# PubChem Database

**PubChem** is a database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). PubChem can be accessed for free through a web user interface. Millions of compound structures and descriptive datasets can be freely downloaded via FTP. PubChem contains multiple substance descriptions and small molecules with fewer than 100 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database.

## Databases

PubChem consists of three dynamically growing primary databases. As of 5 November 2020

- Compounds, 111 million entries (up from 94 million entries in 2017), contains pure and characterized chemical compounds.
- Substances, 293 million entries (up from 236 in 2017 and 163 million entries in Sept 2014), contains also mixtures, extracts, complexes and uncharacterized substances.
- BioAssay, bioactivity results from 1.25 million (up from 6000 in Sept 2014) high-throughput screening programs with several million values.

## Searching

Searching the databases is possible for a broad range of properties including chemical structure, name fragments, chemical formula, molecular weight, XLogP, and hydrogen bond donor and acceptor count. PubChem contains its own online molecule editor with SMILES/SMARTS and InChI support that allows the import and export of all common chemical file formats to search for structures and fragments. Each hit provides information about synonyms, chemical properties, chemical structure including SMILES and InChI strings, bioactivity, and links to structurally related compounds and other NCBI databases like PubMed. In the text search form the database fields can be searched by adding the field name in square brackets to the search term. A numeric range is represented by two numbers separated by a colon. The search terms and field names are case-insensitive. Parentheses and

the logical operators AND, OR, and NOT can be used. AND is assumed if no operator is used.

Example (Lipinski's Rule of Five):

0:500[mw] 0:5[hbdc] 0:10[hbac] -5:5[logp]

**ACS's concerns**

The American Chemical Society has raised concerns about the publicly supported PubChem database, since it appears to directly compete with their existing Chemical Abstracts Service. They have a strong interest in the issue since the Chemical Abstracts Service generates a large percentage of the society's revenue. To advocate their position against the PubChem database, ACS has actively lobbied the US Congress. Soon after PubChem's creation, the American Chemical Society lobbied U.S. Congress to restrict the operation of PubChem, which they asserted competes with their Chemical Abstracts Service.

**Database fields**

**Identification numbers**

| | |
|---|---|
| • Identification number in current database | [UID] |
| • Substance identification number | [SID] |
| • Compound identification number | [CID] |
| • BioAssay identification number | [BAID], [AID] |

**General**

| | |
|---|---|
| • Any database field | [ALL] |
| • Comment | [CMT] |
| • Deposition date | [DDAT], [DEPDAT] |
| • Depositor's external ID | [SRID], [SRCID] |

| | |
|---|---|
| • Source name | [SRC], [SRCNAM], [SRCNAME] |
| • Source release date | [SRD], [SRDAT], [RLSDAT] |
| • Medical Subject Heading (MeSH) term | [MSHT], [MESHT] |
| • MeSH tree node | [MSHN], [MESHTN] |
| • MeSH pharmacological actions | [PHMA], [PHARMA] |

**Substance properties**

| | |
|---|---|
| • Substance synonyms | [SYNO] |
| • IUPAC name | [UPAC], [IUPAC] |
| • International Chemical Identifier (InChI) | [INCHI] |
| • Molecular weight | [MW], [MWT], [MOLWT] |
| • Chemical elements | [ELMT], [EL] |
| • Non-Hydrogen atoms | [HAC], [HACNT] |
| • Isotope count | [IAC], [IACNT] |
| • Total formal charge | [TFC], [CHG], [CHRG] |
| • Chiral atom count | [ACC], [ACCNT] |
| • Defined chiral atom count | [ACDC], [ACDCNT] |
| • Undefined chiral atom count | [ACUC], [ACUCNT] |
| • Hydrogen bond acceptor count | [HBAC], [HBACNT] |
| • Hydrogen bond donor count | [HBDC], [HBDCNT] |
| • Tautomer count | [TC], [TCNT], [TTMC] |
| • Rotatable bond count | [RBC], [RBCNT] |

| | |
|---|---|
| • XLogP | [XLGP], [LOGP] |

**Compound properties**

| | |
|---|---|
| • Compound synonyms | [CSYN], [CSYNO] |
| • Component count | [CC], [CCNT] |
| • Covalent unit (molecule) count | [CUC], [CUCNT] |
| • Total bioactivity count | [TAC] |

Comparative Toxicogenomics Database

The **Comparative Toxicogenomics Database** (**CTD**) is a public website and research tool launched in November 2004 that curates scientific data describing relationships between chemicals/drugs, genes/proteins, diseases, taxa, phenotypes, GO annotations, pathways, and interaction modules. The database is maintained by the Department of Biological Sciences at North Carolina State University.

**Goals and objectives**

One of the primary goals of CTD is to advance the understanding of the effects of environmental chemicals on human health on the genetic level, a field called toxicogenomics.

The etiology of many chronic diseases involves interactions between environmental factors and genes that modulate important physiological processes. Chemicals are an important component of the environment. Conditions such as asthma, cancer, diabetes, hypertension, immunodeficiency, and Parkinson's disease are known to be influenced by the environment; however, the molecular mechanisms underlying these correlations are not well understood. CTD may help resolve these mechanisms. The most up-to-date extensive list of peer-reviewed scientific articles about CTD is available at their publications page.

**Core data**

CTD is a unique resource where biocurators read the scientific literature and manually curate four types of core data:

- Chemical-gene interactions
- Chemical-disease associations
- Gene-disease associations
- Chemical-phenotype associations

**Data integration**

By integrating the above four data sets, CTD automatically constructs putative chemical-gene-phenotype-disease networks to illuminate molecular mechanisms underlying environmentally-influenced diseases.

These inferred relationships are statistically scored and ranked and can be used by scientists and computational biologists to generate and verify testable hypotheses about toxicogenomic mechanisms and how they relate to human health.

Users can search CTD to explore scientific data for chemicals, genes, diseases, or interactions between any of these three concepts. Currently, CTD integrates toxicogenomic data for vertebrates and invertebrates.

CTD integrates data from or hyperlinks to these databases:

- ChemIDplus, a dictionary of more than 400,000 chemicals housed in the US National Library of Medicine
- DrugBank
- Data Infrastructure for Chemical Safety project (diXa) Data Warehouse by the European Bioinformatics Institute which as of November 2015 contained 469 compounds, 188 disease datasets in three sub-categories liver, kidney and cardiovascular disease.
- Gene Ontology Consortium
- KEGG
- NCBI Entrez-Gene
- NCBI PubMed
- NCBI Taxonomy

- NLM Medical Subject Headings
- OMIM
- Reactome

## ChEMBL

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. The database, originally known as StARlite, was developed by a biotechnology company called Inpharmatica Ltd. later acquired by Galapagos NV. The data was acquired for EMBL in 2008 with an award from The Wellcome Trust, resulting in the creation of the ChEMBL chemogenomics group at EMBL-EBI, led by John Overington.

### Scope and access

The ChEMBL database contains compound bioactivity data against drug targets. Bioactivity is reported in Ki, Kd, IC50, and EC50. Data can be filtered and analyzed to develop compound screening libraries for lead identification during drug discovery. ChEMBL version 2 (ChEMBL_02) was launched in January 2010, including 2.4 million bioassay measurements covering 622,824 compounds, including 24,000 natural products. This was obtained from curating over 34,000 publications across twelve medicinal chemistry journals. ChEMBL's coverage of available bioactivity data has grown to become one of the most comprehensive in a public database. In October 2010 ChEMBL version 8 (ChEMBL_08) was launched, with over 2.97 million bioassay measurements covering 636,269 compounds. ChEMBL_10 saw the addition of the PubChem confirmatory assays, in order to integrate data that is comparable to the type and class of data contained within ChEMBL. ChEMBLdb can be accessed via a web interface or downloaded by File Transfer Protocol. It is formatted in a manner amenable to computerized data mining, and attempts to standardize activities between different publications, to enable comparative analysis. ChEMBL is also integrated into other large-scale chemistry resources, including PubChem and the ChemSpider system of the Royal Society of Chemistry.

### Associated resources

In addition to the database, the ChEMBL group have developed tools and resources for data mining. These include Kinase SARfari, an integrated chemogenomics workbench focussed on kinases. The system incorporates and links sequence, structure, compounds and screening data. GPCR SARfari is a similar workbench focused on GPCRs, and ChEMBL-Neglected Tropical Diseases (ChEMBL-NTD) is a repository for Open Access primary screening and medicinal chemistry data directed at endemic tropical diseases of the developing regions of the Africa, Asia, and the Americas. The primary purpose of ChEMBL-NTD is to provide a freely accessible and permanent archive and distribution centre for deposited data. July 2012 saw the release of a new malaria data service, sponsored by the Medicines for Malaria Venture (MMV), aimed at researchers around the globe. The data in this service includes compounds from the Malaria Box screening set, as well as the other donated malaria data found in ChEMBL-NTD. myChEMBL, the ChEMBL virtual machine, was released in October 2013 to allow users to access a complete and free, easy-to-install cheminformatics infrastructure. In December 2013, the operations of the SureChem patent informatics database were transferred to EMBL-EBI. In a portmanteau, SureChem was renamed SureChEMBL. 2014 saw the introduction of the new resource ADME SARfari - a tool for predicting and comparing cross-species ADME targets.

References

Kim S, Thiessen PA, Cheng T, Zhang J, Gindulyte A, Bolton EE. PUG-View: programmatic access to chemical annotations integrated in PubChem. J Cheminform. 2019 Aug.

"PubChem Source Information" - https://pubchem.ncbi.nlm.nih.gov/sources/sources.cgi. The PubChem Project. USA: National Center for Biotechnology Information.

Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, Li Y, Wang R, Lai L. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model. 2007 Nov-Dec

Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan

Mok NY, Brenk R. Mining the ChEMBL database: an efficient chemoinformatics workflow for assembling an ion channel-focused screening library. J Chem Inf Model. 2011.

American Chemical Society Expresses Opposition to NIH's PubChem article retrieved from https://osc.universityofcalifornia.edu/2005/05/american-chemical-society-calls-on-congress-to-shut-down-nihs-pubchem/