Chemical database

A **chemical database** is a database specifically designed to store chemical information. This information is about chemical and crystal structures, spectra, reactions and syntheses, and thermophysical data.

Types of chemical databases

Chemical structures

Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemist, they are unsuitable for computational use and especially for search and storage. Small molecules (also called ligands in drug design applications), are usually represented using lists of atoms and their connections. Large molecules such as proteins are however more compactly represented using the sequences of their amino acid building blocks. Large chemical databases for structures are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory...

Literature database

Chemical literature databases correlate structures or other chemical information to relevant references such as academic papers or patents. This type of database includes STN, Scifinder, and Reaxys. Links to literature are also included in many databases that focus on chemical characterization.

Crystallographic database

Crystallographic databases store X-ray crystal structure data. Common examples include Protein Data Bank and Cambridge Structural Database.

NMR spectra database

NMR spectra databases correlate chemical structure with NMR data. These databases often include other characterization data such as FTIR and mass spectrometry.

Reactions database

Most chemical databases store information on stable molecules but in databases for reactions also intermediates and temporarily created unstable molecules are stored. Reaction databases contain information about products, educts, and reaction mechanisms.

Thermophysical database

Thermophysical data are information about

- phase equilibria including vapor–liquid equilibrium, solubility of gases in liquids, liquids in solids (SLE), heats of mixing, vaporization, and fusion.
- caloric data like heat capacity, heat of formation and combustion,
- transport properties like viscosity and thermal conductivity

Chemical structure representation

There are two principal techniques for representing chemical structures in digital databases

- As connection tables / adjacency matrices / lists with additional information on bond (edges) and atom attributes (nodes), such as:

  MDL Molfile, PDB, CML
- As a linear string notation based on depth first or breadth first traversal, such as:

  SMILES/SMARTS, SLN, WLN, InChI

These approaches have been refined to allow representation of stereochemical differences and charges as well as special kinds of bonding such as those seen inorgano-metallic compounds. The principal advantage of a computer representation is the possibility for increased storage and fast, flexible search.

Search

Substructure

Chemists can search databases using parts of structures, parts of their IUPAC names as well as based on constraints on properties. Chemical databases are particularly different from other general purpose databases in their support for sub-structure search. This kind of search is achieved by looking for subgraph isomorphism (sometimes also called a monomorphism) and is a widely studied application of Graph theory. The algorithms for searching are computationally intensive, often of O (n^3) or O (n^4) time complexity (where n is the number of atoms involved).

The intensive component of search is called atom-by-atom-searching (ABAS), in which a mapping of the search substructure atoms and bonds with the target molecule is sought. ABAS searching usually makes use of the Ullman algorithm[1] or variations of it (*i.e.* **SMSD**). Speedups are achieved by time amortization, that is, some of the time on search tasks are saved

by using precomputed information. This pre-computation typically involves creation of bitstrings representing presence or absence of molecular fragments. By looking at the fragments present in a search structure it is possible to eliminate the need for ABAS comparison with target molecules that do not possess the fragments that are present in the search structure. This elimination is called screening (not to be confused with the screening procedures used in drug-discovery). The bit-strings used for these applications are also called structural-keys. The performance of such keys depends on the choice of the fragments used for constructing the keys and the probability of their presence in the database molecules. Another kind of key makes use of hash-codes based on fragments derived computationally. These are called 'fingerprints' although the term is sometimes used synonymously with structural-keys. The amount of memory needed to store these structural-keys and fingerprints can be reduced by 'folding', which is achieved by combining parts of the key using bitwise-operations and thereby reducing the overall length.

Conformation

Search by matching 3D conformation of molecules or by specifying spatial constraints is another feature that is particularly of use in drug design. Searches of this kind can be computationally very expensive. Many approximate methods have been proposed, for instance BCUTS, special function representations, moments of inertia, ray-tracing histograms, maximum distance histograms, shape multipoles to name a few.

Descriptors

All properties of molecules beyond their structure can be split up into either physico-chemical or pharmacological attributes also called descriptors. On top of that, there exist various artificial and more or less standardized naming systems for molecules that supply more or less ambiguous names and synonyms. The IUPAC name is usually a good choice for representing a molecule's structure in a both human-readable and unique string although it becomes unwieldy for larger molecules. Trivial names on the other hand abound with homonyms and synonyms and are therefore a bad choice as a defining database key. While physico-chemical descriptors like molecular weight, (partial) charge, solubility, etc. can mostly be computed directly based on the molecule's structure, pharmacological descriptors can be derived only indirectly using involved multivariate statistics or experimental (screening, bioassay) results. All of those descriptors can for reasons of computational effort be stored along with the molecule's representation and usually are.

Similarity

There is no single definition of molecular similarity, however the concept may be defined according to the application and is often described as an inverse of ameasure of distance in descriptor space. Two molecules might be considered more similar for instance if their difference in molecular weights is lower than when compared with others. A variety of other measures could be combined to produce a multi-variate distance measure. Distance measures are often classified intoEuclidean measures and non-Euclidean measures depending on whether the triangle inequality holds. Maximum Common Subgraph (MCS) based substructure search [2](similarity or distance measure) is also very common. MCS is also used for screening drug like compounds by hitting molecules, which share common subgraph (substructure).[9]

Chemicals in the databases may be clustered into groups of 'similar' molecules based on similarities. Both hierarchical and non-hierarchical clustering approaches can be applied to chemical entities with multiple attributes. These attributes or molecular properties may either be determined empirically or computationally deriveddescriptors. One of the most popular clustering approaches is the Jarvis-Patrick algorithm .

In pharmacologically oriented chemical repositories, similarity is usually defined in terms of the biological effects of compounds (ADME/tox) that can in turn be semiautomatically inferred from similar combinations of physico-chemical descriptors using QSAR methods.

Registration system

Databases systems for maintaining unique records on chemical compounds are termed as Registration systems. These are often used for chemical indexing, patent systems and industrial databases.

Registration systems usually enforce uniqueness of the chemical represented in the database through the use of unique representations. By applying rules of precedence for the generation of stringified notations, one can obtain unique/'canonical' string representations such as 'canonical SMILES'. Some registration systems such as the CAS system make use of algorithms to generate unique hash codes to achieve the same objective.

A key difference between a registration system and a simple chemical database is the ability to accurately represent that which is known, unknown, and partially known. For example, a chemical database might store a molecule with stereochemistry unspecified, whereas a chemical registry system requires the registrar to specify whether the stereo configuration is unknown, a specific (known) mixture, or racemic. Each of these would be considered a different record in a chemical registry system.

Registration systems also preprocess molecules to avoid considering trivial differences such as differences in halogen ions in chemicals. An example is the Chemical Abstracts Service (CAS) registration system. See also CAS registry number.

PubChem

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 substance descriptions and small molecules with fewer than 1000 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database.

PubChem

PubChem is designed to provide information on biological activities of small molecules, generally those with molecular weight less than 500 daltons(2). PubChem's integration with NCBI's Entrez (3) information retrieval system provides sub/structure, similarity structure, bioactivity data as well as links to biological property information in PubMed and NCBI's Protein 3D Structure Resource.

PubChem Databases

PubChem is comprised of three linked databases --

PubChem Compound,

PubChem Substance and

PubChem Bioassay

PubChem Compound (unique structures with computed properties)

PubChem Compound (4) is a searchable database of chemical structures with validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. PubChem Compound includes over 5M compounds.

- Molecular Name Searches (e.g., Tylenol, Benzene) allow searching with a variety of chemical synonyms,

- Chemical Property Range Searches (e.g., Molecular Weight between 100 and 200, Hydrogen Bond Acceptor Count between 3 and 5) allow searching for compounds with a variety of physical/chemical properties, and descriptors.
- Simple Elemental Searches (all compounds containing Gallium) allow searching with specific element restrictions.

PubChem Substance (deposited structures)

PubChem Substance (5) is a searchable database containing descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results available in PubChem BioAssay. PubChem Substance includes over 8M records. Substances with known content are linked to PubChem Compound.

- Molecule Synonym Searches (e.g. all substances with 'deoxythymidine' as a name fragment, or substances that contain 3'-Azido-3'-deoxythymidine).
- Biology Links Search (e.g. substances with tested, active or inactive bioassays).

- Combined Searches (e.g. substances that are 'Active in any BioAssay' and contain the element Ruthenium).

PubChem BioAssay

PubChem BioAssay (6) is a searchable database containing bioactivity screens of chemical substances described in PubChem Substance. PubChem BioAssay includes over 180 bioassays. Searchable descriptions of each bioassay are provided that include descriptions of screening procedural conditions and readouts.

- To Search for BioAssay Data Sets (e.g. HIV growth inhibition).
- To Browse or Download PubChem BioAssay Results (NCI AIDS Antiviral Assay)

Searching PubChem

PubChem Text Search

PubChem Text Search for searching compound name, synonym or ID that defaults to PubChem Compound. The search results page offers a pull down 'databases' menu that allows searching in PubChem Substance, PubChem BioAssay and a variety of other Entrez databases.

PubChem Chemical Structure Search

PubChem Chemical Structure Search (7) has the following options: Search SMILES (including SMARTS or InChI) or Formula which includes a 'Sketch' link to a drawing program that converts structural diagrams to SMILES(exact), SMARTS(substructure) or InChI(exact) strings for searching.

Clicking 'Done' on the 'structure editor' converts the structural diagram to the appropriate string and transfers it to the search box.

Select Structure File allows importation of standard and common chemical file formats (8).

Specify Search Type allows restriction to: same compound, similar compounds (9), formula or substructure.

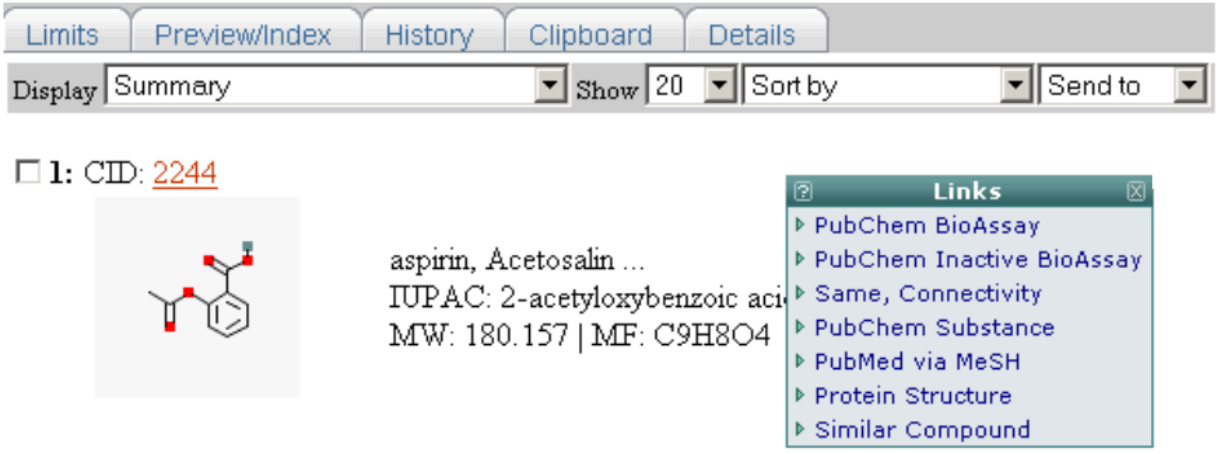
PubChem Indexes and Index Search

PubChem Indexes and Index Search allows fielded/range searching from either the PubChem homepage or Entrez search page. A extensive list of field aliases and examples of range searching is provided

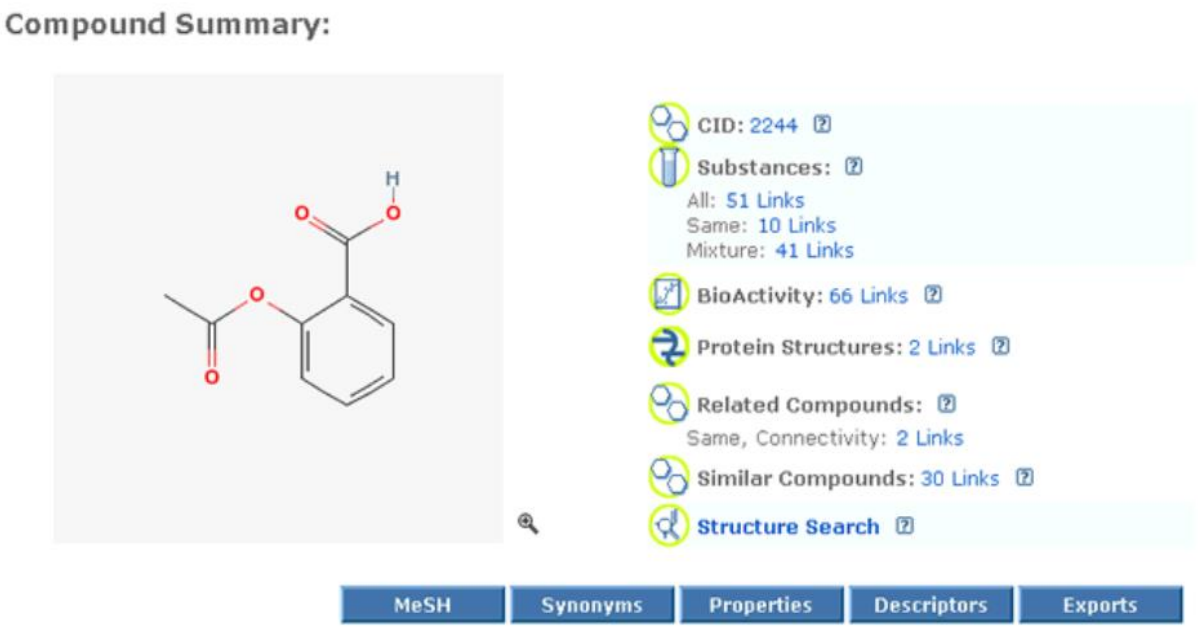
PubChem Search Results

PubChem Compound

PubChem Compound results are derived from PubChem Substance records that provide structures. Since compounds are structurally unique, one compound may link to multiple substances. The default display is a compound summary with thumbnails with cross links(12) to each PubChem database, other NCBI databases, and depositor's databases.



Clicking either the structure or SID link gives the full display which includes the compound's property data, description, related substance information, neighboring structures, and cross links.



PubChem Substance

PubChem Substance has unique records if the structure is not known or supplied. For example, Sulfated polymannuroguluronate, a novel anti-acquired immune deficiency syndrome (AIDS) drug candidate, and other natural products.

The PubChem Substance Summary Record,

Sulfated polymannuroguluronate, AIDS218087 ...

Source: NIAID(218087)

is linked to the full record by clicking on the SID number (PubChem's substance identifier). This displays the full substance record, that includes links: to PubMed and the source; the Medical Subject Annotation (MESH Substance Name) and a MESH PubMed search link; and depositor supplied synonyms and comments.

PubChem BioAssay

The PubChem BioAssay Summary Record,

AID: 179 Links

NCI	AIDS	Antiviral	Assay
Source:			DTP/NCI

15 Readouts, 37678 substances tested

---

is linked to the full record by clicking on the AID number (PubChem's assay (protocol) identifier). This displays the full bioassay record, that includes: links to the substances tested (all, active, inactive, inconclusive) and related PubMed, Protein, Taxonomy, OMIM and related BioAssay records; and a description of the assay possibly with protocols and comments.