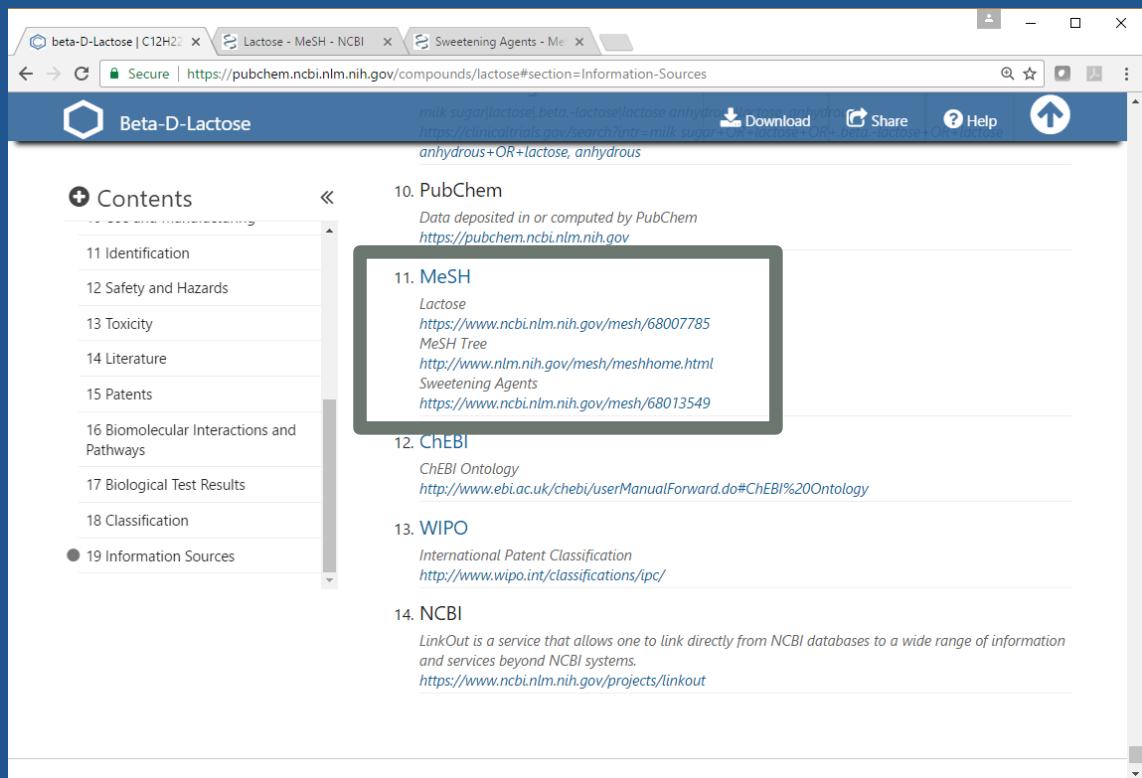


PubChem aggregates content by chemical structure

- Helps users to locate desired content
 - Other websites
 - Categorized annotation
 - Related chemicals (w/ annotation)
- Helps to bridge text world and chemical structure world
- More data means more errors
 - Separation of wheat from chaff important and never-ending battle
- Focus is on researchers (+educators)
- PubChem does not curate data
 - Authoritative and curated sources are very important to PubChem
 - Algorithm consistency checking
 - Not sure who is right
 - Detects errors in curated content
- Concepts change on context
 - Chemical names mean different things to different scientists
 - Can be whole classes of things
 - Much to do here...



Semantics and concepts



The screenshot shows a web browser window with the URL <https://pubchem.ncbi.nlm.nih.gov/compounds/lactose#section=Information-Sources>. The page displays a sidebar with a table of contents and a main content area with several sections. The 'MeSH' section is highlighted with a red box.

Contents

- 11 Identification
- 12 Safety and Hazards
- 13 Toxicity
- 14 Literature
- 15 Patents
- 16 Biomolecular Interactions and Pathways
- 17 Biological Test Results
- 18 Classification
- 19 Information Sources

10. PubChem
Data deposited in or computed by PubChem
<https://pubchem.ncbi.nlm.nih.gov>

11. MeSH
Lactose
<https://www.ncbi.nlm.nih.gov/mesh/68007785>
MeSH Tree
<http://www.nlm.nih.gov/mesh/meshhome.html>
Sweetening Agents
<https://www.ncbi.nlm.nih.gov/mesh/68013549>

12. ChEBI
ChEBI Ontology
<http://www.ebi.ac.uk/chebi/userManualForward.do#ChEBI%20Ontology>

13. WIPO
International Patent Classification
<http://www.wipo.int/classifications/ipc/>

14. NCBI
LinkOut is a service that allows one to link directly from NCBI databases to a wide range of information and services beyond NCBI systems.
<https://www.ncbi.nlm.nih.gov/projects/linkout>

- Why so many different links to MeSH?
 - Annotations and names are wild
 - Cause PubChem many headaches

Semantics and concepts

The screenshot shows the NCBI MeSH search interface for the term "Lactose". The search bar at the top contains "MeSH". Below the search bar, there are "Full", "Limits", and "Advanced" buttons. The main content area displays a detailed description of Lactose: "A disaccharide of GLUCOSE and GALACTOSE in human and cow milk. It is used in pharmacy for tablets, in medicine as a nutrient, and in industry." Under "Subheadings:", there are two columns of terms: administration and dosage, adverse effects, analogs and derivatives, analysis, antagonists and inhibitors, biosynthesis, blood, chemical synthesis, chemistry, deficiency; and economics, etiology, genetics, history, immunology, isolation and purification, metabolism, pharmacokinetics, pharmacology, physiology, radiation effects, secretion, standards, statistics and numerical data, supply and distribution, therapeutic use, toxicity, urine. At the bottom, there are checkboxes for "Restrict to MeSH Major Topic" and "Do not include MeSH terms found below this term in the MeSH hierarchy". Below these are "Tree Number(s)", "MeSH Unique ID", "Registry Number", "Entry Terms" (listing "Anhydrous Lactose" and "Lactose, Anhydrous"), and "Pharmacologic Action". A "PubMed Search Builder" sidebar on the right includes buttons for "Add to search builder" and "Search PubMed", and links to "Related information" like PubMed, PubMed - Major Topic, Clinical Queries, NLM MeSH Browser, MedGen, and PubChem Compound.

The screenshot shows the NCBI MeSH search interface for the term "Sweetening Agents". The search bar at the top contains "MeSH". Below the search bar, there are "Full", "Limits", and "Advanced" buttons. The main content area displays a detailed description of Sweetening Agents: "Substances that sweeten food, beverages, medications, etc., such as sugar, saccharine or other low-calorie synthetic products. (From Random House Unabridged Dictionary, 2d ed)" Under "Subheadings:", there are three columns of terms: administration and dosage, adverse effects, analysis, antagonists and inhibitors, blood, chemical synthesis, chemistry, classification, contraindications; and economics, history, isolation and purification, metabolism, pharmacokinetics, pharmacology, physiology, poisoning, radiation effects, standards, statistics and numerical data, supply and distribution, therapeutic use, toxicity, urine. There are also checkboxes for "Restrict to MeSH Major Topic" and "Do not include MeSH terms found below this term in the MeSH hierarchy". Below these are "Tree Number(s)", "MeSH Unique ID", "Entry Terms" (listing "Agent, Sweetening", "Agents, Sweetening", "Sweetening Agent", "Sweeteners", "Sweetener", and "Sugar Substitutes"), and "Recent Activity" (listing "Sweetening Agents", "Lactose", and "sugar (101)". A "PubMed Search Builder" sidebar on the right includes buttons for "Add to search builder" and "Search PubMed", and links to "Related information" like PubMed, PubMed - Major Topic, Clinical Queries, NLM MeSH Browser, and PubChem Compound.

Semantics and concepts

The screenshot shows the NCBI MeSH database interface. The search term "Lactose" is entered in the search bar. The main content area provides a brief definition: "A disaccharide of GLUCOSE and GALACTOSE in human and cow milk. It is used in pharmacy for tablets, in medicine as a nutrient, and in industry." Below the definition are "PubMed search builder options" and a list of "Subheadings". A large section of checkboxes lists various medical and scientific terms related to lactose. At the bottom, there are two checkboxes for "Restrict to MeSH Major Topic" and "Do not include MeSH terms found below this term in the MeSH hierarchy". Below these are "Tree Number(s)", "MeSH Unique ID", "Registry Number", "Entry Terms" (which includes "Anhydrous Lactose" and "Lactose, Anhydrous"), and "Pharmacologic Action". On the right side, there is a "PubMed Search Builder" panel with sections for "Related information" and "Recent Activity".

The screenshot shows the PubChem compound page for "Beta-D-Lactose". The URL is https://pubchem.ncbi.nlm.nih.gov/compounds/lactose#section=Synonyms. The page title is "Beta-D-Lactose". The main content is organized into sections: "4.4 Synonyms", "4.4.1 MeSH Synonyms", and "4.4.2 Depositor-Supplied Synonyms". The "4.4.1 MeSH Synonyms" section contains three entries: "1. Anhydrous Lactose", "2. Lactose", and "3. Lactose, Anhydrous". The "4.4.2 Depositor-Supplied Synonyms" section contains a list of 20 entries, each with a number and a name, such as "1. beta-D-Lactose" and "20. CHEBI:36218". A callout box highlights the "4.4.1 MeSH Synonyms" section.

Semantics and concepts

The screenshot shows a web browser window with the URL <https://pubchem.ncbi.nlm.nih.gov/compounds/lactose#section=EC-Number>. The page title is "Beta-D-Lactose". On the left, there is a sidebar with a "Contents" section containing numbered links from 1 to 15. The main content area displays two separate boxes under the heading "4.3.2 EC Number".
The first box contains the EC number 200-559-2, with the source noted as "from European Chemicals Agency - ECHA". It includes the record name "lactose", URL "https://echa.europa.eu/", and a description of what an EC Number is.
The second box contains the EC number 227-751-9, also with the source noted as "from European Chemicals Agency - ECHA". It includes the record name "4-O-β-D-galactopyranosyl-β-D-glucopyranose", URL "https://echa.europa.eu/", and a similar description.
Below these boxes, under the heading "4.3.3 UNII", is the UNII code 13Q3A43E05, with the source noted as "from FDA/SPL Indexing Data".

- Authoritative sources have different use cases for the same chemical placed into different concepts
- How can we divide up the annotation variation?
- How can we better represent appropriate content to our users?

Slowly moving towards Concepts-based approach

- Modifying MeSH (to make a PubChem MeSH) by:
 - Expanding names
 - Adding concepts
- Often many concepts per chemical structures [E.g., Benzene, Coal Naptha]
- Use concepts to map annotation
- Give concepts own page to separate out their annotations

The screenshot shows a web browser displaying the PubChem compound page for Benzene (CAS 71-43-2). The page has a blue header with the PubChem logo and the word "Benzene". Below the header is a sidebar with a list of categories: Contents, 1 2D Structure, 2 3D Conformer, 3 Names and Identifiers (which is expanded), 4 Chemical and Physical Properties, 5 Related Records, 6 Chemical Vendors, 7 Drug and Medication Information, 8 Food Additives and Ingredients, 9 Pharmacology and Biochemistry, 10 Use and Manufacturing, 11 Identification, 12 Safety and Hazards, 13 Toxicity, 14 Literature, and 15 Patents. The main content area shows three entries under the heading "3.3.1 CAS":

- 71-43-2: from ILO-ICSC, OSHA Occupational Chemical DB, EPA Chemicals under the TSCA, CA...
Source: OSHA Occupational Chemical DB
Record Name: NAPHTHA (COAL TAR)
URL: http://www.osha.gov/chemicaldata/chemResult.html?RecNo=706
Description: The OSHA Occupational Chemical Database contains over 800 entries with information such as physical properties, exposure guidelines, etc.
- 8030-30-6: from OSHA Occupational Chemical DB, European Chemicals Agency - ECHA, ChemID...
Source: European Chemicals Agency - ECHA
Record Name: Naphtha
URL: https://echa.europa.eu/
Description: The European Community number (EC Number) is a unique seven-digit identifier that was assigned to substances for regulatory purposes within the European Union by the European Commission.
- 26181-88-4: from ChemIDplus
Source: ChemIDplus
Record Name: Petroleum ether
URL: https://chem.nlm.nih.gov/chemidplus/sid/0008030306
Description: The European Community number (EC Number) is a unique seven-digit identifier that was assigned to substances for regulatory purposes within the European Union by the European Commission.

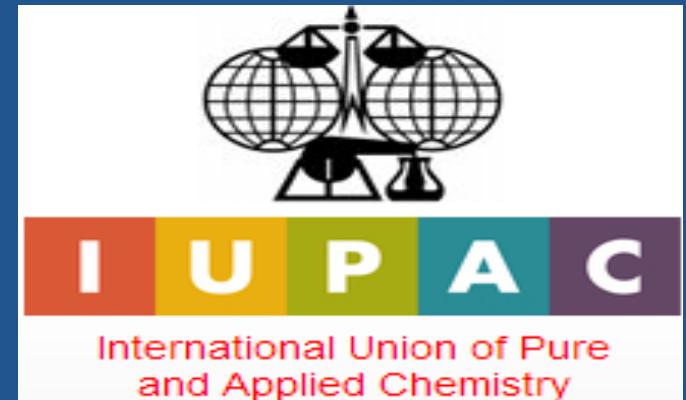
Semantics, terminologies, and standards

- PubChem is increasingly involved with standards
 - IUPAC (International Union of Pure and Applied Chemistry) InChI
 - Chemical structure guidelines, best practices, and normalization approaches
 - Terminology projects (IUPAC Gold Book)
- Computer understanding requires terminologies and semantic approaches
 - Toxicology information, health and safety, experimental procedures, ...
- To make progress in representing scientific information/knowledge, a fair bit of work will be necessary
 - “It takes a village” - a communal effort .. what can the ‘village’ do?



Chemical information standards gaining traction

- InChI ← an acronym
 - International Chemical Identifier
- Standard
 - Initially created by NIST
 - Under auspices of IUPAC
 - Open source, non-proprietary
- Algorithm
 - Normalizes chemical representation
 - Includes ‘hashed’ form called an InChIKey



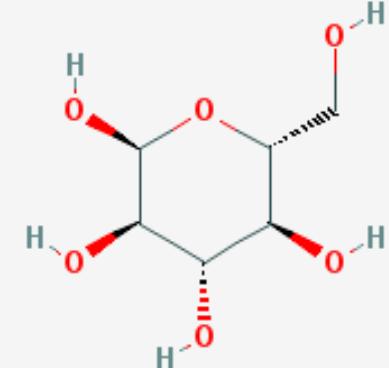
<https://iupac.org/inchi>

InChI is a string

Version Type
Chemical formula
Connectivity
Charge&Proton
Stereochemical
Other (e.g., Isotopic)

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

“layered” line notation



alpha-D-Glucose

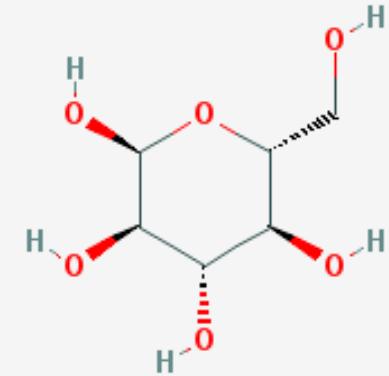
InChIKey is a “hashed” InChI

- Fixed length and search engine friendly InChI
- May allow for ‘secure’ lookup of a chemical

Chemical formula
& Connectivity
Stereochemical &
Hydrogens &
Other (e.g., Isotopic)
Type Version
Charge

WQZGKKKJIJFFOK-DVKNGEFBSA-N

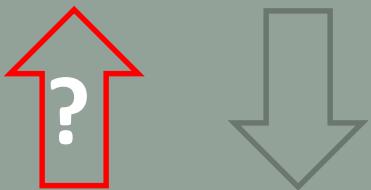
“layered” line notation



alpha-D-Glucose

InChIKey can be a ‘secret’

InChI=**1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1**



WQZGKKKIJFFOK-DVKNGEFBSA-N

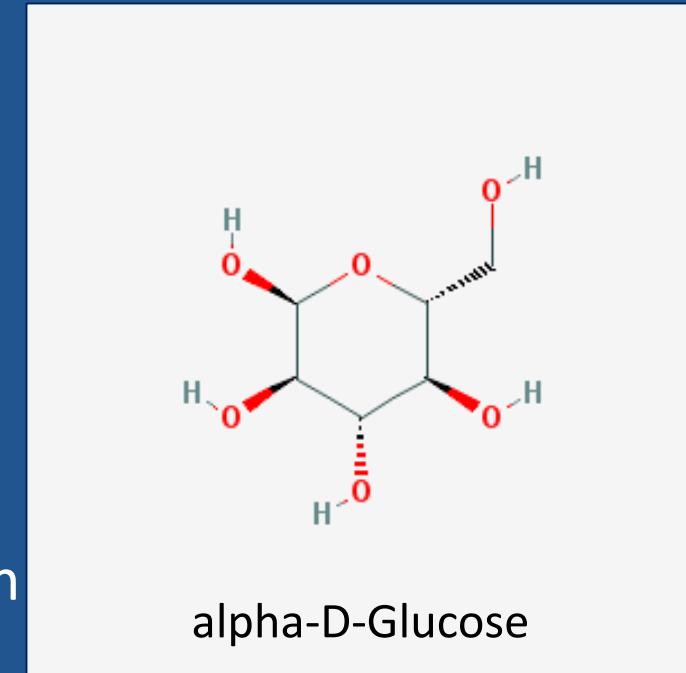
There is no chemical information in an InChIKey ... in that, if you do not know the InChI, you cannot convert the InChIKey back into a chemical structure



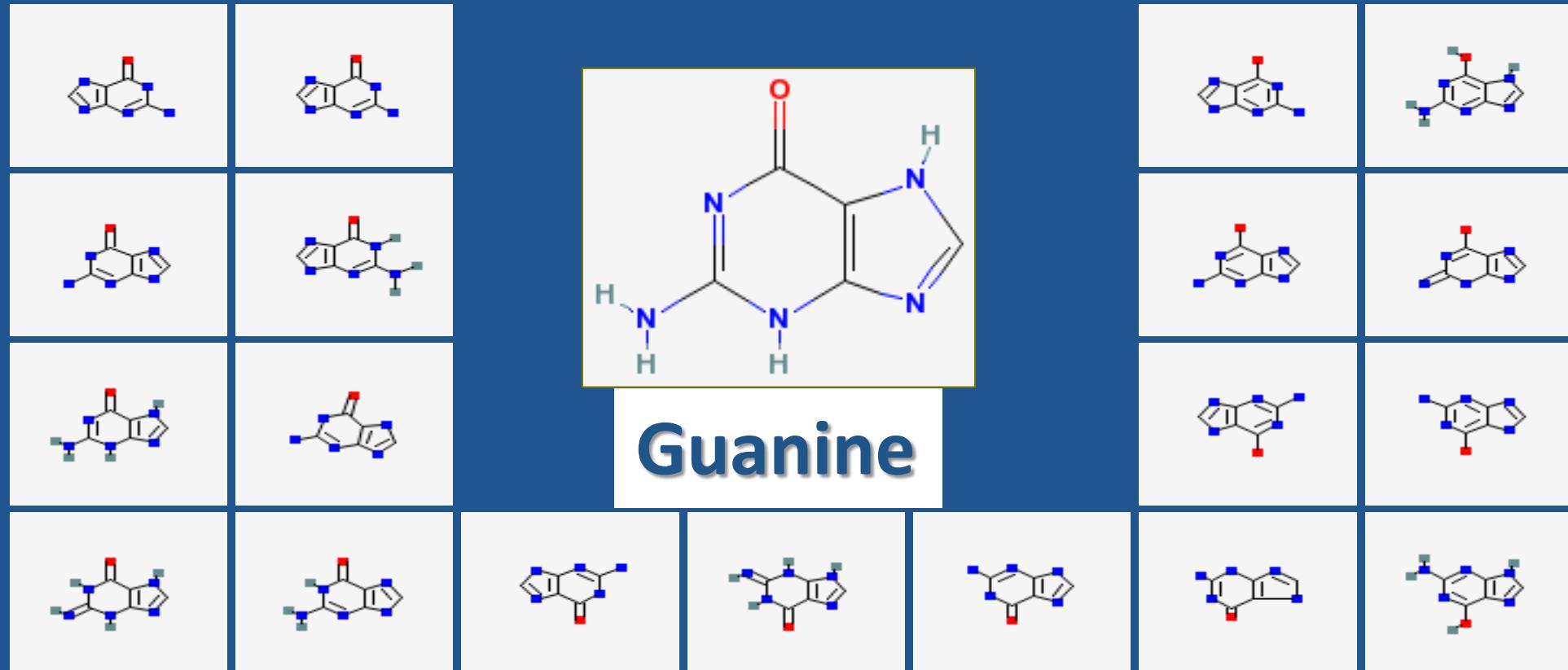
What about SMILES?

C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O)O)O)O)O

- SMILES is a popular “industry standard” line notation
 - But not a published standard
- Every vendor has its own implementation
 - Differences in aromaticity models can lead to structure corruption
- Cannot reliably compare strings
 - Different software packages can make different strings for same structure
- No structure normalization
 - Different structural representations can yield different strings



Different equivalent (tautomer) forms have different SMILES but same InChI



Map your structures to PubChem

- PubChem FTP site provides a complete tab-delimited file containing all ~90M chemicals
 - PubChem Compound CID, InChI, InChIKey
- Easy to get mapping between your structures and PubChem
 - Just compute InChI/Key and then do string comparison (or DB join)



InChI/Key

Your chemical collection

4.5GB file

<ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/CID-InChI-Key.gz>

How does PubChem use InChI/Key?

- Deposition structure format
- Computed descriptor of chemicals
- Text-based chemical structure lookup
- Structure export format
- Structure search input
- Standardization service input/output
- Programmatic service input/output
- Semantic linked data



PubChem is going though some ...

CH-CH-CH-CHANGES!

Image credit:
<http://www.fortemusiceducation.com/wp-content/uploads/2016/06/Ch-ch-ch-changes.png>

Technology revamp continues

The screenshot shows the PubChem Compound Summary for Benzoic Acid (CID 243). The page includes the following details:

- PubChem CID:** 243
- Chemical Names:** Benzoic acid; 65-85-0; Dracylic acid; Benzenecarboxylic acid; Carboxybenzene; Benzeneformic acid
- Molecular Formula:** C₇H₆O₂
- Molecular Weight:** 122.123 g/mol
- InChI Key:** WPYMKLBDIGXBTP-UHFFFAOYSA-N
- Drug Information:** Therapeutic Uses, FDA UNII
- Safety Summary:** Laboratory Chemical Safety Summary (LCSS)

Below the main summary, there are two sections of text:

Benzoic acid is a fungistatic compound that is widely used as a food preservative. It is conjugated to GLYCINE in the liver and excreted as hippuric acid.
from MeSH

Benzoic acid is a Nitrogen Binding Agent. The mechanism of action of benzoic acid is as an Ammonium Ion Binding Activity.

- Continue to refine, improve, add, modify, tweak, remove, expand ...

Technology revamp continues

The screenshot shows the PubChem compound page for Benzoic Acid (C₇H₆O₂). The main content area displays the 2D structure (a benzene ring with a carboxylic acid group) and the 3D conformer (a ball-and-stick model of the molecule). The left sidebar lists 19 major sections, including 1 2D Structure, 2 3D Conformer, 3 Names and Identifiers, 4 Chemical and Physical Properties, 5 Related Records, 6 Chemical Vendors, 7 Drug and Medication Information, 8 Food Additives and Ingredients, 9 Agrochemical Information, 10 Pharmacology and Biochemistry, 11 Use and Manufacturing, 12 Identification, 13 Safety and Hazards, 14 Toxicity, 15 Literature, 16 Patents, 17 Biomolecular Interactions and Pathways, 18 Biological Test Results, and 19 Classification.

- Continue to refine, improve, add, modify, tweak, remove, expand ...
- Adding new major sections as needed .. sometimes with fission/fusion events

Technology revamp continues

The screenshot shows a web browser window for the PubChem compound page of Benzoic Acid (C7H6O2). The left sidebar contains a navigation menu with items like 'Contents', '7 Drug and Medication Information', '8 Food Additives and Ingredients', etc., ending at '20 Information Sources'. The main content area displays the 'Information Sources' section, which lists 20 different databases and resources that provide information about benzoic acid. Each entry includes the name of the source, the chemical name, and a direct link.

Information Source	Description	Link
1. HSDB	BENZOIC ACID	http://toxnet.nlm.nih.gov/cgi-bin/sis/search/r?dbs+hsdb:@term+@m+@rel+65-85-0
2. DrugBank	Benzoinic Acid	http://www.drugbank.ca/drugs/DB03793 http://www.drugbank.ca/drugs/DB03793#targets http://www.drugbank.ca/drugs/DB03793#transporters
3. ILO-ICSC	BENZOIC ACID	http://www.ilo.org/dyn/icsc/showcard.display?p_card_id=0103
4. CAMEO Chemicals	Benzoinic acid	https://cameochemicals.noaa.gov/chemical/2585
5. Human Metabolome Database	Benzoinic acid	http://www.hmdb.ca/metabolites/HMDB01870
6. FDA Pharm Classes	BENZOIC ACID	http://www.accessdata.fda.gov/spl/data/e26cb06e-aa89-4a2c-bb24-e292ebe79230.xml FDA Pharmacological Classification http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162549.htm
7. EPA Chemicals under the TSCA	Benzoinic acid	http://www.epa.gov/chemical-data-reporting
8. European Chemicals Agency - ECHA		

- Continue to refine, improve, add, modify, tweak, remove, expand ...
- Adding new major sections as needed .. sometimes with fission/fusion events
- Information Sources section continues to expand with new data sources but also to enhance linking

Technology revamp continues

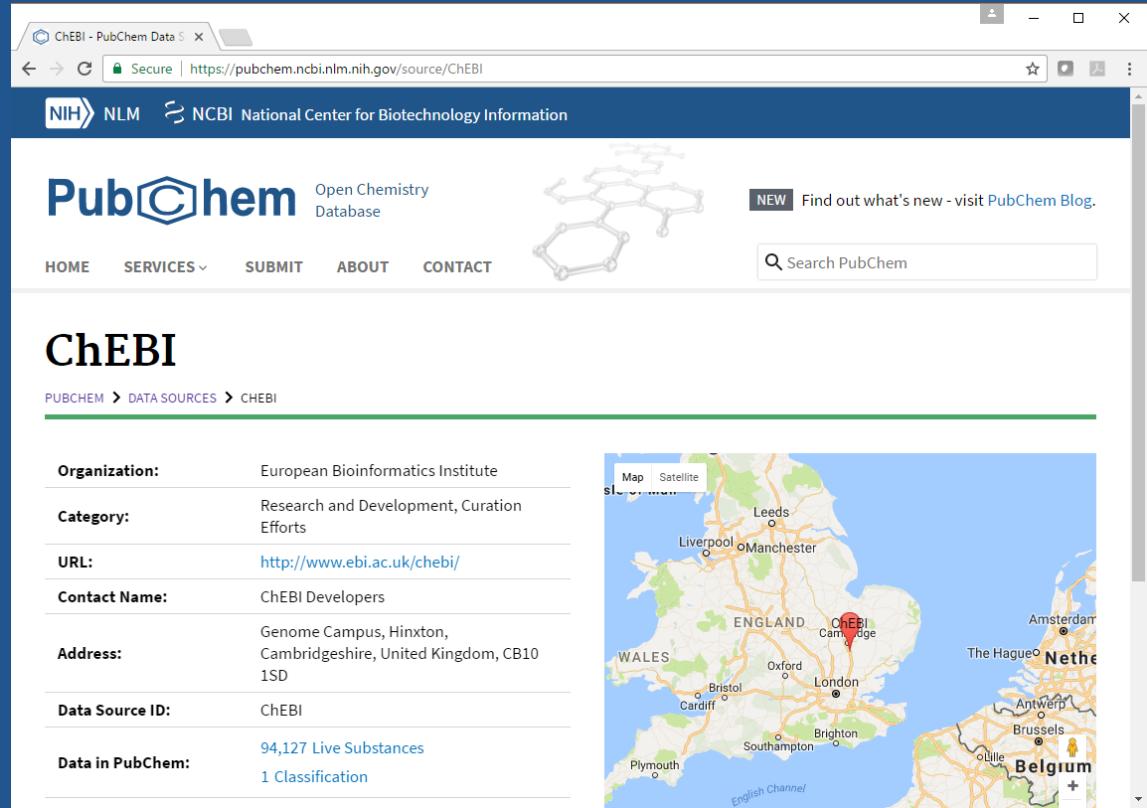
The screenshot shows the PubChem compound page for Benzoic Acid (C7H6O2). The left sidebar contains a table of contents with links to various sections such as Drug and Medication Information, Food Additives and Ingredients, Agrochemical Information, Pharmacology and Biochemistry, Use and Manufacturing, Identification, Safety and Hazards, Toxicity, Literature, Patents, Biomolecular Interactions and Pathways, Biological Test Results, Classification, and Information Sources. The main content area displays several information sources:

- 26. ChEBI
ChEBI Ontology
<http://www.ebi.ac.uk/chebi/userManualForward.do#ChEBI%20Ontology>
- 27. KEGG
Drug
http://www.genome.jp/dbget-bin/www_bget?brite:br08301
JP17
http://www.genome.jp/dbget-bin/www_bget?brite:br08311
Risk category of Japanese OTC drugs
http://www.genome.jp/dbget-bin/www_bget?brite:br08312
Animal drugs
http://www.genome.jp/dbget-bin/www_bget?brite:br08331
Additive
http://www.genome.jp/dbget-bin/www_bget?brite:br08316
Major components of natural products
http://www.genome.jp/dbget-bin/www_bget?brite:br08323
- 28. WIPO
International Patent Classification
<http://www.wipo.int/classifications/ipc/>
- 29. WHO ATC
ATC Code
http://www.whocc.no/atc_dd_index/
- 30. NCBI
LinkOut is a service that allows one to link directly from NCBI databases to a wide range of information and services beyond NCBI systems.
<https://www.ncbi.nlm.nih.gov/projects/linkout>

- Continue to refine, improve, add, modify, tweak, remove, expand ...
- Adding new major sections as needed .. sometimes with fission/fusion events
- Information Sources section continues to expand with new data sources but also to enhance linking
- PubChem highlights quite a number of information sources

Technology revamp continues

- New Data Source pages released in the last year

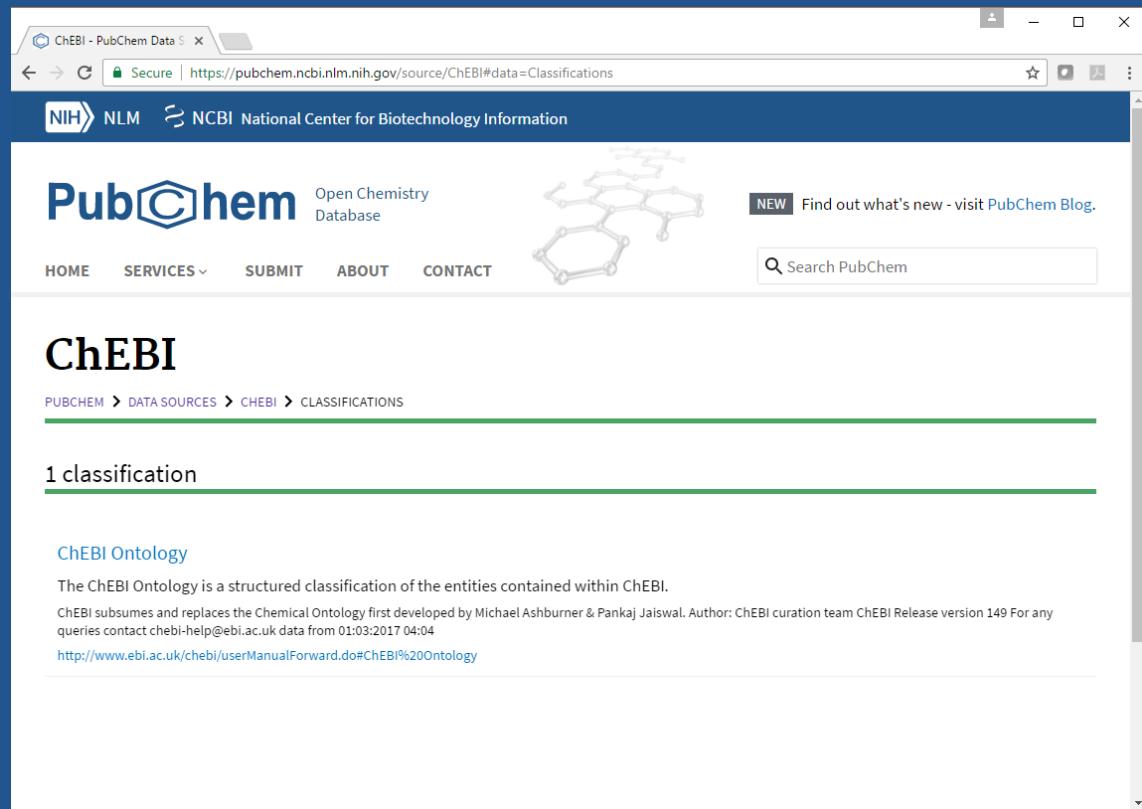


The screenshot shows a web browser window for the ChEBI Data Source page on the PubChem website. The URL in the address bar is <https://pubchem.ncbi.nlm.nih.gov/source/ChEBI>. The page header includes the NIH NLM and NCBI National Center for Biotechnology Information logos. The main content area features the PubChem logo and tagline "Open Chemistry Database". A search bar with the placeholder "Search PubChem" is visible. A banner at the top right says "NEW Find out what's new - visit [PubChem Blog](#)". Below the banner, there's a chemical structure diagram. The main section is titled "ChEBI" and displays the following information:

PUBCHEM > DATA SOURCES > CHEBI
Organization: European Bioinformatics Institute
Category: Research and Development, Curation Efforts
URL: http://www.ebi.ac.uk/chebi/
Contact Name: ChEBI Developers
Address: Genome Campus, Hinxton, Cambridgeshire, United Kingdom, CB10 1SD
Data Source ID: ChEBI
Data in PubChem: 94,127 Live Substances 1 Classification

On the right side of the page, there is a map of the United Kingdom and surrounding regions, with a red dot indicating the location of ChEBI in Cambridge, England.

Technology revamp continues

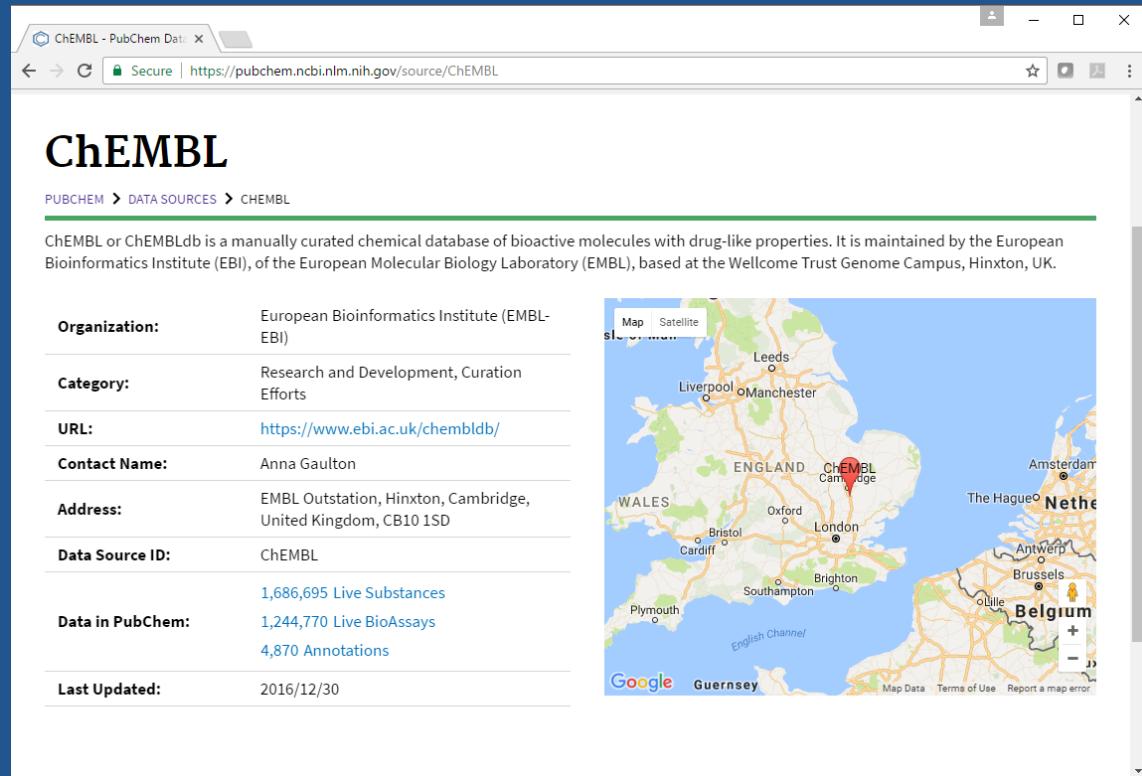


The screenshot shows a web browser window for 'ChEBI - PubChem Data S' at the URL <https://pubchem.ncbi.nlm.nih.gov/source/ChEBI#data=Classifications>. The page is part of the NIH NLM NCBI National Center for Biotechnology Information. The PubChem logo is visible, along with a search bar and a 'Find out what's new - visit PubChem Blog.' link. The main content area is titled 'ChEBI' and shows the path 'PUBCHEM > DATA SOURCES > CHEBI > CLASSIFICATIONS'. A green horizontal bar indicates '1 classification'. Below this, a section titled 'ChEBI Ontology' provides a detailed description of the ontology, mentioning it is a structured classification of entities within ChEBI, developed by Michael Ashburner & Pankaj Jaiswal, and links to the user manual for the ontology.

- New Data Source pages released in the last year
- Provides links to all annotation and classification content from a source



Technology revamp continues



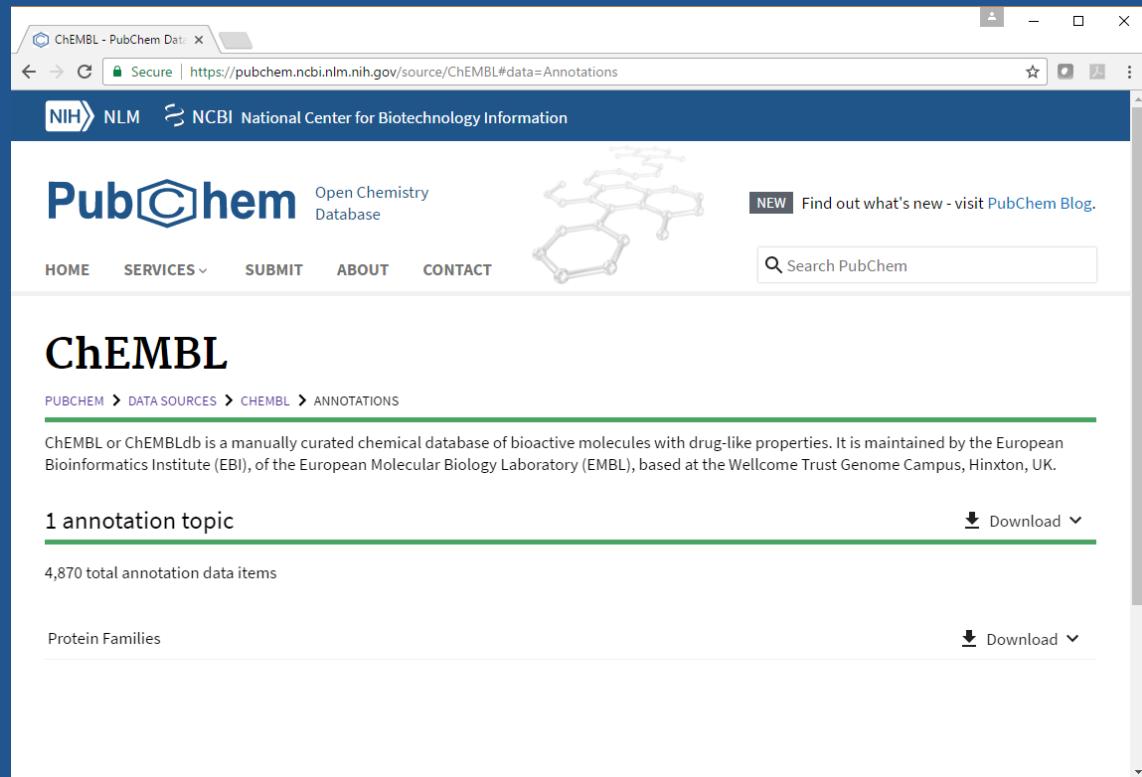
The screenshot shows a web browser window for the ChEMBL Data Source page on PubChem. The URL is https://pubchem.ncbi.nlm.nih.gov/source/ChEMBL. The page title is "ChEMBL". Below the title, it says "PUBCHEM > DATA SOURCES > CHEMBL". A brief description states: "ChEMBL or ChEMBLdb is a manually curated chemical database of bioactive molecules with drug-like properties. It is maintained by the European Bioinformatics Institute (EBI), of the European Molecular Biology Laboratory (EMBL), based at the Wellcome Trust Genome Campus, Hinxton, UK." On the left, there is a table with the following data:

Organization:	European Bioinformatics Institute (EMBL-EBI)
Category:	Research and Development, Curation Efforts
URL:	https://www.ebi.ac.uk/chembl/
Contact Name:	Anna Gaulton
Address:	EMBL Outstation, Hinxton, Cambridge, United Kingdom, CB10 1SD
Data Source ID:	ChEMBL
1,686,695 Live Substances	
Data in PubChem:	1,244,770 Live BioAssays 4,870 Annotations
Last Updated:	2016/12/30

On the right side of the page is a map of the United Kingdom and surrounding regions, with a red dot indicating the location of ChEMBL in Hinxton, Cambridge.

- New Data Source pages released in the last year
- Provides links to all annotation and classification content from a source

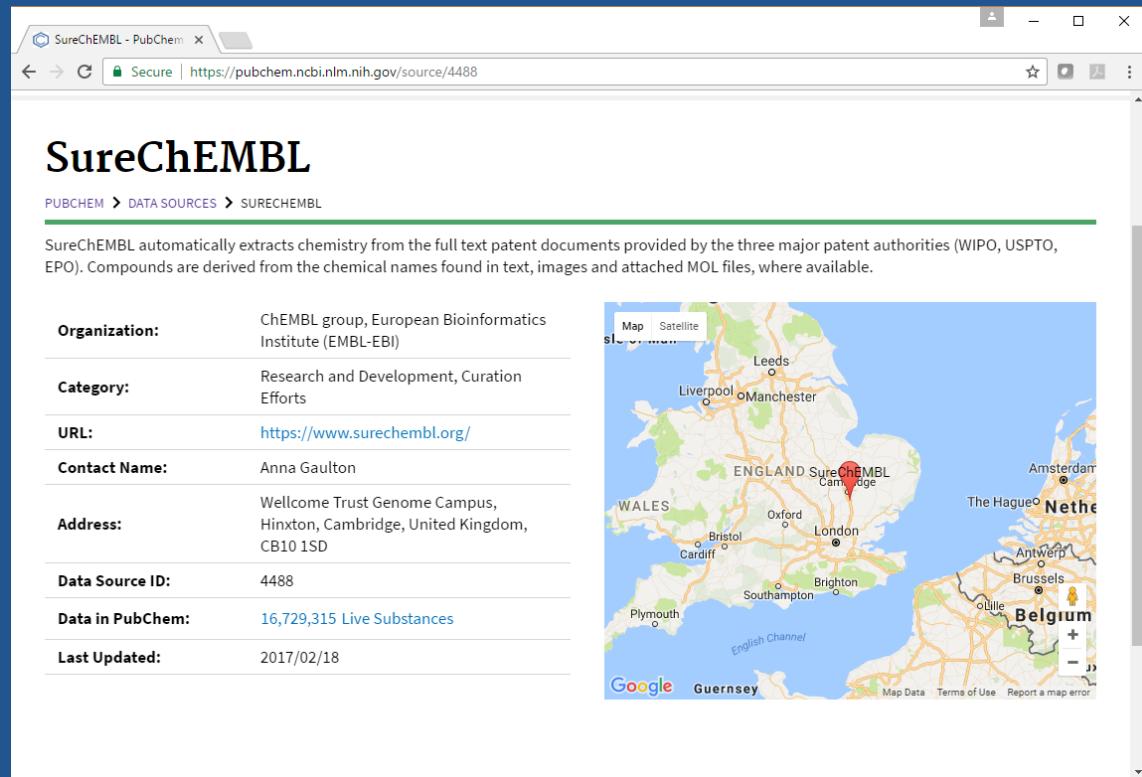
Technology revamp continues



The screenshot shows a web browser window for the PubChem ChEMBL Data Source page. The URL is https://pubchem.ncbi.nlm.nih.gov/source/ChEMBL#data=Annotations. The page header includes the NIH, NLM, and NCBI National Center for Biotechnology Information logos. The main navigation bar has links for HOME, SERVICES, SUBMIT, ABOUT, and CONTACT. A search bar is present. The main content area is titled "ChEMBL" and shows a breadcrumb trail: PUBCHEM > DATA SOURCES > CHEMBL > ANNOTATIONS. It states that ChEMBL or ChEMBLdb is a manually curated chemical database of bioactive molecules with drug-like properties, maintained by the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) at the Wellcome Trust Genome Campus, Hinxton, UK. Below this, there is a section for "1 annotation topic" with a "Download" button, and another section for "Protein Families" with a "Download" button. The total number of annotation data items is 4,870.

- New Data Source pages released in the last year
- Provides links to all annotation and classification content from a source

Technology revamp continues



The screenshot shows a web browser window with the title "SureChEMBL - PubChem". The URL in the address bar is <https://pubchem.ncbi.nlm.nih.gov/source/4488>. The page content includes:

SureChEMBL

PUBCHEM > DATA SOURCES > SURECHEMBL

SureChEMBL automatically extracts chemistry from the full text patent documents provided by the three major patent authorities (WIPO, USPTO, EPO). Compounds are derived from the chemical names found in text, images and attached MOL files, where available.

Organization:	ChEMBL group, European Bioinformatics Institute (EMBL-EBI)
Category:	Research and Development, Curation Efforts
URL:	https://www.surechembl.org/
Contact Name:	Anna Gaulton
Address:	Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, CB10 1SD
Data Source ID:	4488
Data in PubChem:	16,729,315 Live Substances
Last Updated:	2017/02/18

Below the table is a map of the United Kingdom and surrounding regions, with a red marker indicating the location of the "SureChEMBL Cambridge" office. The map also shows major cities like London, Birmingham, and Manchester, along with parts of Wales, Scotland, and the Netherlands.

- New Data Source pages released in the last year
- Provides links to all annotation and classification content from a source
- SureChEMBL is available too!

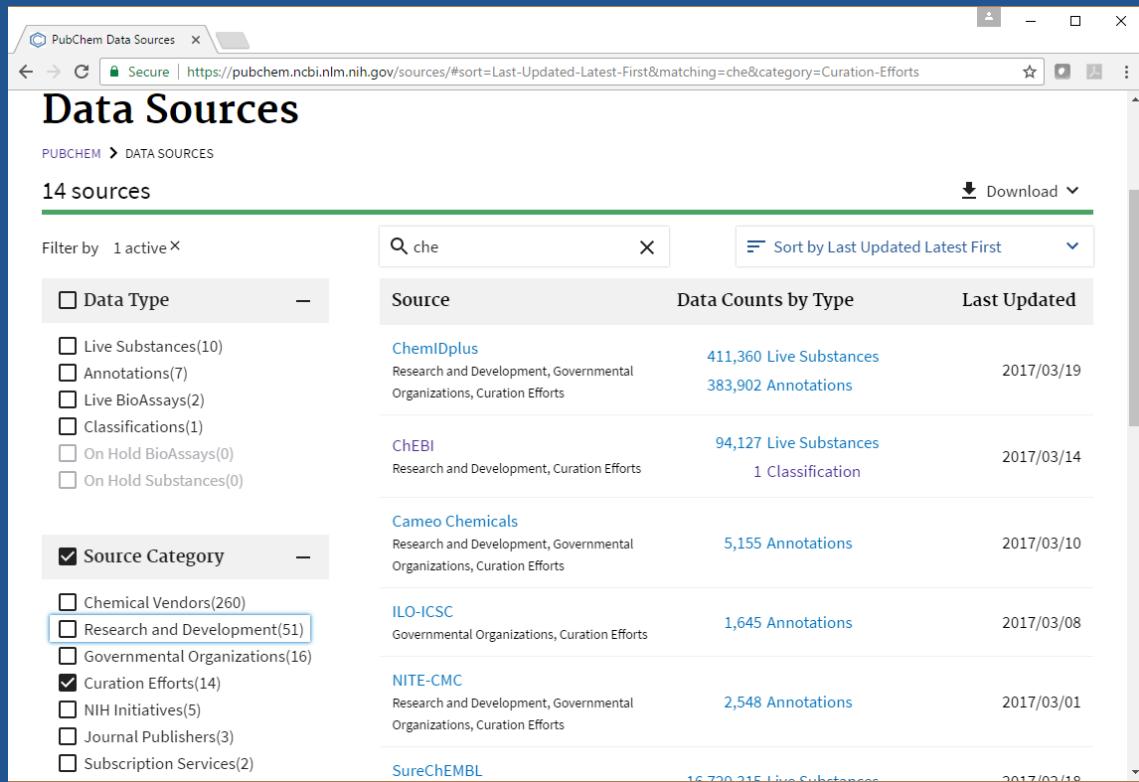
Technology revamp continues

- Data Sources page gives a summary of all data sources

The screenshot shows the 'Data Sources' page of the PubChem website. At the top, there's a navigation bar with links for NIH, NLM, NCBI, and the National Center for Biotechnology Information. Below that is the PubChem logo and a search bar. The main content area has a title 'Data Sources' and a breadcrumb trail 'PUBCHEM > DATA SOURCES'. It displays '507 sources' and includes filters for 'Data Type' (Live Substances, Live BioAssays, Annotations, Classifications, On Hold BioAssays, On Hold Substances) and sorting options ('Search Sources', 'Sort by Last Updated Latest First'). A download button is also present. The data table lists three sources: FDA Pharm Classes, Acorn PharmaTech Product List, and DailyMed, along with their respective counts and last update dates.

Source	Data Counts by Type	Last Updated
FDA Pharm Classes Research and Development, Governmental Organizations, Curation Efforts	2,557 Annotations 1 Classification	2017/03/22
Acorn PharmaTech Product List Chemical Vendors	33,475 Live Substances	2017/03/22
DailyMed Governmental Organizations	1,730 Annotations	2017/03/22

Technology revamp continues



The screenshot shows a web browser window titled "PubChem Data Sources". The URL is https://pubchem.ncbi.nlm.nih.gov/sources/#sort=Last-Updated-Latest-First&matching=che&category=Curation-Efforts. The page title is "Data Sources" under "PUBCHEM > DATA SOURCES". There are 14 sources listed. A search bar at the top contains "che". A dropdown menu says "Sort by Last Updated Latest First". On the left, there are two filter sections: "Data Type" and "Source Category". The "Source Category" section has checkboxes for various categories, with "Research and Development(51)" checked and highlighted with a blue border. The table lists the following data sources:

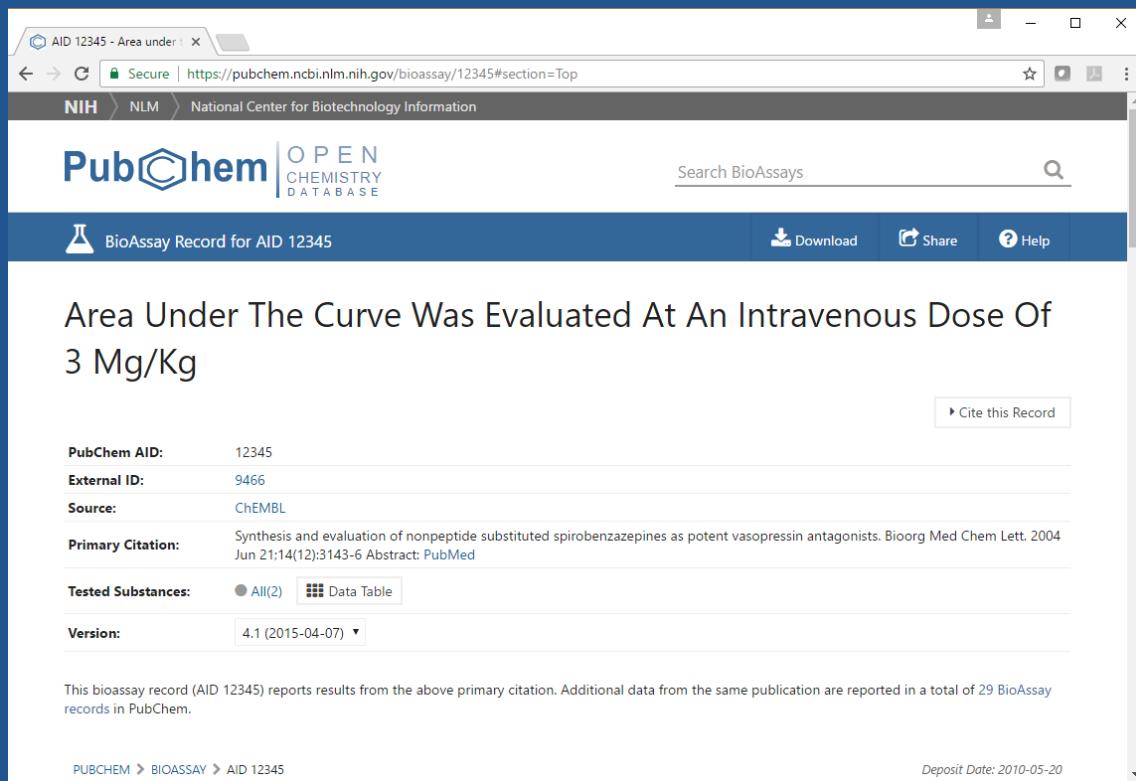
Source	Data Counts by Type	Last Updated
ChemDplus Research and Development, Governmental Organizations, Curation Efforts	411,360 Live Substances 383,902 Annotations	2017/03/19
ChEBI Research and Development, Curation Efforts	94,127 Live Substances 1 Classification	2017/03/14
Cameo Chemicals Research and Development, Governmental Organizations, Curation Efforts	5,155 Annotations	2017/03/10
ILO-ICSC Governmental Organizations, Curation Efforts	1,645 Annotations	2017/03/08
NITE-CMC Research and Development, Governmental Organizations, Curation Efforts	2,548 Annotations	2017/03/01
SureChEMBL	16,720,215 Live Substances	2017/02/18

- Data Sources page gives a summary of all data sources
- One can use faceting to subset by various means
- Uses same/similar approach as Compound Summary page (JSON data delivery followed by rendering in browser .. display is data driven)

Technology revamp continues

- Substance Record page uses same/similar approach as Compound Summary page

Technology revamp continues



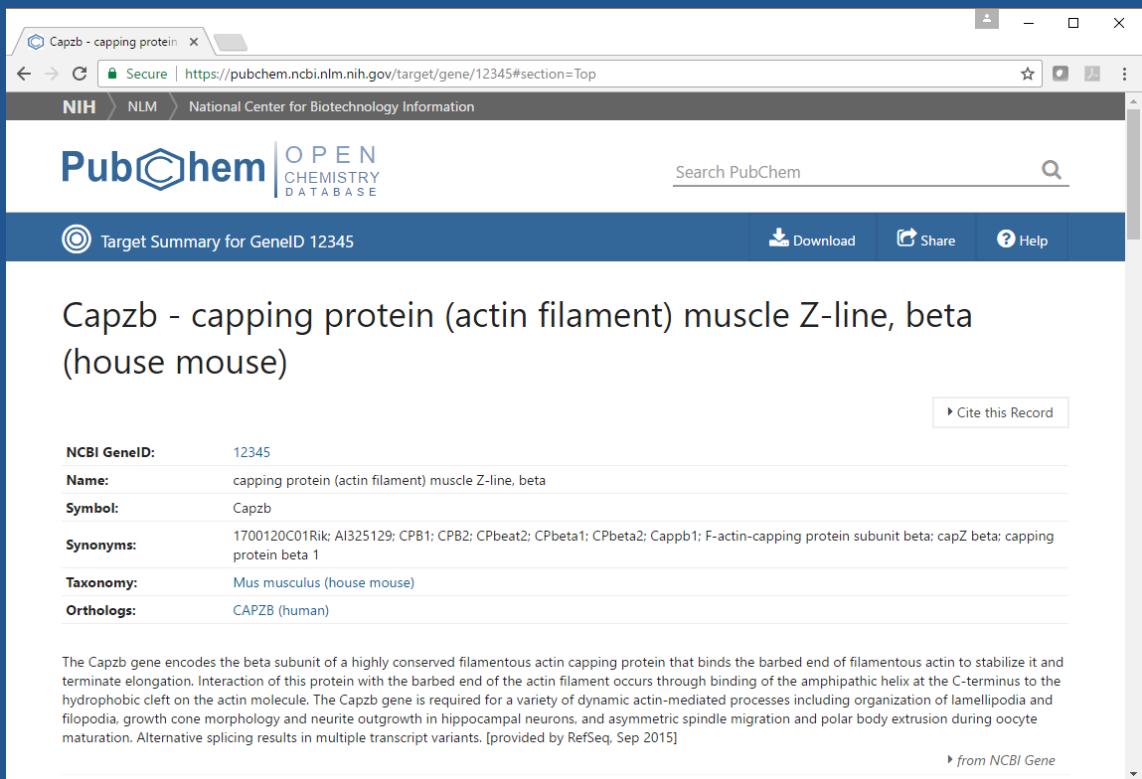
The screenshot shows a web browser displaying the PubChem BioAssay Record for AID 12345. The page title is "AID 12345 - Area under t". The URL is https://pubchem.ncbi.nlm.nih.gov/bioassay/12345#section=Top. The header includes the NIH and NLM logos, and the National Center for Biotechnology Information. The main content area features the PubChem logo and a search bar. Below the search bar, a blue banner displays the text "BioAssay Record for AID 12345" along with download, share, and help buttons. The main content area has a white background and displays the following information:

- Area Under The Curve Was Evaluated At An Intravenous Dose Of 3 Mg/Kg**
- PubChem AID:** 12345
- External ID:** 9466
- Source:** ChEMBL
- Primary Citation:** Synthesis and evaluation of nonpeptide substituted spirobenzazepines as potent vasopressin antagonists. *Bioorg Med Chem Lett.* 2004 Jun 21;14(12):3143-6 Abstract: PubMed
- Tested Substances:** All(2) Data Table
- Version:** 4.1 (2015-04-07)

A note at the bottom states: "This bioassay record (AID 12345) reports results from the above primary citation. Additional data from the same publication are reported in a total of 29 BioAssay records in PubChem." The footer includes navigation links like PUBCHEM > BIOASSAY > AID 12345 and a deposit date of 2010-05-20.

- BioAssay Record page uses same/similar approach as Compound Summary page

Technology revamp continues



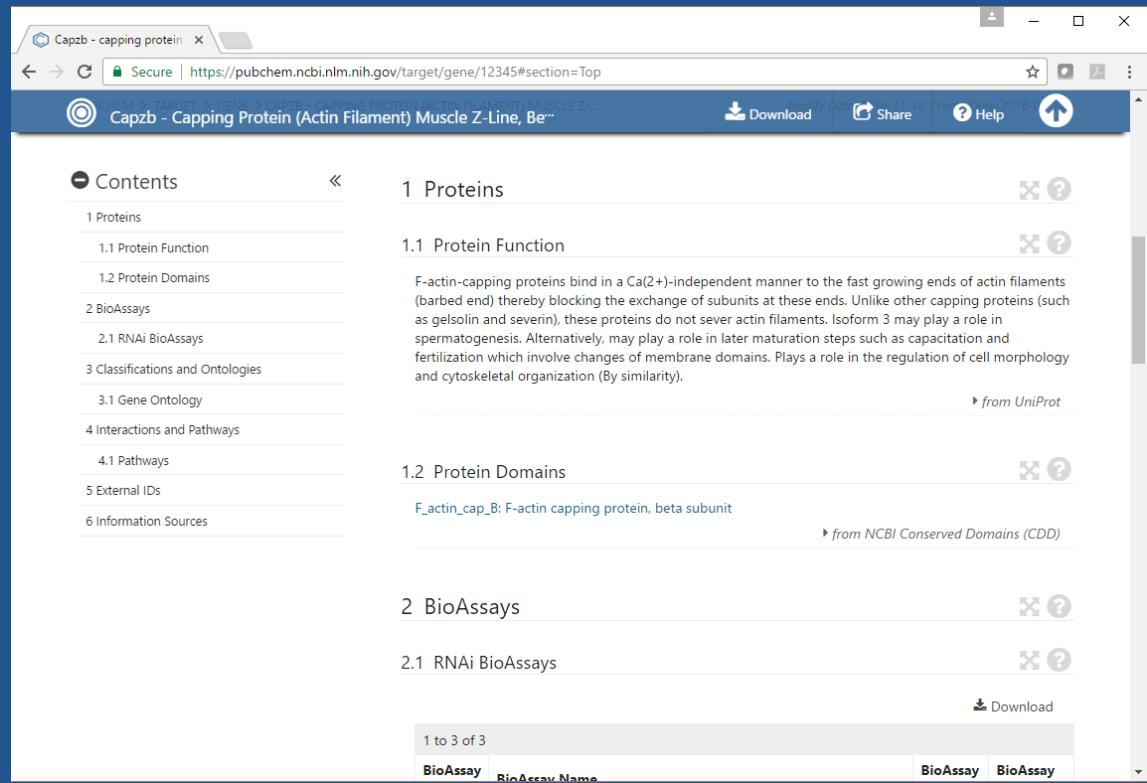
The screenshot shows a web browser window displaying the PubChem Target Summary page for the gene Capzb. The URL in the address bar is https://pubchem.ncbi.nlm.nih.gov/target/gene/12345#section=Top. The page header includes the NIH and NLM logos, followed by "National Center for Biotechnology Information". The main navigation bar features the PubChem logo, an "OPEN CHEMISTRY DATABASE" link, a search bar, and download/share/help buttons. Below the header, a blue banner displays "Target Summary for GenID 12345". The main content area contains the following information:

NCBI GenID:	12345
Name:	capping protein (actin filament) muscle Z-line, beta
Symbol:	Capzb
Synonyms:	1700120C01Rik; AI325129; CPB1; CPB2; CPbeat2; CPbeta1; CPbeta2; Cappb1; F-actin-capping protein subunit beta; capZ beta; capping protein beta 1
Taxonomy:	Mus musculus (house mouse)
Orthologs:	CAPZB (human)

A detailed description of the Capzb gene follows, mentioning its function as a highly conserved filamentous actin capping protein and its role in various biological processes like lamellipodia and filopodia formation.

- Gene Target Summary page uses same/similar approach as Compound Summary page

Technology revamp continues



The screenshot shows the PubChem Compound Summary page for Capzb - capping protein. The URL is https://pubchem.ncbi.nlm.nih.gov/compound/12345#section=Top. The page has a sidebar with links for Contents, Proteins, BioAssays, Classifications and Ontologies, Gene Ontology, Interactions and Pathways, Pathways, External IDs, and Information Sources. The main content area is titled '1 Proteins' and includes sections for '1.1 Protein Function' and '1.2 Protein Domains'. The '1.1 Protein Function' section contains a detailed description of the protein's function in actin filament capping. The '1.2 Protein Domains' section lists 'F_actin_cap_B: F-actin capping protein, beta subunit' with a link to NCBI Conserved Domains (CDD). Below this is a section for '2 BioAssays' under '2.1 RNAi BioAssays'. At the bottom, there is a table header for 'BioAssay' and 'BioAssay Name'.

- Gene Target Summary page uses same/similar approach as Compound Summary page
- Recently announced
- Summarizes chemical and bioactivity content using a Gene Target concept for grouping purposes
- Protein Target page to be released
- Working towards other grouping paradigms

Technology revamp continues

The screenshot shows a web browser window displaying the PubChem compound details page for CID 124220753. The URL in the address bar is <https://pubchem.ncbi.nlm.nih.gov/compound/124220753#section=Biologic-Description>. The page title is "CID 124220753". The main content area is titled "2 Biologic Description". Under this, there is a section "2.1 Biologic Depiction" which contains the SMILES string: H-Val - Ile - Gly - Ala - Lys - Lys-H. Below this is "2.2 Biologic Line Notation" with the IUPAC Condensed string: H-Val-Ile-Gly-Ala-Lys-Lys-al. Further down are sections for "Sequence" (VIGAKK) and "HELM" (PEPTIDE1[V.I.G.A.K.*N[@H](CCCN)C=O \${_R1::::::::\$}]\$\$\$\$). Each section has a "from PubChem" link.

- While delayed by more than a year (for various reasons), progress continues in earnest on Biologics
- Currently a separate section, with naming and descriptor strings
- Annotating amino-acid and saccharides that can be completely described (0.5M out of 10M)

Towards full support of “Biologics”

- Coverage of amino acids, nucleic acids, saccharides, and lipids
 - Standards when mixing or modifying .. naming and line notations need cleanup
 - Lipids absent .. but LipidMaps/ChEBI already did a bit of work here
 - Better line notation approaches needed when chemical modifications are made
 - Many, many ‘biopolymer’ containing structures are in PubChem (10M?)
- Going beyond 1000 atoms/bonds [\leftarrow current structure limit in PubChem]
 - Mastering what is in PubChem first
 - Can we: define what is a biologic? define PTMs? handle small molecule → biologic?
 - Working towards a super-atom approach .. [e.g., Ala = one super atom]
 - Helping lay the ground work with standards ...



Symbol Nomenclature for (Graphical Representations of) Glycans (SNFG)

Table 1. [Monosaccharide symbol nomenclature]. - Essentials of Glycobiology - NCBI Bookshelf - Google Chrome

Secure | https://www.ncbi.nlm.nih.gov/books/NBK310273/table/symbolnomenclature.T.monosaccharide_symb/?report=objectonly

Table 1.
Monosaccharide symbol nomenclature

SHAPE	White (Generic)	Blue	Green	Yellow	Orange	Pink	Purple	Light Blue	Brown	Red
Filled Circle	○ Hexose	● Glc	● Man	● Gal	● Gul	● Alt	● All	● Tal	● Ido	
Filled Square	□ HexNAc	■ GlcNAc	■ ManNAc	■ GalNAc	■ GuNAc	■ AltNAc	■ AllNAc	■ TaINAc	■ IdoNAc	
Crossed Square	■ Hexosamine	■ GlcN	■ ManN	■ GalN	■ GulN	■ AltN	■ AllN	■ TaIN	■ IdON	
Divided Diamond	◇ Hexuronate	◇ GlcA	◇ ManA	◇ GalA	◇ GulA	◇ AltA	◇ AllA	◇ TalA	◇ IdoA	
Filled Triangle	△ Deoxyhexose	▲ Qui	▲ Rha			▲ 6dAlt		▲ 6dTal		▲ Fuc
Divided Triangle	△ DeoxyhexNAc	▲ QuiNAc	▲ RhaNAc							▲ FucNAc
Flat Rectangle	□ Di-deoxyhexose	■ Oli	■ Tyv		■ Abe	■ Par	■ Dig	■ Col		
Filled Star	☆ Pentose		★ Ara	★ Lyx	★ Xyl	★ Rib				
Filled Diamond	◇ Nonulosonate		◆ Kdn			◆ Neu5Ac	◆ Neu5Gc	◆ Neu	◆ Sia	
Flat Hexagon	○ Unknown	● Bac	● LDManHep	● Kdo	● Dha	● DDManHep	● MurNAc	● MurNGc	● Mur	
Pentagon	◇ Assigned	● Api	● Fru	● Tag	● Sor	● Psi				

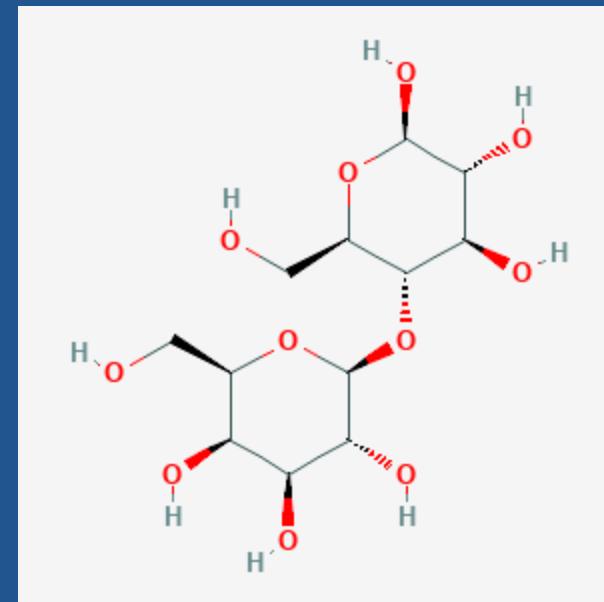
- Glycans use symbolic approach
- How to combine these with small molecule changes?
- How to combine these with other biopolymers? (glycoproteins, glycolipids, etc.)
- Helping to organize the glycobiology community (@ Biohackathon)
- PubChem helped launch the Glycans website at NCBI

<https://www.ncbi.nlm.nih.gov/glycans/>

Lactose

The screenshot shows the PubChem page for Beta-D-Lactose (C₁₂H₂₂O₁₁). The left sidebar contains a table of contents with items 1 through 18. The main content area is titled '3 Biologic Description' and '3.1 Biologic Depiction'. It displays a 2D structure diagram where a yellow circle labeled 'Gal' is connected to a blue circle labeled 'Glc' at the 4-position. Below the diagram is an IUPAC condensed notation: Gal(β1-4)β-Glc. The LINUS section is visible at the bottom.

- Symbolic representation in action .. makes for pretty pictures that are way easier to interpret .. if you know what you are looking at ...



PubChem Data Sources

<https://pubchem.ncbi.nlm.nih.gov/sources/>

NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

Find out what's new - visit PubChem Blog.

Search PubChem

Data Sources

PUBCHEM > DATA SOURCES

522 sources

Download

Source	Data Counts by Type	Last Updated
WHO ATC Governmental Organizations	3,941 Annotations 1 Classification	2017/06/06
ChEBI Research and Development, Curation Efforts	95,010 Live Substances 1 Classification	2017/06/06
A1 BioChem Labs Chemical Vendors	4,473 Live Substances	2017/06/06
European Chemicals Agency - ECHA Governmental Organizations	261,683 Annotations	2017/06/06
ChEMBL Research and Development, Curation Efforts	1,735,442 Live Substances 1,244,770 Live BioAssays 5,031 Annotations	2017/06/06
Acorn PharmaTech Product List Chemical Vendors	48,280 Live Substances	2017/06/06

Filter by

- Data Type —
 - Live Substances(469)
 - Live BioAssays(86)
 - Annotations(57)
 - On Hold Substances(8)
 - Classifications(8)
 - On Hold BioAssays(6)
- Source Category —
 - Chemical Vendors(266)
 - Research and Development(160)
 - Governmental Organizations(62)
 - Curation Efforts(44)
 - NIH Initiatives(23)
 - Subscription Services(8)
 - Journal Publishers(8)
 - siRNA Reagent Vendors(4)

Search Sources

Sort by Last Updated Latest First

NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

Find out what's new - visit PubChem Blog.

Search PubChem

Data Sources

PUBCHEM > DATA SOURCES

57 sources

Download

Source	Data Counts by Type	Last Updated
WHO ATC Governmental Organizations	3,941 Annotations 1 Classification	2017/06/06
KEGG Research and Development, Curation Efforts	39,388 Live Substances 2,705 Annotations 55 Classifications	2017/06/06
FDA Pharm Classes Research and Development, Governmental Organizations, Curation Efforts	2,572 Annotations 1 Classification	2017/06/06
ChEMBL Research and Development, Curation Efforts	1,735,442 Live Substances 1,244,770 Live BioAssays 5,031 Annotations	2017/06/06
European Chemicals Agency - ECHA Governmental Organizations	261,683 Annotations	2017/06/06

Filter by 1 active X

- Data Type —
 - Live Substances(469)
 - Live BioAssays(86)
 - Annotations(57)
 - On Hold Substances(8)
 - Classifications(8)
 - On Hold BioAssays(6)
- Source Category —
 - Governmental Organizations(44)
 - Curation Efforts(24)
 - Research and Development(21)
 - NIH Initiatives(1)
 - Chemical Vendors(0)
 - Subscription Services(0)
 - siRNA Reagent Vendors(0)
 - Journal Publishers(0)

Search Sources

Sort by Last Updated Latest First

PubChem Data Sources – what data comes from whom?

NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

NEW Find out what's new - visit PubChem Blog.

Search PubChem

European Chemicals Agency – ECHA

PUBCHEM > DATA SOURCES > EUROPEAN CHEMICALS AG...

Organization:	European Chemicals Agency
Category:	Governmental Organizations
URL:	https://echa.europa.eu/
Contact Name:	ECHA Staff
Address:	Annankatu 18, P.O. Box 400, Helsinki, Finland, FI-00121
Data Source ID:	11946
Data in PubChem:	261,683 Annotations
Last Updated:	2017/06/06



NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

NEW Find out what's new - visit PubChem Blog.

Search PubChem

European Chemicals Agency – ECHA

PUBCHEM > DATA SOURCES > EUROPEAN CHEMICALS AG... > ANNOTATIONS

3 annotation topics

[Download](#)

261,683 total annotation data items

CAS [Download](#)

EC Number [Download](#)

GHS Classification [Download](#)

<https://pubchem.ncbi.nlm.nih.gov/source/ECHA>

PubChem Data Sources – data downloadable

NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

Find out what's new - visit PubChem Blog.

Search PubChem

European Chemicals Agency - ECHA

PUBCHEM > DATA SOURCES > EUROPEAN CHEMICALS AGENCY > ANNOTATIONS

3 annotation topics

261,683 total annotation data items

CAS

EC Number

GHS Classification

Download

Download

Download

JSON Save Display

XML Save Display

ASNT Save Display

```
{
  "Annotations": [
    "Annotation": [
      {
        "SourceName": "EU REGULATION (EC) No 1272/2008",
        "SourceID": "001-001-00-9",
        "Name": "hydrogen",
        "Description": "Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures. Ref: OJ L 353, 31.12.2008, p. 1-1355 Special edition in Croatian: Chapter 13 Volume 020 P. 3-1357",
        "URL": "http://ec.europa.eu/growth/sectors/chemicals/classification-labelling/index_en.htm",
        "LinkToPubChemBy": {
          "CID": [
            783
          ]
        },
        "Data": [
          {
            "TOCHeading": "GHS Classification",
            "Name": "GHS Classification",
            "Value": [
              "String": [
                "div class=\"pc-thumbnail-container\"<img src=\"/images/ghs/GHS02.svg\" title=\"GHS02: Flammables\"></div><b>Signal: </b> <span class=\"fred\">Danger</span><br><div class=\"ghs-hazards\"><b>H220: Extremely flammable gas </b><span class=\"fred\">Danger</span> Flammable gases - Category 1<br></div><br><div class=\"ghs-precautionary\"><b>Precautionary Statement Codes</b><br>P210, P377, P381, and P403<br>(The corresponding statement to each P-code can be found <a href=\"https://pubchem.ncbi.nlm.nih.gov/ghs/#_prec\">here</a>.)</div>"
              ]
            ]
          }
        ],
        "SourceName": "EU REGULATION (EC) No 1272/2008",
        "SourceID": "001-002-00-4",
        "Name": "aluminium lithium hydride",
        "Description": "Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures. Ref: OJ L 353, 31.12.2008, p. 1-1355 Special edition in Croatian: Chapter 13 Volume 020 P. 3-1357",
        "URL": "http://ec.europa.eu/growth/sectors/chemicals/classification-labelling/index_en.htm",
        "LinkToPubChemBy": {
          "CID": [
            21226445
          ]
        },
        "Data": [
          {
            "TOCHeading": "GHS Classification",
            "Name": "GHS Classification",
            "Value": [
              "String": [
                "div class=\"pc-thumbnail-container\"<img src=\"/images/ghs/GHS02.svg\" title=\"GHS02: Flammables\"><img src=\"/images/ghs/GHS05.svg\" title=\"GHS05: Corrosives\"></div><b>Signal: </b> <span class=\"fred\">Danger</span><br><div class=\"ghs-hazards\"><b>H260: In contact with a class \"pubchem-internal-link CID-962\"</b><span class=\"fred\">Danger</span> Substances And Mixtures Which, In Contact With <a class=\"pubchem-internal-link CID-962\" href=\"https://pubchem.ncbi.nlm.nih.gov/compound/water\">water</a> releases flammable gases which may ignite spontaneously [<span class=\"fred\">Danger</span> Substances And Mixtures Which, In Contact With <a class=\"pubchem-internal-link CID-962\" href=\"https://pubchem.ncbi.nlm.nih.gov/compound/Water\">Water</a>, Emit Flammable Gases - Category 1] <br>H314: Causes severe skin burns and eye damage <br>H315: Causes serious eye damage <br>H317: Causes damage to organs (Skin) <br>H318: Causes damage to organs (Eye) <br>H319: Causes serious damage to the environment"
              ]
            ]
          }
        ]
      }
    ]
  ]
}
```

Classification Browser enables finding annotations

PubChem Classification E X

Secure | https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72

NCBI

PubChem Classification Browser

Help

Browse PubChem data using a classification of interest, or search for PubChem records annotated with the desired classification/term (e.g., MeSH: phenylpropionates, or Gene Ontology: DNA repair). More...

Select classification Search selected classification by

PubChem: PubChem Compound TOC ▾ Keyword ▾ Enter desired search term Search

Classification description (from PubChem)

This classification was created automatically from the PubChem Compound TOC on 2017/03/20.

Note that in some cases a number of highly populated nodes - those for which all or nearly all IDs have information - have been left out of the tree.

The sections, along with their child subsections, that are not shown in this tree are: Computed Properties, Substances by Category, Computed Descriptors, Molecular Formula, Depositor-Supplied Synonyms, Removed Synonyms, Create Date, Modify Date, Record Title, Related Compounds, Related Compounds with Annotation, Related Substances, 2D Structure, 3D Conformer, and Chemical Vendors. More...

Data type counts to display Display zero count nodes?

None Compound Yes No

Browse PubChem: PubChem Compound TOC Tree

- ▼ PubChem Compound TOC ? 28,792,472
 - ▶ Agrochemical Information ? 1,945
 - ▶ Biologic Description ? 538,830
 - Biologic Depiction ? 518,804
 - Biologic Line Notation ? 536,941
 - ▶ Biological Test Results ? 2,400,077
 - ▶ Biomolecular Interactions and Pathways ? 48,475
 - ▶ Chemical and Physical Properties ? 391,336

Give me all “Biologics” found in PubChem

Classification Trees are a form of hierarchical annotation

The screenshot shows the PubChem Classification Browser interface for the MeSH classification. At the top, there are dropdown menus for "Select classification" (set to "MeSH") and "Search selected classification by" (set to "Keyword"). Below these are search fields for "Enter desired search term" and a "Search" button. A note explains that MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. Under "Data type counts to display", the "Compound" tab is selected. The main area is titled "Browse MeSH Tree" and displays a hierarchical tree structure. The root node is "MeSH Tree" with a count of 120,881. It branches into categories like "Chemicals and Drugs Category" (120,881), "Amino Acids, Peptides, and Proteins" (12,849), "Biological Factors" (3,806), "Biomedical and Dental Materials" (583), "Carbohydrates" (9,417), "Chemical Actions and Uses" (15,510), "Complex Mixtures" (326), "Enzymes and Coenzymes" (1,034), "Heterocyclic Compounds" (49,138), and "Inorganic Chemicals" (4,241). Each node has a question mark icon for details.

The screenshot shows the PubChem Classification Browser interface for the FDA Pharm Classes classification. The layout is identical to the MeSH version, with "FDA Pharm Classes" selected in the "Select classification" dropdown and "Keyword" selected in the "Search selected classification by" dropdown. The main area is titled "Browse FDA Pharm Classes Tree" and displays a hierarchical tree structure. The root node is "FDA Pharmacological Classification" with a count of 1,226. It branches into "Chemical Structure [Chemical/Ingredient]" (411), "Established Pharmacologic Classes [EPC]" (1,204), "Mechanism of Action [MoA]" (730), and "Physiological Effects [PE]" (231). "Physiological Effects [PE]" further branches into "Generalized Systemic Effects [PE]" (78), "Organ System Specific Effects [PE]" (211), "Cardiovascular Activity Alteration [PE]" (20), "Dermatologic Activity Alteration [PE]" (8), "Digestive/GI System Activity Alteration [PE]" (12), "Endocrine Activity Alteration [PE]" (2), and "Hemic/Immunologic Activity Alteration [PI]" (75). Each node has a question mark icon for details.

Classification Browser – PubChem Compound TOC

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72>

NCBI

PubChem Classification Browser

Browse PubChem data using a classification of interest, or search for PubChem records annotated with the desired classification/term (e.g., MeSH: phenylpropionates, or Gene Ontology: DNA repair). [More...](#)

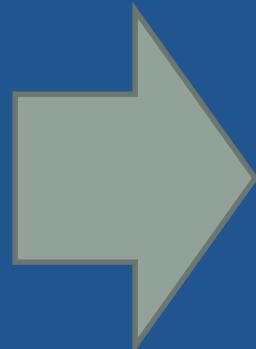
Select classification Search selected classification by
PubChem: PubChem Compound TOC Keyword Enter desired search term Search

Classification description (from PubChem)
This classification was created automatically from the PubChem Compound TOC on 2017/05/22.
Note that in some cases a number of highly populated nodes - those for which all or nearly all IDs have information - have been left out of the tree.
The sections, along with their child subsections, that are not shown in this tree are: Computed Properties, Substances by Category, Computed Descriptors, Molecular Formula, Depositor-Supplied Synonyms, Removed Synonyms, Create Date, Modify Date, Record Title, Related Compounds, Related Compounds with Annotation, Related Substances, 2D Structure, 3D Conformer, and Chemical Vendors. [More...](#)

Data type counts to display Display zero count nodes?
None Compound Yes No

Browse PubChem: PubChem Compound TOC Tree

- ▶ PubChem Compound TOC [?] 29,048,135
 - ▶ Agrochemical Information [?] 1,943
 - ▶ Biologic Description [?] 472,404
 - ▶ Biological Test Results [?] 2,428,866
 - ▶ Biomolecular Interactions and Pathways [?] 48,977
 - ▶ Chemical and Physical Properties [?] 399,735
 - ▶ Classification [?] 16,454,686
 - ▶ Drug and Medication Information [?] 12,065
 - ▶ Food Additives and Ingredients [?] 3,213
 - ▶ Identification [?] 5,663
 - ▶ Information Sources [?] 16,963,480
- ▶ Names and Identifiers [?] 482,242
- ▶ Patents [?] 19,993,556
- ▶ Pharmacology and Biochemistry [?] 44,427
- ▶ Related Records [?] 5,477,966
- ▶ Safety and Hazards [?] 105,220
- ▶ Toxicity [?] 10,657
- ▶ Use and Manufacturing [?] 10,005
- ▶ 3D Status [?] 4,973,572
- ▶ LCSS [?] 102,728



Information Sources	
?	16,963,480
Burnham Center for Chemical Genomics	57,854
CAMEO Chemicals	4,757
CDC-ATSDR Toxic Substances Portal	237
ChEBI	87,593
ChemIDplus	324,770
ClinicalTrials.gov	7,802
DailyMed	2,919
DOT Emergency Response Guidebook	986
DrugBank	8,692
EPA Air Toxics	169
EPA Chemicals under the TSCA	5,774
EPA Office of Pesticide Programs	1,229
EU Pesticides Database	1,187
EU REGULATION (EC) No 1272/2008	2,986
European Chemicals Agency - ECHA	167,159

And many more....

Classifications include the Compound Table of Contents (TOC)

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72>

PubChem Compound TOC ? 29,263,861

- ▶ Agrochemical Information ? 1,949
- ▶ Biologic Description ? 475,667
- ▶ Biological Test Results ? 2,424,328
- ▶ Biomolecular Interactions and Pathways ? 49,922
- ▶ Chemical and Physical Properties ? 453,484
- ▶ Classification ? 16,839,286
- ▶ Drug and Medication Information ? 12,316
- ▶ Food Additives and Ingredients ? 4,087
- ▶ Identification ? 5,665
- ▶ Information Sources ? 17,675,749
- ▶ Literature ? 350,742
- ▶ Names and Identifiers ? 1,195,508
- ▶ Patents ? 20,323,664
- ▶ Pharmacology and Biochemistry ? 40,160
- ▶ Related Records ? 5,114,012
- ▶ Safety and Hazards ? 102,128
- ▶ Toxicity ? 10,701
- ▶ Use and Manufacturing ? 13,815
- 3D Status ? 4,964,429
- LCSS ? 99,531

Classification ? 16,839,286

- ▶ Ontologies ? 16,839,286
 - ChEBI Ontology ? 88,271
 - ChemIDplus ? 315,693
 - EPA Safer Choice ? 812
 - FDA Pharm Classes ? 1,226
 - KEGG: Additive ? 368
 - KEGG: Animal Drugs ? 280
 - KEGG: Antiinfectives ? 953
 - KEGG: ATC ? 4,209
 - KEGG: Carcinogen ? 749
 - KEGG: Crude Drug ? 2
 - KEGG: CYP ? 1,011
 - KEGG: Drug ? 1,475
 - KEGG: EDC ? 89
 - KEGG: JP15 ? 916
 - KEGG: Lipid ? 1,903

- KEGG: Major components of natural products ? 234
- KEGG: Metabolite ? 521
- KEGG: Natural Toxins ? 281
- KEGG: OTC drugs ? 306
- KEGG: Peptide ? 22
- KEGG: Pesticides ? 917
- KEGG: Phytochemical Compounds ? 2,845
- KEGG: Phytochemicals Used as Drugs ? 296
- KEGG: Risk Category of Japanese OTC Drugs ? 619
- KEGG: Target-based Classification of Compounds ? 312
- KEGG: Target-based Classification of Drugs ? 3,247
- KEGG: USP ? 1,597
- LIPID MAPS Classification ? 36,095
- MeSH Tree ? 120,860
- WHO ATC Classification System ? 13,809
- WIPO IPC ? 16,513,480

Many annotation information sources

Information Sources ? 17,675,749

Burnham Center for Chemical Genomics	57,854
CAMEO Chemicals	4,746
CDC-ATSDR Toxic Substances Portal	180
ChEBI	88,271
ChemIDplus	360,516
ClinicalTrials.gov	7,854
DailyMed	3,036
DOT Emergency Response Guidebook	992
DrugBank	8,704
DTP/NCI	260,248
EPA Air Toxics	170
EPA Chemicals under the TSCA	5,775
EPA DSStox	681,956
EPA Office of Pesticide Programs	1,229
EPA Safer Choice	812
EU Food Improvement Agents	3,254
EU Pesticides Database	1,188

EU REGULATION (EC) No 1272/2008	2,820
European Chemicals Agency - ECHA	163,363
FAO/WHO Food Additive Evaluations - JECFA	2,504
FDA Medication Guides	601
FDA Orange Book	4,467
FDA Pharm Classes	1,227
FDA/SPL Indexing Data	61,895
Flavor & Extract Manufacturers Association - FEMA	2,781
HSDB	8,256
Human Metabolome Database (HMDB)	20,705
ILO-ICSC	1,720
KEGG	11,876
LIPID MAPS	36,095
LiverTox	1,591
MeSH	120,931
NCBI	11,685
NCI Investigational Drugs	212
NCIt	5,907

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72>

NIOSH Manual of Analytical Methods	572
NIST	262,518
NITE-CMC	2,875
NJDOH RTK Hazardous Substance List	2,289
OSHA Chemical Sampling Information	1,479
OSHA Occupational Chemical DB	842
PubMed Health	2,737
Safe Work Australia - HCIS	3,399
SpectraBase	84,376
The Cambridge Structural Database	60,727
The National Institute for Occupational Safety and Health - NIOSH	2,258
UN Globally Harmonized System of Classification and Labelling of Chemicals (GHS)	101,562
USDA Pesticide Data Program	530
USGS Columbia Environmental Research Center	380
WHO ATC	13,812
Wikipedia	168,048
WIPO	16,513,480

Many chemicals, many properties (Overall +500 types of information)

Experimental Properties		
Acid Value	?	160,573
Auto-Ignition	?	990
Boiling Point	?	3,803
Caco ₂ Permeability	?	120
Color	?	6,807
Corrosivity	?	766
Decomposition	?	4,217
Density	?	4,219
Dissociation Constants	?	2,136
Flash Point	?	2,138
Heat of Combustion	?	572
Heat of Vaporization	?	719
Hydrophobicity	?	21
Ionization Potential	?	323
Isoelectric Point	?	17
Kovats Retention Index	?	79,656
LogP	?	8,040

LogS	?	245
Melting Point	?	11,676
Odor	?	3,060
Odor Threshold	?	468
Optical Rotation	?	184
pH	?	1,062
Physical Description	?	19,848
pKa	?	547
Polymerization	?	184
Refractive Index	?	11
Relative Evaporation Rate	?	120
Solubility	?	68,276
Stability	?	3,540
Surface Tension	?	584
Taste	?	1,064
Vapor Density	?	1,366
Vapor Pressure	?	4,877
Viscosity	?	741

Literature		
Depositor Provided PubMed Citations	?	350,742
General References	?	260,864
Metabolite References	?	2,220
NLM Curated PubMed Citations	?	12,430
Synthesis References	?	116,087
Spectral Properties		
C13-NMR	?	39,636
GC-MS	?	234,509
H1-NMR	?	46,946
Infrared Spectra	?	65,238
Mass	?	953
MS-MS	?	16,370
Raman	?	10,831
UV	?	14,032
1D NMR	?	1,632
2D NMR	?	1,024
EI-MS	?	1,007
GC	?	2
HPLC	?	202

Much more to do... and lots in the pipeline

- Dramatic quality improvements
 - Classification of chemical names (e.g., to hide/remove ‘bad’ names)
 - Cross-validation of billions/trillions of links (e.g., PubTator, LeadMine)
- Construct chemical concept map (e.g., to capture all chemical concepts in use)
- Improve integration with patents (e.g., USPTO, EPO, WIPO, JPO, KPO, CPO)
- Reinvent PubChem search (e.g., handle natural language questions)
- Consider more chemical biology use cases (e.g., spectra, reactions, expand annotations)
- And many, many more things we should be doing... especially with text



PubChem Crew ...

Evan Bolton

Jie Chen

Tiejun Cheng

Gang Fu

Asta Gindulyte

Jane He

Siqian He

Sunghwan Kim

Ben Shoemaker

Paul Thiessen

Bo Yu

Leonid Zaslavsky

Jian Zhang

Special thanks to the NCBI Help Desk, especially Rana Morris, and past PubChem group members.

Special thanks

- PubChemRDF Collaborators
- Chemical Health and Safety collaborators
 - Especially: Leah McEwen (Cornell U.), Ralph Stuart (Keene State College)
- Software collaborators
 - NextMove Software (Roger Sayle, Noel O'Boyle, John May) and Daniel Lowe
 - Xemistry GmbH (Wolf D. Ihlenfeldt)
 - OpenEye Scientific Software
- All PubChem Contributors and Collaborators
- PubChem is supported by the Intramural Research Program of the NIH, National Library of Medicine.

Have any
questions?

