D1202–D1213 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 22 September 2015*
doi: 10.1093/nar/gkv951

PubChem Substance and Compound databases

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton*, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang and Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20894, USA

Received August 14, 2015; Revised September 10, 2015; Accepted September 11, 2015

ABSTRACT

PubChem (https://pubchem.ncbi.nlm.nih.gov) is a public repository for information on chemical substances and their biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH). For the past 11 years, PubChem has grown to a sizable system, serving as a chemical information resource for the scientific research community. PubChem consists of three inter-linked databases, Substance, Compound and BioAssay. The Substance database contains chemical information deposited by individual data contributors to PubChem, and the Compound database stores unique chemical structures extracted from the Substance database. Biological activity data of chemical substances tested in assay experiments are contained in the BioAssay database. This paper provides an overview of the PubChem Substance and Compound databases, including data sources and contents, data organization, data submission using PubChem Upload, chemical structure standardization, web-based interfaces for textual and non-textual searches, and programmatic access. It also gives a brief description of PubChem3D, a resource derived from theoretical three-dimensional structures of compounds in PubChem, as well as PubChemRDF, Resource Description Framework (RDF)-formatted PubChem data for data sharing, analysis and integration with information contained in other databases.

INTRODUCTION

PubChem (https://pubchem.ncbi.nlm.nih.gov) (1–6) is a public repository for information on chemical substances and their biological activities. Since launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH), PubChem has rapidly grown to a key chemical information resource that serves scientific communities in many areas such as cheminformatics, chemical biology, medicinal chemistry and drug discovery.

PubChem contains one of the largest corpus of publicly available chemical information. As of September 2015, it has more than 157 million depositor-provided chemical substance descriptions, 60 million unique chemical structures and 1 million biological assay descriptions, covering about 10 thousand unique protein target sequences. PubChem organizes this vast amount of data into three interlinked databases: Substance, Compound and BioAssay (see Figure 1). The Substance database (https://www.ncbi.nlm.nih.gov/pcsubstance) stores depositor-contributed information. Unique chemical structures are extracted from the Substance database and stored in the Compound database (https://www.ncbi.nlm.nih.gov/pccompound). The BioAssay database (https://www.ncbi.nlm.nih.gov/pcassay) stores descriptions of biological assays on chemical substances. The primary identifiers for the Substance, Compound and BioAssay databases are SID (SubstanceID), CID (CompoundID) and AID (AssayID), respectively.

Herein we describe the PubChem Substance and Compound databases, as well as related tools and services (see Table 1). While these databases were covered in our previous paper (1) published in early 2008, considerable changes have been made to these resources to keep up with rapidly advancing technology. Therefore, it is necessary to provide updated information on these databases. In this paper we will discuss various aspects of the two databases, including data contents and organization, data uploading and downloading, access through web interfaces, programmatic access and other relevant tools and services. The BioAssay database has been well described in our previous papers (2–5), including those published in Nucleic Acids Research Database issues (3–5).

DATA SOURCES AND CONTENTS

PubChem's data are provided by more than 350 contributors (https://pubchem.ncbi.nlm.nih.gov/sources/), includ-

[*] To whom correspondence should be addressed. Tel: +1 301 451 1811; Fax: +1 301 480 4559; Email: bolton@ncbi.nlm.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2015.

Table 1. Tools and services in the PubChem Compound and Substance databases

- **Chemical structure search (https://pubchem.ncbi.nlm.nih.gov/search/search.cgi)**
  Allows users to query the PubChem Compound database by chemical structure or chemical structure pattern.
- **Chemical structure sketcher (https://pubchem.ncbi.nlm.nih.gov/edit/)**
  A platform-independent 2-D molecule drawer, compatible with major web browsers.
- **PubChem Upload (https://pubchem.ncbi.nlm.nih.gov/upload/)**
  A data submission system that enables one to contribute substance or assay data to PubChem.
- **Standardization service (https://pubchem.ncbi.nlm.nih.gov/standardize/)**
  Validates and normalizes an input chemical structure in the same way as PubChem standardization process.
- **Classification browser (https://pubchem.ncbi.nlm.nih.gov/classification/)**
  Allows users to browse PubChem data using a classification of interest, or search for records annotated with the desired classification/term.
- **Identifier exchange service (https://pubchem.ncbi.nlm.nih.gov/idexchange/)**
  Converts one type of identifiers for a given set of chemical structures into a different type of identifiers for identical or similar chemical structures.
- **Score matrix service (https://pubchem.ncbi.nlm.nih.gov/score_matrix/)**
  Computes matrices of 2-D and 3-D similarity scores for a given set of compounds.
- **Structure clustering (https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=clustering)**
  Clusters compounds/substances based on their structural similarity using the Single Linkage Algorithm.
- **Widgets (https://pubchem.ncbi.nlm.nih.gov/widget/docs/)**
  Provides a rapid way to display some commonly requested PubChem data views.
- **Web-based 3D viewer (https://pubchem.ncbi.nlm.nih.gov/vw3d/)**
  An interactive web-based viewer for 3-D conformations of molecules, which visualizes 3-D information available within PubChem.
- **Pc3D viewer (https://pubchem.ncbi.nlm.nih.gov/pc3d/)**
  An interactive 3-D molecular viewer that can be downloaded and installed on local machines.
- **PubChem FTP Site (ftp://ftp.ncbi.nlm.nih.gov/pubchem/)**
  Enables users to bulk download PubChem Data.
- **Structure download (https://pubchem.ncbi.nlm.nih.gov/pc_fetch/)**
  Downloads a set of substance or compound records in PubChem.
- **Power User Gateway (PUG) (https://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html)**
  Provides programmatic access to PubChem services via a single common gateway interface (CGI), called "pug.cgi".
- **PUG-REST (https://pubchem.ncbi.nlm.nih.gov/pug_rest/)**
  A Representational State Transfer (REST)-full style web service access layer to PubChem.
- **PUG-SOAP (https://pubchem.ncbi.nlm.nih.gov/pug_soap/)**
  A web service access method that uses the simple object access protocol (SOAP).
- **PubChemRDF (https://pubchem.ncbi.nlm.nih.gov/rdf/)**
  the RDF-based resource compatible with Semantic Web standards and technologies.

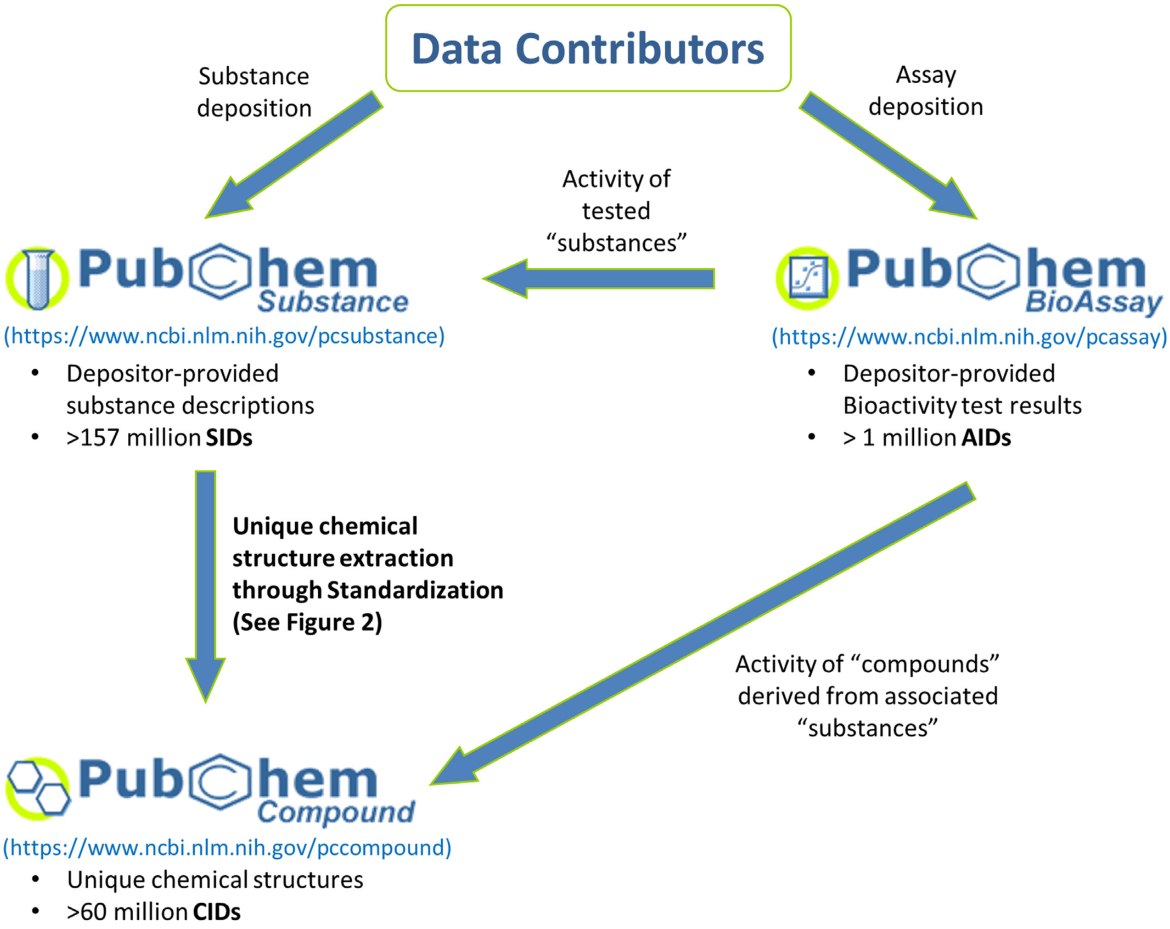

Data Contributors
Substance deposition
Assay deposition
Activity of tested "substances"
PubChem Substance
(https://www.ncbi.nlm.nih.gov/pcsubstance)
• Depositor-provided substance descriptions
• >157 million SIDs
PubChem BioAssay
(https://www.ncbi.nlm.nih.gov/pcassay)
• Depositor-provided Bioactivity test results
• > 1 million AIDs
Unique chemical structure extraction through Standardization (See Figure 2)
Activity of "compounds" derived from associated "substances"
PubChem Compound
(https://www.ncbi.nlm.nih.gov/pccompound)
• Unique chemical structures
• >60 million CIDs

Figure 1. Data organization in PubChem. SID, CID and AID are the identifiers for the Substance, Compound and BioAssay databases, respectively.

ing university labs, government agencies, pharmaceutical companies, chemical vendors, publishers and a number of chemical biology resources. The data provided by these contributors are not just limited to small molecules, but also include siRNAs, miRNAs, carbohydrates, lipids, peptides, chemically modified macromolecules and many others.

Although PubChem was initially conceived as a central repository for the NIH's Molecular Libraries Program (MLP) it includes data contributed by numerous non-MLP organizations. For example, PubChem contains a substantial amount of literature-derived bioactivity data of chemical substances, manually extracted from tens of thousands of scientific articles by data contributors such as ChEMBL (7) and BindingDB (8). In addition, through integration with data from DrugBank (9), the Hazardous Substances Data Bank (10) and other databases, the annotation of chemical records includes pharmacology, drug target information, toxicology, safety and handling information.

PubChem also hosts data from important regulatory agencies. For instance, the US Food and Drug Administration (FDA) (http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/) contributes information on chemicals of interest to the FDA, including their Unique Ingredient Identifiers (UNIIs) and pharmacological classifications. Similarly, the US Environmental Protection Agency (EPA) Substance Registry Services (SRS) (http://www.epa.gov/srs/) contributes chemical substance information on chemicals of interest to the EPA.

Links to patent information are provided thanks to data contributions from a growing number of organizations, including IBM (http://www.almaden.ibm.com/) and SureChEMBL (formerly known as SureChem) (https://www.surechembl.org/). As a result, PubChem offers links between about 6 million patent documents and more than 16 million unique chemical structures, with over 329 million chemical substance-patent links covering US, European and World Intellectual Property Organization patent documents published since 1800.

The data contents of PubChem and other databases complement each other. For example, while ChEMBL focuses on bioactivity data of molecules determined from various types of assays (including binding assays, functional assays and ADMET assays), the primary focus of BindingDB is to collect experimentally determined binding affinities of protein–ligand complexes, and therefore it excludes bioactivity data from functional assays and ADMET assays, which ChEMBL may have. PubChem contains a large amount of bioactivity data, primarily from high-throughput screening experiments. It should be noted that data exchange and integration between PubChem and other databases are very common. For instance, ChEMBL contributes their bioactivity data to PubChem, and at the same time, a small portion of PubChem's bioactivity data (from confirmatory and panel assays with dose-response endpoints) are integrated into ChEMBL.

DATA ORGANIZATION

Figure 1 shows PubChem's three primary databases (i.e., Substance, Compound and BioAssay) and the data flow between them. Each data contributor provides descriptions on chemical samples using PubChem Upload (see the 'Data submission using PubChem Upload' section). PubChem calls these community-provided chemical sample descriptions 'substances' and stores them in the Substance database (https://www.ncbi.nlm.nih.gov/pcsubstance). Different substance records (provided by the same or different contributors) may contain different kinds of information for the same molecule. For example, one substance record may provide information on biological functions of glucose, while another may describe characteristics of a research grade sample of glucose. The Substance database stores these different descriptions about the same molecule as separate records that are independent of each other. The Substance database maintains the provenance of substance records, helping users see who provided what information to PubChem.

Inevitably, information on a given molecule may be scattered across many records in the Substance database, presenting a problem for users who are interested in an aggregated view of information on a single chemical structure. To address this issue, PubChem extracts unique chemical structures from the Substance database through a process called 'standardization' (see 'Standardization' section) and stores them in the Compound database (https://www.ncbi.nlm.nih.gov/pccompound). This allows substance records from different data sources about the same molecule to be aggregated through a common 'compound' record in the Compound database.

The bioactivity of a chemical sample may be measured using various experimental techniques and conditions. These experimental factors may substantially affect the accuracy and precision of the measured bioactivity of a molecule. Especially, considering that the bioactivity data in PubChem are collected from many different data sources, these data would have a very limited use unless detailed information on experimental protocols is provided together with the data. Therefore, the descriptions of biological experiments on chemical substances are stored in a separate database called the BioAssay database (https://www.ncbi.nlm.nih.gov/pcassay). The depositor-provided information on an assay may have multiple bioactivity outcomes determined with different samples (i.e., different *substances*) of the same molecule. Therefore, PubChem provides a comprehensive overview of the biological activity profile of the tested *compounds*, by generating associations between the test results and tested compounds (via tested substances).

Data submission using PubChem upload

PubChem Upload (https://pubchem.ncbi.nlm.nih.gov/upload/) is PubChem's data submission system that allows data contributors to provide substance descriptions, assay experiment descriptions, and the results of substances tested in assays. This system can also be used for updating or revoking existing PubChem records. To reduce the time and effort required to make data submissions, it offers streamlined upload procedures and includes an extensive set of wizards, inline help tips and templates. Data providers can make a quick submission with a simple decision-tree set of wizards, which guides them through the process of publishing their data in PubChem. Alternatively, one can avoid the wizards and use the interfaces directly. For large and/or frequent data uploads, PubChem supports File Transfer Protocol (FTP)-based depositions using a private FTP account, which enables completely automated data upload into PubChem.

PubChem Upload has many important features to assist data contributors. It allows users to enter data and descriptive information by web form or by file. The supported file formats for substance deposition are Structure-Data file (SDF) (11) and Comma Separated Values (CSV). Assay data may be provided in either widely-used spreadsheet formats [XLSX (for Microsoft Excel), ODS (for OpenOffice) or CSV] or eXtended Markup Language (XML)-based data specification (XML or ASN). PubChem Upload offers an expanded ability to edit data directly in the browser. For example, the spreadsheet editor allows contributors to upload large spreadsheets with minimal reformatting and to edit those large datasets online. PubChem Upload provides an automated suite of validation checks that help contributors identify potential issues before the data are made public. In addition, it has a 'Preview' function, which allows contributors to check how the incoming data will appear in PubChem.

Contributors can specify a 'hold-until date', until which the contributed data will not be visible to the public. This option is useful when contributors need to time data release with the publication of a paper or the filing of a patent or in coordination with a grant administrator. PubChem Upload allows contributors to create private Uniform Resource Locators (URLs) to the on-hold data. These URLs can be used for contributors to share the on-hold data with reviewers and collaborators while the data are still not publicly available.

Standardization

Figure 2 shows a schematic diagram for PubChem's standardization method, which is an automated process that extracts unique chemical structures from the Substance database to construct the Compound database. Roughly, this process takes two steps: validation of chemical contents and normalization of chemical representations.

The first step begins with checking the validity of the depositor-provided chemical structure, by verifying the atomic number and isotope for each atom. The number of implicit hydrogens attached to each non-hydrogen atom is adjusted to an appropriate value according to a simple valence bond model and non-standard representations of functional groups are modified to a preferred one (e.g., nitro groups represented by N(= O) = O or [N2+]([O−])[O−] are standardized to [N+](= O)[O−]). Then, the valence and formal charge for each atom are examined by comparing it

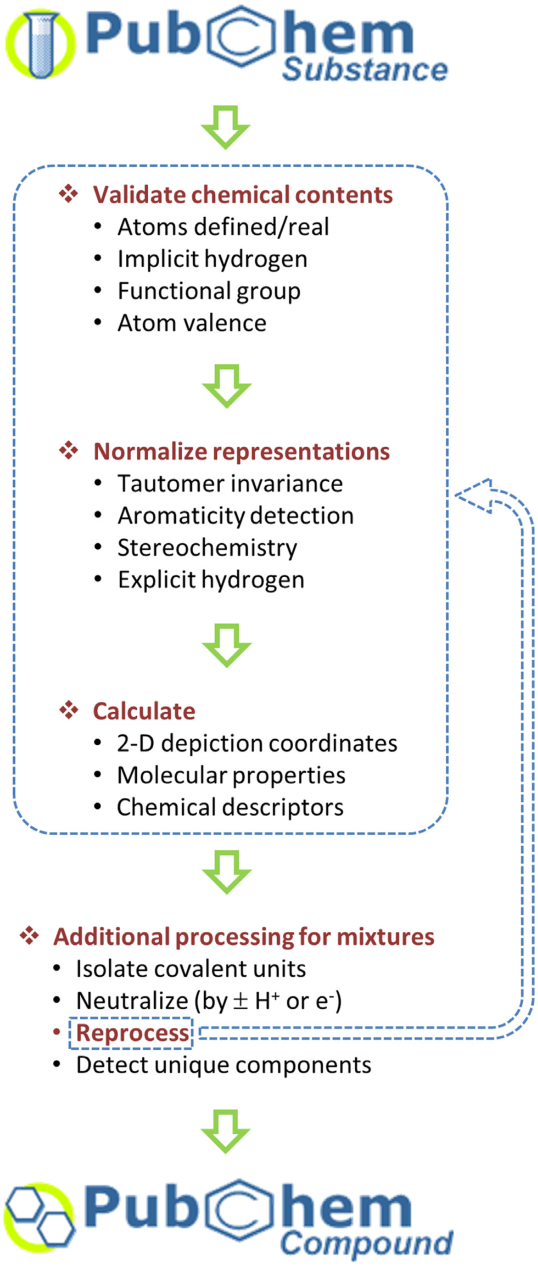


Figure 2. PubChem standardization process in which unique chemical structures are extracted from the Substance database and stored in the Compound database.

to an extensive list of allowed configurations for each element type (with respect to the formal charge, the number of σ bonds, the number of π bonds and the maximum allowed number of implicit hydrogen atoms).

In the second step, equivalent or alternative valence bond structures (i.e., tautomeric and/or resonance forms) are normalized into a single representation. If a structure has an aromatic substructure, its canonical Kekulé form is determined. For a structure with stereocenters, their configurations will be annotated as 'clockwise' or 'counterclockwise' for atom stereocenters, or as 'cis' or 'trans' for bond stereocenters. If the configuration of a stereocenter cannot be resolved, it is annotated as 'undefined'. Lastly, implicit hydrogens are converted into explicit ones.

Subsequent processing of each standardized structure involves computation of 2-D depiction coordinates, basic molecular properties (e.g., molecular weight, molecular formula, etc.) and chemical descriptors (e.g., International Union of Pure and Applied Chemistry (IUPAC) name, canonical and isomeric Simplified Molecular-Input Line-Entry System (SMILES) (12–14), IUPAC International Chemical Identifier (InChI) (15–18), etc.). If the standardized structure is a mixture with multiple covalent units, unique covalent structures are isolated and reprocessed using the procedures described above.

The mapping from substances to entries in the Compound database is made based on structural hash codes (19,20) calculated for the standardized structures. If the hash code of a standardized structure is not present in the Compound database, a new entry with a new CID is created. If a CID with an identical hash code already exists, an association is created between this CID and the SID of the substance that the standardized structure was generated from. Contributed substance descriptions that do not include a chemical structure or that fail the PubChem chemical structure standardization procedure do not enter nor have links to the PubChem Compound database.

PubChem contributors and users often need to understand the modifications made to chemical structures through the standardization process, for example, when attempting to integrate external resources with PubChem. For this reason, PubChem provides the standardization service (https://pubchem.ncbi.nlm.nih.gov/standardize), which allows one to see how their structures would be handled during the standardization process. Although this service itself can process only a single structure at a time, one may process multiple structures by submitting multiple Standardization service requests programmatically through the Power User Gateway (PUG), PUG-SOAP or PUG-REST (21) (discussed later in the 'Programmatic Access' section).

ACCESS TO PUBCHEM

Web interfaces for textual search

Entrez is the search and retrieval system used for PubChem's three primary databases and other major NCBI databases (6), including PubMed, Nucleotide and Protein Sequences, Protein Structures, Genome, Taxonomy, BioSystems, Gene Expression Omnibus (GEO) and many others. One can search the PubChem databases through Entrez by initiating a search from the PubChem home page (https://pubchem.ncbi.nlm.nih.gov), which also provides launch points to various PubChem services, tools, help documents and more. Alternatively, one can begin the search from the NCBI home page (https://www.ncbi.nlm.nih.gov). By default, if a specific database is not selected in the search menu, Entrez searches all Entrez databases available and lists the number of records in each database that are returned for this 'global query'. Simply by selecting one of the three PubChem database from the global query result page, one can see the query result specific to that database.

If an Entrez search returns multiple records, they are displayed in a document summary (DocSum) report. Figure 3 shows an example of the DocSum page from an Entrez search in the PubChem Compound database. The DocSum page for a search in the Substance database is very similar to Figure 3 in layout and format. For each record in the DocSum page, some data-specific information is provided with a link to the Summary page for that record (Figure 4) (see below for the Summary page). The DocSum page contains controls to change the display type, to sort the results by various means, or to export the page to a file or printer. In addition, the icons and links on the right column of the DocSum page allow users to perform further analysis on the query result, to download the corresponding records, to refine or modify the search, to obtain associated records in other databases and so on.

If a search against the Compound database returns a single record, the Compound Summary page for that record is displayed (Figure 4). The Compound Summary page provides a comprehensive view that recaps all information known about a particular chemical, collected from different data sources. If a search against the Substance database returns a single substance, the Substance Summary page for that substance is displayed, which shows information provided by the data contributor for that record.

Non-textual search using the chemical structure search tool

Because Entrez is primarily a text-based search system, it cannot be used for searching that involves data types specific to PubChem, such as chemical structures. The Chemical Structure Search tool (https://pubchem.ncbi.nlm.nih.gov/search/search.cgi) (Figure 5) enables one to query and subset the Compound database using various chemical structure search types, including identity search, substructure/superstructure search, molecular formula search and 2-D and 3-D similarity searches.

In 2-D similarity search, the similarity between chemical structures is quantified using the Tanimoto equation (22–24) in conjunction with the PubChem substructure fingerprint (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf). In 3-D similarity search, the shape-Tanimoto (ST) and color-Tanimto (CT) are used to evaluate similarity between conformers in 3-D shape and functional group orientation, respectively. The two 3-D similarity measures (ST and CT) are calculated using the Gaussian-shape overlay method by Grant and Pickup (25–28), implemented in the Rapid Overlay of Chemical Structures (ROCS) (29). The 3-D similarity search is implemented as part of the PubChem3D project (https://pubchem.ncbi.nlm.nih.gov/release3d.html), which is discussed later in the 'PubChem3D' section.

The Chemical Structure Search tool supports a variety of query formats, including SMILES (12–14), SMARTS, InChI (15–18), CID, molecular formula and SDF (11). One can also manually draw a query chemical structure using the PubChem Chemical Structure Sketcher (https://pubchem.ncbi.nlm.nih.gov/edit/) (30). This JavaScript-based structure editor is platform-independent and compatible with major web browsers, and does not require the user to download or install special software. In addition, it contains import and export features such as support for chemical structure files.

The Chemical Structure Search tool allows users to narrow a search to the result from a previous Entrez or chemical structure search or to the set of CIDs uploaded in a file. Optional filters may be applied to limit the search result, based on various properties, such as molecular weight, heavy atom count, presence or absence of stereochemistry, depositor name or category and so on. A query can be exported to an XML file, which allows one to import the query from the XML file and to repeat the search without filling out the search form again. This XML file can also serve as an example for constructing queries for the PUG interface, which is described later.

Download

PubChem data are available for bulk download on the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem). In addition, one may use PubChem Structure Download service (https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi) to download a subset of substance or compound records in PubChem, rather than all PubChem records. The records can be exported in several formats, including SDF (11), large and small images, SMILES (12–14), InChI (15–18), XML and either text or binary ASN.1. The files may be optionally compressed in standard gzip (.gz) or bzip2 (.bz2) formats.

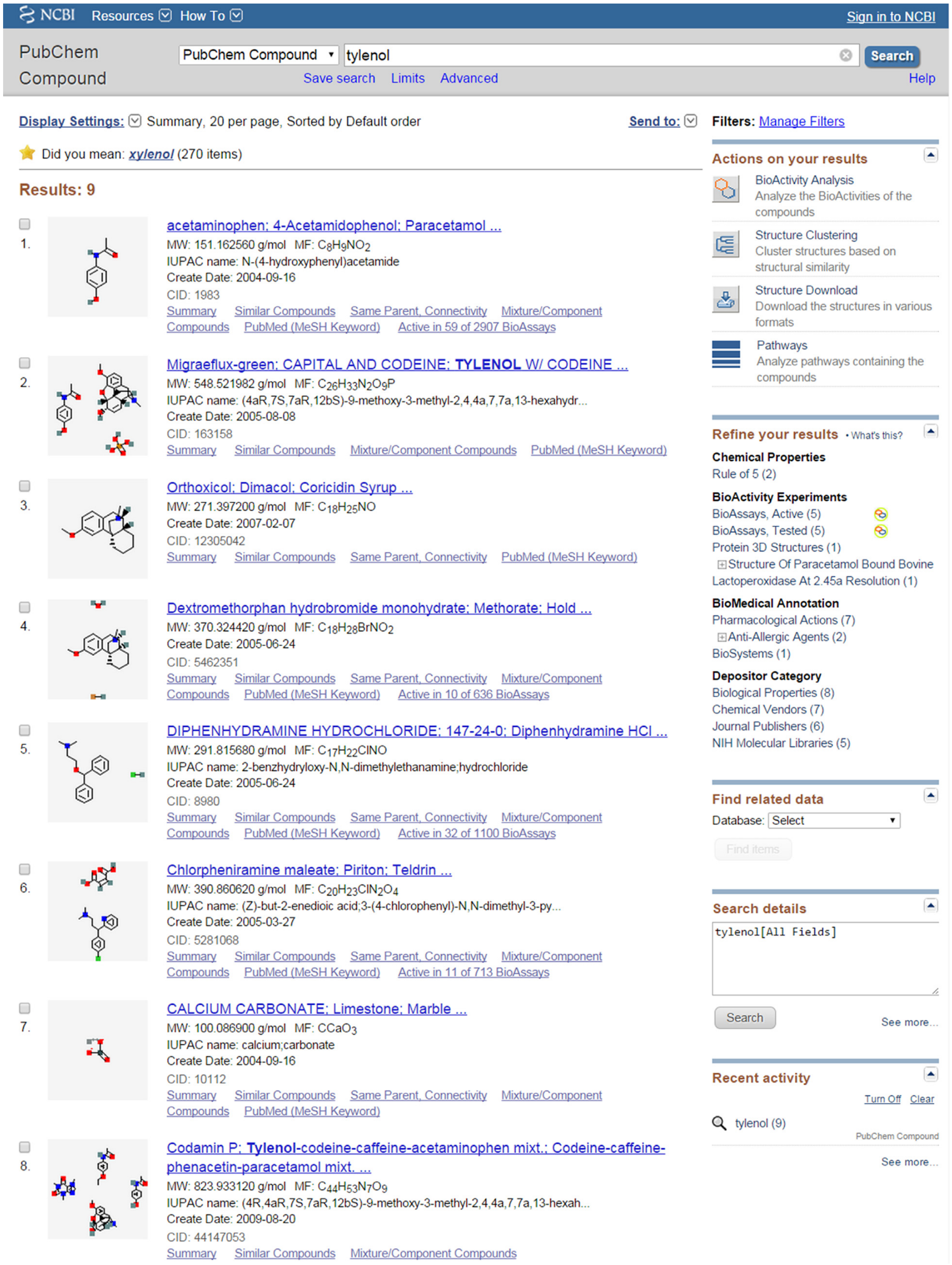
A list of CIDs or SIDs for records to download may be provided directly into the web page form or uploaded from a local file. Alternatively, they may be provided by using Entrez history, which stores a list of CIDs or SIDs returned from a previous Entrez search. Results from an Entrez or PubChem-specific search can also be used with the Structure Download tool, by clicking the download link (as indicated by a button with a disk icon) on the top-right side of the DocSum page (Figure 3).

Downloads through the structure download tool are limited to a maximum of 500 000 records per request, with an exception of image downloads, which are limited to 50 000 per request. These limits keep the download file sizes practical. The structure download service is accessible using the PubChem Power User Gateway (PUG), which allows for a programmatic access to PubChem. One can download more than this limit, through multiple interactive or programmatic requests.

Programmatic access

PubChem provides multiple programmatic data access routes, including:

(i) Entrez Utilities (also called E-Utilities or E-Utils) (https://www.ncbi.nlm.nih.gov/books/NBK25501/)
(ii) Power User Gateway (PUG) (https://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html)
(iii) PUG-SOAP (https://pubchem.ncbi.nlm.nih.gov/pug_soap/pug_soap_help.html)
(iv) PUG-REST (https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST.html)



tylenol

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Did you mean: *xylenol* (270 items)

Results: 9

□ 1. acetaminophen; 4-Acetamidophenol; Paracetamol ...
MW: 151.162560 g/mol MF: $C_8H_9NO_2$
IUPAC name: N-(4-hydroxyphenyl)acetamide
Create Date: 2004-09-16
CID: 1983
Summary Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword) Active in 59 of 2907 BioAssays

□ 2. Migraeflux-green; CAPITAL AND CODEINE; **TYLENOL** W/ CODEINE ...
MW: 548.521982 g/mol MF: $C_{26}H_{33}N_2O_9P$
IUPAC name: (4aR,7S,7aR,12bS)-9-methoxy-3-methyl-2,4,4a,7,7a,13-hexahydr...
Create Date: 2005-08-08
CID: 163158
Summary Similar Compounds Mixture/Component Compounds PubMed (MeSH Keyword)

□ 3. Orthoxicol; Dimacol; Coricidin Syrup ...
MW: 271.397200 g/mol MF: $C_{18}H_{25}NO$
Create Date: 2007-02-07
CID: 12305042
Summary Similar Compounds Same Parent, Connectivity PubMed (MeSH Keyword)

□ 4. Dextromethorphan hydrobromide monohydrate; Methorate; Hold ...
MW: 370.324420 g/mol MF: $C_{18}H_{28}BrNO_2$
Create Date: 2005-06-24
CID: 5462351
Summary Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword) Active in 10 of 636 BioAssays

□ 5. DIPHENHYDRAMINE HYDROCHLORIDE; 147-24-0; Diphenhydramine HCl ...
MW: 291.815680 g/mol MF: $C_{17}H_{22}ClNO$
IUPAC name: 2-benzhydryloxy-N,N-dimethylethanamine;hydrochloride
Create Date: 2005-06-24
CID: 8980
Summary Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword) Active in 32 of 1100 BioAssays

□ 6. Chlorpheniramine maleate; Piriton; Teldrin ...
MW: 390.860620 g/mol MF: $C_{20}H_{23}ClN_2O_4$
IUPAC name: (Z)-but-2-enedioic acid;3-(4-chlorophenyl)-N,N-dimethyl-3-py...
Create Date: 2005-03-27
CID: 5281068
Summary Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword) Active in 11 of 713 BioAssays

□ 7. CALCIUM CARBONATE; Limestone; Marble ...
MW: 100.086900 g/mol MF: $CCaO_3$
IUPAC name: calcium;carbonate
Create Date: 2004-09-16
CID: 10112
Summary Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword)

□ 8. Codamin P; **Tylenol**-codeine-caffeine-acetaminophen mixt.; Codeine-caffeine-phenacetin-paracetamol mixt. ...
MW: 823.933120 g/mol MF: $C_{44}H_{53}N_7O_9$
IUPAC name: (4R,4aR,7S,7aR,12bS)-9-methoxy-3-methyl-2,4,4a,7,7a,13-hexah...
Create Date: 2009-08-20
CID: 44147053
Summary Similar Compounds Mixture/Component Compounds

Filters: Manage Filters

Actions on your results

BioActivity Analysis
Analyze the BioActivities of the compounds

Structure Clustering
Cluster structures based on structural similarity

Structure Download
Download the structures in various formats

Pathways
Analyze pathways containing the compounds

Refine your results • What's this?

Chemical Properties
Rule of 5 (2)

BioActivity Experiments
BioAssays, Active (5)
BioAssays, Tested (5)
Protein 3D Structures (1)
⊞Structure Of Paracetamol Bound Bovine Lactoperoxidase At 2.45a Resolution (1)

BioMedical Annotation
Pharmacological Actions (7)
⊞Anti-Allergic Agents (2)
BioSystems (1)

Depositor Category
Biological Properties (8)
Chemical Vendors (7)
Journal Publishers (6)
NIH Molecular Libraries (5)

Find related data
Database: Select
Find items

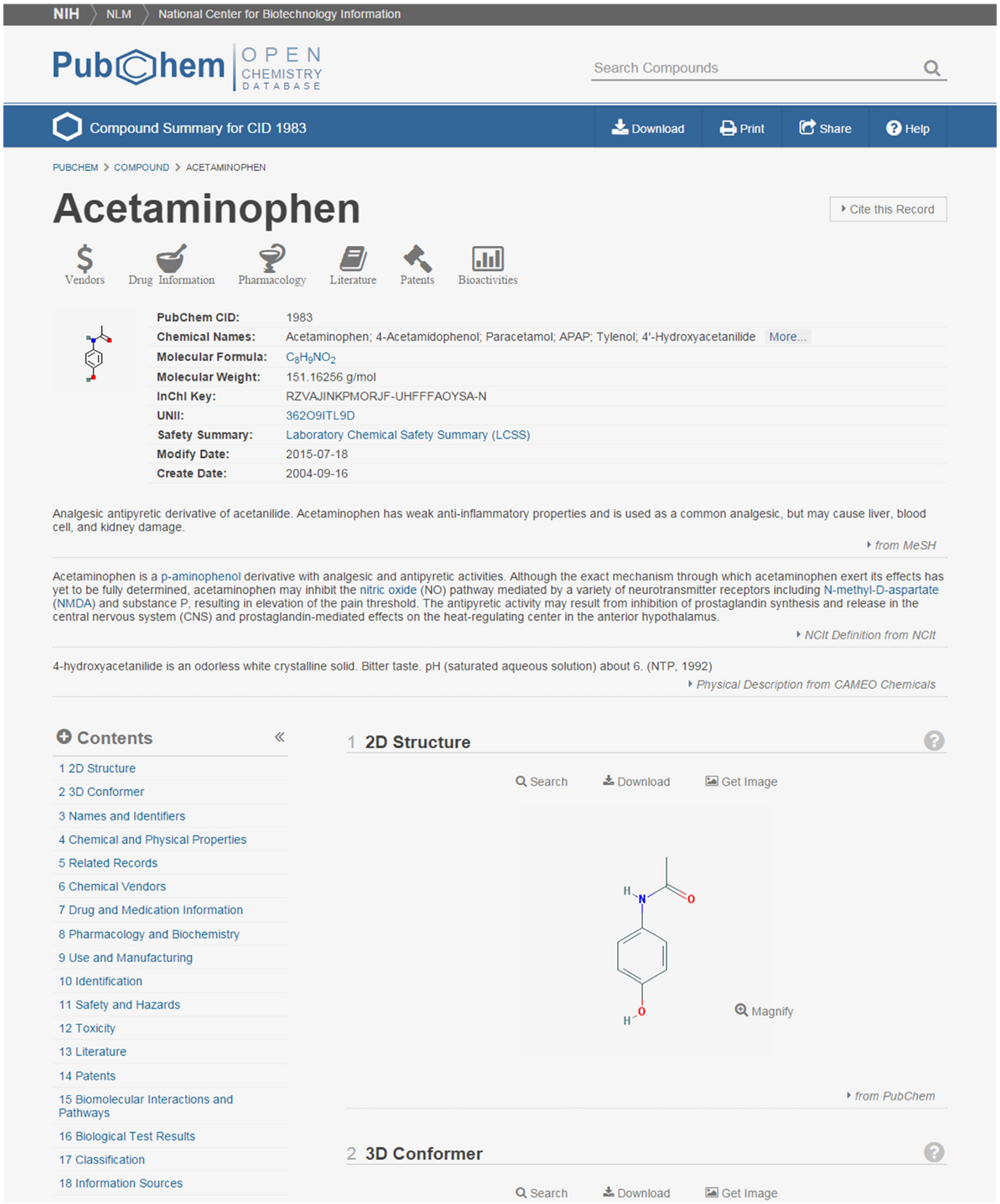
Search details
tylenol[All Fields]
Search
See more...

Recent activity
Turn Off Clear
tylenol (9)
PubChem Compound
See more...

Figure 3. A snapshot of the Document Summary (DocSum) page returned from an Entrez Search for 'tylenol' against the PubChem Compound database.

E-Utilities, used for programmatic access to information contained in the Entrez system (31,32), are suited for accessing text or numeric-fielded data, but cannot deal with more complex types of data specific to PubChem, such as chemical structures and tabular bioactivity data. Thus, PubChem provides additional programmatic access routes specialized for PubChem data: PUG, PUG-SOAP and PUG-REST.

PUG is a common gateway interface (CGI) available at https://pubchem.ncbi.nlm.nih.gov/pug/pug.cgi. It serves as the central gateway to several PubChem services. While suitable for low-level programmatic access to PubChem, PUG exchanges data through a complex XML schema that could require some expertise to use. For the sake of user-friendliness and for integration with a variety of third party tools, PubChem provides two easier-to-use web service ac-



Acetaminophen

PubChem CID:	1983
Chemical Names:	Acetaminophen; 4-Acetamidophenol; Paracetamol; APAP; Tylenol; 4'-Hydroxyacetanilide More...
Molecular Formula:	$C_8H_9NO_2$
Molecular Weight:	151.16256 g/mol
InChI Key:	RZVAJINKPMORJF-UHFFFAOYSA-N
UNII:	362O9ITL9D
Safety Summary:	Laboratory Chemical Safety Summary (LCSS)
Modify Date:	2015-07-18
Create Date:	2004-09-16

Analgesic antipyretic derivative of acetanilide. Acetaminophen has weak anti-inflammatory properties and is used as a common analgesic, but may cause liver, blood cell, and kidney damage.

from MeSH

Acetaminophen is a p-aminophenol derivative with analgesic and antipyretic activities. Although the exact mechanism through which acetaminophen exert its effects has yet to be fully determined, acetaminophen may inhibit the nitric oxide (NO) pathway mediated by a variety of neurotransmitter receptors including N-methyl-D-aspartate (NMDA) and substance P, resulting in elevation of the pain threshold. The antipyretic activity may result from inhibition of prostaglandin synthesis and release in the central nervous system (CNS) and prostaglandin-mediated effects on the heat-regulating center in the anterior hypothalamus.

NCIt Definition from NCIt

4-hydroxyacetanilide is an odorless white crystalline solid. Bitter taste. pH (saturated aqueous solution) about 6. (NTP, 1992)

Physical Description from CAMEO Chemicals

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 Pharmacology and Biochemistry
9 Use and Manufacturing
10 Identification
11 Safety and Hazards
12 Toxicity
13 Literature
14 Patents
15 Biomolecular Interactions and Pathways
16 Biological Test Results
17 Classification
18 Information Sources

1 2D Structure

from PubChem

2 3D Conformer

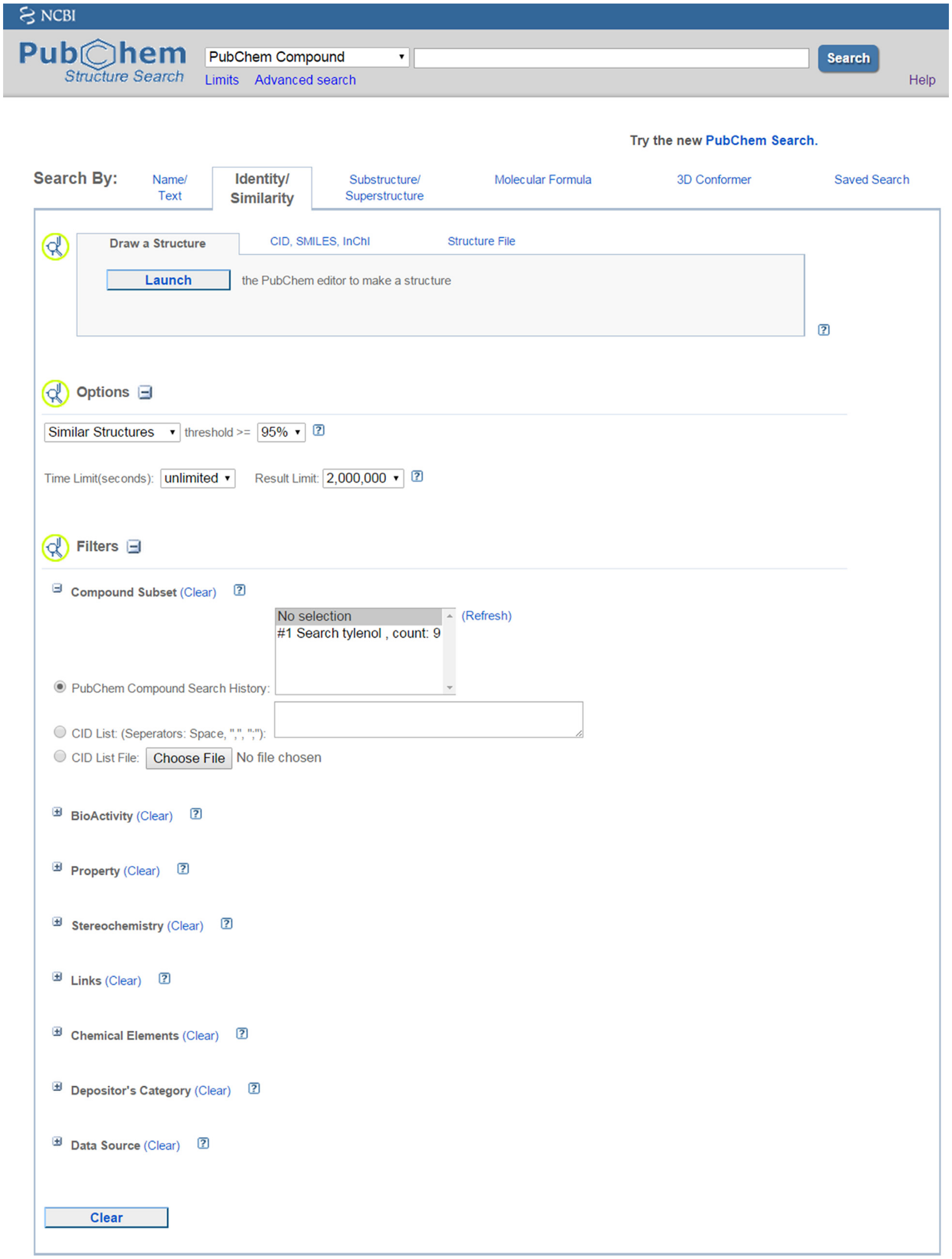
Figure 4. A snapshot of the top portion of the Compound Summary page for CID 1983 (Tylenol).

cess methods: PUG-SOAP, which uses the simple object access protocol (SOAP) (http://www.w3.org/TR/soap), and PUG-REST, which is a Representational State Transfer (REST)-like interface (33,34).

PUG-SOAP provides an easier programmatic access route to much of the same functionality as PUG. It breaks down operations into simpler functions, as defined via the web service definition language (WSDL; http://www.w3.org/TR/wsdl), using SOAP-formatted message envelopes for information exchange. PUG-SOAP is most suitable for SOAP-aware GUI workflow applications (e.g., Taverna and Pipeline Pilot) and most programming/scripting languages (e.g., C, C++, C#, .NET, Perl, Python and Java).

PUG-REST is the simplest programmatic access tool to use and learn because it does not require the overhead of XML and SOAP envelopes. Information necessary to make a PUG-REST request can be encoded into a single URL and readily incorporated into web pages or complex work flows. PUG-REST also provides convenient access to information on PubChem records not possible with the other PUG services.

Note that PubChem web services are not designed for very large volumes of requests (e.g., millions of requests). Users should limit programmatic web requests to no more than three per second in order to avoid overloading PubChem servers. Some programmatic access approaches are



PubChem Compound

Similar Structures threshold >= 95%

Time Limit(seconds): unlimited Result Limit: 2,000,000

● PubChem Compound Search History: No selection; #1 Search tylenol , count: 9

○ CID List: (Seperators: Space, ",", ";"):

○ CID List File: Choose File No file chosen

Figure 5. A snapshot of the Chemical Structure Search tool.

queued (e.g., PUG and PUG-SOAP). More detailed information on programmatic access to PubChem can be found in a paper recently published in the Nucleic Acids Research Web Server issue (21).

PubChemRDF

PubChemRDF (https://pubchem.ncbi.nlm.nih.gov/rdf/) (35) encodes PubChem's data using the Resource Description Framework (RDF) (http://www.w3.org/RDF/), which is a core part of Semantic Web standards. PubChemRDF harnesses ontological frameworks to help facilitate PubChem data sharing, analysis and integration with resources external to the NCBI and across scientific domains. In particular, the chemical and drug ontologies, [e.g., National Drug File—Reference Terminology (NDF-RT), National Cancer Institute (NCI) Thesaurus and ChEBI], are used to annotate PubChem compounds and substance; and the

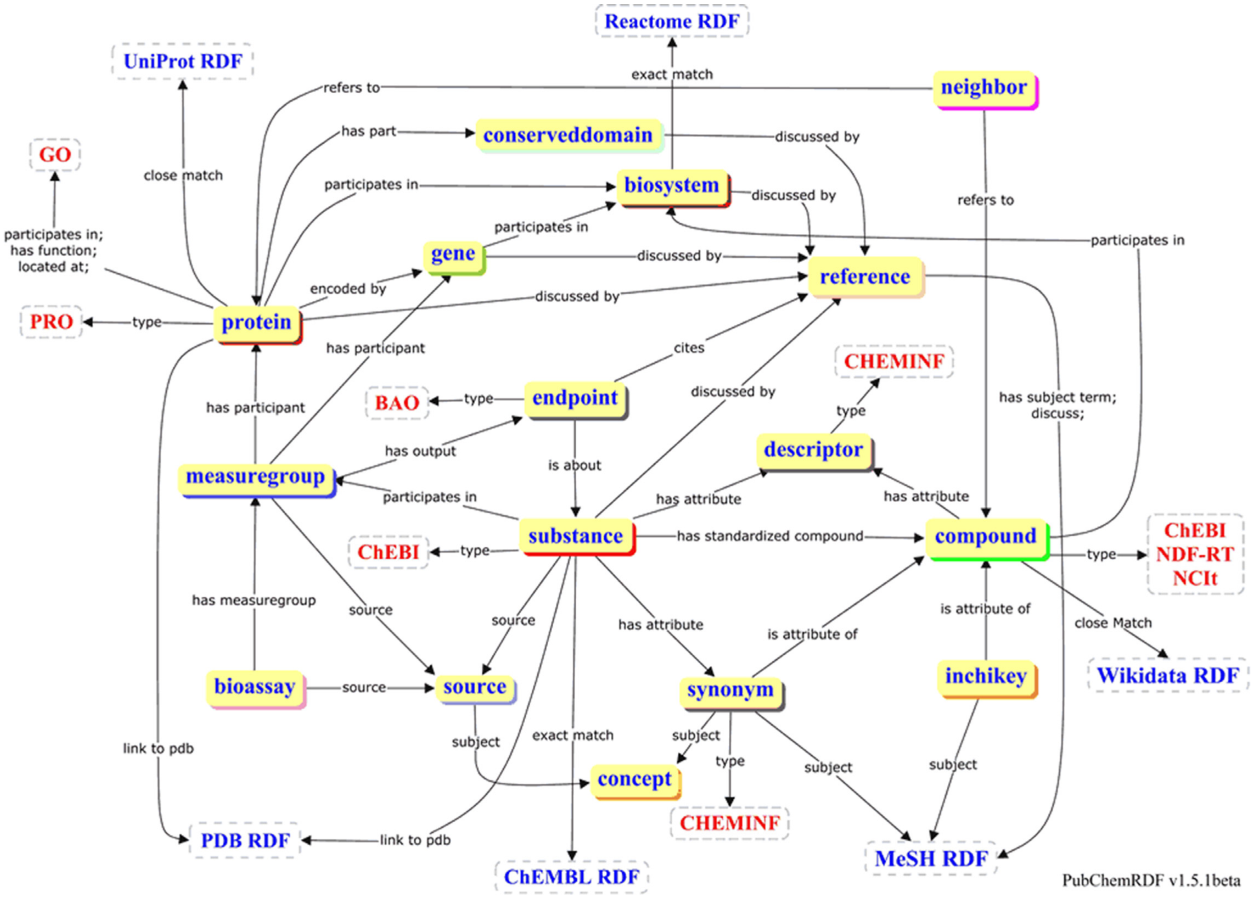


Figure 6. Diagram showing the high-level overview of PubChemRDF semantic relationships.

biomedical ontologies [e.g., Protein ontology (PRO) and gene (GO)], are used to annotate the bioassay molecular targets.

As shown in Figure 6, the PubChemRDF exposes a number of semantic relationships among compounds, substances, synonyms, bioassays, endpoints, proteins, genes, biosystems and so on. PubChemRDF enhances cross-integration by providing direct links to available authorative RDF-based resources within applicable subdomains, including reference, synonym and InChIKey (17,18) to MeSH RDF (http://id.nlm.nih.gov/mesh/); protein to UniProt RDF (https://www.ebi.ac.uk/rdf/services/uniprot/) (36); protein and substance to PDB RDF (http://pdbj.org/help/rdf) (37); biosystem to Reactome RDF (https://www.ebi.ac.uk/rdf/services/reactome/) (36); substance to ChEMBL RDF (https://www.ebi.ac.uk/rdf/services/chembl/) (36,38); and compound to WikiData RDF (https://meta.wikimedia.org/wiki/Wikidata/Development/RDF) (39).

PubChemRDF aims to help researchers work with PubChem data on local computing resources using Sematic Web technologies. The selected PubChemRDF data files in any subdomains can be downloaded from the PubChem FTP site, and imported into a RDF triple/quad store (such as Apache Jena TDB or OpenLink Virtuoso), which usually provides SPARQL query interface. Alternatively, PubChemRDF data can also be loaded into RDF-aware graph databases such as Neo4j, and the graph traversal algorithms can be used to query the PubChem knowledge graphs. In addition to bulk download via FTP, PubChemRDF also provides programmatic data access through a REST-full interface. In addition to dereferencing Uniform Resource Identifiers (URIs), the PubChemRDF REST-like interface provides simple SPARQL-like query capabilities for grouping and filtering relevant results.

PubChem3D

PubChem3D (https://pubchem.ncbi.nlm.nih.gov/release3d.html) (40–48) is an information resource derived from theoretical 3-D structures of molecules contained in the Compound database. PubChem generates a 3-D conformer model for each compound in PubChem, if the compound satisfies the following conditions: (i) neither too large nor too flexible (with ≤50 non-hydrogen atoms and ≤15 rotatable bonds) (ii) has only a single covalent unit (i.e., neither a salt nor mixture), (iii) consists of only supported elements (H, C, N, O, F, Si, P, S, Cl, Br and I), (iv) contains only atom types supported by the Merck Molecular Force Field (MMFF94s) and (v) has five or fewer undefined stereocenters. About 90% of compounds in PubChem satisfy these conditions and have computationally-generated conformer models, which contain up to 500 conformers for each compound.

PubChem3D delivers tools and services that use these conformer models, such as 3-D conformer search, 3-D structure clustering and web-based 3-D viewer. These tools and services exploit the notion of 3-D molecular similarity between conformers, which is quantified using the Gaussian-shape comparison method implemented in ROCS (25–29). PubChem3D pre-computes compounds similar to each applicable compound in PubChem in terms of 3-D similarity, and provides immediate access to these '3-D neighbors' as well as their respective superpositions.

SUMMARY

In the present paper, we described the PubChem Substance and Compound databases. The Substance database contains information submitted by individual data contributors. Through the standardization process, unique chemical structures are extracted from the Substance database and stored in the Compound database, which provides an aggregated view of information on a given chemical structure.

PubChem Upload allows for seamless data submission to PubChem, by providing streamlined procedures for data submission via an extensive set of wizards, inline help tips and templates. In addition to text-based search through Entrez, PubChem also enables users to perform various non-textual searches (such as identity search, molecular formula search, substructure/superstructure search, 2-D and 3-D similarity searches) using the Chemical Structure Search tool.

For efficient use of its data, PubChem provides various tools and services. Many of these exploit the notion of molecular similarity. Although widely used, fingerprint-based 2-D similarity methods often fail to detect structural similarity between molecules with similar biological activities, which can readily be detected by 3-D similarity methods. PubChem3D is an extension to PubChem data, which provides users with 3-D alternatives to existing 2-D similarity-based tools and services.

PubChem also provides robust programmatic access routes through PUG-REST, as well as through other programmatic protocols (such as E-Utilities, PUG and PUG-SOAP). In addition, PubChem provides its data in RDF to facilitate data sharing, analysis and integration with other databases. The PubChemRDF data are also programmatically accessible through a REST-full interface.

PubChem is committed to continue serving as a key chemical information resource not only to the biomedical research community but also to the scientific community as a whole. PubChem continues to adapt, developing new tools and services to exploit ever-advancing technologies. As opportunities arise, PubChem will continue to improve the breadth and depth of data.

ACKNOWLEDGEMENTS

Special thanks to the entire PubChem team and the NCBI staff. We also thank the hundreds of data contributors for making their data openly accessible within PubChem.

FUNDING

Intramural Research Program of the National Library of Medicine, National Institutes of Health (in part). Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler,RA and Spellmeyer,DC (eds). *Annual Reports in Computational Chemistry*. Elsevier, Amsterdam, Vol. **4**, pp. 217–241.
2. Wang,Y.L., Xiao,J.W., Suzek,T.O., Zhang,J., Wang,J.Y. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
3. Wang,Y.L., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek,T.O., Wang,J.Y., Xiao,J.W., Zhang,J. and Bryant,S.H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.
4. Wang,Y.L., Xiao,J.W., Suzek,T.O., Zhang,J., Wang,J.Y., Zhou,Z.G., Han,L.Y., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.
5. Wang,Y.L., Suzek,T., Zhang,J., Wang,J.Y., He,S.Q., Cheng,T.J., Shoemaker,B.A., Gindulyte,A. and Bryant,S.H. (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, D1075–D1082.
6. Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bourexis,D., Brister,J.R., Bryant,S.H., Canese,K. *et al.* (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
7. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
8. Liu,T.Q., Lin,Y.M., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
9. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y.F., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
10. Fonger,G.C., Hakkinen,P., Jordan,S. and Publicker,S. (2014) The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): background, recent enhancements and future plans. *Toxicology*, **325**, 209–216.
11. Dalby,A., Nourse,J.G., Hounshell,W.D., Gushurst,A.K.I., Grier,D.L., Leland,B.A. and Laufer,J. (1992) Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.*, **32**, 244–255.
12. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
13. Weininger,D., Weininger,A. and Weininger,J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101.
14. Weininger,D. (1990) SMILES. 3. Depict - graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.*, **30**, 237–243.
15. McNaught,A. (2006) The IUPAC International Chemical Identifier: InChI — A New Standard for Molecular Informatics. *Chem. Int.*, **28**, 12–14.
16. Frey,J.G. (2006) Using InChI. *Chem. Int.*, **28**, 14–15.
17. Heller,S., McNaught,A., Stein,S., Tchekhovskoi,D. and Pletnev,I. (2013) InChI - the worldwide chemical structure identifier standard. *J. Cheminform.*, **5**, 7.
18. Heller,S., McNaught,A., Pletnev,I., Stein,S. and Tchekhovskoi,D. (2015) InChI, the IUPAC International Chemical Identifier. *J. Cheminform.*, **7**, 23.

19. Ihlenfeldt,W.D. and Gasteiger,J. (1994) Hash codes for the identification and classification of molecular-structure elements. *J. Comput. Chem.*, **15**, 793–813.
20. Ihlenfeldt,W.D., Takahashi,Y., Abe,H. and Sasaki,S. (1994) Computation and management of chemical-properties in cactvs - an extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.*, **34**, 109–116.
21. Kim,S., Thiessen,P.A., Bolton,E.E. and Bryant,S.H. (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.*, **43**, W605–W611.
22. Holliday,J.D., Hu,C.Y. and Willett,P. (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen*, **5**, 155–166.
23. Chen,X. and Reynolds,C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **42**, 1407–1414.
24. Holliday,J.D., Salim,N., Whittle,M. and Willett,P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **43**, 819–828.
25. Grant,J.A. and Pickup,B.T. (1995) A Gaussian description of molecular shape. *J. Phys. Chem.*, **99**, 3503–3510.
26. Grant,J.A., Gallardo,M.A. and Pickup,B.T. (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, **17**, 1653–1666.
27. Grant,J.A. and Pickup,B.T. (1996) A Gaussian description of molecular shape (vol 99, pg 3505, 1995). *J. Phys. Chem.*, **100**, 2456–2456.
28. Grant,J.A. and Pickup,B.T. (1997) Gaussian shape methods. In: van Gunsteren,WF, Weiner,PK and Wilkinson,AJ (eds). *Computer Simulation of Biomolecular Systems*. Kluwer Academic Publishers, Dordrecht, pp. 150–176.
29. Hawkins,P.C.D., Skillman,A.G. and Nicholls,A. (2007) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, **50**, 74–82.
30. Ihlenfeldt,W.D., Bolton,E.E. and Bryant,S.H. (2009) The PubChem chemical structure sketcher. *J. Cheminform.*, **1**, 9.
31. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
32. McEntyre,J. (1998) Linking up with Entrez. *Trends Genet.*, **14**, 39–40.
33. Fielding,R.T. and Taylor,R.N. (2000) Principled design of the modern Web architecture. In: *Proceedings of the 22nd international conference on Software engineering*. Limerick, pp. 407–416.
34. Fielding,R.T. (2000) Representational State Transfer (REST). In: *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine.
35. Fu,G., Batchelor,C., Dumontier,M., Hastings,J., Willighagen,E. and Bolton,E. (2015) PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *J. Cheminform.*, **7**, 34.
36. Jupp,S., Malone,J., Bolleman,J., Brandizi,M., Davies,M., Garcia,L., Gaulton,A., Gehant,S., Laibe,C., Redaschi,N. *et al.* (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, **30**, 1338–1339.
37. Kinjo,A.R., Suzuki,H., Yamashita,R., Ikegawa,Y., Kudou,T., Igarashi,R., Kengaku,Y., Cho,H., Standley,D.M., Nakagawa,A. *et al.* (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.*, **40**, D453–D460.
38. Willighagen,E.L., Waagmeester,A., Spjuth,O., Ansell,P., Williams,A.J., Tkachenko,V., Hastings,J., Chen,B. and Wild,D.J. (2013) The ChEMBL database as linked open data. *J. Cheminform.*, **5**, 12.
39. Erxleben,F., Günther,M., Krötzsch,M., Mendez,J. and Vrandečić,D. (2014) Introducing Wikidata to the Linked Data Web. In: Mika,P, Tudorache,T, Bernstein,A, Welty,C, Knoblock,C, Vrandečić,D, Groth,P, Noy,N, Janowicz,K and Goble,C (eds). *The Semantic Web – ISWC 2014*. Springer International Publishing, Cham, Vol. **8796**, pp. 50–65.
40. Bolton,E.E., Kim,S. and Bryant,S.H. (2011) PubChem3D: conformer generation. *J. Cheminform.*, **3**, 4.
41. Bolton,E.E., Kim,S. and Bryant,S.H. (2011) PubChem3D: diversity of shape. *J. Cheminform.*, **3**, 9.
42. Bolton,E.E., Kim,S. and Bryant,S.H. (2011) PubChem3D: similar conformers. *J. Cheminform.*, **3**, 13.
43. Kim,S., Bolton,E.E. and Bryant,S.H. (2011) PubChem3D: shape compatibility filtering using molecular shape quadrupoles. *J. Cheminform.*, **3**, 25.
44. Kim,S., Bolton,E.E. and Bryant,S.H. (2011) PubChem3D: biologically relevant 3-D similarity. *J. Cheminform.*, **3**, 26.
45. Bolton,E.E., Chen,J., Kim,S., Han,L., He,S., Shi,W., Simonyan,V., Sun,Y., Thiessen,P.A., Wang,J. *et al.* (2011) PubChem3D: a new resource for scientists. *J. Cheminform.*, **3**, 32.
46. Kim,S., Bolton,E. and Bryant,S. (2012) Effects of multiple conformers per compound upon 3-D similarity search and bioassay data analysis. *J. Cheminform.*, **4**, 28.
47. Kim,S., Bolton,E.E. and Bryant,S.H. (2013) PubChem3D: conformer ensemble accuracy. *J. Cheminform.*, **5**, 1.
48. Kim,S., Han,L., Yu,B., Hähnke,V.D., Bolton,E.E. and Bryant,S.H. (2015) PubChem structure-activity relationship (SAR) clusters. *J Cheminform*, **7**, 33.