Web-accessible Chemical Compound Information

Dana L. Roth

ABSTRACT. Web-accessible chemical compound information resources are widely available. In addition to fee-based resources, such as SciFinder Scholar and Beilstein, there is a wide variety of freely accessible resources such as ChemSpider and PubChem. The author provides a general description of various fee-based and free chemical compound resources. The free resources generally offer an acceptable alternative to fee-based resources for quick retrieval. It is assumed that readers will be familiar with *The Merck Index*, *Handbook of Chemistry and Physics,* and *Knovel Critical Tables*.

KEYWORDS. Chemistry sources, chemical information sources, chemical structure searching, PubChem, SciFinder Scholar, ChemSpider, Beilstein CrossFire, DiscoveryGate for Academics, SPRESIweb, Combined Chemical Dictionary, Wikipedia, ChemFinder.com, e-Molecules, SIS Chemical Information portal, Chemical Structure Lookup Service

AUTHOR. Dana L. Roth (dzrlib@library.caltech.edu) is Chemistry Librarian at Millikan Library at California Institute of Technology (Caltech), 1200 E. California Blvd. Pasadena, CA 91125.

INTRODUCTION

Chemical compound information resources are widely available on the World Wide Web. In addition to fee-based resources, there is a wide variety of freely accessible resources and both types are described below. The fee-based resources are curated more accurately, and generally, are linked to the peer-reviewed literature.

Many of the freely available resources are "impure," in that they are derived from a variety of secondary sources (e.g. chemical supplier catalogs and research laboratory databases). These free resources, however, are a generally accepted alternative to fee-based resources for quick retrieval of structure diagrams and basic physical properties. This article describes the features of both fee-based and freely accessibly resources of chemical compound information.

FEE-BASED RESOURCES

Fee-based resources for chemical compound information include SciFinder/SciFinder Scholar, Beilstein CrossFire, DiscoveryGate for Academics, SPRESIweb, and Combined Chemical Dictionary.

Chemical Abstract Service (CAS) SciFinder/SciFinder Scholar

<http://www.cas.org/SCIFINDER/SCHOLAR/index.html >

SciFinder/SciFinder Scholar[1,2] includes the Chemical Abstracts Service (CAS) Regfile, which contains both descriptive (chemical names, CAS Registry Numbers, synonyms, and structural diagrams) and both fully referenced experimental information (physical constants; spectra; lethal dose 50% (LD50), etc.), and Available Chemicals Directory (ACD) predicted (Lipinski values, Koc, etc.) property data. SciFinder/SciFinder Scholar can be searched with a

chemical name, synonym, CAS Registry Number, molecular formula, or a partial or complete structural diagram.

The more than 34 million compounds and more than 60 million biological sequences featured in the Regfile are linked to Chemical Abstracts (abstracts of research articles, patents, etc.); CASREACT (database of organic/organometallic reactions); CHEMCATS (database of chemicals and their suppliers); and CHEMLIST (database of regulated compounds and associated information). CAS Registry Numbers are also used extensively in the retrieval of chemical compound information from the National Library of Medicine's TOXNET and PubMed databases.

It is very important to recognize that CAS has a greatly expanded concept of chemistry, which includes the chemical aspects of astronomy, biology, education, engineering, economics, geology, history, mathematics, medicine, and physics. The broadness of this coverage is exemplified by the fact that organic, inorganic, physical, and analytical chemistry combined comprise only 36% of the over 25 million records (with over 1 million currently being added each year). Biochemistry, biological chemistry, and medicinal chemistry comprise about 34%, while applied chemistry (chemical engineering, materials science and engineering (including polymers), and environmental engineering make up 30%. In addition, Chemical Abstracts uniquely indexes a wide variety of publication formats. Articles from journals and regularly published conference proceedings account for 73% of the records, articles from one-time or first-time conference proceedings (7%), dissertations (2%), technical reports (1%), patents (16%), and books/edited research monograph chapters (1%).

Beilstein CrossFire - <http://www.beilstein.com/>

Beilstein CrossFire[3] is an organic compound and reaction database that complements the CAS Regfile. It offers the same searching features and is fairly comprehensive for organic compounds known between 1779 and 1980. Beilstein lists approximately 10 million compounds and 10 million reactions. Since 1980, compound selection is limited to about 180 chemistry journals that report organic synthetic techniques. This is in contrast with the approximately 9,000 journals indexed by CAS for SciFinder/SciFinder Scholar.

DiscoveryGate for Academics -

<http://www.mdl.com/solutions/solutions_for/academics/dg_academics.jsp>

DiscoveryGate for Academics[4] is a multi-database system of more than 27 million structures, 17 million reactions, and 500 million calculated and reported physical properties, and also provides access to Beilstein. Its MDL Compound Index serves as a structure searchable index for twenty databases (bioactivity, chemical sourcing, and synthetic methodology), with thirteen of these databases being separately searchable with structures and additional text terms.

SPRESIweb - <http://www.spresiweb.com>

SPRESIweb[5] is a structure and reaction database containing 6 million compounds and about 4 million reactions from 627,000 references including 164,000 patents. SPRESIweb offers both text and structure searching, with links to chemical supplier catalogs. SPRESIweb version 2.5 covers literature from 1974 to 2005.

Combined Chemical Dictionary (CCD) - <http://www.chemnetbase.com/scripts/ccdweb.exe>

The Combined Chemical Dictionary[6] includes all the compounds contained in *Dictionary of Organic Compounds* (266,000), *Dictionary of Inorganic/Organometallic Compounds* (102,000), *Dictionary of Natural Products* (181,000), *Dictionary of Drugs* (45,000), and *Dictionary of Analytical Reagents* (14,000). The CCD provides descriptive and numerical data on chemical, physical, and biological properties; systematic and common names; literature references, structural diagrams (for example, click on the "Benzene" ring); and connection tables (for structure searching). The browse feature provides a ranked list of each data set. In the Molecular Formula dictionary, there are sub-sets for "all metals" and "all nonmetals" as well as all compounds with a given element (e.g., "all compounds with Technetium" would be expressed as "ALL-Tc") and a variety of other element combinations.

The compounds selected include fundamental organic and inorganic compounds; virtually every known natural product; all currently marketed drugs; compounds with an established use (e.g. catalysts, solvents, reagents, etc.); important coordination compounds; organometallic compounds representative of all important structural types; important biochemicals and minerals; and miscellaneous compounds of active research interest.

FREELY ACCESSIBLE RESOURCES

Google and Yahoo are very effective search engines for locating many Web resources. However, their inability to accommodate structure searching, coupled with the widespread use of synonyms for chemical names, generally precludes their use for comprehensive searching of chemical compound information. Freely accessible resources described in this section include Wikipedia, ChemFinder, e-Molecules, ChemSpider, PubChem, Chemical Structure Lookup Service, and SIS Chemical Information portal.

Wikipedia **-** <http://en.wikipedia.org>

Wikipedia lists a variety of widely known chemical elements and compounds and is useful as a popular information resource. One example is the article on Aspirin <http://en.wikipedia.org/wiki/Aspirin>, which displays the chemical structure diagram (see Figure 1), and provides fairly extensive descriptions of its history, trademark issues, synthesis, therapeutic uses, adverse effects, etc., and includes an extensive list of literature references. Aspirin is one example of the chemical elements and compounds listed on the WikiProject Chemicals/Organization Web page <http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Chemicals/Organization>.

[INSERT FIGURE 1 HERE]

LEGEND: FIGURE 1. Wikipedia Entry - Aspirin

ChemFinder.com **-** <http://chemfinder.cambridgesoft.com/>

ChemFinder[7,8] indexes a wide variety of unusual Web resources that are not covered in the traditional databases (see Figure 2). While focusing on access to health and safety data, both physical property data and Material Safety Data Sheets (MSDS) sources are also provided. Biological macromolecules are not covered since they are well indexed in GenBank <http://www.ncbi.nlm.nih.gov/Genbank/> and Protein Databank <http://www.rcsb.org/>, for example.

Text searching for specific chemicals is facilitated by a proprietary normalization scheme for chemical names. This process strips chemical names of spaces, parentheses, and other

punctuation, as well as accommodating British/American spelling differences, inverting Chemical Abstracts Index Names, and including a variety of chemical synonyms. The ChemFinder interface also allows searching with structural diagrams, CAS Registry Numbers, and ranges of molecular formulas, molecular weights, boiling points, and melting points.

Each substance record provides a property listing, including comments on color/odor/sensitivity and natural occurrence, if available. This is followed by a listing of Web sites (e.g. NIST Chemistry WebBook) which contain additional information (e.g. spectra). ChemFinder.com provides links to ChemACX Net (a supplier database), a 3-dimensional model view, and CAS Registry Number links to both *The Merck Index* (if subscribed) and to the CambridgeSoft version of the National Cancer Institute's Developmental Therapeutics Program (DTP) database that includes screening results and chemical structural data for cancer and AIDS treatments.

[INSERT FIGURE 2 HERE]

LEGEND: FIGURE 2. ChemFinder Example - Benzene

e-Molecules - <http://www.emolecules.com/>

E-Molecules[9] is a freely accessible, and regularly updated, database of 8 million unique chemical structures related to 19 million records from 150 suppliers, including more than 4 million commercially available screening compounds, building blocks, and pharmaceutical intermediates.

E-Molecules can be searched by drawing chemical structures or substructures using ISIS/ Draw, ChemDraw, ChemSketch, or Java Molecule Editor (JME). It links to publicly-available

databases include NIST Chemistry WebBook, National Cancer Institute, DrugBank, and PubChem, which provide spectra, physical properties, and biological data. Links to proprietary fee-based sources allow viewing and purchasing of chemical spectra, as well as price quotes for chemicals from suppliers.

[INSERT FIGURE 3 HERE]

LEGEND: FIGURE 3. e-Molecules Example - Benzene

ChemSpider - <http://www.chemspider.com>

ChemSpider[10,11] is a database of more than 18 million chemical structures from a wide variety of both free and fee-based proprietary sources of literature data, chemical vendor catalogs, molecular properties, environmental, toxicity, and analytical data.[12] There are over 80 data sources including, for example, ChemIDplus, Environmental Protection Agency's Distributed Structure-Searchable Toxicity (DSSTox), DiscoveryGate, Developmental Therapeutics Program (DTP) of the National Cancer Institute, Food and Drug Administration (FDA), National Institute of Standards and Technology (NIST), PubChem, and Thomson Pharma. A complete list is available at <http://www.chemspider.com/DataSources.aspx>.

ChemSpider offers text-based, chemical structure, and physical property searching (see Figure 4). The default is text-based (Systematic Names, Synonyms, Trade Names, Registry Numbers, SMILES (typographical representation of molecules with letters and numbers), or International Chemical Identifier). A ChemSpider textword search "add-in" is available for Internet Explorer (version 7.0) and Firefox browsers.

The advanced searching feature offers searching with structures (exact or substructure), identifier, elements, properties, calculated properties, data source, and similarity (singly or in combination). ChemSpider also provides an "input text to convert to molecular structure" program <http://www.chemspider.com/Services.aspx?>. Search results include a structural diagram, basic properties (molecular formula, molecular weight) and identifiers; data sources; an extensive list of chemical names, database IDs and synonyms; ACD predicted physical property and Lipinski values; and a small, but growing, number of spectra (nuclear magnetic resonance**,** carbon nuclear magnetic resonance, near infrared). ChemSpider has enhanced structural diagrams with an optional 3-dimensional molecule viewer (Jmol) display and provides Web links for many of its data sources. A ChemSpider blog provides a running commentary on new developments and improvements. ChemSpider manuals, technical notes, and newsletters are also offered.

[INSERT FIGURE 4 HERE]

LEGEND: FIGURE 4. ChemSpider – Inherent Properties, Identifiers, and References for Benzene

PubChem <http://pubchem.ncbi.nlm.nih.gov/>

PubChem[4,12] is a component of the National Institutes of Health's Molecular Libraries Roadmap Initiative, and is integrated with Entrez, the primary search engine for National Center for Biotechnology Information (NCBI). PubChem is designed to provide substance information, chemical structures, and bioactivity data for small molecules (having a molecular weight that is less than 500 Daltons). PubChem currently contains more than 38 million substances, 18 million

compounds, and 800 bioassays, and is comprised of three linked databases: PubChem Compound (unique structures); PubChem Substance (deposited structures); and PubChem Bioassay.

PubChem Compound is a searchable database of chemical structures with validated name/structure/property information. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. Searching options include:

- *Molecule Name Searches* (e.g., Tylenol, Benzene) allow searching with a variety of chemical synonyms (see Figure 5).
- *Chemical Property Range Searches* (e.g. molecular weight between 100 and 200 Daltons, or Hydrogen Bond Acceptor Count between three and five) 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 contains descriptions of chemical samples, from a wide variety of sources, with links to PubMed citations, protein 3-dimensional structures, and biological screening results available in PubChem BioAssay. Substances with known content are linked to PubChem Compound. Search options include:

- Molecule Synonym Searches (e.g. all substances with “deoxythymidine” as a name fragment, or substances that contain 3'-Azido-3'-deoxythymidine); (see Figure 6).
- 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).

[INSERT FIGURE 5 HERE]

LEGEND: FIGURE 5. PubChem Compound – Molecule Name Search - Tylenol

[INSERT FIGURE 6 HERE]

LEGEND: FIGURE 6. PubChem Substance – Molecule Synonym Search - Deoxythymidine

[INSERT FIGURE 7 HERE]

LEGEND: FIGURE 7. PubChem Substance Summary - Thymidine

PubChem BioAssay is a searchable database containing bioactivity screens of chemical substances described in PubChem Substance. Searchable descriptions of each bioassay are provided that include descriptions of procedural conditions and readouts. Options include searching for BioAssay Data Sets (e.g. HIV growth inhibition), and browsing or downloading PubChem BioAssay Results (National Cancer Institute Antiviral Assay).

Chemical Structure Lookup Service (CSLS) - <http://cactus.nci.nih.gov/cgi-bin/lookup/search>

The CSLS searches a database of more than 40 million compounds from 80 (commercial and public) databases. The interface accepts a variety of input formats (e.g., SMILES strings, International Chemical Identifier, structure data (SD) files, drawn structures). For example, from the search page, draw a structure and click “transfer” (which converts the structure to a SMILES

string). A listing of database categories is given with the option to specify individually or accept the "All" default. Clicking "search" retrieves a listing of all appropriate database links. Each database, in addition to listing the chemical identifiers, also may have unique properties and/or information. Examples of the database categories include:

- Bioactivity screening databases.
- Compounds claimed/mentioned in patents.
- Drugs or compounds in drug development.
- Imaging/contrast agent databases.
- Ligand/binding/crystal-structure databases.
- Natural products.

SIS – Specialized Information Services/Chemical Information -

<http://sis.nlm.nih.gov/chemical.html>

The SIS Chemical Information portal[13,14] defaults to ChemIDplus,[15] a searchable database of nearly 400,000 chemicals. The portal has two flavors:

- ChemIDplus Lite (for searching with chemical names or CAS Registry Numbers).
- ChemIDplus Advanced (for searching with chemical names, CAS Registry Numbers, molecular formulae, structures, and physical/toxicological data.

Individual chemical compound records in ChemIDplus provide basic information (e.g., names and synonyms, toxicity, physical properties) and link to records in the full range of NLM's TOXNET databases,[16] plus PubChem, PubMed, TOXLINE, and TOXMAP, as well as other U.S. government-sponsored databases (e.g., EPA Envirofacts, Syracuse Biodeg File) and a

SuperList Locator (e.g., California Proposition 65, NTP Report on Carcinogens, Workplace Hazardous Materials).

CONCLUSION

The wide variety of chemical compound databases is both a blessing and a curse. Each database has unique strengths and weaknesses, which should be determined on an individual basis. It is important for librarians to not only be aware of these databases, but also to experiment with them to search for common compounds (e.g., Benzene) and compare and contrast the results. Routine use of multiple databases by researchers is strongly recommended, especially those that are freely available, since these databases are in a constant flux.

Received: February 14, 2008

Revised: March 1, 2008

Accepted: March 3, 2008

REFERENCES

1. Meier, J.J. "SciFinder Scholar." *CHOICE: Current Reviews for Academic Libraries* 44, no. 12(August 1, 2007): 2078-9.

2. Bolek, A. "SciFinder Scholar." *Issues in Science and Technology Librarianship* 50 (2000), Available: <http://www.istl.org/00-fall/databases.html>. Accessed: March 3, 2008.

3. Meehan, P. and Schofield, H. "CrossFire: A Structural Revolution for Chemists." *Online Information Review* 25, no. 4(2001): 241-9.

4. Baykoucheva, S. "A New Era in Chemical Information: PubChem, DiscoveryGate, and Chemistry Central." *Information Today* 31, no. 5(September/October 2007): 16.

Available: <http://www.infotoday.com/online/sep07/Baykoucheva.shtml>. Accessed: March 3, 2008.

5. Roth, D.L. "SPRESIweb 2.1: A Selective Chemical Synthesis and Reaction Database." *Journal of Chemical Information and Modeling* 45, no. 5 (2005): 1470-3.

6. Lafferty, M. "ChemNetBase." *Issues in Science and Technology Librarianship* vol. 50 (2007). Available: <http://www.istl.org/07-spring/electronic3.html>. Accessed: March 3, 2008.

7. Svee, M. "ChemFinder." *ChemBioNews* 11, no. 2(2001). Available: <http://chembionews.cambridgesoft.com/Articles/Default.aspx?articleID=437>. Accessed: March 3, 2008.

8. Roth, Dana L. "ChemINDEX Database: The Professional's Version of ChemFinder.com." *ChemBioNews* 13, no. 1(2003). Available: <http://caltechlib.library.caltech.edu/54/>. Accessed: March 3, 2008.

9. Wikipedia. "eMolecules." Available: <http://en.wikipedia.org/wiki/EMolecules>. Accessed: March 3, 2008.

10. Wikipedia. "ChemSpider." Available: <http://en.wikipedia.org/wiki/ChemSpider>. Accessed: March 3, 2008.

11. Depth-First (Thirty-Two Free Chemistry Databases). Available: <http://depth-first.com/articles/2007/01/24/thirty-two-free-chemistry-databases>. Accessed: March 3, 2008.

12. Wikipedia. "PubChem." Available: <http://en.wikipedia.org/wiki/PubChem>. Accessed: March 3, 2008.

13. National Library of Medicine. Fact Sheet - Division of Specialized Information Services. Available: < <http://www.nlm.nih.gov/pubs/factsheets/sis.html>. Accessed: March 3, 2008.

14. Brown, M.C. "Chemical Information: SIS - Specialized Information Services." *CHOICE: Current Reviews for Academic Libraries* 44, no. 5(2007): 859.

15. National Library of Medicine. Fact Sheet – ChemIDplus. Available: <http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html>. Accessed: March 3, 2008.

16. National Library of Medicine. Fact Sheet - Toxicology and Environmental Health Information Program. Available: <http://www.nlm.nih.gov/pubs/factsheets/tehipfs.html>. Accessed: March 3, 2008.

Figure 1 – Wikipedia Entry - Aspirin

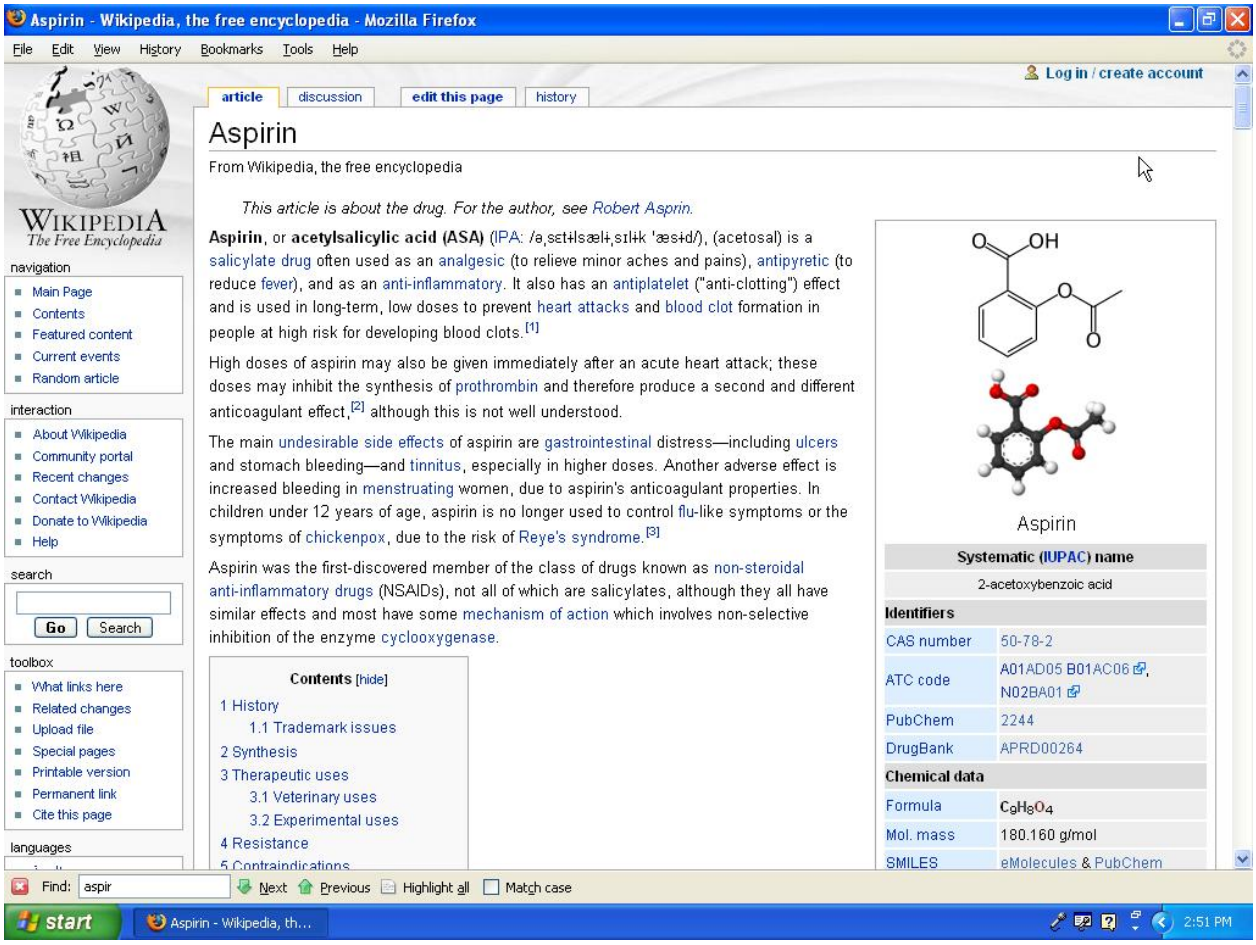

Figure 2 – ChemFinder - Benzene

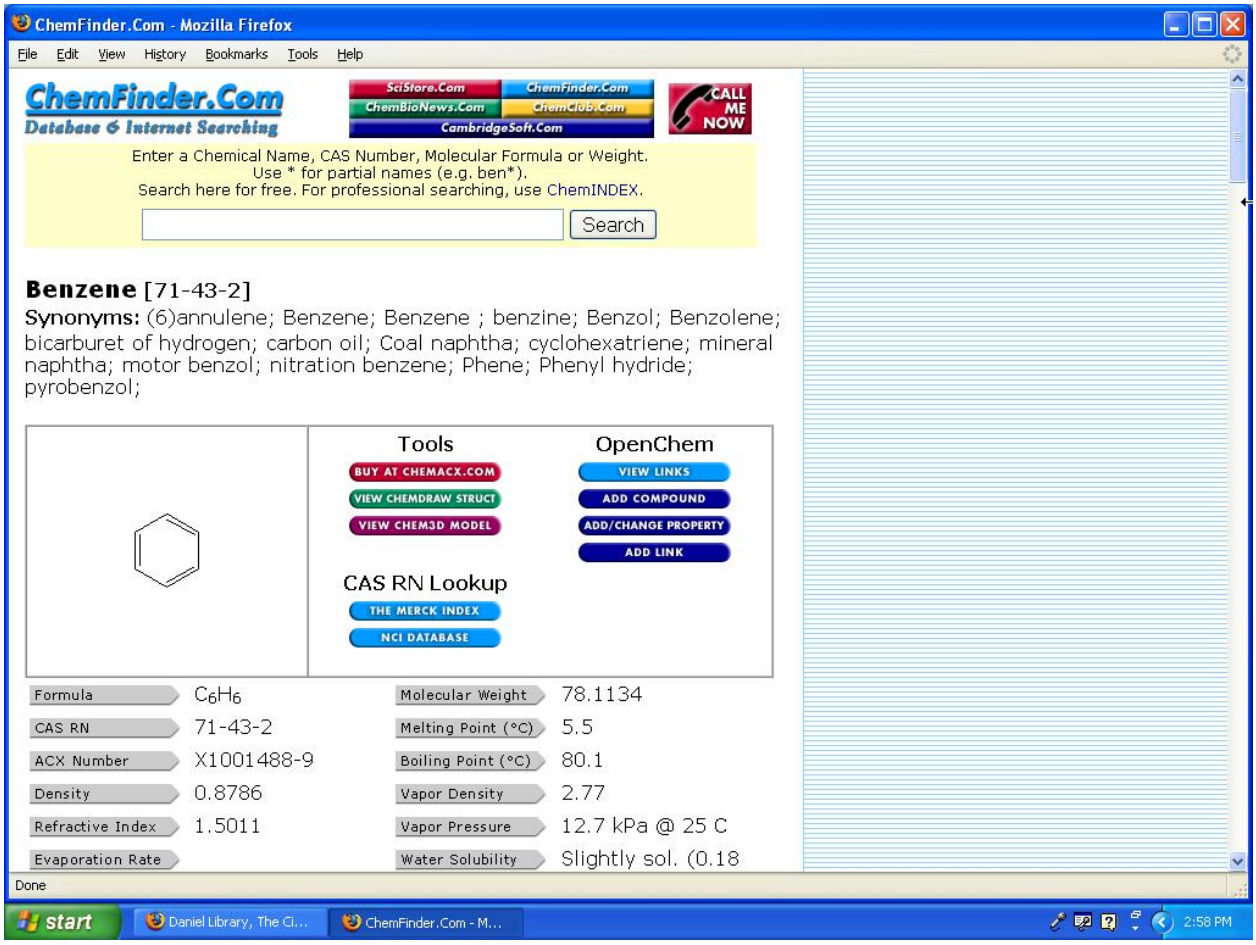

Figure 3 – eMolecules - Benzene

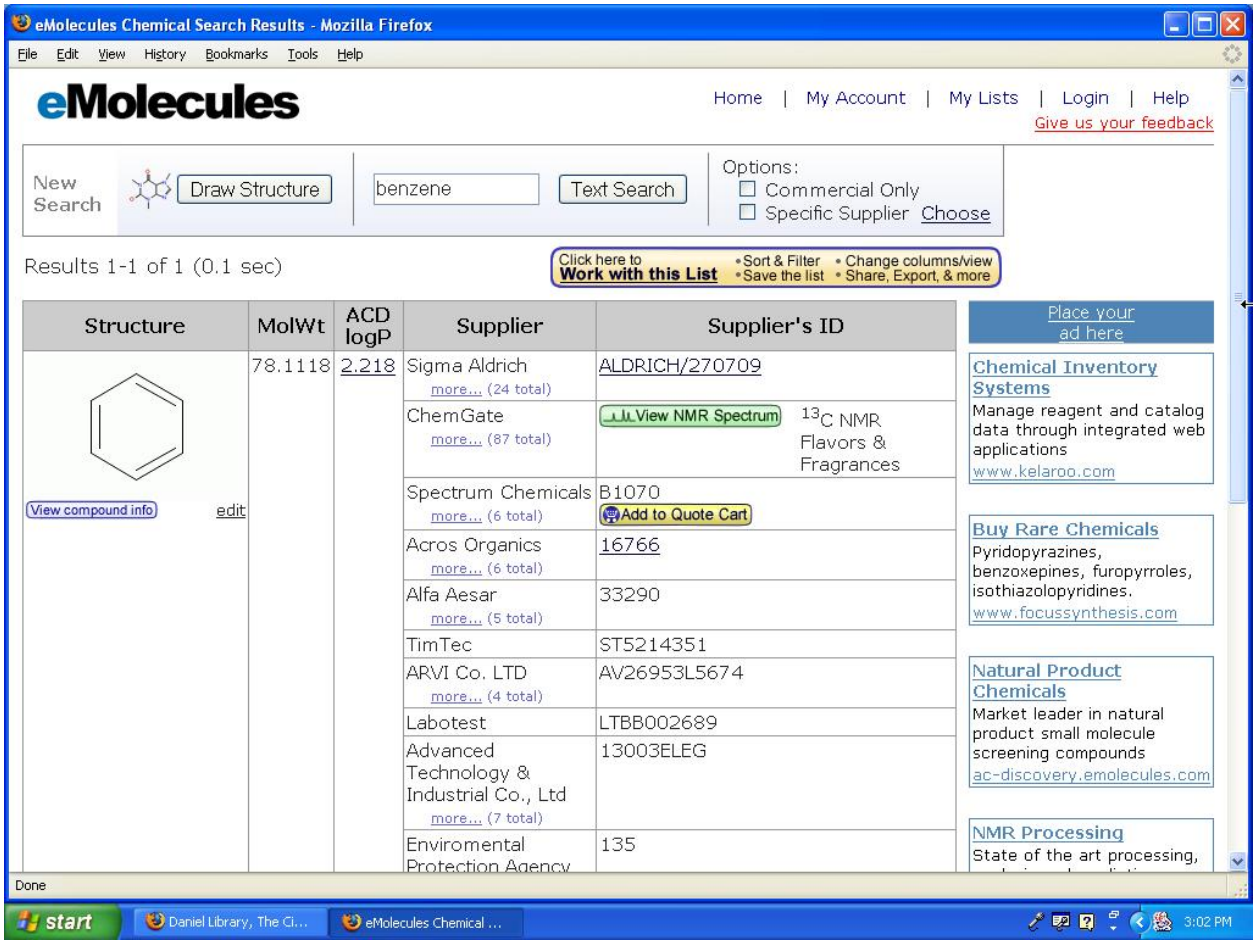

Figure 4 – ChemSpider – Inherent Properties, Identifiers, and References for Benzene

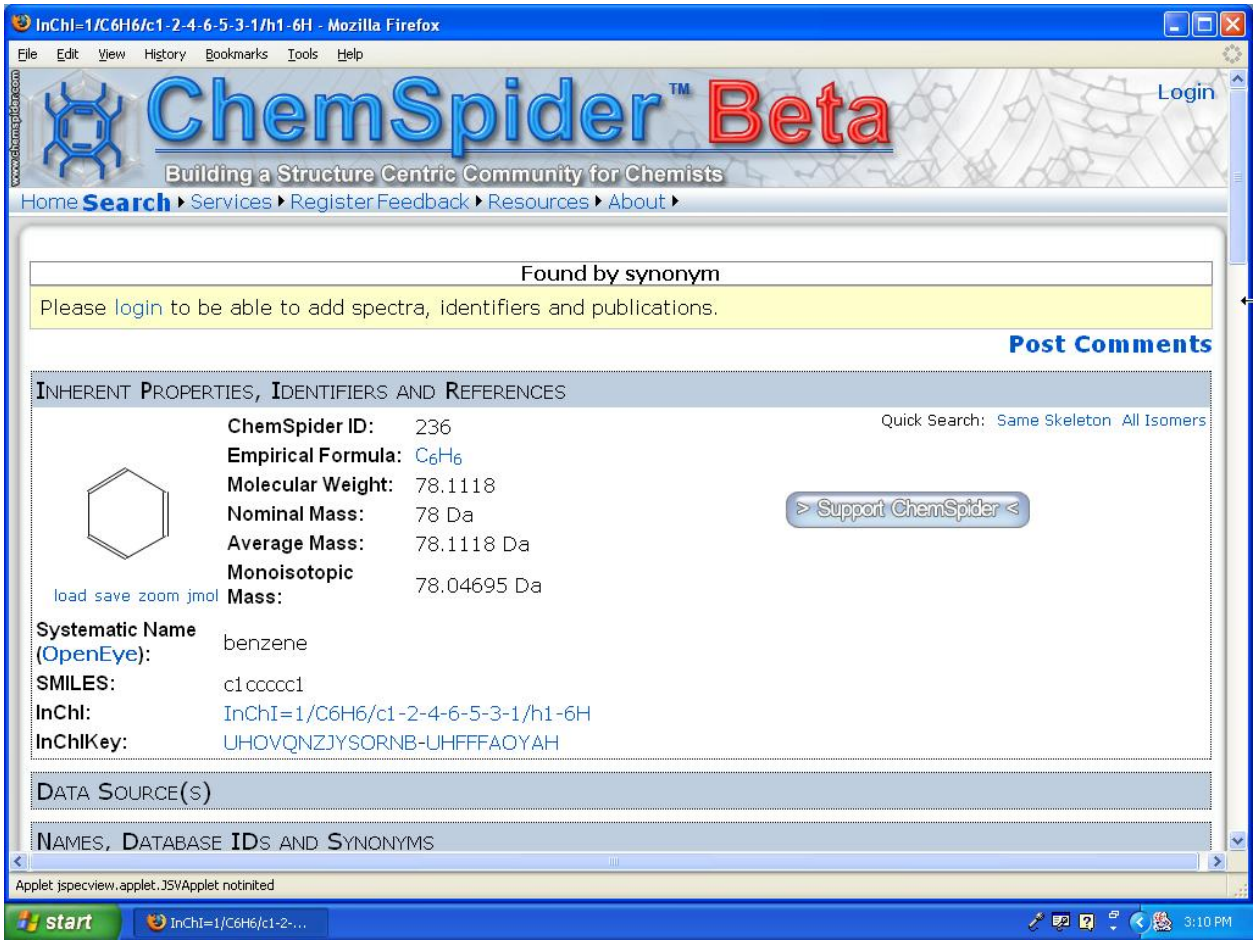

Figure 5 – PubChem – Tylenol (Acetaminophen)

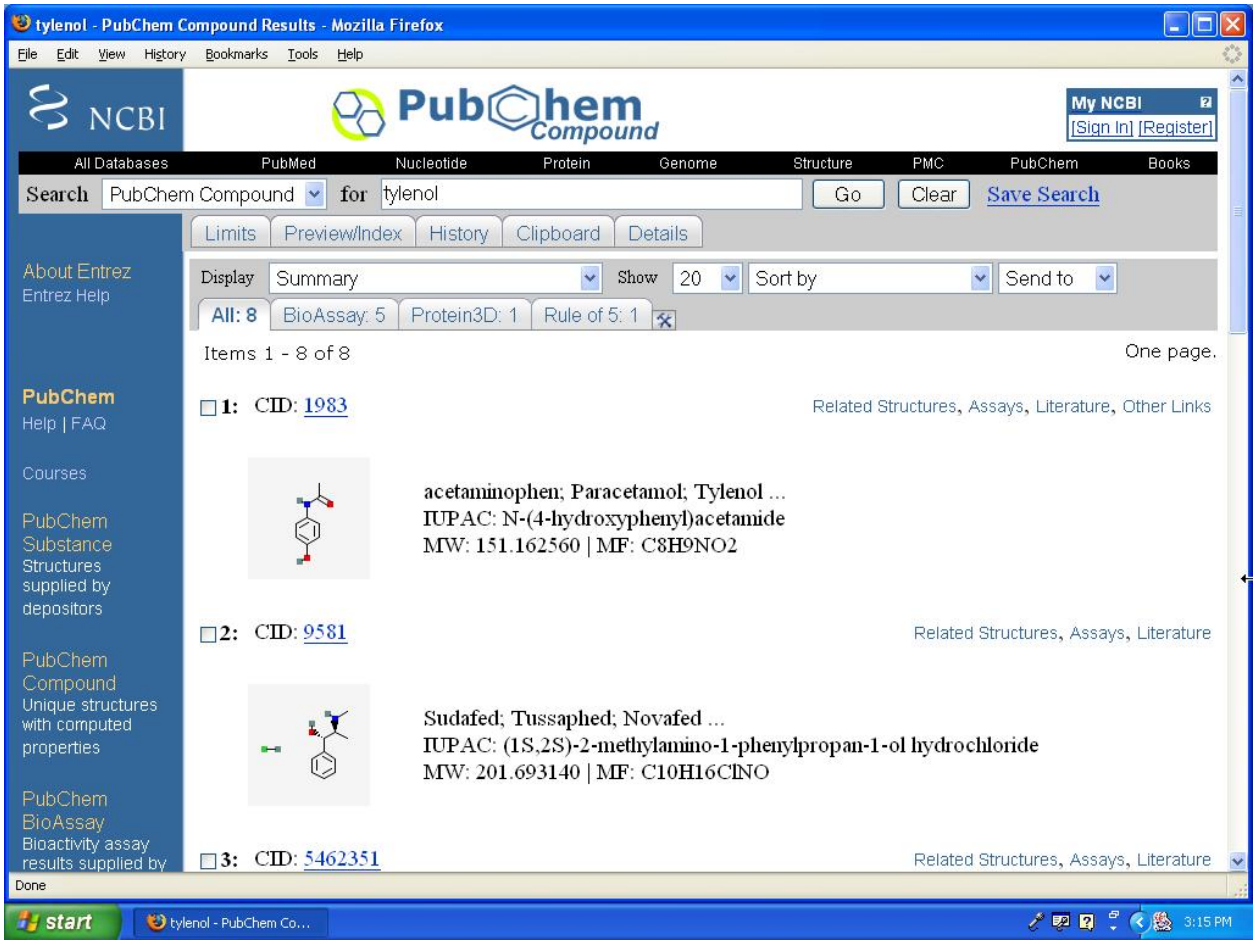

Figure 6 – PubChem – Molecule Synonym Search - Deoxythymidine

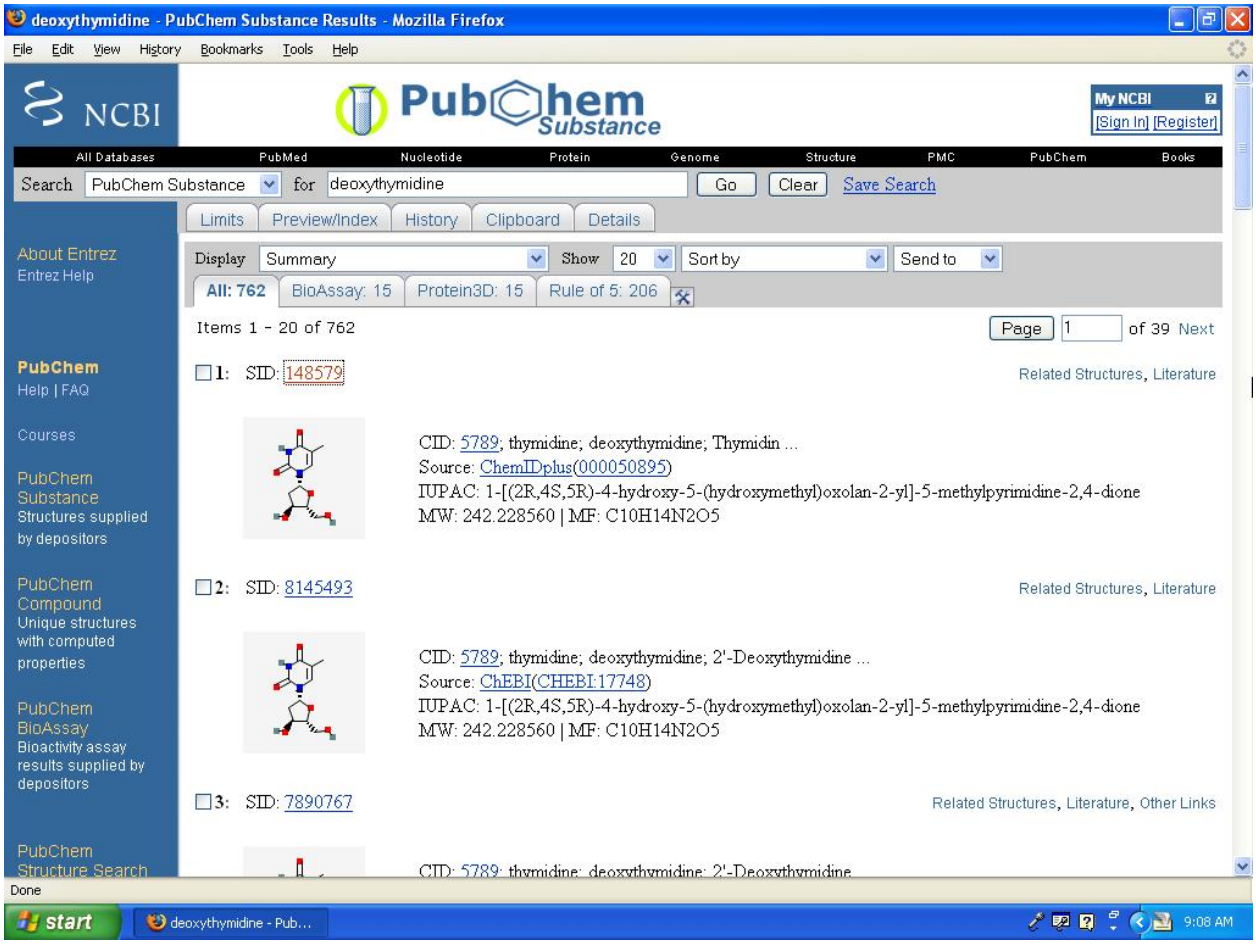

Figure 7 – PubChem Substance Summary – Thymidine