



Using Text-Mining and Crowd-Sourcing to Build a Structure Centric Community for Chemists

Antony Williams

ACS Philadelphia 2008



Imagine a time when

- Chemistry articles are indexed by a free online service
- Authors have tools for automated markup according to community standards
- Structure-article connections go live the day articles are published



ChemSpider - A Search Engine for Chemists

- Questions a chemist might ask...
 - What is the melting point of n-butanol?
 - What is the chemical structure of Xanax?
 - Chemically, what is viagra?
 - What are the stereocenters of cholesterol?
 - Where can I find publications about Taxol?
 - What are the different trade names for Ketoconazole?
 - What is the NMR spectrum of Aspirin?
 - What are the safety handling issues for Thymol Blue?
- ChemSpider can answer all of these questions



ChemSpider Data Content

- Over 21.5 million unique chemical structures from >150 data sources
 - Online Databases – PubChem, Drugbank, HMDB, Wikipedia
 - Chemical Vendors – over 40 different vendors and growing
 - Personal Depositions – individual contributions
 - Journal Publishers
 - Content database vendors
 - Analytical data collections
 - Patents (9 MILLION Structures being deposited now)
 - Web scraping

Content is generally linked back to the original data sources



Tell me about Aspirin

1 hit(s) found in 0.11 seconds

Search term: Aspirin

Found by synonym

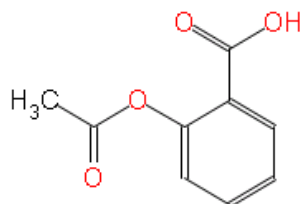
Please [login](#) to be able to add spectra, identifiers, links and publications.

Add: [Comments](#)

INHERENT PROPERTIES, IDENTIFIERS AND REFERENCES

2D 3D Cell

Quick Links: [Permalink](#) [Similar](#) [Isomers](#)



ChemSpider ID: [2157](#)
Empirical Formula: $C_9H_8O_4$
Molecular Weight: 180.1574
Nominal Mass: 180 Da
Average Mass: 180.1574 Da
Monoisotopic Mass: 180.042259 Da

*Place Your Ad Here
or
Claim this Molecule*
[Click for Details](#)

load save zoom

Systematic Name: 2-acetoxybenzoic acid
SMILES: O=C(Oc1ccccc1C(=O)O)C
InChI: [InChI=1/C9H8O4/c1-6\(10\)13-8-5-3-2-4-7\(8\)9\(11\)12/h2-5H,1H3,\(H,11,12\)](#)
InChIKey: [BSYNRYMUTXBXSQ-UHFFFAOYAW](#)

[ORIGINAL REFERENCE\(S\)](#)

[LICENSE](#)

Aspirin was the first-discovered member of the class of drugs known as [non-steroidal anti-inflammatory drugs](#) (NSAIDs), not all of which are salicylates, although they all have similar effects and most have some [mechanism of action](#) which involves non-selective inhibition of the enzyme [cyclooxygenase](#). Today, aspirin is one of the most widely used medications in the world, with an estimated 40,000 [metric tons](#) of it being consumed each year. [Read more...](#) or [Edit at Wikipedia...](#)



Tell me about Aspirin

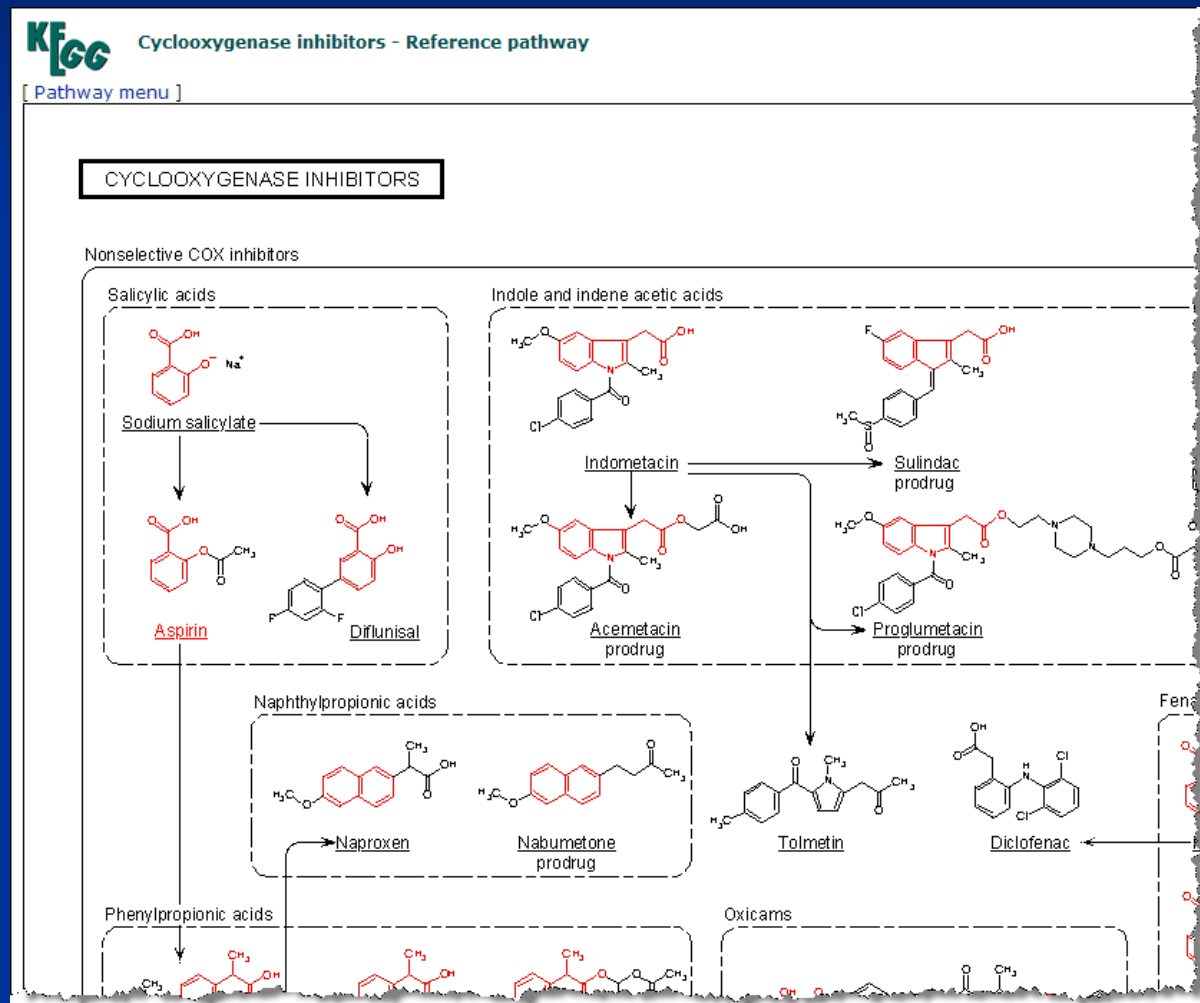
DiscoveryGate	187063 , 2244 , 24666 , 3042924 , 6247 , 51869 , 68375 , 71089 , 3454979 , 3047224 , ...
DrugBank	2244 , APRD00264
DTP/NCI	27223 , 406186
EINECS	N/A
Emory University Molecular Libraries Screening Center	EU-0100038
EPA DSSTox	111 CPDBAS_v5b , 446 NTPHTS_v2b , 726 NTPBSI_v2b , 447 NTPHTS_v2b , 83 FDAMDD_v3b
FDA	16030
Human Metabolome Database	HMDB01879
Journal of Heterocyclic Chemistry	19930997_X2 , 20060813_3D
KEGG	C01405 , D00109
MDPI	20129
Microsource	01500130 , 01500130 , 01500130
MMDB	27242.2 , 27954.2 , 38393.15
NCGC	NCGC00015067-01 , NCGC00090977-01 , NCGC00090977-02
NIAID	006788
NINDS Approved Drug Screening Program	01500130
NIOSH	VO0700000
NIST	2512372342
NIST Chemistry WebBook	2512372342
NMRShiftDB	20038075
Oxford University Chemical Safety Data	Link to Record
PubChem	2244



www.chemspider.com

Links out to KEGG Kyoto Encyclopedia of Genes and Genomes

Entry	D00109	Drug
Name	Aspirin (JP15/USP); Acetylsalicylic acid; Easprin (TN)	
Formula	C ₉ H ₈ O ₄	
Mass	180.0423	
Structure	 D00109 Mol file KCF file DB search Jmol KegDraw	
Target	cyclooxygenase-1 (COX-1) inhibitor [HSA:5742] [EC:1.14.99.1]; cyclooxygenase-2 (COX-2) inhibitor [HSA:5743] [EC:1.14.99.1]	
Activity	Analgesic; Antipyretic; Antirheumatic	
Remark	Same as: C01405 Therapeutic category: 1143 3399 ATC code: A01AD05 B01AC06 N02BA01 BRITe hierarchy	
Comment	Name previously used: Acetylsalicylic acid Component of Bufferin (TN), Percodan (TN), Darvon compound-65 (TN), E.A.C (TN)	
Pathway	PATH: map07110 Benzoic acid family PATH: map07219 Cyclooxygenase inhibitors	
Other DBs	CAS: 50-78-2 PubChem: 7847177 ChEBI: 15365 DrugBank: DB00945 FDB-CCD: AIN DailyMed: aspirin LigandBox: D00109	
LinkDB	All DBs	
KCF data	Show	





Tell me about Aspirin

SUPPLEMENTAL INFORMATION Disclaimer

User Data

- Miscellaneous**
 - Appearance:** white crystalline powder or tablets
 - Appearance:** Odorless, colorless to white, crystal-line powder. [aspirin] [Note: Develops the vinegar-like odor of acetic acid on contact with moisture.]
 - Stability:** Stable. Keep dry. Incompatible with strong oxidizing agents, strongbases, strong acids, various other compounds such as iodides, iron salts,quinine salts, etc.
 - Toxicity:** ORL-RAT LD50 200 mg kg-1 , SKN-RBT LD50 > 7940 mg kg-1 , ORL-MAM LD50 1750 mg kg-1 , ORL-MAN LD50 (estimated) 400 mg kg-1
 - Safety:** Safety glasses.
 - First Aid:** Eye: Irrigate immediately Skin: Soap wash Breathing: Respiratory support Swallow: Medical attention immediately
 - Exposure Routes:** inhalation, ingestion, skin and/or eye contact
 - Symptoms:** Irritation eyes, skin, upper respiratory system; increased blood clotting time; nausea, vomiting; liver, kidney injury
 - Target Organs:** Eyes, skin, respiratory system, blood, liver, kidneys
 - Incompatibilities And Reactivities:** Solutions of alkali hydroxides or carbonates, strong oxidizers, moisture [Note: Slowly hydrolyzes in moist air to salicyclic & acetic acids.]
 - Personal Protection And Sanitation:** Skin: Prevent skin contact Eyes: Prevent eye contact Wash skin: When contaminated Remove: No recommendation Change: Daily Provide: Eyewash, Quick drench
 - Exposure Limits:** NIOSH REL : TWA 5 mg/m 3 OSHA PEL ? : none
- Experimental Physchem Properties**
 - Melting Point:** 138 - 140 C
 - Boiling Point:** 284F (Decomposes)
 - Boiling Point:** ca. 140 C (decomposes)
 - Specific Gravity:** 1.35
 - Vapor Pressure:** 0 mmHg (approx)



Tell me About Aspirin

⊗ NAMES AND SYNONYMS

Validated by Experts, Validated by Users, Non-Validated, Removed by Users, Redirected by Users, Redirect Approved by Experts

11126-35-5 [\[RN\]](#)

11126-37-7 [\[RN\]](#)

2-(Acetyloxy)benzoic acid

200-064-1 [\[EINECS/ELINCS\]](#)

2349-94-2 [\[RN\]](#)

26914-13-6 [\[RN\]](#)

4-10-00-00138 [\[Beilstein\]](#)

50-78-2 [\[RN\]](#)

98201-60-6 [\[RN\]](#)

A.S.A.

Acenterine

Acesal

Acesan

Acetard

Aceticyl

Acetilum acidulatum

Acetisal

Acetophen

Acetosal

Acetosalic acid

Acetoxybenzoic acid

Acetylsal

Acetyl-SAL

Acetyonyl

Acetysal

Acetylsalicylic acid

Acide acetylsalicylique

Acido acetilsalicilico [\[Italian\]](#)

Acimetten

Acisal

Acylpyrin

Asagran

Asatard

Aspirin (JP15/USP)

Aspirin (VAN)

Aspirin [BAN:JAN]



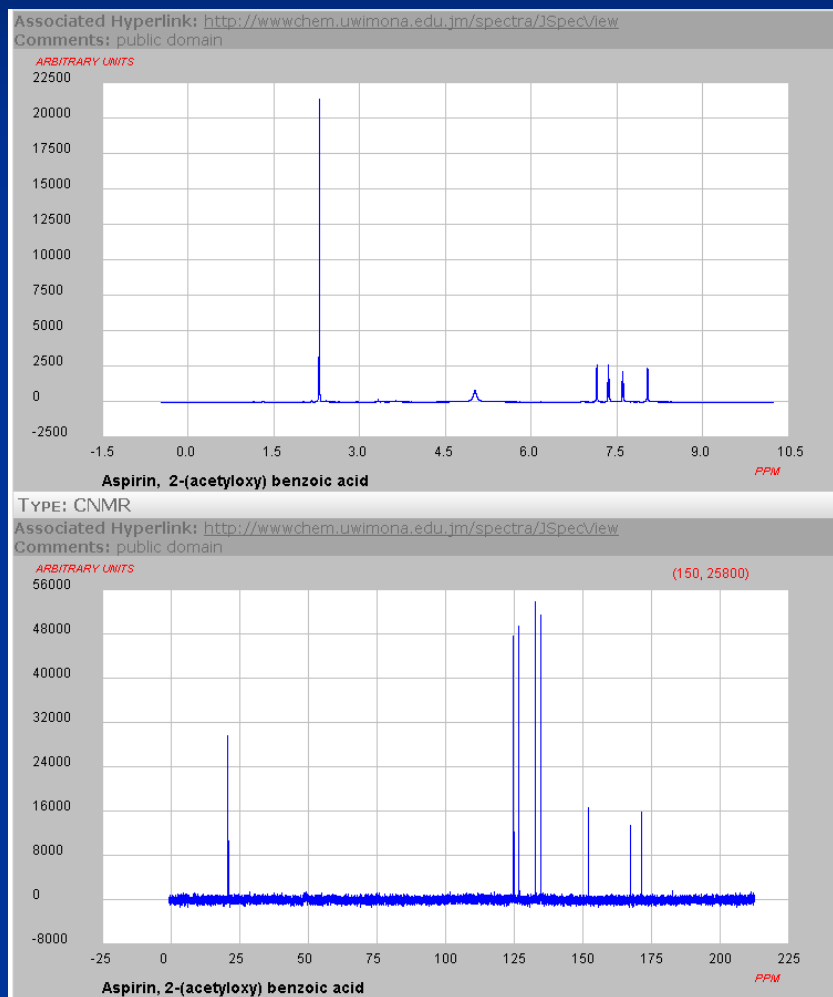
Tell me about Aspirin

⊗ PREDICTED PROPERTIES

LogP:	ACD/LogP: 1.19 XLogP: 1.40 ALOGPS: <u>1.43</u>	# of Rule of 5 Violations:	0
ACD/LogD (pH 5.5):	-0.8	ACD/LogD (pH 7.4):	-1.89
ACD/BCF (pH 5.5):	1	ACD/BCF (pH 7.4):	1
ACD/KOC (pH 5.5):	1.08	ACD/KOC (pH 7.4):	1
#H bond acceptors:	4	#H bond donors:	1
#Freely Rotating Bonds:	3	Polar Surface Area:	52.6 Å ²
Index of Refraction:	1.55	Molar Refractivity:	44.52 cm ³
Molar Volume:	139.5 cm ³	Polarizability:	17.65 10 ⁻²⁴ cm ³
Surface Tension:	49.8 dyne/cm	Density:	1.29 g/cm ³
Flash Point:	131.2 °C	Enthalpy of Vaporization:	59.45 kJ/mol
Boiling Point:	321.4 °C at 760 mmHg	Vapour Pressure:	0.000124 mmHg at 25°C



Tell me about Aspirin





Text-Indexing and ChemSpider?

- ChemSpider text-indexes almost 500,000 Open Access and Free Access articles

Indexed Sources	
<input checked="" type="checkbox"/> Association of Clinical Biochemists of India	<input checked="" type="checkbox"/> Hindawi Publishing Corporation
<input checked="" type="checkbox"/> International Journal of Electrochemical Science	<input checked="" type="checkbox"/> International Union of Crystallography
<input checked="" type="checkbox"/> International Union of Pure and Applied Chemistry	<input checked="" type="checkbox"/> Journal of Biological Chemistry
<input checked="" type="checkbox"/> Libertas Academica	<input checked="" type="checkbox"/> Medknow Publications
<input checked="" type="checkbox"/> Molecular Diversity Preservation International	<input checked="" type="checkbox"/> Proceedings of National Academy of Sciences
<input checked="" type="checkbox"/> PubMed Central	<input checked="" type="checkbox"/> RepositoriUM
<input checked="" type="checkbox"/> Royal Society of Chemistry	

- Collection is growing weekly and more publishers have already agreed



Open Access Literature Search

Taxol, a molecule for all seasons (Chem. Commun. 2001 Issue 10 Page 867-880) - Royal Society of Chemistry

David G. I. Kingston

... CHEMCOMM www.rsc.org/chemcomm Feature Article **Taxol**, a molecule for all seasons David G. I. Kingston Department of Chemistry, M/C 0212, Virginia ...

Nitric oxide, cell death and increased **taxol** recovery (BMC Plant Biology 2005 Volume 5 Issue Suppl 1 Page S12) - PubMed Central Open Archives Service

Don J Durzan

... -5-S1-S12 Meeting Abstract Nitric oxide, cell death and increased **taxol** recovery Durzan Don J 1 djdurzan@ucdavis.edu 1Department of Plant Sciences ...

Induction of Survivin Expression by **Taxol** (Paclitaxel) Is an Early Event, Which Is Independent of **Taxol**-mediated G2/M Arrest (J. Biol. Chem. 2004 Volume 279 Issue 15 Page 15196) - American Society for Biochemistry and Molecular Biology

Xiang Ling, Ralph J. Bernacki, Michael G. Brattain, Fengzhi Li

... 15196 15203, 2004 Printed in U.S.A. Induction of Survivin Expression by **Taxol** (Paclitaxel) Is an Early Event, Which Is Independent of **Taxol**-mediated G2 ...

Studies on the chemistry of **Taxol** (Pure Appl. Chem. 1998 Volume 70 Issue 2 Page 331-334) - IUPAC

D.G.I. Kingston

... 1998. Printed in Great Britain. Q 1998 IUPAC Studies on the chemistry of **Taxol**@ David G. I. Kingston Department of Chemistry, Virginia Polytechnic Institute ...

Taxol Induces Caspase-10-dependent Apoptosis (J. Biol. Chem. 2004 Volume 279 Issue 49 Page 51057) - American Society for Biochemistry and Molecular Biology

Soo-Jung Park, Ching-Haung Wu, John D. Gordon, Xiaoling Zhong, Armaghan Emami, Ahmad R. Safa

... December 3, pp. 51057 51067, 2004 Printed in U.S.A. **Taxol** Induces Caspase-10-dependent Apoptosis* Received for publication, June 11, 2004 ...

Fast Kinetics of **Taxol** Binding to Microtubules. EFFECTS OF SOLUTION VARIABLES AND MICROTUBULE-ASSOCIATED PROTEINS (J. Biol. Chem. 2003 Volume 278 Issue 10 Page 8407) - American Society for Biochemistry and Molecular Biology

Jose Fernando Diaz, Isabel Barasoain, Jose Manuel Andreu

... pp. 8407 8419, 2003 Printed in U.S.A. Fast Kinetics of **Taxol** Binding to Microtubules EFFECTS OF SOLUTION VARIABLES AND MICROTUBULE-ASSOCIATED PROTEINS* Received for ...



Search PubMed – ChemSpider

Search Term:

NCBI Entrez

1650 hits found in 30.14 seconds

["paclitaxel"\[MeSH Terms\]](#) OR [taxol\[Acknowledgments\]](#) OR [taxol\[Figure/Table Caption\]](#) OR [taxol\[Section Title\]](#) OR [taxol\[Body - All Words\]](#) OR [taxol\[Title\]](#) OR [taxol\[Abstract\]](#)

1 2 3 4 5 6 7 8 9 10 ...

ZHOU J, O'BRATE A, ZELNAK A, GIANNAKAKOU P. [Survivin Deregulation in \$\beta\$ -Tubulin Mutant Ovarian Cancer Cells Underlies Their Compromised Mitotic Response to Taxol.](#) *Cancer Res.* 2004 Dec 1; **64** : 8708-8714.

PAREKH HK, ADIKARI M, VENNAPUSA B. [Differential partitioning of Gai1 with the cellular microtubules: a possible mechanism of development of Taxol resistance in human ovarian carcinoma cells.](#) *J Mol Signal.* 2006; **1** : 3.

BOEHMERLE W, ZHANG K, SIVULA M, HEIDRICH FM, LEE Y, JORDT SE, EHRLICH BE. [Chronic exposure to paclitaxel diminishes phosphoinositide signaling by calpain-mediated neuronal calcium sensor-1 degradation.](#) *Proc Natl Acad Sci U S A.* 2007 Jun 26; **104** : 11103-11108.

BOEHMERLE W, SPLITTGERBER U, LAZARUS MB, MCKENZIE KM, JOHNSTON DG, AUSTIN DJ, EHRLICH BE. [Paclitaxel induces calcium oscillations via an inositol 1,4,5-trisphosphate receptor and neuronal calcium sensor 1-dependent mechanism.](#) *Proc Natl Acad Sci U S A.* 2006 Nov 28; **103** : 18356-18361.

GUPTA ML JR, BODE CJ, GEORG GI, HIMES RH. [Understanding tubulin-Taxol interactions: Mutations that impart Taxol binding to yeast tubulin.](#) *Proc Natl Acad Sci U S A.* 2003 May 27; **100** : 6394-6397.

SHANNON KB, CANMAN JC, MOREE CB, TIRNAUER JS, SALMON ED. [Taxol-stabilized Microtubules Can Position the Cytokinetic Furrow in Mammalian Cells.](#) *Mol Biol Cell.* 2005 Sep; **16** : 4423-4436.



Structure-Centric

- We want to search Open-Access articles by structure, substructure, similarity of structure
- Standard approaches would be:
 - Identify chemical names “entity extraction”
 - Convert chemical names to structures and index
- ChemSpider has a validated dictionary of structure-name pairs
- Use name extraction, name-conversion and dictionary look-up. THEN curate.



“Entity Extraction”

- Rule-based recognition of systematic names:
 - Use a lexeme of name fragments
 - Rules for identifying bounds of a name

- Look-up dictionary:
 - Drug Names
 - Trivial Names
 - Numbers : Registry IDs, EINECS/ELINCS/Beilstein IDs



Name Recognition

- Azo aldehyde **2** was synthesized according to a reported method [17]. To a stirred solution of azo aldehyde **2** (1.08 g, 3.76 mmol) in dry CH₂Cl₂ (30.00 mL) at 0 °C were successively added (3,4-diaminophenyl)phenyl methanone **1** (0.40 g, 1.88 mmol) and an excess of anhydrous MgSO₄ (2.00 g, 16.67 mmol). The resulting mixture was stirred for 6 hours at room temperature [18]. The mixture was filtered and washed with dichloromethane. Then the solvent was evaporated under reduced pressure to give azo Schiff base **3** as a red solid which was recrystallized from ethanol 95% (1.28 g, 91 %)



Name Recognition

- Azo aldehyde **2** was synthesized according to a reported method [17]. To a stirred solution of azo aldehyde **2** (1.08 g, 3.76 mmol) in dry **CH₂Cl₂** (30.00 mL) at 0 °C were successively added **(3,4-diaminophenyl)phenyl methanone 1** (0.40 g, 1.88 mmol) and an excess of anhydrous **MgSO₄** (2.00 g, 16.67 mmol).

The resulting mixture was stirred for 6 hours at room temperature [18]. The mixture was filtered and washed with **dichloromethane**. Then the solvent was evaporated under reduced pressure to give azo Schiff base **3** as a red solid which was recrystallized from **ethanol** 95% (1.28 g, 91 %)



How Many Chemical Names?

“She had the drive to derive success in any venture and was well versed in Karate. When the man in the tartan shirt approached her with a dagger in his hand she spat in his face, took the stance of a commando and took advantage of his shock to release the dagger from his grip and causing him to recoil. He went home and took an aspirin after the beating.”



How Many Chemical Names?

“She had the drive to derive success in any venture and was well versed in Karate. When the man in the tartan shirt approached her with a dagger in his hand she spat in his face, took the stance of a commando and took advantage of his shock to release the dagger from his grip and causing him to recoil. He went home and took an aspirin after the beating.”



Daily Med

- “DailyMed provides high quality information about marketed drugs. This information includes FDA approved labels (package inserts).”





Search Aspirin on Daily Med

Search results for

aspirin

Total Results Found: [Count:19]

[Acetaminophen, Aspirin And Codeine Phosphate \(acetaminophen, aspirin and codeine phosphate\) Capsule](#)

[Aggrenox \(aspirin and dipyridamole\) Capsule](#)
[Boehringer Ingelheim Pharmaceuticals Inc.]

[Butalbital, Aspirin, And Caffeine \(butalbital, aspirin, and caffeine\) Capsule](#)
[Lannett Company, Inc.]

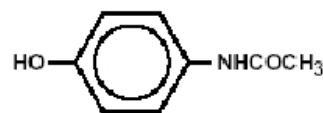
[Butalbital, Aspirin, And Caffeine \(butalbital, aspirin, and caffeine\) Capsule](#)
[MUTUAL PHARMACEUTICAL COMPANY, INC.]

[Butalbital, Aspirin, And Caffeine \(butalbital, aspirin, and caffeine\) Capsule](#)
[Watson Laboratories, Inc.]



Daily Med

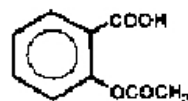
Acetaminophen, 4'-hydroxyacetanilide, is a non-opiate, non-salicylate analgesic and antipyretic which occurs as a white, odorless, crystalline powder, possessing a slightly bitter taste. Its structure is as follows:



$C_8H_9NO_2$

M.W. 151.16

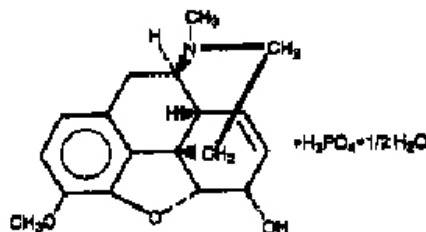
Aspirin, salicylic acid acetate, is a non-opiate analgesic, anti-inflammatory and antipyretic agent. It occurs as a white, crystalline tabular or needle-like powder and is odorless or has a faint odor, Its structure is as follows:



$C_9H_8O_4$

MW. 180.16

Codeine is an alkaloid, obtained from opium or prepared from morphine by methylation. Codeine phosphate occurs as fine, white, needle-shaped crystals, or white, crystalline powder. It is affected by light. Its chemical name is: 7,8-didehydro-4,5 α -epoxy-3-methoxy-17-methylmorphinan-6 α -ol phosphate (1:1) (salt) hemihydrate. Its structure is as follows:





Analysis of Daily Med

- Identify and extract chemical names – combination of name extraction and dictionary lookup
- Convert names to structures using name to structure conversion software
- Manually curate
 - Roundtrip validation of names in Daily Med, on ChemSpider and in the dictionary of the Name to Structure engine
 - For ease of review implement “structure balloons” for viewing



What did we find?

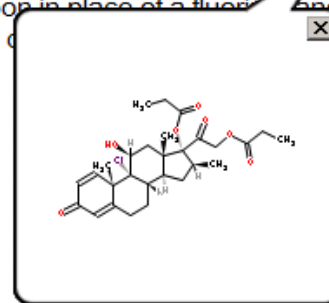
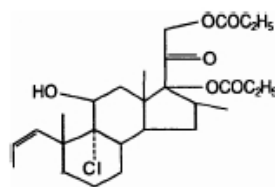
- There were errors in ChemSpider name-structure pairs. KNOWN situation and all curated.
- There were errors in the Name to Structure dictionary – stereochemistry, hydrates/non-hydrates, full structures. All errors reported to vendor.
- There were errors in Daily Med. Stereochemistry, full structures. Over 100 out of 4000 structures in error.



Qvar (beclomethasone dipropionate) Aerosol, Metered [IVAX Laboratories, Inc.]

Description

The active component of QVAR 40 mcg Inhalation Aerosol and QVAR 80 mcg Inhalation Aerosol is beclomethasone dipropionate, USP, an anti-inflammatory corticosteroid having the chemical name 9-chloro-11 β , 17,21-trihydroxy-16 β -methylpregna-1,4-diene-3,20-dione 17,21-dipropionate. Beclomethasone dipropionate is a diester of beclomethasone, a synthetic corticosteroid chemically related to dexamethasone. Beclomethasone differs from dexamethasone in having a chlorine at the 9-alpha carbon in place of a fluorine and in having a 16 beta-methyl group instead of a 16 alpha-methyl group. Beclomethasone dipropionate is a white to off-white powder with a molecular formula of C₂₈H₃₇ClO₇ and a molecular weight of 521.1. Its chemical structure is:



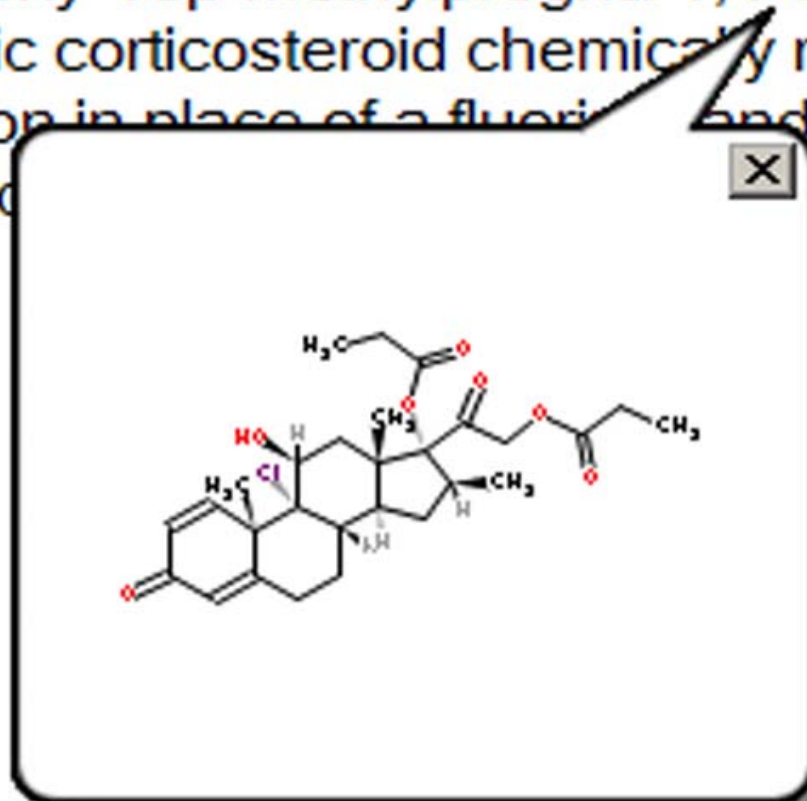
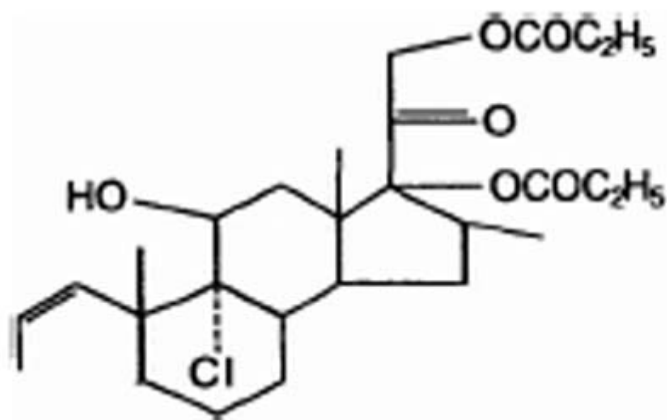
Beclomethasone dipropionate is slightly soluble in water, very soluble in chloroform and freely soluble in acetone and in alcohol.

QVAR is a pressurized, metered-dose aerosol intended for oral inhalation only. Each unit contains a solution of beclomethasone dipropionate in propellant HFA-134a (1,1,1,2 tetrafluoroethane) and ethanol. QVAR 40 mcg delivers 40 mcg of beclomethasone dipropionate from the actuator and 50 mcg from the valve. QVAR 80 mcg delivers 80 mcg of beclomethasone dipropionate from the actuator and 100 mcg from the valve. This product delivers 50 microliters (59 milligrams) of solution formulation from the valve with each actuation. Each canister provides 100 inhalations. QVAR should be "primed" or actuated twice prior to taking the first dose from a new canister, or when the inhaler has not been used for more than ten days. Avoid spraying in the eyes or face while priming QVAR. This product does not contain chlorofluorocarbons (CFCs).



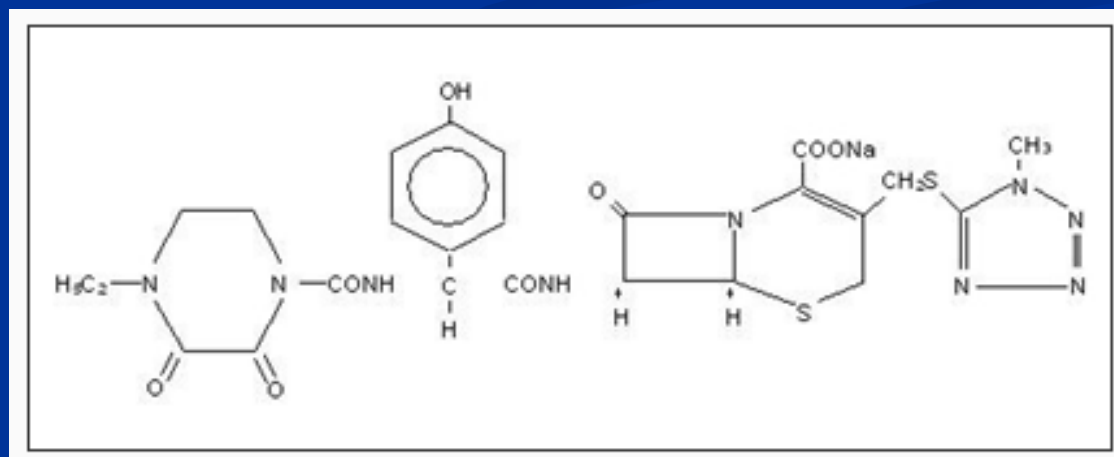
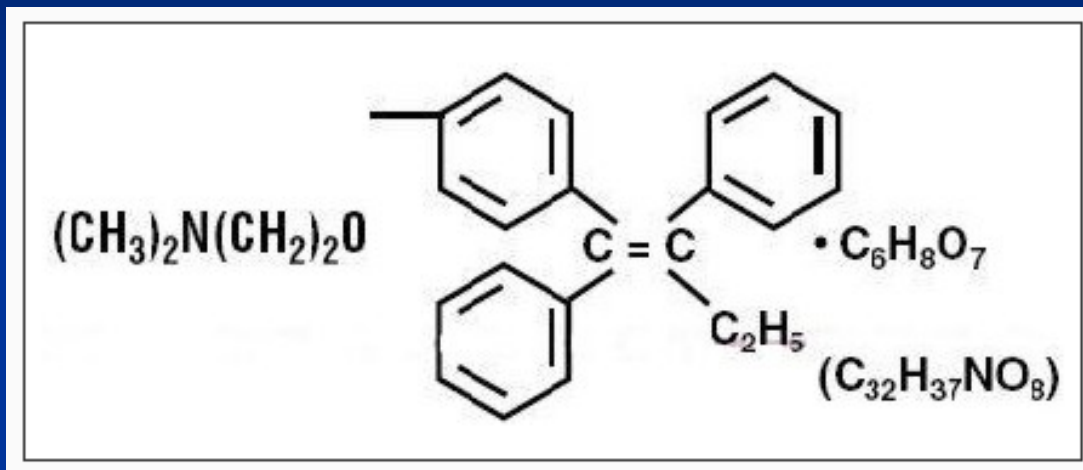
QVAR

9-chloro-11 β , 17,21-trihydroxy-16 β -methylpregna-1,4-diene
of beclomethasone, a synthetic corticosteroid chemically related to
chlorine at the 9-alpha carbon in place of a fluorine atom. The sodium
dipropionate is a white to off-white powder. Its chemical structure is:





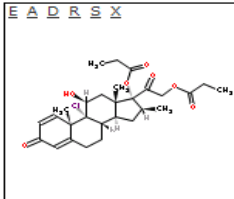
Other Daily Med Structures





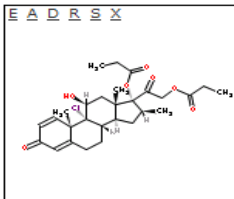
Full Markup of the Document

Manufactured Medicine



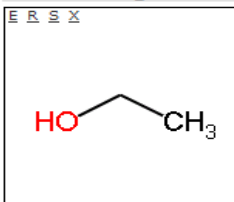
qvar
qvar
QVAR

Generic Medicine



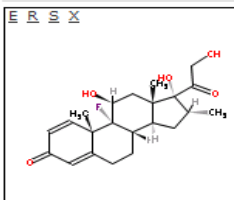
beclomethasone dipropionate
beclomethasone dipropionate
BECLOMETHASONE
DIPROPIONATE

Inactive Ingredient(s)

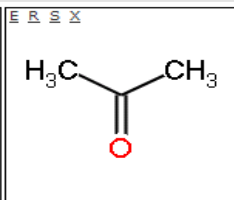


ethanol
ethanol
Ethanol

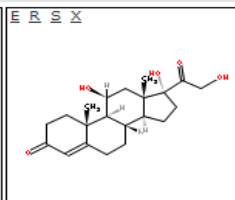
Other



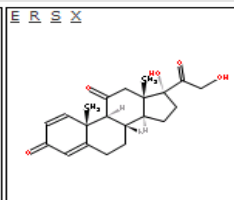
dexamethasone
dexamethasone
Dexamethasone



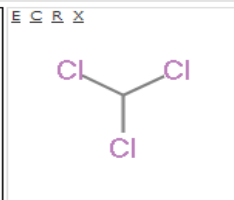
acetone
acetone
Acetone



cortisol
cortisol
Cortisol



prednisone
prednisone
Prednisone



chloroform
chloroform
Chloroform



Making Open Access Articles Searchable Proof of Concept

- Can we HOST Chemistry Open Access articles on ChemSpider and add-value
- Can we identify chemical names in Open Access articles in a user-friendly manner
- Can we convert names to structures in Open-Access articles and expand ChemSpider and provide structure searching of Open Access chemistry articles?
- Can we provide an environment for chemists to mark-up their own articles and crowd-source markup of an archive?



Document markup

- ChemSpider now hosting Open Access articles from MDPI, Molecular Diversity Preservation International
- Hosting the Molbank collection at present

<http://www.mdpi.org/molbank>

MolBank

One Compound-per-Paper Short Notes



A Standard for Document Markup?

- NLM-DTD: National Library of Medicine; Document Type Definition
- Approved markup definitions to apply to journal articles – extended as necessary for our purposes

NLM Journal Archiving and Interchange Tag Suite

National Center for Biotechnology Information

National Library of Medicine



Archiving and
Interchange Tag Set

Journal Publishing
Tag Set

Article Authoring
Tag Set

NCBI Book
Tag Set



NLM/DTD markup

NLM/DTD [Chemistry](#) [Biology](#)

Article Front [Article Front 2](#) [Article Body](#) [Article Back](#)

Title Contributor Submission Date Acceptance Date Publication Date URI Abstract Keyword

NLM/DTD [Chemistry](#) [Biology](#)

Article Front [Article Front 2](#) [Article Body](#) [Article Back](#)

Subtitle Product Supp. Material Permission Related Contract Number Contract Sponsor

Grant Number Grant Sponsor Conference Number

NLM/DTD [Chemistry](#) [Biology](#)

Article Front [Article Front 2](#) [Article Body](#) [Article Back](#)

Title Text Figure Graphic Media Formula Speech Statement

NLM/DTD [Chemistry](#) [Biology](#)

Article Front [Article Front 2](#) [Article Body](#) [Article Back](#)

Glossary Term Reference Appendix Acknowledgment

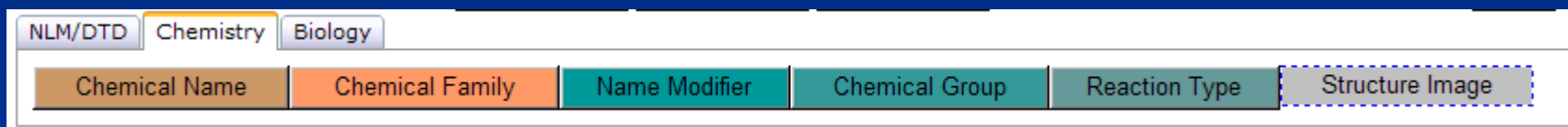


Chemistry and Biology

NLN/DTD	Chemistry	Biology		
Chemical Name	Chemical Family	Name Modifier	Chemical Group	Reaction Type



Chemistry and Biology



- Menus can be extended as necessary



Document markup



NLM/DTD Chemical Data Analytical Data PhysChem Properties Biological Data

Chemical Name Chemical Formula Chemical Family Name Modifier Chemical Group Reaction Type Structure Image

Molbank 2004, M372 <http://www.mdpi.net/molbank/>

3,3'-Bis(*N,N*-dimethylamino)-5,5'-bi-1,2,4-triazine and 6,6'-dibromo-3,3'-bis(*N,N*-dimethylamino)-5,5'-bi-1,2,4-triazine

Danuta Branowska*, Beata Iwańska and Andrzej Rykowski

Institute of Chemistry, University of Podlasie, ul. 3 Maja 54, PL-08-110 Siedlce, Poland
E-mail: dankab@ap.siedlce.pl

Received: 18 November 2003 / Accepted: 18 February 2004 / Published: 24 February 2004

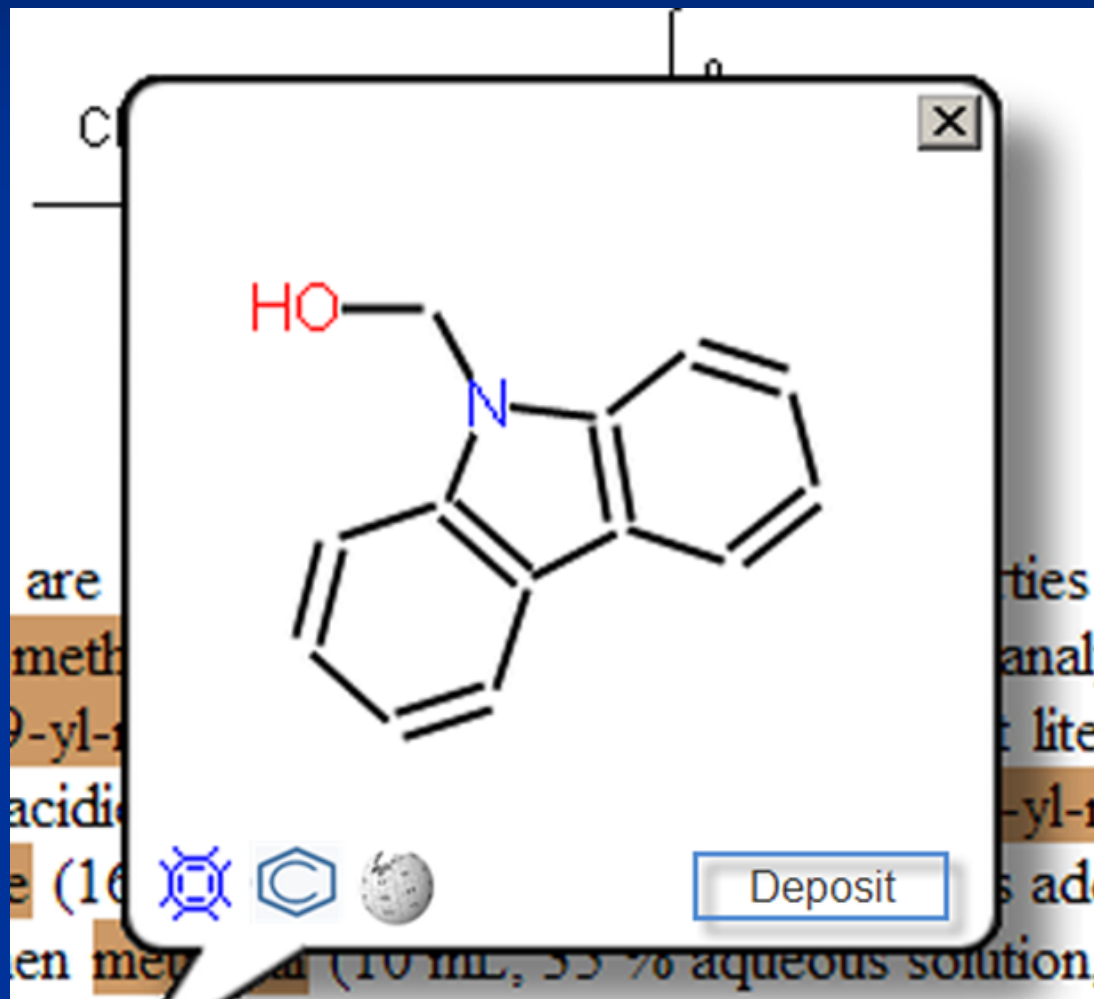
Keywords: 5,5'-bi-1,2,4-triazine, nucleophilic substitution, bromination

Continuing our study on the application of 1,2,4-triazines in organic synthesis [1] we prepared the title compounds as valuable intermediates for metalation reactions leading to functionalized 5,5'-bi-1,2,4-triazines [2].

The mixture of 3,3'-bis(methylsulfany)-5,5'-bi-1,2,4-triazine (1) [3] (756 mg, 3.0 mmol) and dimethylamine, 40 wt. % solution in water (40 g), was stirred at room temperature for 20 hrs, and then was heated at 70 °C during a period of 30 min. The precipitate was filtered off and it was purified by column chromatography on silica gel (Merck type 60, 230-400 mesh) using a mixture of chloroform/acetone (100:1) as eluent to give 709 mg (96 %) of 3,3'-bis(*N,N*-dimethylamino)-5,5'-bi-1,2,4-triazine of (2) as a yellow solid. To a solution 2 (246 mg, 1.0 mmol) in acetic acid (8 mL) the bromine (1.6 g, 10 mmol) was added. The reaction mixture was refluxed for 2 hrs. After that time the reaction mixture was cooled to 20 °C, diluted with water (50 mL) and extracted with chloroform (5 x 25 mL). The organic extract was washed with water (125 mL) and dried over MgSO₄. Removal of the solvent in vacuum and purification of the residue by column chromatography on silica gel (Merck type 60, 230-400 mesh) using a mixture of chloroform/acetone (100:1) as eluent gave 222 mg (55 %) of 6,6'-dibromo-3,3'-bis(*N,N*-dimethylamino)-5,5'-bi-1,2,4-triazine (3) as a yellow solid.



Searching from the Structure Balloon





A Platform for Markup

- Can we provide a platform for document markup for chemists?
- Workflow:
 - Upload word docs, RTF files or point to HTML and load
 - Apply entity extraction, convert names to structures, mark-up automatically and ask for user participation
 - Publish final version with NLM-DTD markup
 - Deposit all structures on ChemSpider under embargo and wait for article DOI to release



Online Markup

Load Revision... Save Revision Auto Markup Export Markup Mark All Entries Ask Confirmation Clear

NLM/DTD Chemistry Biology

Chemical Name Chemical Family Name Modifier Chemical Group Reaction Type Structure Image

Utilizing Long-Range ^1H - ^{15}N 2D NMR Spectroscopy for Chemical Structure Elucidation and Confirmation

Gary E. Martin* and Antony J. Williams[†]

Rapid Structure Characterization Laboratory
Schering-Plough Research Institute
Summit, NJ 07901

and

[†]ChemZoo, Inc.
Wake Forest, NC 27587

Inverse- or proton-detected heteronuclear 2D NMR methods have opened many avenues of investigation that would otherwise be closed. In particular, beginning in the mid-1990's, reports began to appear in the literature regarding preliminary investigations of long-range ^1H - ^{15}N 2D NMR at natural abundance. The subsequent development and continually-increasing access to cryogenic NMR probe technology has further facilitated the acquisition of ^1H - ^{15}N heteronuclear shift correlation data. Since 2000, the field has been reviewed several times,¹⁻⁴ and chapters have also appeared surveying the application of these methods in the field of alkaloid chemistry.^{5,6} The literature representing this area of investigation now comprises several hundred publications.

There is also a rich body of literature associated with the more difficult direct detection of ^{15}N spectra that has been the subject of several monographs⁷⁻¹⁰ and very extensive reviews by Webb and co-authors.¹¹⁻¹⁶ Direct observation of ^{15}N is fraught with difficulty as a result of the relatively low natural abundance of 0.37%, the low



Automated markup

Later the same year, Quéguiner and co-workers³⁰ utilized long-range ^1H - ^{15}N 2D NMR methods in the determination of the site of metalation in the benzene portion of some benzodiazines. The site of reaction with electrophiles was established based on the analysis of the 3 Hz long-range optimized ^1H - ^{15}N GHMBC spectra. The authors began by unequivocally assigning the N1 and N4 shifts of 2-methoxy- and 2-phenylquinoxaline (17). With these assignments in hand it was possible to unequivocally establish the identities of the H5 and H8 resonances of 2-methoxy-3-phenylquinoxaline, 18. These data were finally employed to establish the site of electrophilic substitution of 19 based on the long-range correlations of the H7 and H8 protons to N1.

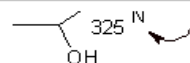
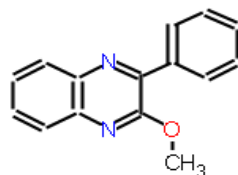
A somewhat similar and interesting study was reported by Gawinecki and co-workers³¹ in which they employed long-range ^1H - ^{15}N 2D methods to characterize pyrazoloquinoxaline products (20) formed from the condensation of pyrazoline-4,5-diones (21) with aromatic 1,2-diamines (22). A further comprehensive study of an extensive series of di- and tri-substituted pyrazines was reported in 2000 by Sommer and co-workers.³²



Name to Structure Conversion

15

Later the same year, Quéguiner and co-workers³⁰ employed N 2D NMR methods in the determination of the site of metalation in the benzene portion of some benzodiazines. The site of reaction with electrophiles was determined by unequivocally assigning the N1 and N4 shifts of the H5 and H8 resonances of 2-methoxy-3-phenylquinoxaline, **18**. These data were finally employed to establish the site of electrophilic substitution of **19** based on the long-range correlations of the H7 and H8 protons to N1.



...N 2D NMR methods in the determination of the site of metalation in the benzene portion of some benzodiazines. The site of reaction with electrophiles was determined by unequivocally assigning the N1 and N4 shifts of the H5 and H8 resonances of 2-methoxy-3-phenylquinoxaline, **18**. These data were finally employed to establish the site of electrophilic substitution of **19** based on the long-range correlations of the H7 and H8 protons to N1.

A somewhat similar and interesting study was reported by Gawinecki and co-workers³¹ in which they employed long-range ^1H - ^{15}N 2D methods to characterize pyrazoloquinoxaline products (**20**) formed from the condensation of pyrazoline-4,5-diones (**21**) with aromatic 1,2-diamines (**22**). A further comprehensive study of an extensive series of di- and tri-substituted pyrazines was reported in 2000 by Sommer and co-workers.³²



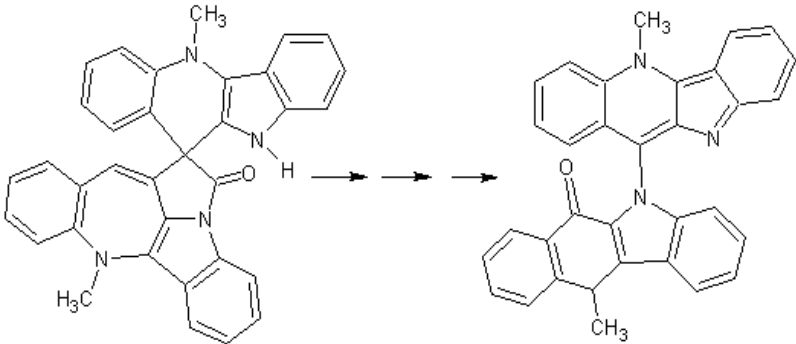
Conversion of Structure Images

- Not all compounds have a “name”
- Structure images can be converted to connection tables and databased using commercial or Open Source software
 - CLiDe – Commercial
 - OSRA – Open Source

Cryptomisine



Chemical Name	Molecular Formula	Chemical Family	Name Modifier	Chemical Group	Reaction Type	Structure Image
---------------	-------------------	-----------------	---------------	----------------	---------------	-----------------

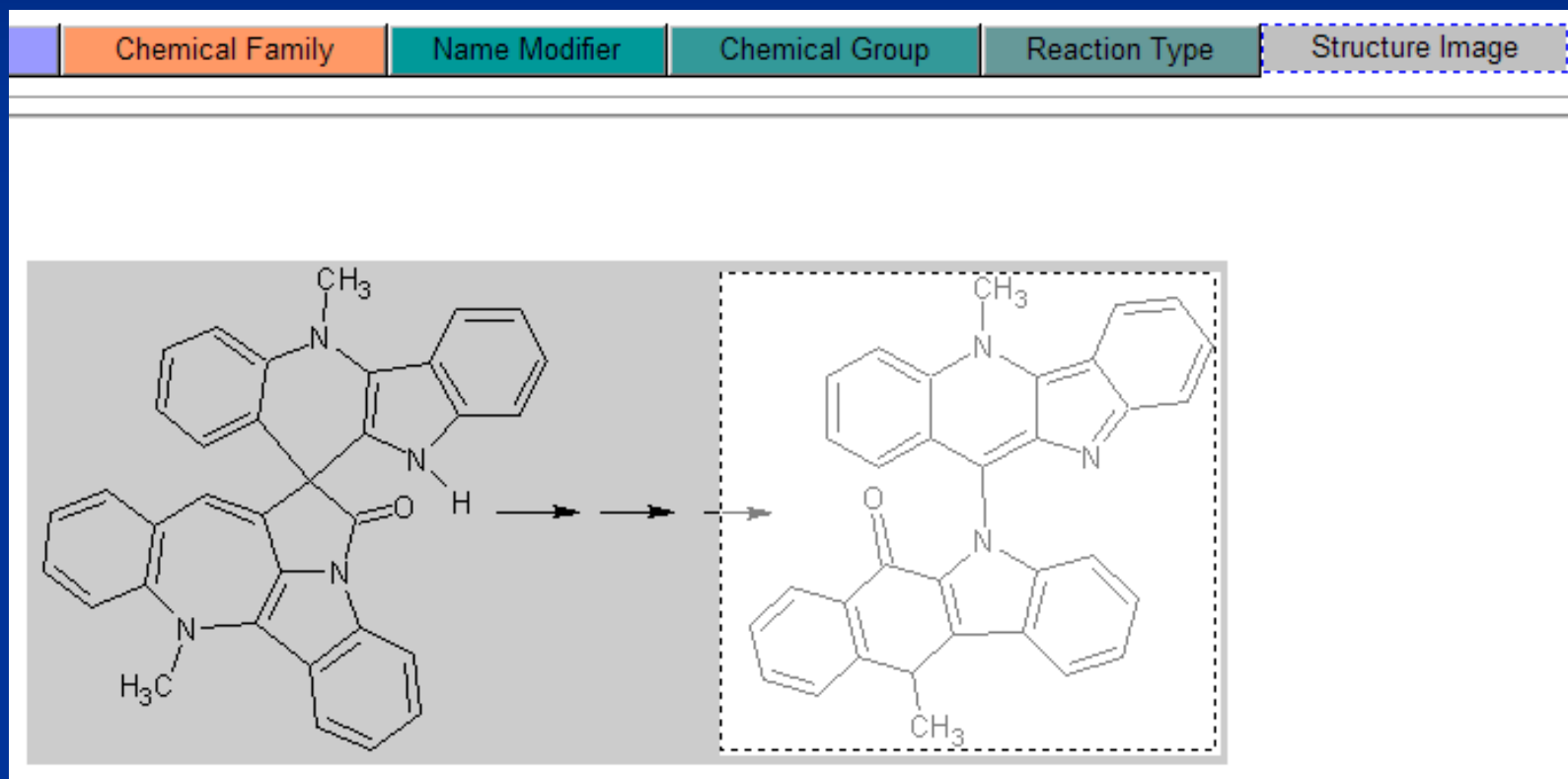


9 10

Another application of partial ^1H - ^{15}N chemical shift correlation data was seen in the application of the Structure Elucidator CASE program to the dimeric indoloquinoline alkaloid cryptomisine, 11. Without any ^1H - ^{15}N chemical shift correlation data, the program ran for 210 h generating >75 million structures, of which >22,000 remained after filtration and removal of duplicates. In contrast, when only simple ^1H - ^{15}N HMQC direct correlation data were added to the input data, only 5 structures were generated in ~1 m, the correct structure among them.²⁶

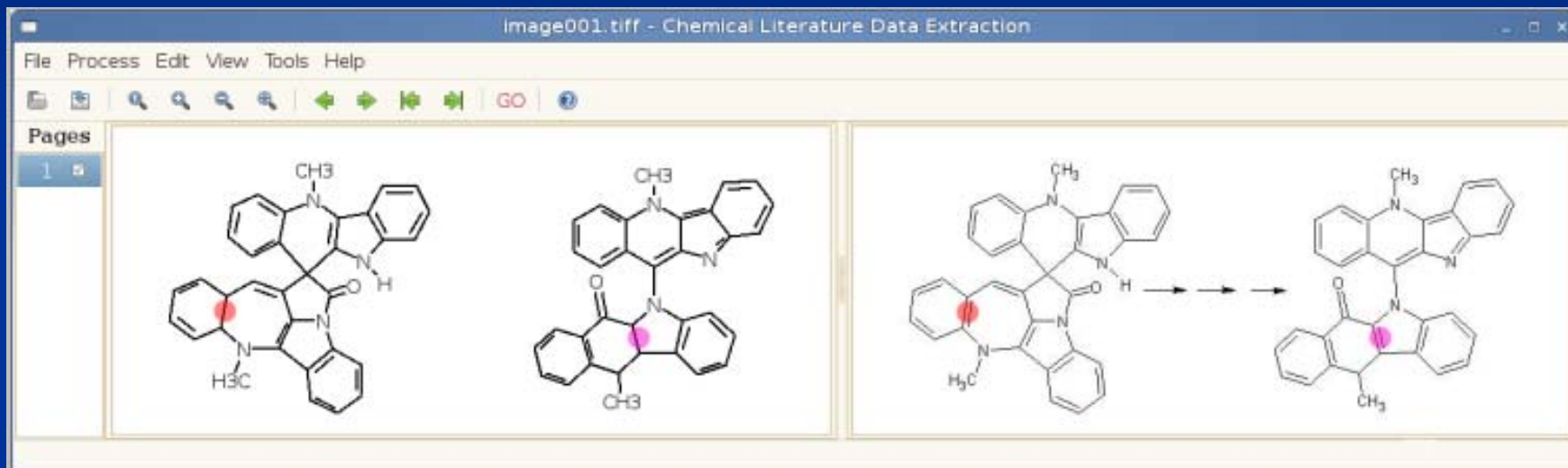


Structure Conversion from Images - OSRA





Structure Conversion from Images-CLiDE



- Conversion dependent on zoom-factor can give perfect conversion!

Supports Word .DOC, HTML, RTF



NLM/DTD Chemical Data Analytical Data PhysChem Properties Biological Data

Article Front Article Body Article Back

Title Text Figure Graphic Media Formula Speech Statement

home about index aup faq

totallysynthetic*

Formaldehyde Hydrolysis
Wittig
Grignard
Reductive Amination
Iminium Ion NH Cyclization
Grignard
Wittig
OTfPS

Cortistatin A Pt. II

Cortistatin A

Nicolaou, Chen, Sun, Peng and Polet. *ACIEE*, 2008, *EarlyView*. DOI: [10.1002/anie.200803550](https://doi.org/10.1002/anie.200803550).

A second showing for this natural product on these pages, the first synthesis was by a former student of Nicolaou, [Phil Baran, three months ago](#). However, this is quite a different piece of work, as Nicolaou completes the synthesis from scratch - whereas the Baran approach was more a [semi-synthesis](#). I covered the (few) biological detail in the previous post, so we'll move on to the synthetic action directly.

Now, I said that Nicolaou started the synthesis from scratch - that not quite true. Sure, the students in the labs definitely began by cracking-open the Aldrich bottles (a good feeling, for some reason), but the discussion starts with compound **8**, which is the work of quite a few steps. First up is the [HPESW](#) reaction, which builds the 6,5-system using that proto-organocatalysis. The reference given for this, usefully in some senses, is the [Organic Syntheses prep](#), which suggests a 70–76% yield for the three steps.



Extensible Markup Process

- Markup process is easily extendable
- Configurable from one XML file
- NLM/DTD is incorporated but is easy to extend
- Markup supports ontologies – presently working to support the Medical Subject Headings Ontology - MeSH.



Patents and PubMed



SureChem

HOME

SEARCH PATENTS

CHEMICAL SEARCH

DATA SERVICES

[Back to search](#)

Results for search 'chemical:(valium)'

USPTO Granted 4595 hits ([view all](#))

▶ Show top results:

USPTO Applications 4965 hits ([view all](#))

▶ Show top results:

European Granted 1346 hits ([view all](#))

▶ Show top results:

European Applications 1277 hits ([view all](#))

▶ Show top results:

WO/PCT 4725 hits ([view all](#))

▶ Show top results:


MedLine 15197 hits ([view all](#))

▶ Show top results:



What's Coming...today..

- Pubmed made structure searchable...collaboration with SureChem provides structure searchable access to Pubmed abstracts

<u>PubChem</u>	<u>2118</u>
<u>PubMed</u>	<u>10631626, 10648728, 10653202, 10663429, 10664927, 10668858, 10688619, 10691246, 10696114, 10698361, 10709776, 10758169, 10760357, 10770452, 10770483, 10782977, 16563516, 16707239, 16791582, 16991019, 17020957, 17058100, 17124639, 17158196, 17175036, 17219217, 17244765, 6140317, 6141159, 6141930, 6142907, 6143507, 6143633, 6143649, 6143766, 6144090, 6144110, 6144136, 6145358, 6145360, 6145365, 6145726, 6701100, 6733175</u>
Sigma-Aldrich 	611026 ALDRICH, A6551 SIGMA, A8800 SIGMA

- Necessary extension – text and structure-based integrated searching.



Working on

- Balloon pop-up over every link to reduce clicks..

Comparison of the frequency of behavioral disinhibition on alprazolam, clonazepam, or no benzodiazepine in hospitalized psychiatric patients.

[Rothschild AJ](#), [Shindul-Rothschild](#), [Viguera A](#), [Murray M](#), [Brewster S](#).

Department of Psychiatry, University of Massachusetts Medical School, Worcester 01655, USA. rothscha@ummc.org

Several case reports have suggested that treatment with the benzodiazepine alprazolam can result in behavioral disinhibition. To address this question, the authors reviewed the medical records (blinded to all pharmacologic treatments the patients received) of 323 psychiatric inpatients treated with alprazolam (108 patients), clonazepam (111 patients), or no benzodiazepine (104 patients) between January 1989 and June 1990. During benzodiazepine treatment, there were no significant differences among the three groups on the following measures: (1) acts of self-injury (alprazolam, 1.9%; clonazepam, 1.8%; no benzodiazepine, 2.9%); (2) assaults on staff or other patients (alprazolam, 0%; clonazepam, 0.9%; no benzodiazepine, 1.0%); (3) need for seclusion or restraints (alprazolam, 3.7%;

PubChem	2118
PubMed	10631626 , 10648728 , 10653202 , 10663429 , 10664927 , 10668858 , 10688619 , 10691246 , 10696114 , 10698361 , 10709776 , 10758169 , 10760357 , 10770452 , 10770483 , 10782977 , 16563516 , 16707239 , 16791582 , 16991019 , 17020957 , 17058100 , 17124639 , 17158196 , 17175036 , 17219217 , 17244765 , 6140317 , 6141159 , 6141930 , 6142907 , 6143507 , 6143633 , 6143649 , 6143766 , 6144090 , 6144110 , 6144136 , 6145358 , 6145360 , 6145365 , 6145726 , 6701100 , 6733175
Sigma-Aldrich	611026 ALDRICH, A6551 SIGMA, A8800 SIGMA



What's Coming?

- Agreement with Royal Society of Chemistry that we can add their structure-based RSS feeds to ChemSpider
- Agreement with Nature Publishing Group to add their Nature Chemical Biology structure collections to ChemSpider as they issue
- Presently indexing Acta Chemica Scandinavica, 1947-1999 PDF backfile – our first foray into OCR
- Presently indexing PLoS journals directly
- More publishers have agreed...

Conclusions

- The quality of structure-based data online should always be questioned – that includes ChemSpider
- Robots and software algorithms can help but eyeballs are necessary
- Data on ChemSpider are being added and curated on a daily basis but we need more eyeballs helping always
- ChemSpider now has a large validated structure-name dictionary

Conclusions

- Text indexed journal articles can be text-mined and linked to chemical structures for searching
- Proof of concept document markup of chemistry documents is proven
- A public facility for document markup and indexing is in development – chemists can submit their documents for markup and nomenclature validation prior to submission to Open Access journals





Acknowledgements

- Aniko Valko, KeyModule
- Matt Stahl and Joe Corkery, OpenEye
- Andrey Yerin and Andrew Anderson, ACD/Labs
- Nicko Goncharoff and Michael Faust, SureChem
- The OSRA development team
- The many ChemSpider users adding and cleaning data on a daily basis



Further reading

- www.chemspider.com/blog
- Internet-based tools for communication and collaboration in chemistry, *Drug Discovery Today*, Volume 13, Numbers 11/12, June 2008 502-506, [doi:10.1016/j.drudis.2008.03.015](https://doi.org/10.1016/j.drudis.2008.03.015)
- A perspective of publicly accessible/open-access chemistry databases, *Drug Discovery Today*, Volume 13, Numbers 11/12, June 2008, 495-501, [doi:10.1016/j.drudis.2008.03.017](https://doi.org/10.1016/j.drudis.2008.03.017)