

Activities at the Royal Society of Chemistry to Gather, Extract and Analyze Big Datasets in Chemistry

RSC-CICAG Meeting
April 22nd 2015





1 **NEW**
DEFINITION
IS ADDED ON
URBAN

1,600+
READS ON
Scribd

13,000+ HOURS
MUSIC
STREAMING ON
PANDORA

12,000+
NEW ADS
POSTED ON
craigslist

370,000+ MINUTES
VOICE CALLS ON
skype

98,000+
TWEETS

320+
NEW
twitter
ACCOUNTS

100+
NEW
Linked in
ACCOUNTS

1 associatedcontent
NEW
ARTICLE IS
PUBLISHED

6,600+
NEW
PICTURES ARE
UPLOADED ON
flickr

50+
WORDPRESS
DOWNLOADS

695,000+
facebook
STATUS
UPDATES

125+
PLUGIN
DOWNLOADS

79,364
WALL
POSTS

510,040
COMMENTS

694,445
SEARCH
QUERIES

168 MILLION
EMAILS
ARE SENT

60+
NEW
BLOGS

1,500+
BLOG
POSTS

70+
DOMAINS
REGISTERED

600+
NEW
VIDEOS

100+
Answers.com
40+
YAHOO! ANSWERS

QUESTIONS
ASKED ON THE
INTERNET...

25+ HOURS
TOTAL
DURATION



What of the World of Chemistry?

CAS REGISTRY - The gold standard for chemical substance information

CAS REGISTRYSM contains more than 95 million unique organic and inorganic chemical substances, such as alloys, coordination compounds, minerals, mixtures, polymers and salts, and more than 66 million sequences—more than any other database of its kind.



What of the World of Chemistry?



A global team of scientists is continually adding substance information from the world's disclosed chemistry to the [CAS REGISTRYSM](#), the gold standard for chemical substance information.

Database Counter

In addition to organic and inorganic substances, REGISTRY has:

66,076,852 sequences

CAS RN 1687856-73-0 is the most recent CAS Registry Number

CAS also provides specialized databases of chemical reactions, regulated chemicals, commercially available chemicals and Markush substance information.

Specialized Substance Collections Count

[CASREACT^{\(1\)}](#) 79,420,997 Single and multi-step reactions, and synthetic preparations

[CHEMLIST](#) 343,713 Inventoried/regulated substances

[CHEMCATS](#) 100,844,238 Commercially available chemicals

[MARPAT](#) 1,076,732 Searchable Markush structures



Prophetic Enumeration

Paul Peters (CAS) discussing CAS recent practice of indexing prophetic compounds from patents, included the Information Conference in Biarritz, France (Nov. 3-5, 2009).

"The Hall of Shame: Top 5 patents with the most indexed exemplified prophetics

Patent number	Company	Number of pages	Number of records	Number of prophetics
US2008004323	Sumitomo Chem.	177p	123	778,674
WO2001056358	Rohm and Haas	1646p	96	621,062
WO2009028727	Sumitomo Chem.	288p	31	174,085
WO2008135413	BASF	106p	27	154,481
WO2009019005	Syngenta	201p	25	149,881

"Most patents in pharma and agro contain less than 20 exemplified prophetic compounds."



What of the World of Chemistry?

PubChem

Compounds:	68,281,028
Tested Compounds:	2,083,015
Substances:	196,542,951
Tested Substances:	3,141,494
BioAssays:	1,154,327
RNAi BioAssays:	64
BioActivities:	228,499,734
Protein Targets:	9,853
Gene Targets:	57,039

What of the World of Chemistry?



Published: 11 February 2013

InChI in the wild: an assessment of InChIKey searching in Google

Christopher Southan

“The InChIKey indexing has therefore turned Google into a *de-facto* open global chemical information hub by merging links to most significant sources, including over **50 million** PubChem and ChemSpider records.”



What of the World of Chemistry?



What does the Reaxys Chemistry Discovery Engine offer?

Essential and relevant chemical data

Gain access to over 16,000 periodicals containing 500 million experimentally verified facts.



RSC's ChemSpider

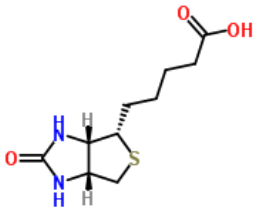
>34 million chemicals from >500 sources and
>40,000 users per day

ChemSpider

Search and share chemistry

[About](#) | [More Searches](#) | [Web APIs](#) | [Help](#)

Search term: **vitamin H** (Found by approved synonym) ?



The image shows the chemical structure of Biotin, which consists of a fused bicyclic system (a thiazolidine ring fused to an imidazole ring) with a side chain containing a carboxylic acid group. The structure is shown in a 2D representation with stereochemistry indicated by wedges and dashes.

Biotin

ChemSpider ID: **149962**

Molecular Formula: $C_{10}H_{16}N_2O_3S$

Average mass: 244.310593 Da

Monoisotopic mass: 244.088165 Da

▼ Systematic name

5-[(3aS,4S,6aR)-2-Oxohexahydro-1H-thieno[3,4-d]imidazol-4-yl]pentanoic acid

► SMILES and InChI

► Cite this record

? 2D 3D Save Zoom

3 of 3 defined stereocentres

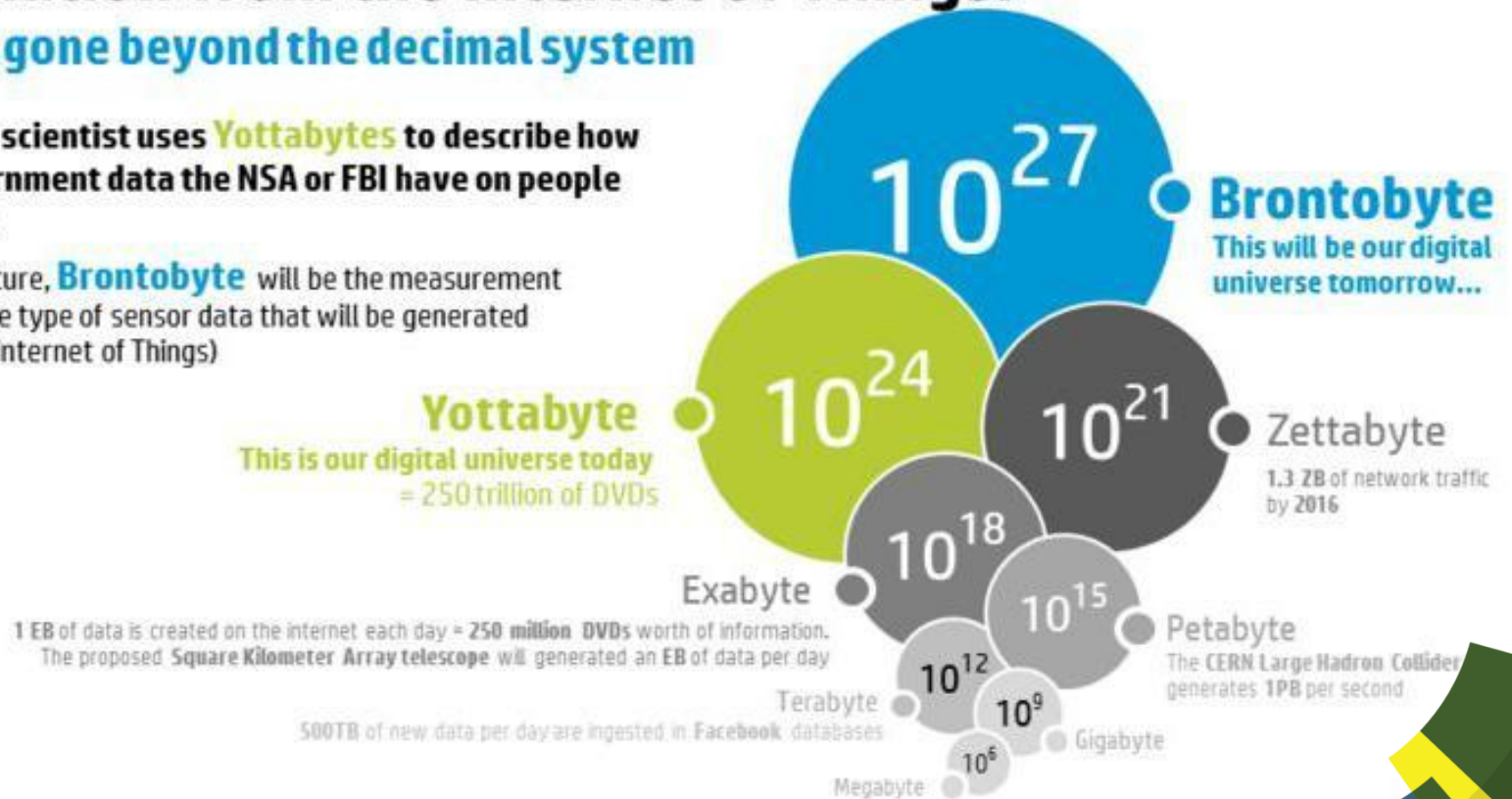
Not Dealing With Big Data...

Information from the Internet of Things:

We have gone beyond the decimal system

Today data scientist uses **Yottabytes** to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)





Open Access/Data Mandates

Open Access funder mandates...

UK Funding Bodies Mandate Open Access



Research funders' open access policies



Office of Science and Technology Policy

**Expanding Public Access to the Results of Federally
Funded Research**

Posted by Michael Stebbins on February 22, 2013 at 12:04 PM EDT



RESEARCH DATA ALLIANCE



Chemistry Open Data???

- Where are all of the Open Chemistry Data?
 - Is there a willingness to contribute more?
 - Can we harvest more?
- 



Chemistry Open Data???

- Where are all of the Open Chemistry Data?
 - *Not that much showing up yet from scientists*
 - Is there a willingness to contribute more?
 - Can we harvest more?
- 



Chemistry Open Data???

- Where are all of the Open Chemistry Data?
 - *Not that much showing up yet from scientists*
 - Is there a willingness to contribute more?
 - *Many concerns about IP and much lip service*
 - Can we harvest more?
- 



Chemistry Open Data???

- Where are all of the Open Chemistry Data?
 - *Not that much showing up yet from scientists*
 - Is there a willingness to contribute more?
 - *Many concerns about IP and much lip service*
 - Can we harvest more?
 - Yes
- 

There are Efforts...

ContentMine

[Get involved](#) ▾

[Resources](#) ▾

[Software](#) ▾

[Events](#)

[Blog](#)

[Login](#)



The ContentMine
uses machines to
liberate 100,000,000
facts from the
scientific literature



RSC >36,000 Articles in 2015

- Consider articles published by RSC in 2015
 - How many compounds?
 - How many reactions?
 - How many figures?
 - How many properties?
 - How many spectra?
 - How many, how many, how many?
- 

The Graph of Relationships is Lost



The flexibility of querying...

Chemical genetics reveals a complex functional ground state of neural stem cells

Phedias Diamandis¹⁻⁴, Jan Wildenhain⁴, Ian D Clarke^{1,2}, Adrian G Sacher^{1,2}, Jeremy Graham^{1,2}, David S Bellows³, Erick K M Ling^{1,2,5}, Ryan J Ward^{1,2,5}, Leanne G Jamieson^{1,2,5}, Mike Tyers^{3,4} & Peter B Dirks^{1,2,5,6}

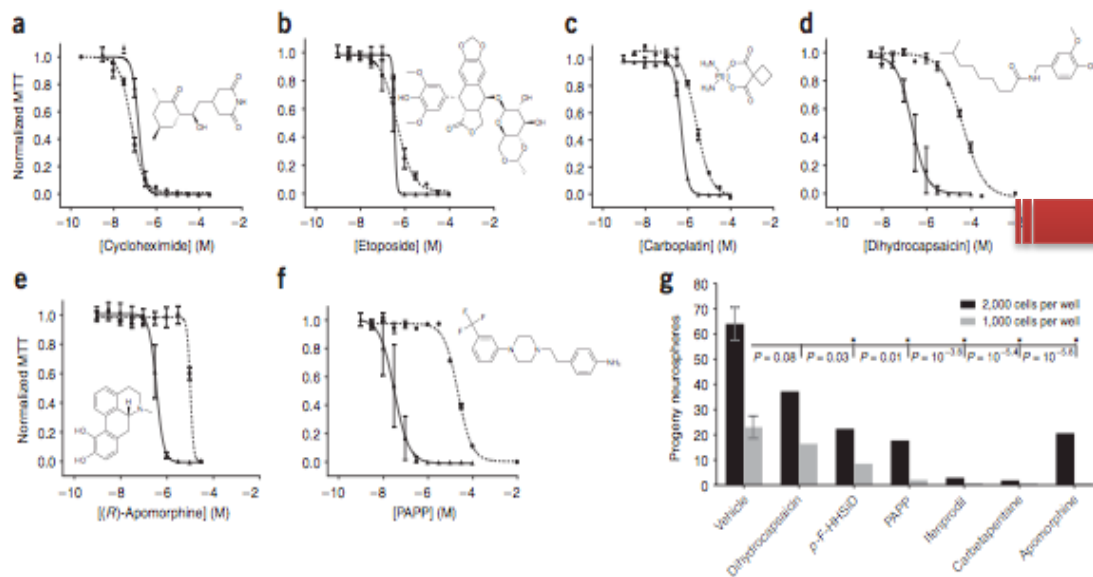


Figure 2 Identification of potent NPC-specific compounds. (a-f) Dose-response curves and chemical structures of controls: cycloheximide (a), etoposide (b) and carboplatin (c), and of selected newly identified compounds: dihydrocapsaicin (d), apomorphine (e) and PAPP (f). Each plot shows the fitted sigmoidal logistic curve to MTT proliferation assay readings of both astrocytes (-●-) and neurosphere cultures (-▲-). Values represent the mean and





Publications-**summary** of work

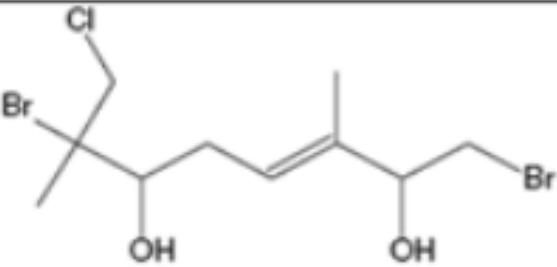
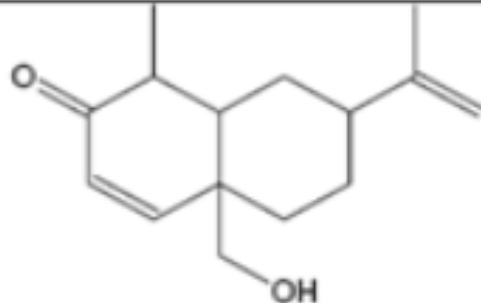
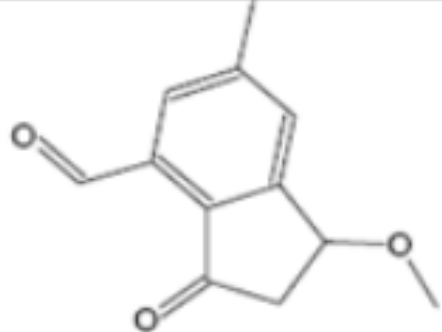
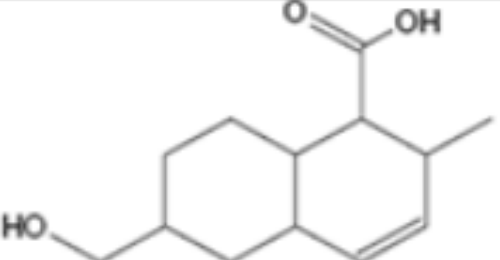
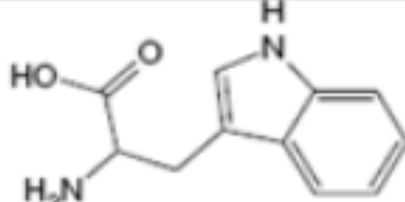
- Scientific publications are a summary of work
 - Is all work reported?
 - How much science is lost to pruning?
 - What of value sits in notebooks and is lost?
 - Publications offering access to “real data”?
 - How much **data** is lost?
 - How many compounds never reported?
 - How many syntheses fail or succeed?
 - How many characterization measurements?
- 



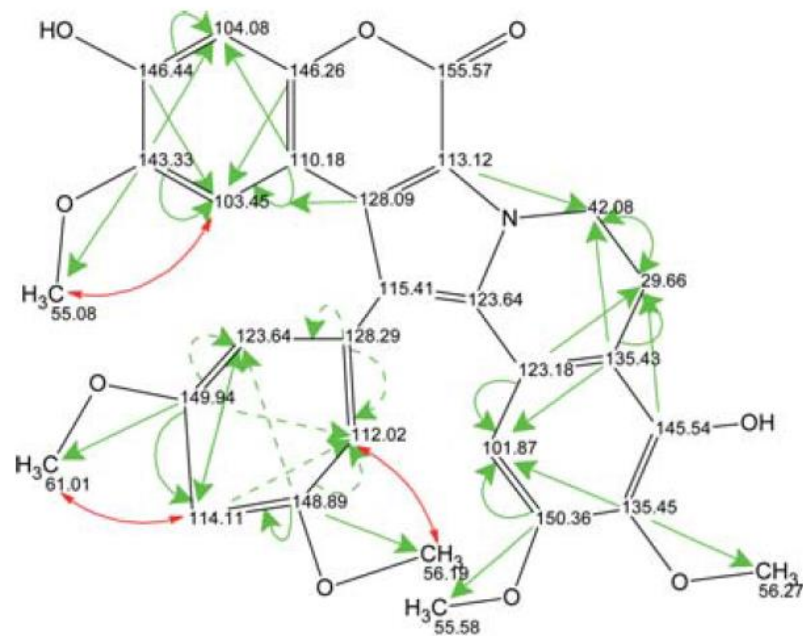
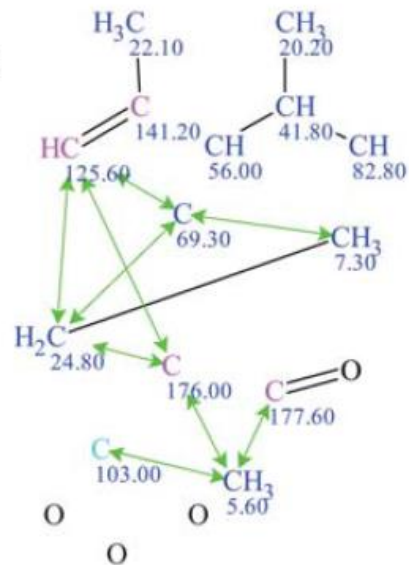
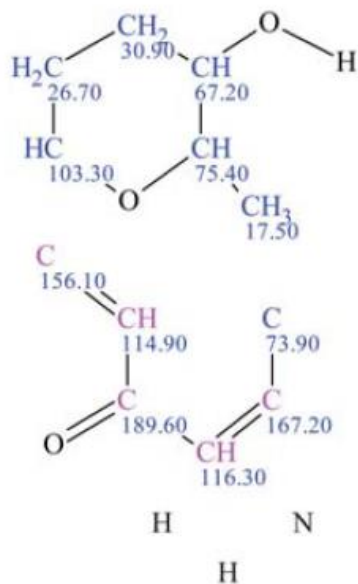
If I wanted to share data...

- I've performed a few dozen chemical syntheses
 - I've run thousands of analytical spectra
 - I've generated thousands of NMR assignments
 - I've probably published <5% of all work..most lost
 -
 - Things can be different today in terms of sharing
 - I would like to share more data, would like at least provenance traced to me and somehow to be acknowledged for the contribution
- 

How Many Structures Can You Generate From a Formula?

 <chem>C10H17Br2ClO2</chem>, 50,502,293	 <chem>C15H22O2</chem>, 138,136,211,624
 <chem>C15H20O1</chem>, 37,568,150,635	 <chem>C12H12O3</chem>, 68,930,547,646
 <chem>C13H20O3</chem>, 14,431,269,166	 <chem>C11H12N2O2</chem>, $3 \cdot 10^{11} < n \rightarrow 10^{12}$

My research...in this CASE



Some NMR...

REVIEW

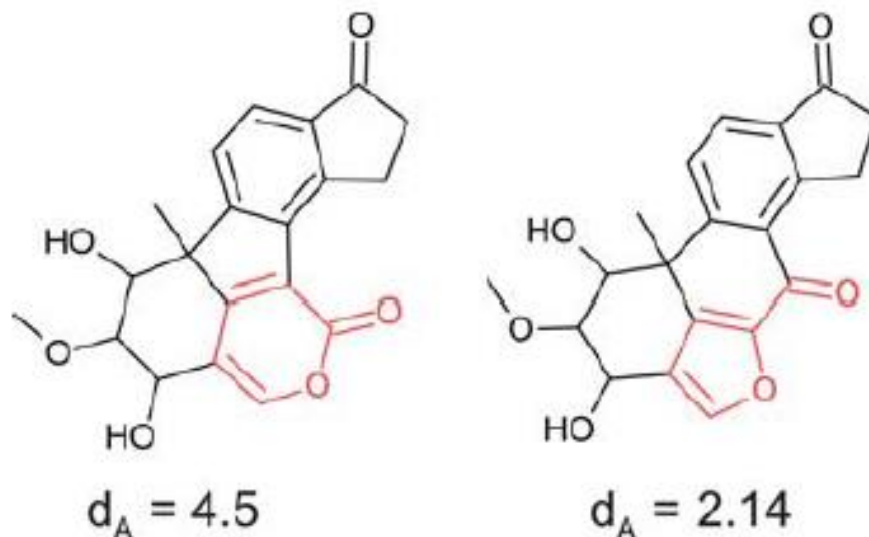
www.rsc.org/npr | Natural Product Reports

Structural revisions of natural products by Computer-Assisted Structure Elucidation (CASE) systems

Mikhail Elyashberg,^a Antony J. Williams^{†*b} and Kirill Blinov^a

Received 2nd February 2010

DOI: 10.1039/c002332a





In researcher mode...

- I want to access and use data
 - I want to:
 - Download molecules
 - Download tables
 - Download spectra
 - Download figures
 - Then reprocess, replot, repurpose
- 



The Challenge of Data Analysis

- NO access to raw data files – in binary or even standard file formats for processing
 - Figures are close to USELESS for 2D NMR – representative not accurate shifts
 - Tabulated shifts are in PDF files and needed transcribing – where are CSV files???
 - TORTUROUS WORK!!!!
 - What if we wanted to do this for all manuscripts submitted to RSC? Of course it is Feasible...
- 



Community Norms

- Some wonderful community norms & mandates!
 - Deposit crystal structures in CSD
 - Deposit Proteins in PDB
 - Deposit gene sequences in Genbank
 - Increasingly deposit bioassay data in Pubchem





But what of general chemistry?

- We publish into document formats
 - *Could* publishers help drive a community norm for:
 - Chemical compound registration
 - Spectral data
 - Property data
 - What else?
 - Who would host it? How would it be funded?
- 

Not even a References Standard

6 References

- 1 C. Steinbeck, V. Spitzer, M. Starosta and G. von Poser, *J. Nat. Prod.*, 1997, **60**, 627–628.
- 2 G. N. Belofsky, M. Anguera, P. R. Jensen, W. Fenical and M. Köck, *Chem. Eur. J.*, 2000, **6**, 1355–1360.
- 3 N. Lysek, E. Rachor and T. Lindel, *Z. Naturforsch., C*, 2002, **57**, 1056–1061.
- 4 D. Mulholland, M. Randrianarivelosoa, C. Lavaud, J.-M. Nuzillard and S. L. Schwikkard, *Phytochemistry*, 2000, **53**, 115–118.
- 5 D. Mulholland, S. L. Schwikkard, P. Sandor and J.-M. Nuzillard, *Phytochemistry*, 2000, **53**, 465–468.
54. Clark AM: **Rendering Molecular Sketches for Publication Quality Output.** *Mol Inf* 2013, **32**:291–301. [Publisher Full Text](#)
55. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, et al.: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *J Chem Inf Com Sci* 1992, **32**:244. [Publisher Full Text](#)
56. Townsend JA, Murray-Rust P: **CMLite: a design philosophy for CML.** *J Cheminf* 2011, **3**:39. [BioMed Central Full Text](#)
57. Rzepa HS, Murray-Rust P, Whitaker BJ: **The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World Wide Web Information Exchange.** *J Chem Inf Comp Sci* 1998, **38**:976–82. [Publisher Full Text](#)
8. Schulz, W. G. *Chem. Eng. News* **2009**, 87 (37), 3.
9. Hettne, K. M.; Williams, A. J.; van Mulligen, E. M.; Kleinjans, J.; Tkachenko, V.; Kors, J. A. *J. Cheminform.* **2010**, 2 (1), 3; DOI: 10.1186/1758-2946-2-3. <http://www.jcheminf.com/content/2/1/3> (accessed Aug 2010).
10. The ChemSpider Blog <http://www.chemspider.com/blog/hacking-pubchem-technology-easy-quality-difficult.html> (accessed Aug 2010).
11. Bradley, J.-C.; Lang, A.; Williams, A. Spectral Game Home Page. <http://www.spectralgame.com/> (accessed Aug 2010).
12. Bradley, J.-C. Open Notebook Science. <http://usefulchem.wikispaces.com/> (accessed Aug 2010).





We can solve for Authors...

Will it be used though??? **YES!**

Authoring Tools

Click on one of the links below:

Author Templates

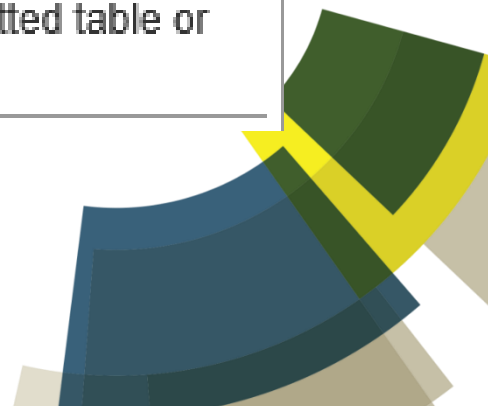
Format your text and structures in the RSC style - we provide downloadable templates and instructions for their use.

Experimental Data Checker and OSCAR Toolkit

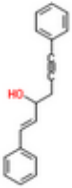
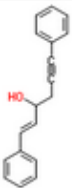
Use our stand-alone application to analyse the experimental data in your manuscript

CIF Data Importer

The CIF data importer is a tool available in the RSC Word templates for authors to import crystallographic structure data from a CIF file as a formatted table or footnote.



Moves in Supplementary Info

1 2 3 4 5 6 7 8 9 10 ...					
ID	Structure	Type File	Submitted	Open Data	Comments
9054739		CNMR 2a-C-NMR.jdx	20/03/2015 15:58:37	False	Compound 2a - 13C NMR made available from PtI2-Catalyzed Cyclization of 3-Acyloxy-1,5-enynes with the Elimination of HOAc and a benzyl Shift: Synthesis of Unsymmetrical m-Terphenyls, Lingyan liu, Jing Li, Xiufang xu, Kaimeng Huang, Hongkai Wang, Junying Wang, Xiaona Ke and Chenchen Zhou, Org. Biomol. Chem., 2015
9054739		HNMR 2a-H-NMR.jdx	20/03/2015 15:57:58	False	Compound 2a - 1H NMR made available from PtI2-Catalyzed Cyclization of 3-Acyloxy-1,5-enynes with the Elimination of HOAc and a benzyl Shift: Synthesis of Unsymmetrical m-Terphenyls, Lingyan liu, Jing Li, Xiufang xu, Kaimeng Huang, Hongkai Wang, Junying Wang, Xiaona Ke and Chenchen Zhou, Org. Biomol. Chem., 2015
34485662		FNMR 3a_19F.jdx	20/03/2015 13:59:09	False	Author supplied fluorine NMR spectrum - See article Electronic Supplementary information
34485662		CNMR 3a_13C.jdx	20/03/2015 13:58:33	False	Author supplied carbon NMR spectrum - See article Electronic Supplementary information
34485662		HNMR 3a_1H.jdx	20/03/2015 13:58:19	False	Author supplied proton NMR spectrum - See article Electronic Supplementary information

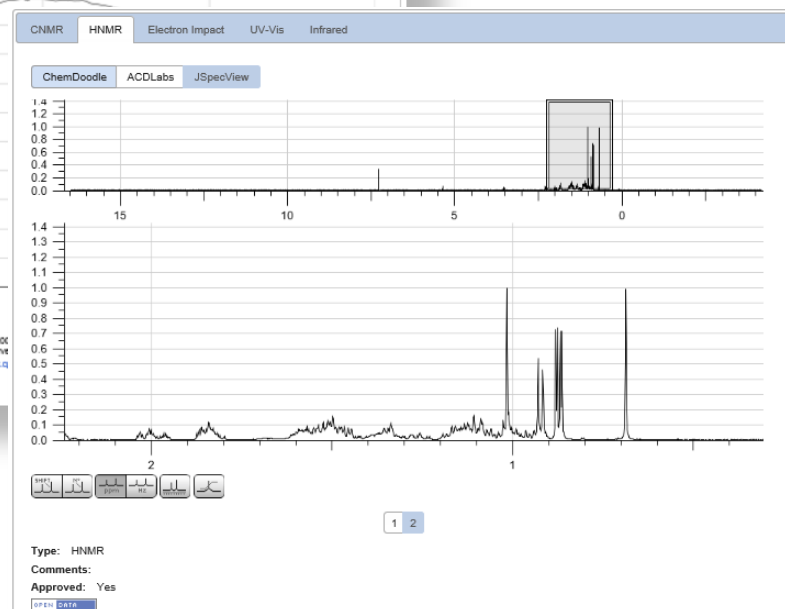
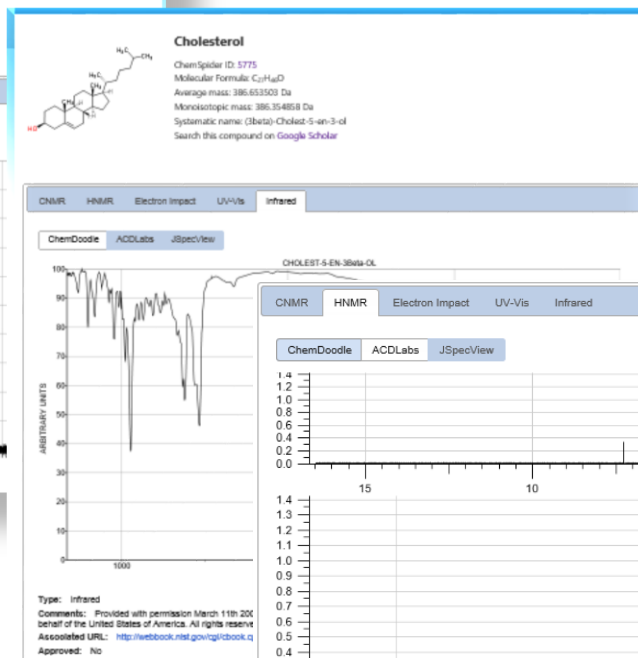
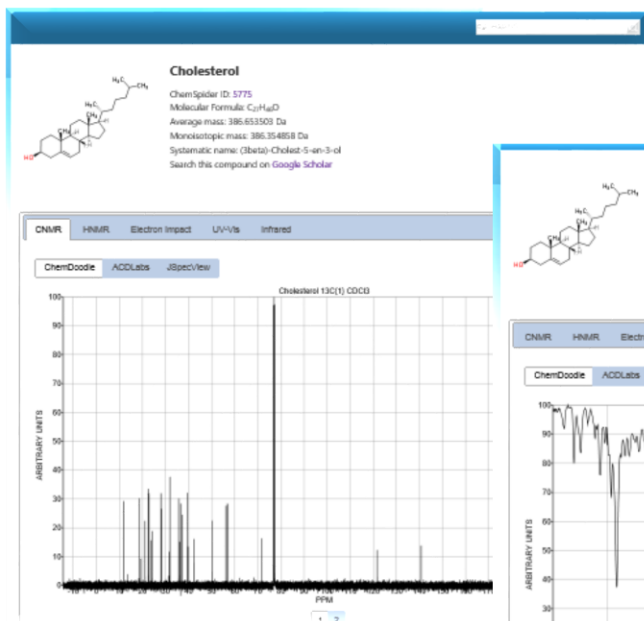


The challenges of analytical data

- Vendors produce complex proprietary data formats and standard formats are required (JCAMP, NetCDF, AniML)
 - ChemSpider already hosts thousands of JCAMP spectra
 - Data validation approaches understood
 - There are a myriad of analytical data types...



Analytical data





Encouraging data deposition

- Open Data mandates don't offer solutions
 - We would like to host:
 - Compounds, Reactions, Spectra, Images, Figures, Graphs etc.
 - We will offer embargoing, collaborative sharing and public release of data
 - Integration to Electronic Lab Notebooks and Institutional Repositories for deposition
- 

RSC Repository Architecture

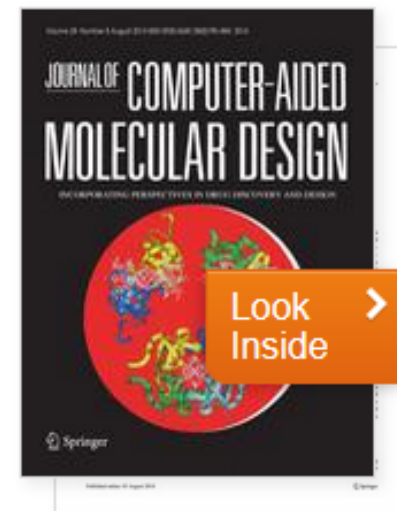
doi: 10.1007/s10822-014-9784-5

Journal of Computer-Aided Molecular Design
August 2014

Date: 03 Aug 2014

The Royal Society of Chemistry and the delivery of chemistry data repositories for the community

Antony Williams, Valery Tkachenko





Registering of Data

- We hear...“We need standards”





There are Standards!

Pure Appl. Chem., Vol. 80, No. 2, pp. 277–410, 2008.

doi:10.1351/pac200880020277

© 2008 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION*

GRAPHICAL REPRESENTATION STANDARDS FOR CHEMICAL STRUCTURE DIAGRAMS**

(IUPAC Recommendations 2008)





There are Standards!

Pure Appl. Chem., Vol. 78, No. 10, pp. 1897–1970, 2006.

doi:10.1351/pac200678101897

© 2006 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION*

GRAPHICAL REPRESENTATION OF STEREOCHEMICAL CONFIGURATION

(IUPAC Recommendations 2006)





There are Standards!

Food and Drug Administration Substance Registration System Standard Operating Procedure

Title: Substance Definition Manual¹	Version: 5c
Effective Date: June 10, 2007	Supersedes: Version 5b
Approved By: SRS Chemistry Review Committee; FDA Data Standards Council's Vocabulary Standards Working Group	



There are standards

- JCAMP, NetCDF, SPC, AnIML for analytical data
- Plus newer efforts in development – Allotrope Foundation efforts



There are Ontologies in Use

The ChEBI Ontology (1)

The ChEBI [ontology](#) is used to classify entities according to their structural and biological properties.

► Name Reaction Ontology (RXNO)

The name reaction ontology connects organic name reactions to their roles in an organic synthesis

► Chemical Methods Ontology (CMO)

The Chemical Methods Ontology describes the methods, instruments and some material artefacts used in chemical experiments.

► Molecular Processes Ontology (MOP)

The molecular processes ontology describes chemical processes that take place at the molecular level, such as methylations and electron transfer

Chemical Information Ontology

[Summary](#) [Classes](#) [Properties](#) [Notes](#) [Mappings](#) [Widgets](#)

MeSH Ontology in OWL format



Registering of Data

- We hear...“We need standards”
 - Many standards exist already!
 - GREAT progress can be made with
 - Data checking and “warnings”
 - Normalization and standardization
 - SIMPLE checks would help databases
 - “High-quality databases” have rigorous checks in place
- 

Data Quality Issues

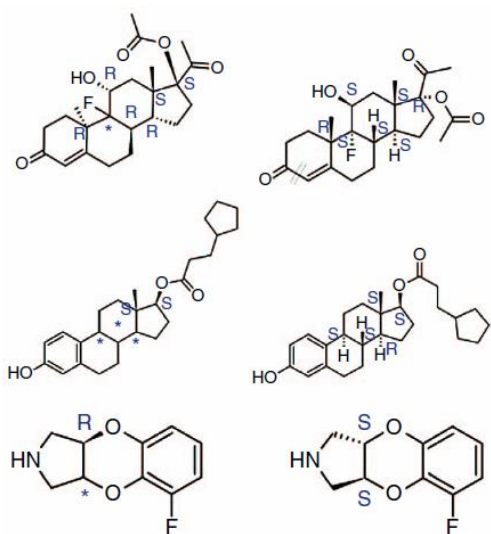
Williams and Ekins, DDT, 16: 747-750 (2011)

PERSPECTIVE

PHARMACOLOGY

The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics

Ruili Huang,* Noel Southall,* Yuhong Wang, Adam Yasgar, Paul Shinn,
Ajit Jadhav, Dac-Trung Nguyen, Christopher P. Austin†



Drug Discovery Today

FIGURE 2

Stereochemistry issues with the content of the NPC browser database: left hand side 'NCGC Structures' and right hand side 'Correct Structures'.

editorial



Antony J. Williams



Sean Ekins

medicine and now drug repositioning or repurposing efforts. Their utility depends on the quality of the underlying molecular structures used. Unfortunately, the quality of much of the chemical structure-based data introduced to the public domain is poor. As an example we describe some of the errors found in the recently released NIH Chemical Genomics Center 'NPC browser' database as an example. There is an urgent need for government funded data curation to improve the quality of internet chemistry and to limit the proliferation of errors and wasted efforts.

A quality alert and call for improved curation of public chemistry databases

In the last ten years, public online databases have rapidly become trusted valuable resources upon which researchers rely for their chemical structures and data for use in cheminformatics, bioinformatics, systems biology, translational

US funding agencies have been investing in the development of public domain chemistry platforms with the primary attention being given to the informatics platform itself rather than the quality of the data content. This is clearly exemplified by the recently released NPC browser from the NIH Chemical Genomics Center (NCGC) [1]. Public online databases such as PubChem, ChemID-Plus [2] and the EPA's ACToR [3], to name just a few, have rapidly become trusted valuable resources which researchers rely on for downloadable chemical structures and associated data. While online chemistry databases can certainly be of value, we feel the reader should be immediately alerted to consider issues of data quality when using these resources and we call into question both their status and the trust we place in them. To our knowledge the issues we raise, using the example of a recently released database, have not been described elsewhere and the user community, and funding agencies, should not ignore them any longer. The development of cheminformatics platforms without due care given to the data quality they contain, is a poor strategy for long term science.

In the last decade numerous attempts have been made to expand our understanding of biological mechanisms by producing vast ligand and protein-protein interaction databases and by the application of computational methods to mine the data and, where possible, develop computational models. These approaches have enabled the clustering of biological activity spectra similarity

Data quality is a known issue

The screenshot displays a chemical database interface with a search bar at the top right containing the text "neomycin". Below the search bar, there are tabs for "Molecular", "Topology", "Framework", and "Ungr". The "Molecular" tab is selected. Below the tabs, there is a row of six chemical structures, each with a label underneath. The first, third, and fifth structures are highlighted with red boxes. The labels for these highlighted structures are "Neomycin", "Neomycin", and "Neomycin" respectively. The second structure is labeled "Neomycin undec..." and the sixth is labeled "Framycetin". The fourth structure is labeled "Triamcinolone h...".

Search results for "neomycin" under the "Molecular" tab:

- Neomycin (highlighted)
- Neomycin undec...
- Neomycin (highlighted)
- Triamcinolone h...
- Neomycin (highlighted)
- Framycetin

Data quality is a known issue

1/14

Identifiers Regulatory Status Therapeutics Properties

Name
Selenium oxide, cadmium salt (1:1)

PubChem CID

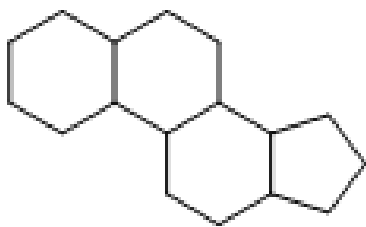
CAS
[13446-34-9](#) [7785-87-7](#)
[10034-96-5](#) [10101-68-5](#)
[13444-72-9](#) [21005-91-4](#)
[12640-89-0](#) [13768-86-0](#)
[13814-59-0](#) [15702-34-8](#)
[29121-98-0](#) [7446-08-4](#)

Synonyms **Identifiers**
Selenium oxide, manganese (+2)
Manganese chloride (mncl2)
Manganese
Selenium oxide, copper (+2) salt
Tetraborate
Strontium tetraborate (1:1)
Barium tetraborate
Manganese proteinate
Hydrogen tetraborate
Manganese hyp chelate
Sodium selenium oxide
Manganese(2+) sulfate (1:1)
Manganese sulfate monohydrate
Manganese sulfate tetrahydrate
Sodium selenium trioxide
Manganese(2+) sulfate (1:1) mor
Manganese sulfate
Manganese succinate
Manganese hypophosphite

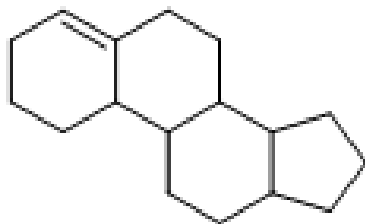
Chemical Structure
 Mn^{2+}
Selenium oxide, cadmium salt (1:1)

Targets (0) Clinical Trials (0) Documents (0) Product Labels (1) Fragments (0) Topologies (0)

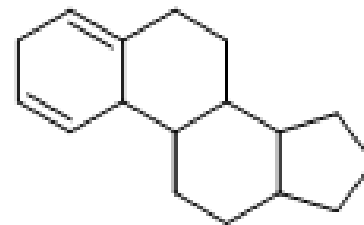
One of the "drugs" from the HTC Screening Set in the NPC Browser



gonane



gon-4-ene

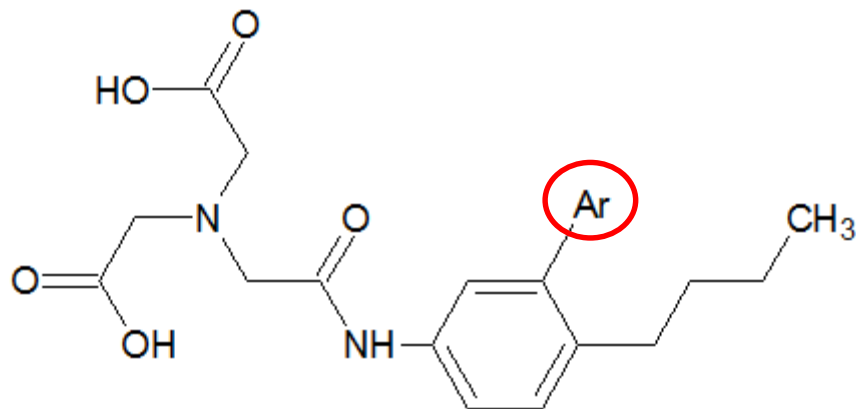
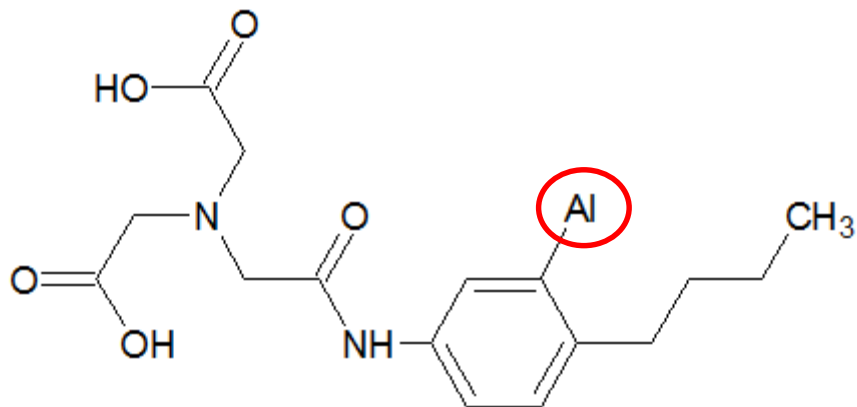


gona-1,4-diene

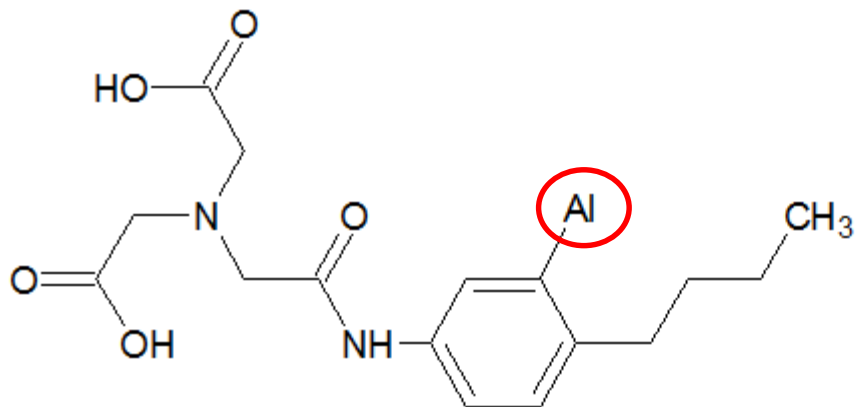
Substructure	# of Hits	# of Correct Hits	No stereochemistry	Incomplete Stereochemistry	Complete but incorrect stereochemistry
Gonane	34	5	8	21	0
Gon-4-ene	55	12	3	33	7
Gon-1,4-diene	60	17	10	23	10

Only 34 out of 149 structures were correct!

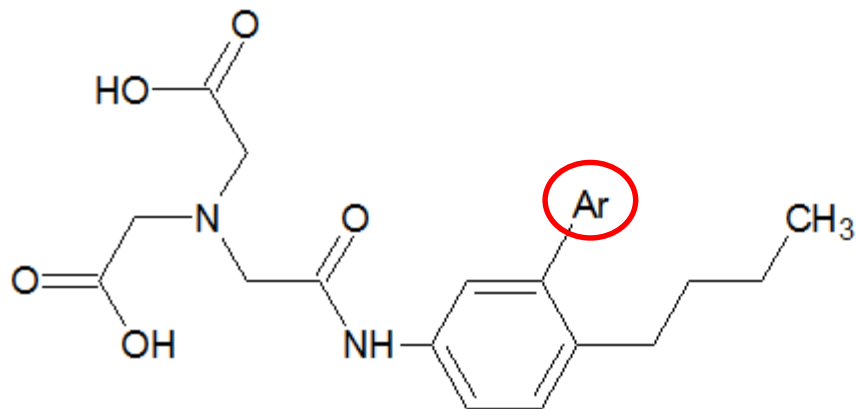
Patent data in public databases



Patent data in public databases

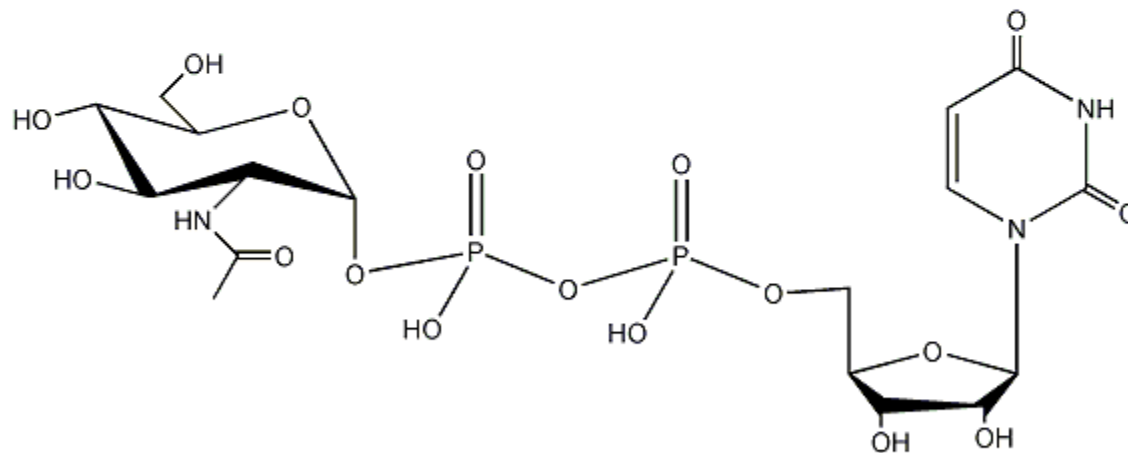


Al= Me, Et, i-Pr



Ar= Ph, Naphthyl

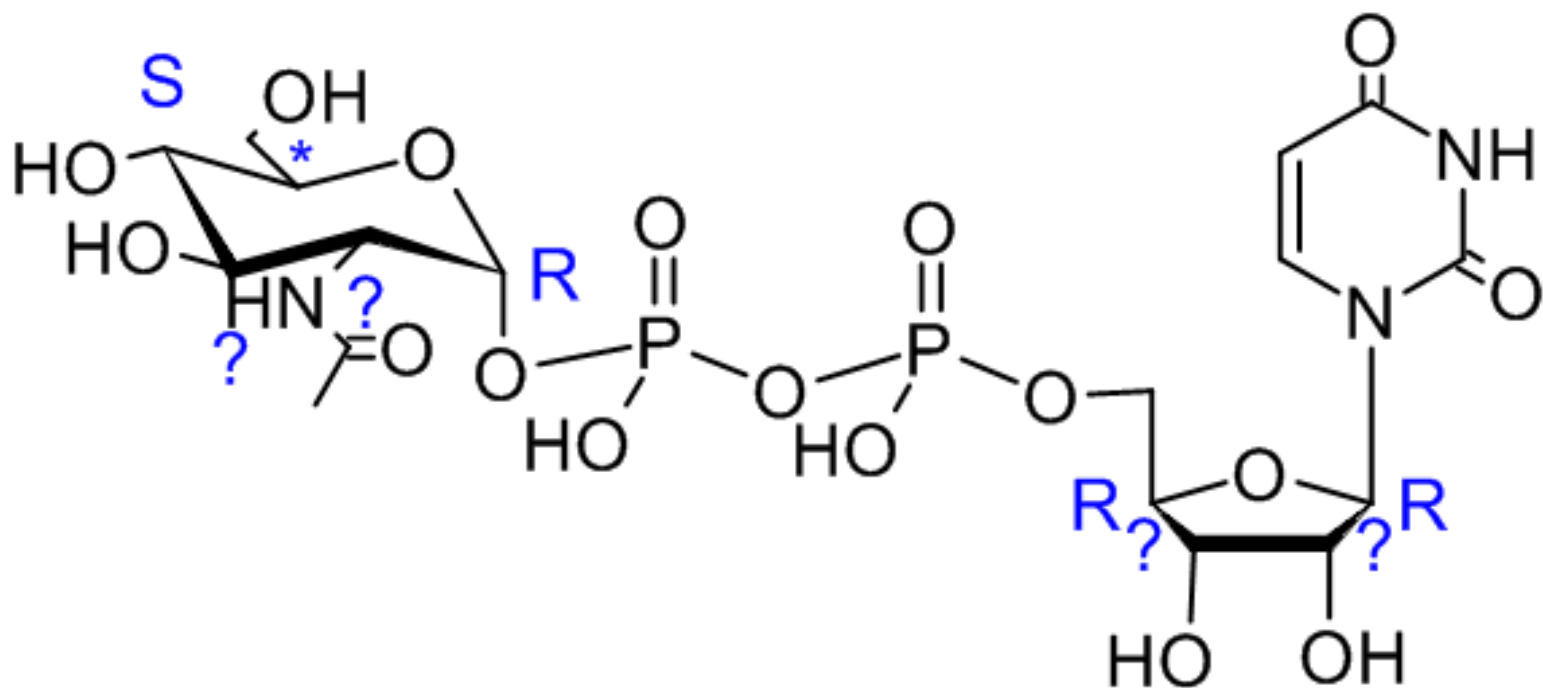
Structure database (LMSD)



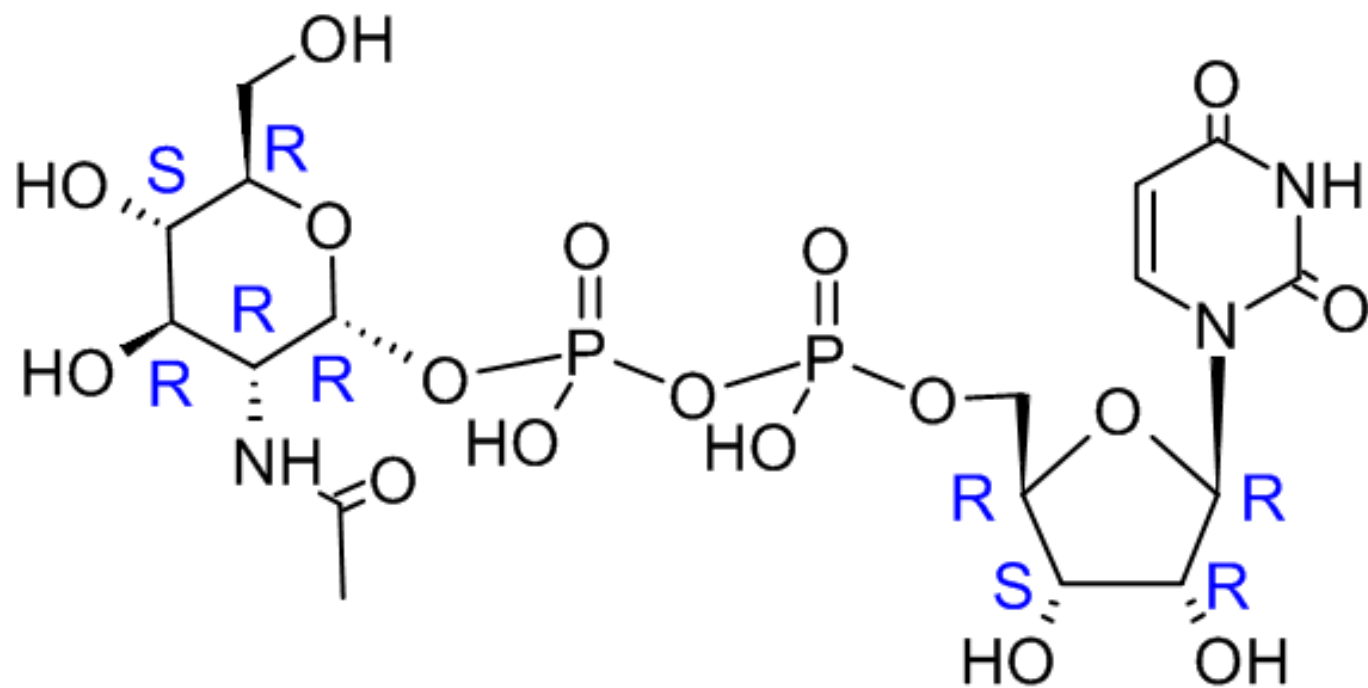
LMSL01010002

UDP-GlcNAc

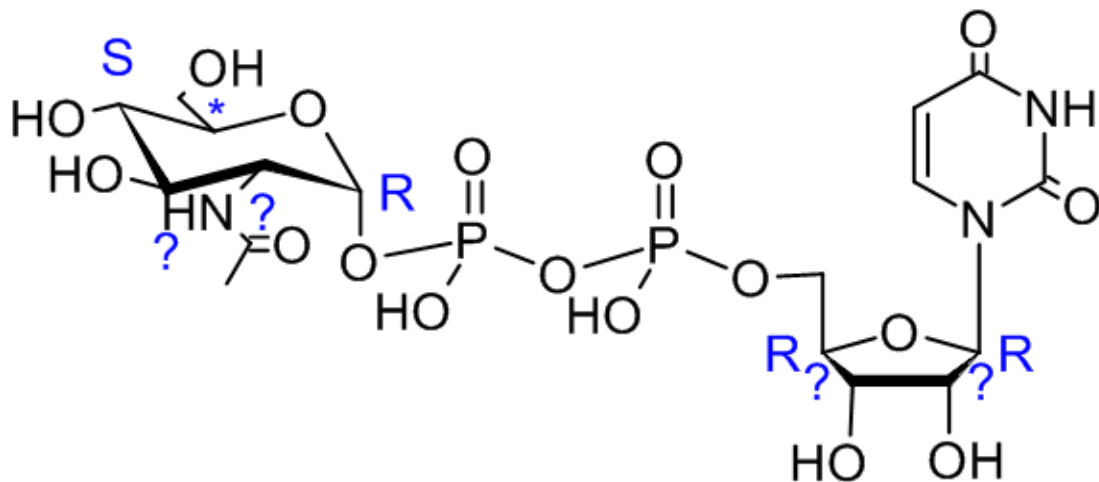
UDP-N-acetyl- α D-glucosamine



LFTYTUAZOPRMMI-OIYKPMIKSA-N



LFTYTUAZOPRMMI-CFRASDGPSA-N



LFTYTUAZOPRMMI-OIYKPMIKSA-N



LFTYTUAZOPRMMI-OIYKPMIKSA-N

Everything

Images

Videos

News

Shopping

More

Show search tools

Your search - **LFTYTUAZOPRMMI-OIYKPMIKSA-N** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

EXPERTS must get it right?!



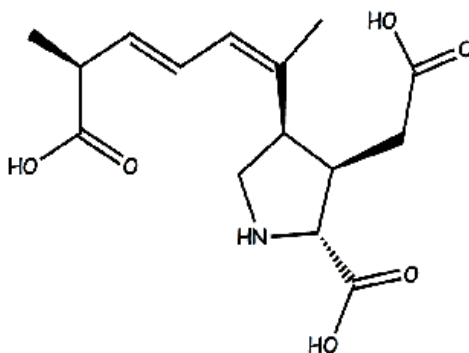
[Home](#) • [HABs & Biotoxins](#) • [Research](#) • [Outreach](#) • [Links](#) • [Search](#)

Domoic Acid Poisoning

Domoic acid has been responsible for several deaths and both permanent and transitory illness in over a hundred people. The toxin is produced by marine diatoms which are members of the genus *Pseudo-nitzschia*. Both shellfish and fish can accumulate this toxin without apparent ill effects; however, in humans the toxin crosses into the brain and interferes with nerve signal transmission. People poisoned with very high doses of the toxin can die, while lower doses can cause permanent brain damage (short term memory loss). When this toxin was discovered in certain West coast fish and shellfish, both recreational and commercial fisheries were briefly closed. This closure, though relatively short, had serious economic impacts on those communities dependent on these fisheries.

Harmful effects of Domoic Acid

The first reported outbreak of domoic acid poisoning occurred in 1987 when shellfish from Prince Edward Island Canada were consumed. In that outbreak, 3 people died and over 100 people developed various toxic symptoms. Domoic acid was found to be produced by the diatom *Pseudo-nitzschia multiseries*. The most unusual, and most serious toxic symptom, was a loss of short term memory--hence the initial designation of the syndrome in humans as amnesiac shellfish poisoning (ASP). However, since the toxin has been found in fin-fish and the chemical structure of the toxin is now known, a more accurate term is Domoic Acid Poisoning. In 1991, along the beaches of Monterey Bay, CA, dead and dying seabirds



Domoic Acid

HABs & Biotoxins



OVERVIEW



PHYTOPLANKTON



MARINE BIOTOXINS

- Detection & Analyses
- Domoic Acid Poisoning
- Paralytic Shellfish Poisoning
- Diarrhetic Shellfish Poisoning

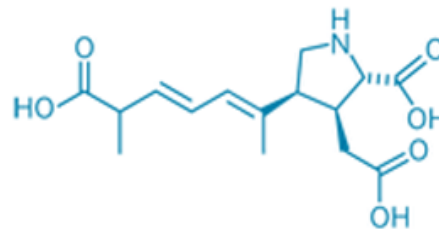


WEST COAST HABs

The value of a validated dictionary

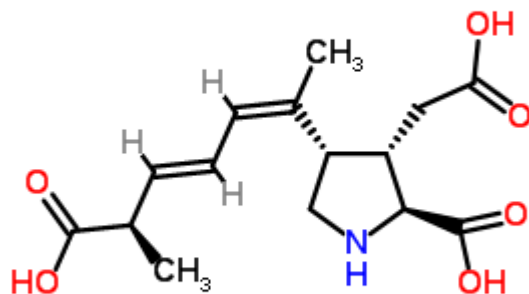
oic acid
ae can
ore
ans than
rding to a
nce (DOI:
sonal
chia

arge amounts of domoic acid, which can cause
y loss in people who ingest it in
one, whales, and marine birds

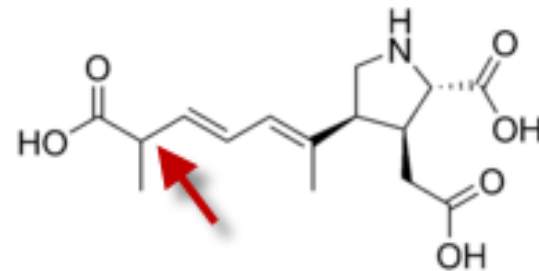


Domoic acid

» S
» M
Mas
moi
dis
» In
Scie
lipic



Domoic acid



Compounds are challenging...



Journal of
Cheminformatics

IMPACT
FACTOR
4.54

Research article

Open Access

Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data

Alex M Clark^{1*}, Antony J Williams² and Sean Ekins^{3,4}



The Open PHACTS community ecosystem





Open PHACTS

- Innovative Medicines Initiative EU project
 - 16 Million Euros, 3 years – meshing chemistry and biology *Open Data* primarily
 - Semantic web project and driven by ODOSOS – Open Data, Open Source, Open Standards
 - RSC developed the chemistry registration system and “CVSP”
- 

CVSP: Validate and Standardize

[Home](#)[Submit](#)[Depositions](#)[Profile](#)

Compounds

Supported formats and extensions of structure files:

CDX (*.cdx, *.cdx.gzip, *.cdx.zip)

MOL (*.mol, *.mol.gzip, *.mol.zip)

SDF (*.sdf, *.sdf.gzip, *.sdf.zip)

Files up to 40Mb are allowed. Zip files should not contain folders, only files.

No file selected.

CVSP Rules Sets

XML Content

```
<?xml version="1.0" encoding="utf-8" ?>
<rules>

  <!-- CVSP modules (value should be from the case-sensitive list of supported modules
list below):

    Dearomatize
    Aromatize
    ConvertDoubleBondWithAttachedEitherSingleBondStereo2EitherDoubleBond
    Layout
    StandardizeHexagons
    Disconnect_Metals_from_NonMetals
    Disconnect_Metals_From_NOF
    Ionize_Neutral_Alkaline_Metals_With_Carboxylic_Acids
    Apply_CVSP_AcidBase_SMIRKS
    Remove_Free_Metals
    Standard_InChI_Normalization
    CanonicalizeTautomers
    Retain_Largest_Organic_Fragment
    NeutralizeCharges

    ConvertUpOrDownBondAdjacentToDoubleBondToNoStereoSingleBondAndCrossedDoubleBond
    Remove_Water
    TreatAmmonia
    Remove_Neutral_Inorganic_Residue
```

Validation Rules

Acid-Base Rules

Standardization Rules

CVSP Filtering of DrugBank

Errors (27)

Warnings (745)

Information (4830)

All Records (6764)

Advanced Filtering And Download

« ‹ 1 2 3 › »

Ordinal

REGID

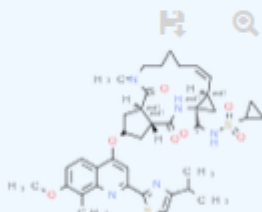
Original

Issues

Standardized

1364

DB06290



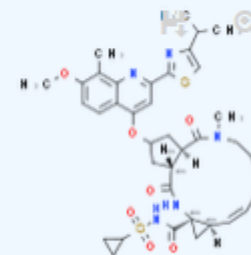
1 molfile loader (indigo): direction of bond #26 makes no sense

2 Contains relative stereo: sp3 stereo bonds are present with no chiral flag

3 Contains partially undefined stereo – epimers

4 Non-standard InChI can't be validated

5 V2000 relative stereo treated as absolute: chiral flag set to ON



CVSP Filtering of DrugBank

2026

DB01706

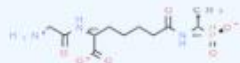


e Non unique dearomatization (indigo): Dearomatization is not unique



2173

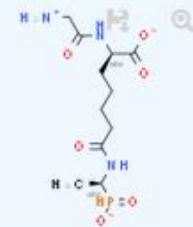
DB01868



e molfile loader (indigo): direction of bond #18 makes no sense

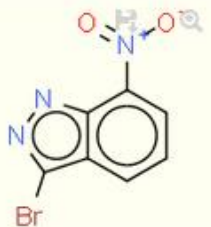
i Contains non-neutral component

w Not an overall neutral system

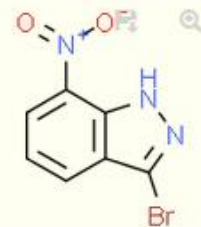


2288

DB01997

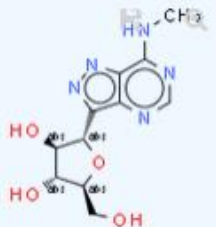


e Non unique dearomatization (indigo): Dearomatization is not unique

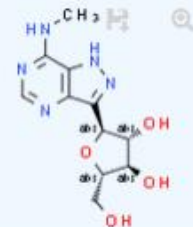


2351

DB02066



e Non unique dearomatization (indigo): Dearomatization is not unique



CVSP is Open to Anyone!

ID	File	Submission Time	User	Records				
32	Drugbank_4.2all.sdf	12/03/2015 20:48	Andy Fant	6764	☒	Review		
28	Fourches_data_4.sdf	03/03/2015 14:10	Eleni Kotsampasakou	947	☒	Review		
27	Fourches_data_4.sdf	03/03/2015 13:45	Eleni Kotsampasakou	947	☒	Review		
26	chembl3cutdown.sdf	15/01/2015 15:12	Louisa Bellis	14000	☒	Review		
25	NonNeutralTest.mol	15/01/2015 14:54	Louisa Bellis	1	☒	Review		
24	NonNeutralTest2.mol	15/01/2015 14:52	Louisa Bellis	1	☒	Review		
23	UndefinedStereo Mixtures.mol	15/01/2015 14:45	Louisa Bellis	1	☒	Review		
22	Morethanoneinstanceofsame.mol	15/01/2015 14:41	Louisa Bellis	1	☒	Review		
21	ChiralBondButNot Chiral.mol	12/01/2015 18:30	Louisa Bellis	1	☒	Review		
15	molecule (2).mol	04/01/2015 09:12	NIKHIL ARORA	1	☒	Review		



What if...

- CVSP was used to check molecular files before submitting to publishers or databases?
 - Publishers used CVSP to check their data?
 - All rules were openly available for adoption?
 - Standards, a community norm, access to data
- 



What if we could do the same...

- Check/validate procedures:
 - File format checking (think CIF checker)
 - Nomenclature checking
 - Compare experimental vs. predicted data and flag suspicious data for inspection
 - Physchem parameter comparisons
 - NMR shift prediction (and assignment)
- 



Building a **BIG Data** Repository

- We have validation procedures in place:
 - Compound validation
 - Reaction checking
 - Analytical data formats (in development)
 - But **how long** to get to a Big Data Repository?
 - Users want to get data more than contribute!
 - Where can we find data???
- 

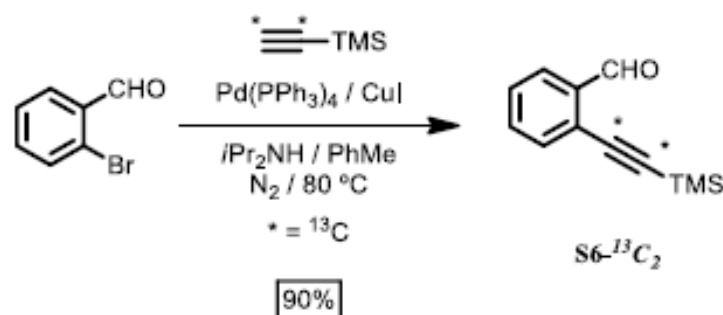


The RSC Archive

- Over 300,000 articles containing chemistry
 - Compounds, reactions, property data, spectral data, the usual....
 - Document formats to analyze and extract
 - Previous experience with “Prospecting” compounds
- 

Electronic Supplementary Info

Scheme S13. Synthesis of **S6-¹³C₂**



Synthesis of S6-¹³C₂ 2-bromobenzaldehyde (0.526 g 2.84 mmol), CuI (5 mg, 0.028 mmol), and Pd(PPh₃)₄ (15 mg, 0.013 mmol) were loaded in a 25 mL Schlenk flask equipped with a magnetic stirrer. The flask was evacuated under dynamic vacuum to 150 mtorr and backfilled with N₂ three times. Anhydrous PhMe (3 mL) and anhydrous *i*Pr₂NH (1 mL) were added via cannula under N₂. The mixture was bubbled with N₂ for 20 min and trimethylsilylacetylene-¹³C₂ (99% atom ¹³C, 0.435 mL, 300 mg, 2.994 mmol) was added dropwise with stirring. The mixture was heated to 80 °C and stirred for 12 h, after which it was quenched with saturated NH₄Cl (aq), and extracted with CH₂Cl₂ (3 x 10 mL). The combined organic extracts were rinsed with saturated NH₄Cl (aq), water, and brine. The solution was dried over anhydrous MgSO₄, filtered over celite, and concentrated to dryness. The obtained dark residue was purified by column chromatography (SiO₂, 5% v/v THF/hexanes) to provide **S6-¹³C₂** (0.554 g, 90% yield) as a yellow oil. ¹H NMR (400 MHz, CDCl₃) δ 10.56 (d, *J*_{CH} = 0.76 Hz, 1H), 7.91 (d, *J* = 7.95 Hz, 1H), 7.57 (m, *J* = 7.85, 1.71, 0.70 Hz, 1H), 7.54 (m, *J* = 7.75, 0.68 Hz, 1H), 7.43 (m, *J* = 7.75, 0.68 Hz, 1H), 0.28 (d, *J*_{CH} = 2.48 Hz, 9H). ¹³C NMR (100 MHz, CDCl₃) δ 102.74 (d, *J*_{CC} = 136.2 Hz), 100.02 (d, *J*_{CC} = 136.3 Hz). EI-MS: |calcd for [C₁₀¹³C₂H₁₄OSi]⁺ 204.09, found 203.15.

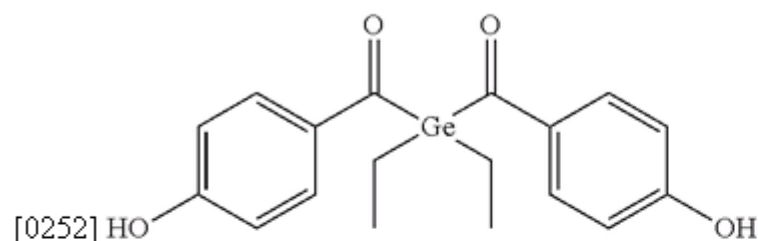


What was our NextMove?

- Daniel Lowe worked on text-mining and named-entity recognition at University of Cambridge
 - Extracted millions of chemical reactions from US Patents
 - Working with NextMove products (LeadMine and CaffeineFix) and optimization by Daniel
- 

What could we get?

Step 4: Bis-(4-hydroxybenzoyl)-diethylgermanium



[0253] Bis-(4-methoxybenzoyl)diethylgermanium (8.0 g, 20.0 mmol) was dissolved in anhydrous toluene (200 ml) under protective gas and had Celatom (10 g) and aluminium chloride (9.6 g, 72.0 mmol) added to it. The reaction mixture was heated for 2 h under reflux. After cooling, water (10 ml) was added and the suspension was stirred for 10 min at room temperature. The solvent was removed on the rotary evaporator. The residue had ethyl acetate (300 ml) added to it. The suspension was stirred for 16 h at room temperature and filtered over a thin layer of silica gel. The filtrate was concentrated on the rotary evaporator. The oily brown residue had chloroform (200 ml) added to it. The suspension was stirred for 16 h at room temperature and filtered. The filtration residue was washed with chloroform (80 ml) and dried in the vacuum drying oven (50° C., 125 mbar). 4.23 g (11.3 mmol, 57% yield) of a light yellow solid was obtained (mp: 167-168° C.).

[0254] ¹H-NMR (DMSO-d₆, 400 MHz): δ=1.04 (t, 6H, J=7.9 Hz, —CH₃), 1.38 (q, 4H, J=7.9 Hz, Ge—CH₂—), 6.88 (d, 4H, J=8.5 Hz, Ar—H^{3,5}), 7.58 (d, 4H, J=8.5 Hz, Ar—H^{2,6}), 10.53 (s, 2H, OH). ¹³C-NMR (DMSO-d₆, 100.6 MHz): δ=6.1 (—CH₃), 8.9 (Ge—CH₂—), 115.7 (Ar—C^{3,5}), 130.3 (Ar—C^{2,6}), 133.3 (Ar—C¹), 162.7 (Ar—C⁴), 225.5 (C=O).

PhysChem first: Melting Points

- Melting/sublimation/decomposition points extracted for **287,635** distinct compounds from 1976-2014 USPTO patent applications/grants
- Sanity checks used to flag dubious values – probably 130-4°C

title compound (4.09 g, 90%) as a tan crystalline solid. mp 1304° C. ¹H NMR (CDCl₃; 300 MHz) δ 3.52 (2H, bs, NH),

- Non-melting outcomes recorded e.g. mp 147-150°C. (subl.)
- What models could be built?

QSPR/QSAR modelling in OCHEM

<http://ochem.eu>

Select model template

Training set (*required*): [...]

Validation set (*optional*): [...]

Choose template for the model:

- ☒ ASNN (ASsociative Neural Networks) 
- ☐ Consensus model (experimental) 
- ☐ FSMLR (Fast Stagewise Multiple Linear Regression) 
- ☐ KNN (K-Nearest Neighbors) 
- ☐ KPLS_mathematica 
- ☐ KRR (Kernel Ridge Regression) 
- ☐ LibSVM wrapper with grid-search parameter optimisation 
- ☐ LogP-LIBRARY 
- ☐ MLR (Multiple Linear Regression) 
- ☐ PLS (Partial Least Square) 
- ☐ SVM (Support Vector Machines) 
- ☐ WEKA-J48 (Weka-based implementation of C4.5 decision tree) 
- ☐ WEKA-RF (Weka-based implementation of Random Forest) 

Model validation

Validation method: N-Fold cross-validation 

Number of folds: 5

☐ Stratified cross-validation

Select descriptor blocks

Please select the MOLECULAR descriptors:

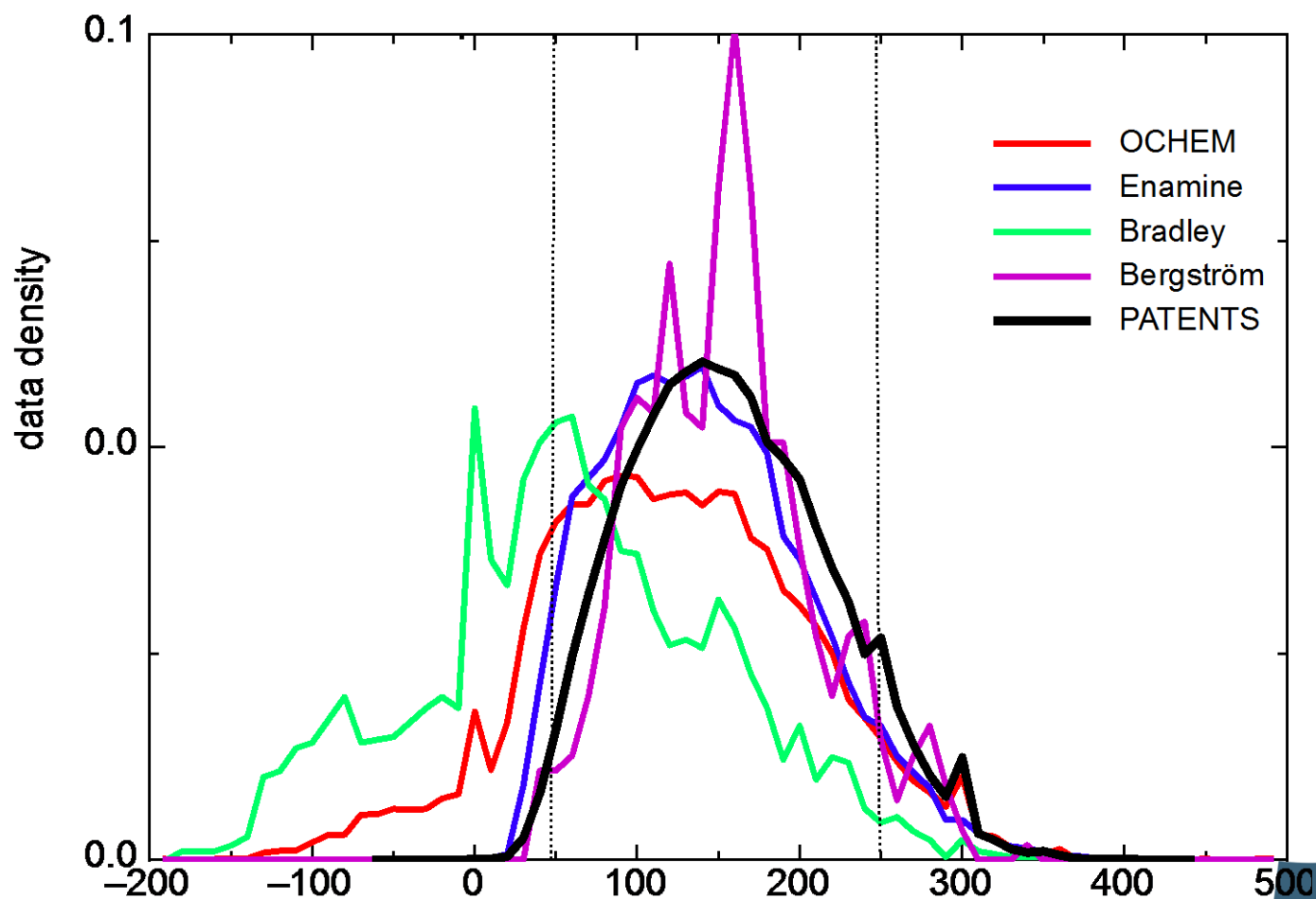
- ☐ E-state 
- ☐ OEState 
- ☐ ALogPS (2) 
- ☐ AMBIT Descriptors 
- ☐ MolPrint 
- ☐ GSFragment (1138) 
- ☐ Dragon v. 5.4 (1630/3D) 
- ☐ Dragon v. 5.5 (3190/3D) 
- ☐ Dragon v. 6.0 (4885/3D) 
- ☐ ISIDA fragments 
- ☐ ISIDA fragments (2011) 
- ☐ MOPAC descriptors (21/3D) 
- ☐ ADRIANA.Code (211/3D) 
- ☐ CDK descriptors (246/3D) 
- ☐ QNPR 
- ☐ ShapeSignatures (3D) 
- ☐ 'Inductive' descriptors (54/3D) 
- ☐ MERA descriptors (529/3D) 
- ☐ MERSY descriptors (42/3D) 
- ☐ Vina Docking descriptors (alfa version)(3D) 
- ☐ Chemaxon descriptors (499/3D) 
- ☐ Chiral Descriptors (/3D) 
- ☐ ETM descriptors 
- ☐ Spectrophores (144/3D) 



Modeling “BIG data”

- Melting point models developed with ca. 300k compounds
 - Required 34Gb memory and about 400MB disk space (zipped)
 - Matrix with **2×10^{11}** entries (300k molecules x 700k descriptors)
 - >12k core-hours (>600 CPU-days) for parameter optimization
 - Parallelized on > 600 cores with up to 24 cores per one task
 - Consensus model as average of individual models
 - Accuracy of consensus model is ~ 33.6 °C for drug-like region compounds
 - Models publicly available at <http://ochem.eu>
- 

Distribution of MPs in the analyzed sets



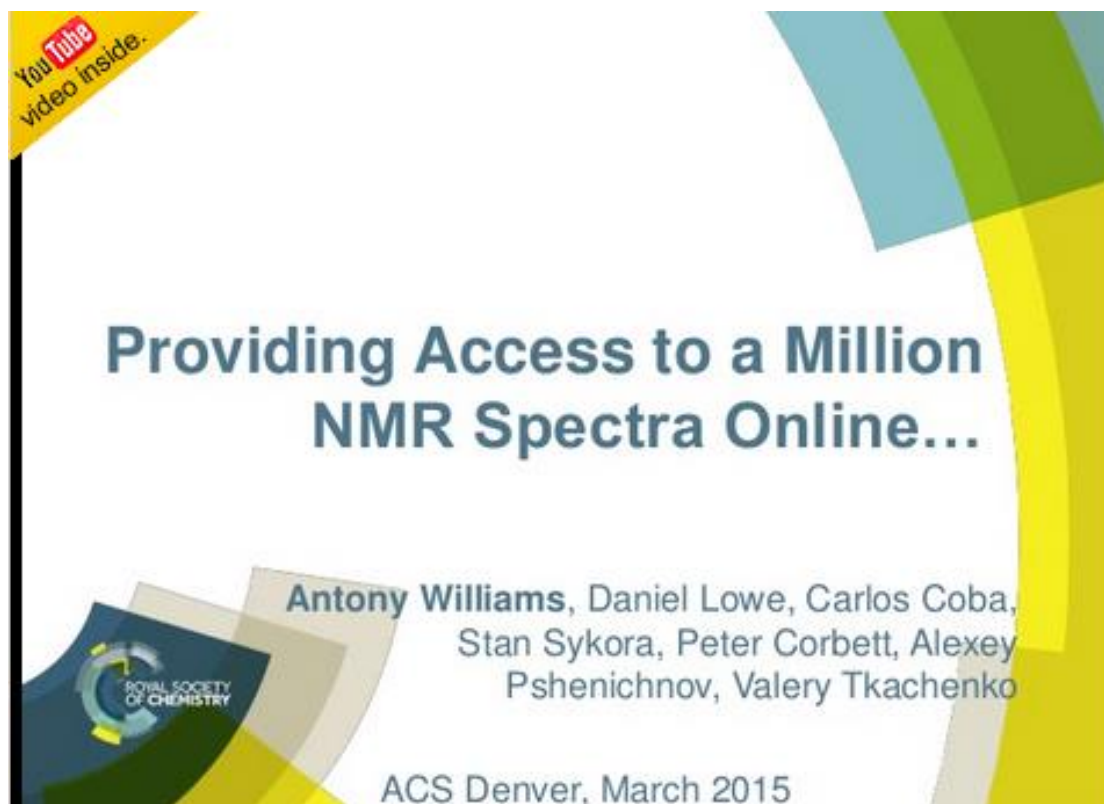


PhysChem parameters

- Melting point model and data – good data extracted and filtered “automagically”
 - Boiling point data next – pressure dependence
 - What next – logP, pKa, aq/non-aq. Solubility
 - Prove the algorithms on US Patent Collection then apply to RSC archive
 - Ideally plumb the algorithms for all new papers
 - More ideal – authors submit DATA!
- 

A Recent Talk at ACS/Denver

<http://www.slideshare.net/AntonyWilliams/>



Spectral Data

▼ Spectra

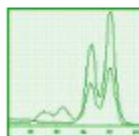


- **Type:** HNMR

Associated Hyperlink: http://blogs.chem.soton.ac.uk/whitby_group/44949/Compound_191_6Bromo22bismethoxymethylhexahydro1Hinden56Hone.html
Comments: ¹H NMR: (300 MHz, CDCl₃) collected by A.Henderson as part of his thesis, under the supervision of Prof. Richard Whitby at the University of Southampton

Approved: No

Submitted by: [Alexandra Hartke](#)



OPEN DATA

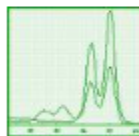
Download

- **Type:** CNMR

Associated Hyperlink: http://blogs.chem.soton.ac.uk/whitby_group/44949/Compound_191_6Bromo22bismethoxymethylhexahydro1Hinden56Hone.html
Comments: ¹³C NMR: (75 MHz, CDCl₃) collected by A.Henderson as part of his thesis, under the supervision of Prof. Richard Whitby at the University of Southampton

Approved: No

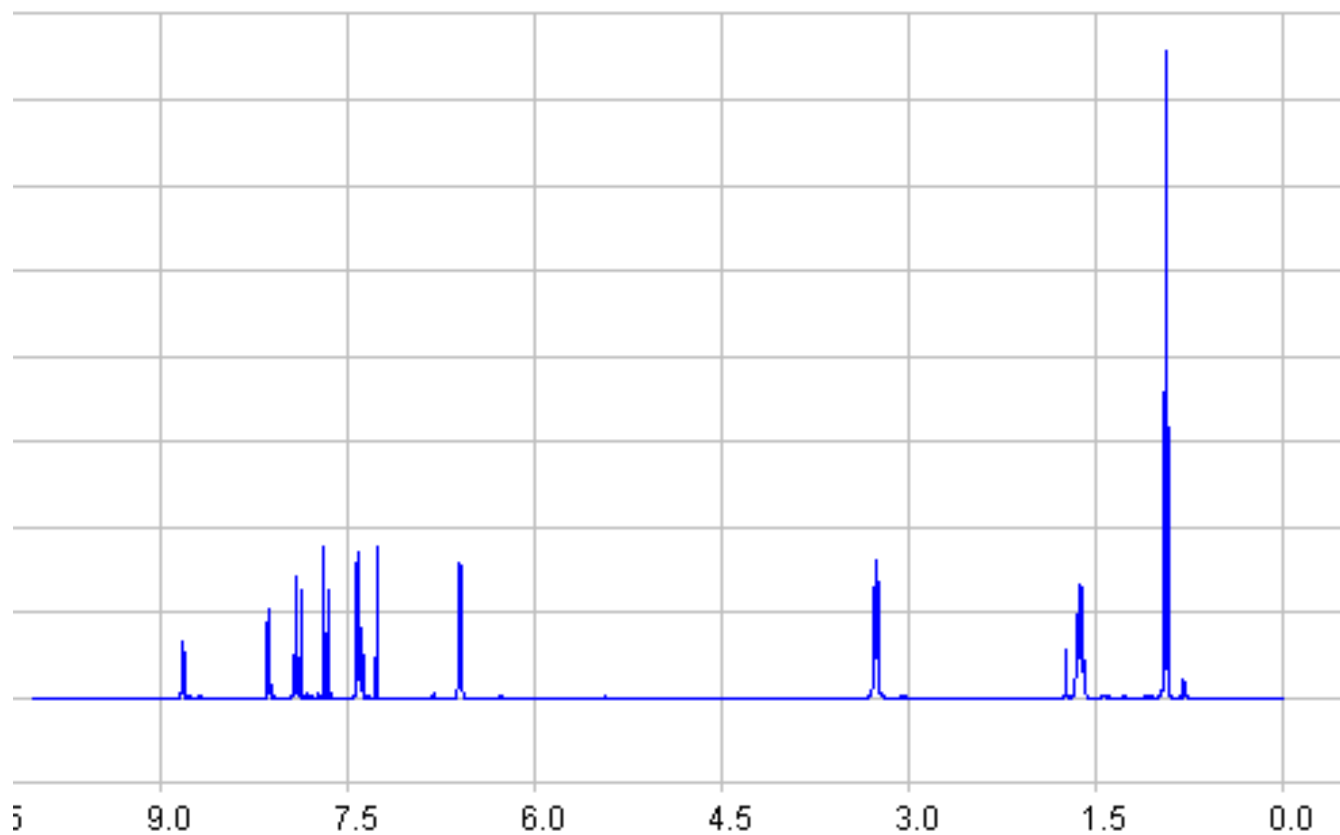
Submitted by: [Alexandra Hartke](#)



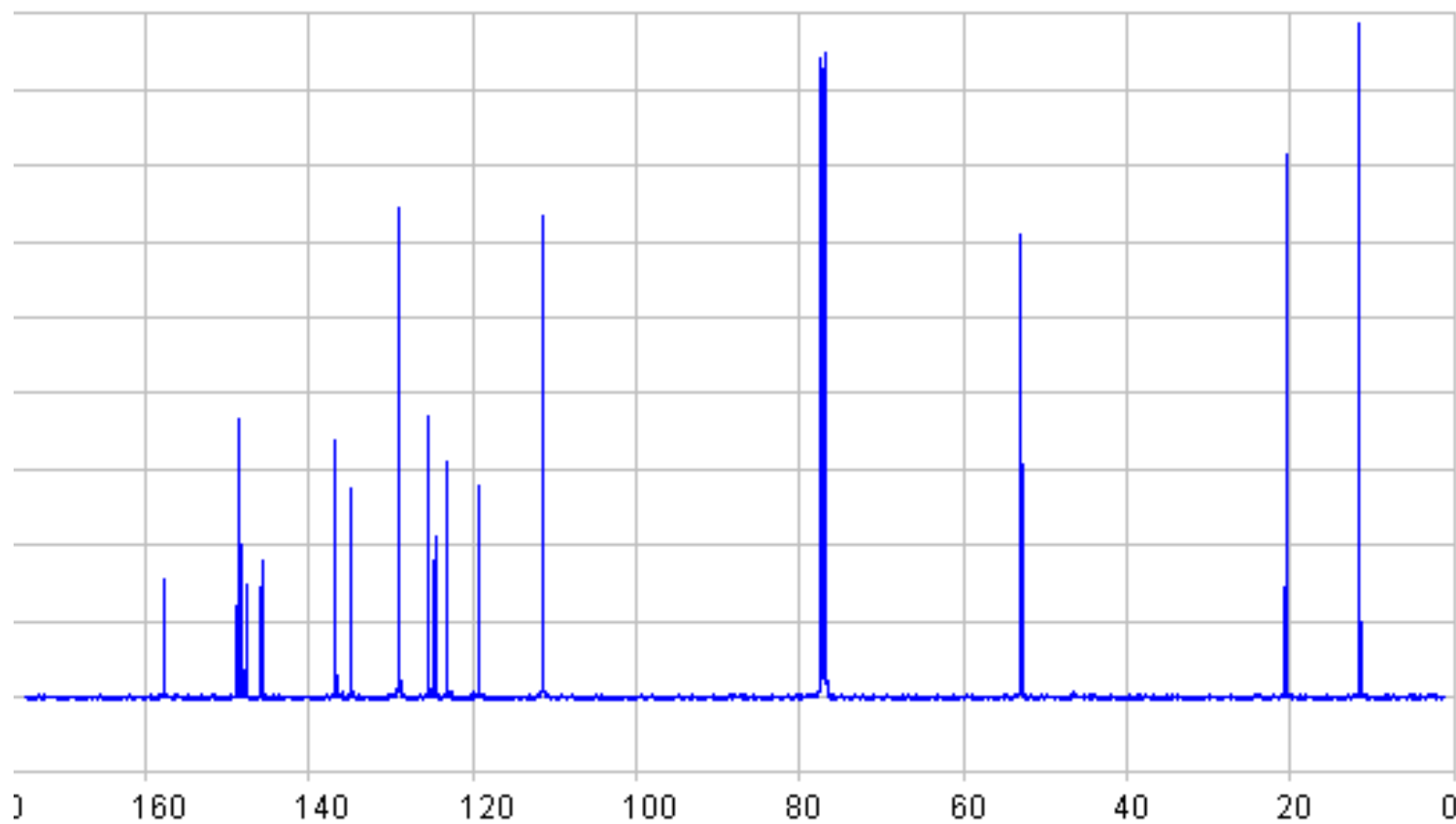
OPEN DATA

Download

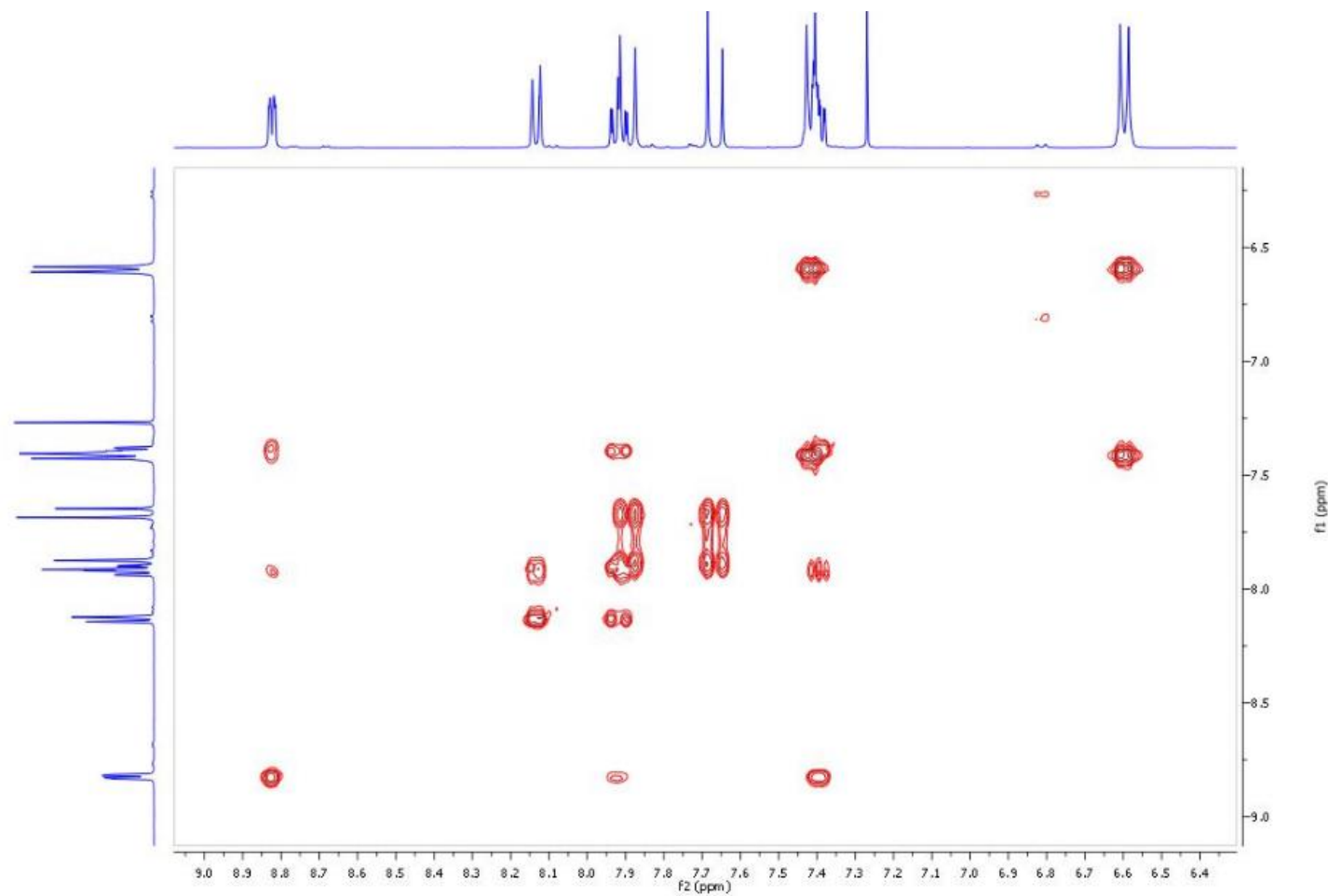
ChemSpider ID 24528095 H1 NMR



ChemSpider ID 24528095 C13 NMR

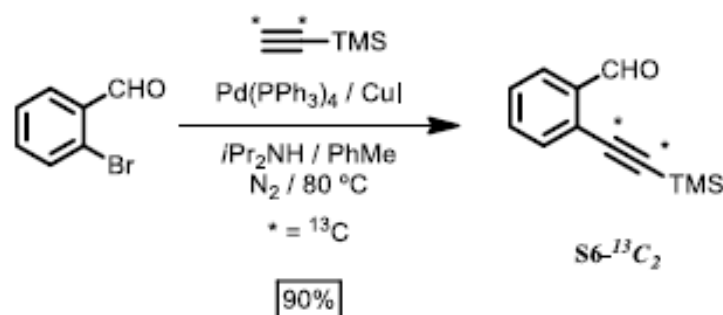


ChemSpider ID 24528095 HHCOSY



ESI – Text Spectra

Scheme S13. Synthesis of **S6-¹³C₂**



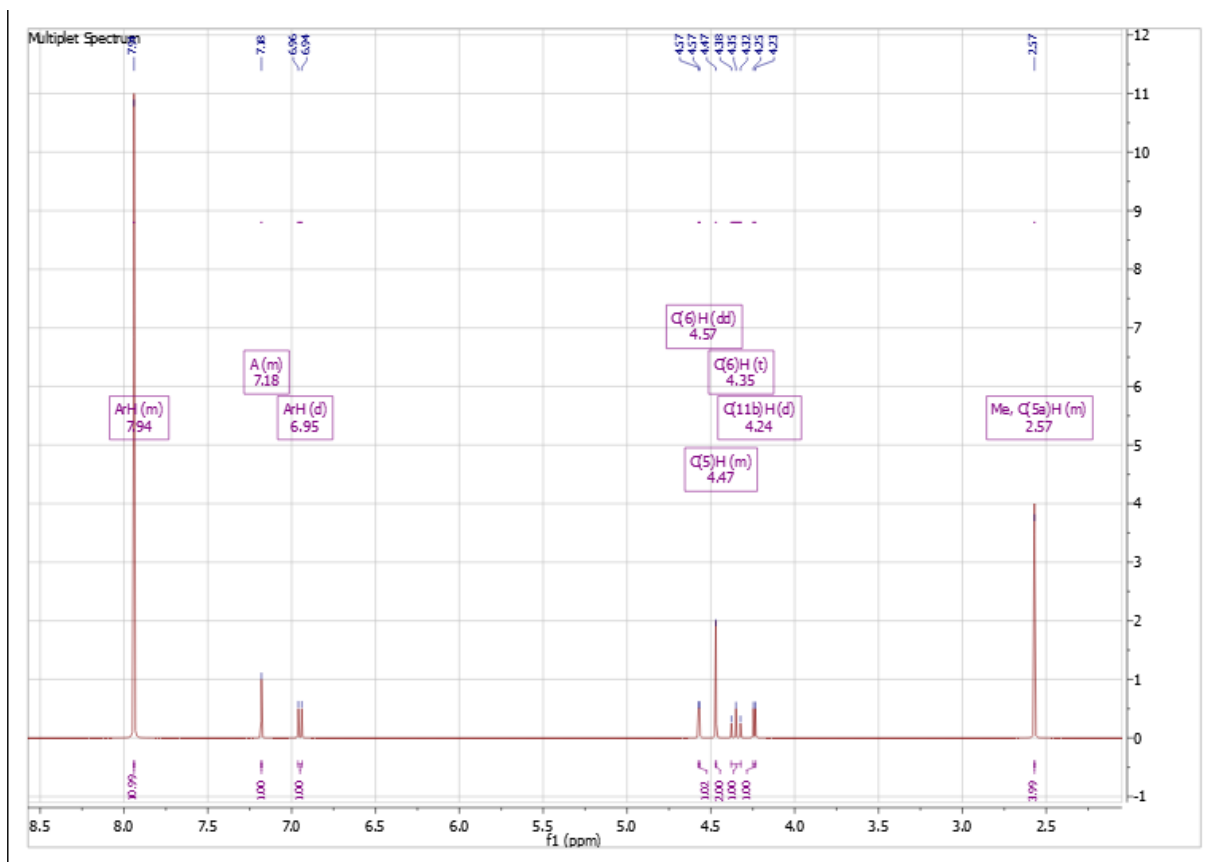
Synthesis of S6-¹³C₂ 2-bromobenzaldehyde (0.526 g 2.84 mmol), CuI (5 mg, 0.028 mmol), and Pd(PPh₃)₄ (15 mg, 0.013 mmol) were loaded in a 25 mL Schlenk flask equipped with a magnetic stirrer. The flask was evacuated under dynamic vacuum to 150 mtorr and backfilled with N₂ three times. Anhydrous PhMe (3 mL) and anhydrous *i*Pr₂NH (1 mL) were added via cannula under N₂. The mixture was bubbled with N₂ for 20 min and trimethylsilylacetylene-¹³C₂ (99% atom ¹³C, 0.435 mL, 300 mg, 2.994 mmol) was added dropwise with stirring. The mixture was heated to 80 °C and stirred for 12 h, after which it was quenched with saturated NH₄Cl (aq), and extracted with CH₂Cl₂ (3 x 10 mL). The combined organic extracts were rinsed with saturated NH₄Cl (aq), water, and brine. The solution was dried over anhydrous MgSO₄, filtered over celite, and concentrated to dryness. The obtained dark residue was purified by column chromatography (SiO₂, 5% v/v THF/hexanes) to provide **S6-¹³C₂** (0.554 g, 90% yield) as a yellow oil. ¹H NMR (400 MHz, CDCl₃) δ 10.56 (d, *J*_{CH} = 0.76 Hz, 1H), 7.91 (d, *J* = 7.95 Hz, 1H), 7.57 (m, *J* = 7.85, 1.71, 0.70 Hz, 1H), 7.54 (m, *J* = 7.75, 0.68 Hz, 1H), 7.43 (m, *J* = 7.75, 0.68 Hz, 1H), 0.28 (d, *J*_{CH} = 2.48 Hz, 9H). ¹³C NMR (100 MHz, CDCl₃) δ 102.74 (d, *J*_{CC} = 136.2 Hz), 100.02 (d, *J*_{CC} = 136.3 Hz). EI-MS: calcd for [C₁₀¹³C₂H₁₄OSi]⁺ 204.09, found 203.15.

We want to find text spectra?

- We can find and index text spectra: ^{13}C NMR (CDCl₃, 100 MHz): δ = 14.12 (CH₃), 30.11 (CH, benzylic methane), 30.77 (CH, benzylic methane), 66.12 (CH₂), 68.49 (CH₂), 117.72, 118.19, 120.29, 122.67, 123.37, 125.69, 125.84, 129.03, 130.00, 130.53 (ArCH), 99.42, 123.60, 134.69, 139.23, 147.21, 147.61, 149.41, 152.62, 154.88 (ArC)
- What would be better are spectral figures – and include assignments where possible!

^1H NMR (CDCl_3 , 400 MHz):

δ = 2.57 (m, 4H, Me, C(5a)H), 4.24 (d, 1H, J = 4.8 Hz, C(11b)H), 4.35 (t, 1H, J_b = 10.8 Hz, C(6)H), 4.47 (m, 2H, C(5)H), 4.57 (dd, 1H, J = 2.8 Hz, C(6)H), 6.95 (d, 1H, J = 8.4 Hz, ArH), 7.18–7.94 (m, 11H, ArH)



MestreLabs Mnova NMR

Mnova NMR



Overview

NMR processing, analysis, simulation and reporting at your fingertips.

This vendor independent ([check supported formats](#)) NMR software is ideal both for the expert user requiring advanced functionality and for organic or synthetic chemists focused on routine processing with maximum productivity.



NMR Spectra

- 2,316,005 distinct spectra in 2001-2015 USPTO

Nucleus	Count
H	1993384
C	173970
Unknown	107439
F	22158
P	16333
B	980
Si	715
Pt	275
N	170
V	101

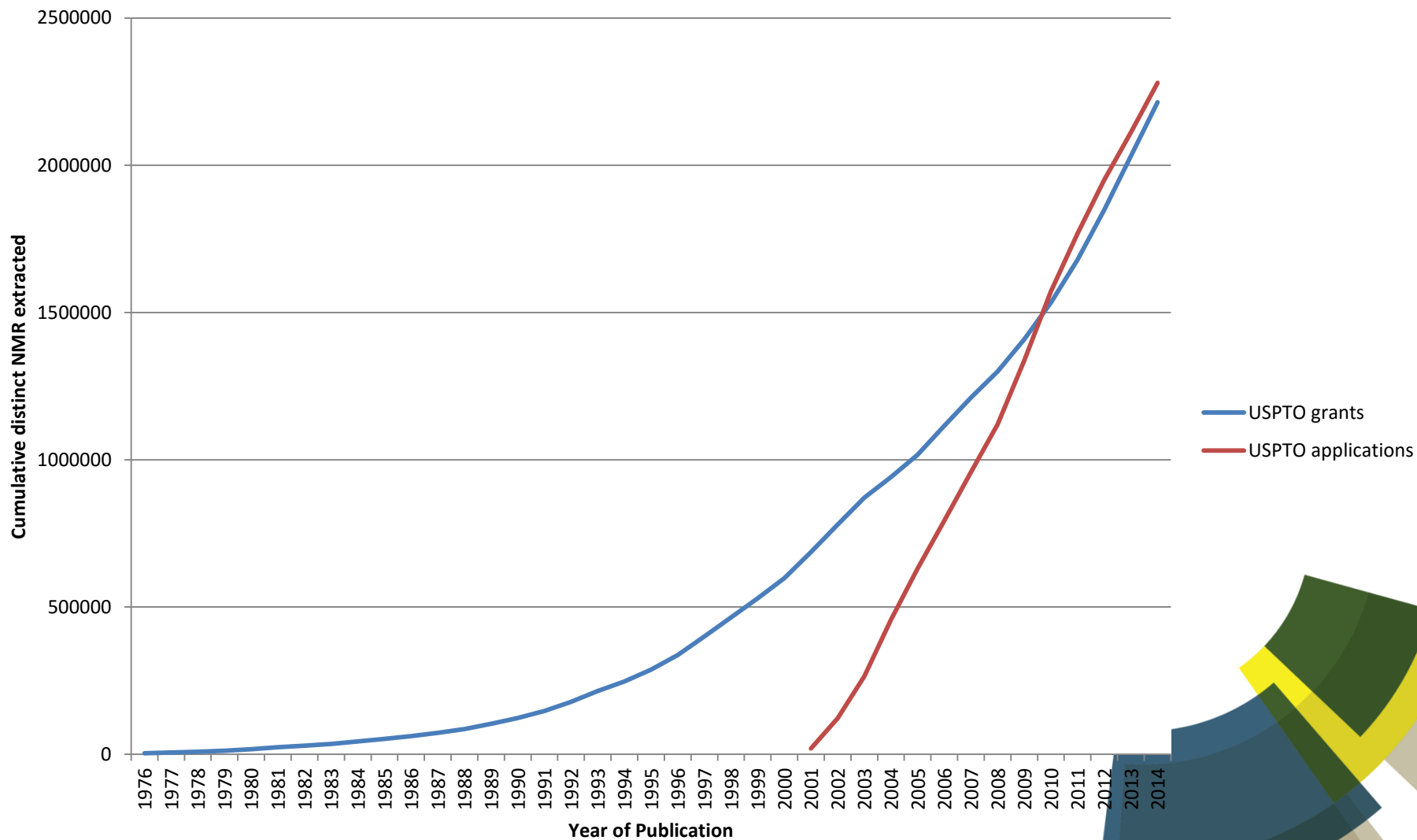


1H-NMR (DMSO-d6, 400 MHz): δ =1.04 (t, 6H; J=7.9 Hz, -CH3), 1.38 (q, 4H; J=7.9 Hz, Ge-CH2-), 6.88 (d, 4H; J=8.5 Hz, Ar-H3,5), 7.58 (d, 4H; J=8.5 Hz, Ar-H2,6), 10.53 (s, 2H, OH)

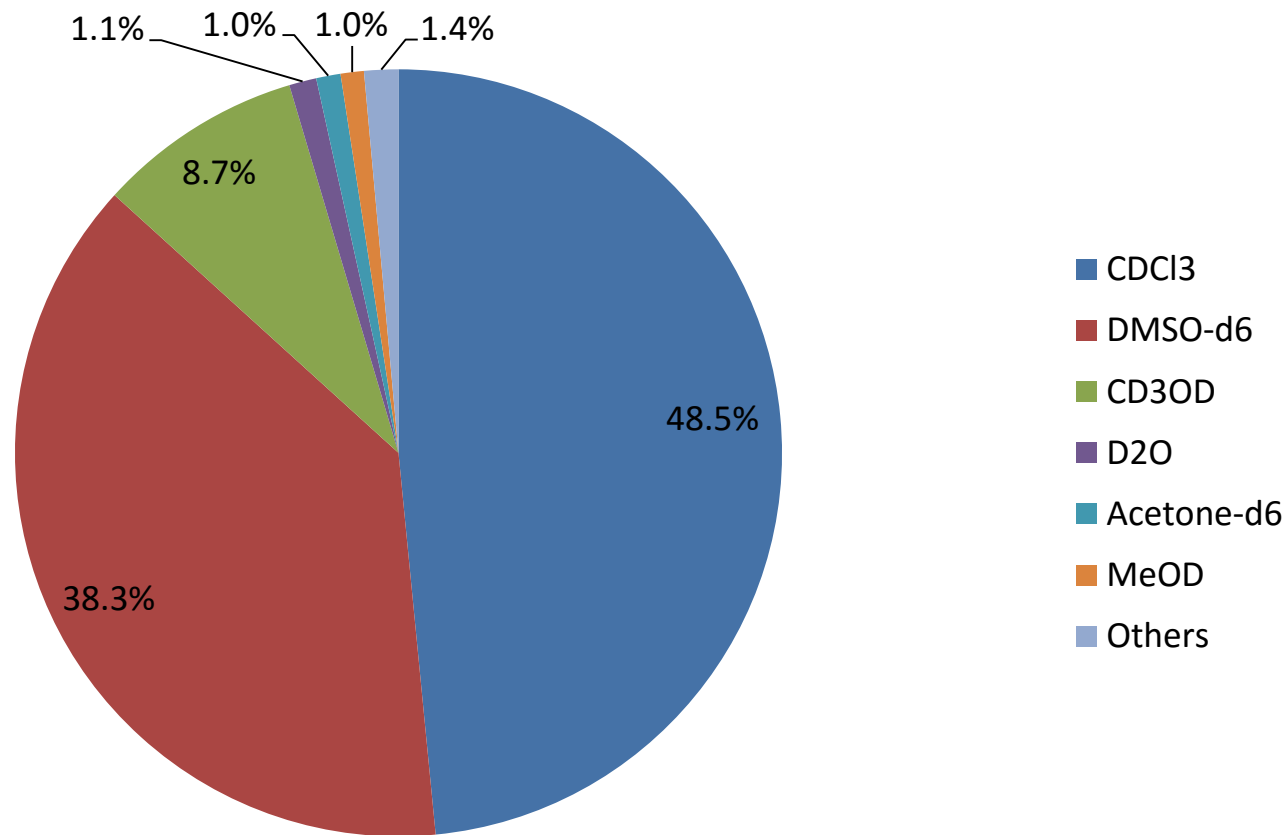
```
<parse>
  <nmrElement isotope="1" element="H">1H</nmrElement>
  <nmrMethodAndSolvent>DMSO-d6, 400 MHz</nmrMethodAndSolvent>
  <peak>
    <peakValue>1.04</peakValue>
    <peakAnnotation>t, 6H; J=7.9 Hz, -CH3</peakAnnotation>
  </peak>
  <peak>
    <peakValue>1.38</peakValue>
    <peakAnnotation>q, 4H; J=7.9 Hz, Ge-CH2-</peakAnnotation>
  </peak>
  <peak>
    <peakValue>6.88</peakValue>
    <peakAnnotation>d, 4H; J=8.5 Hz, Ar-H3,5</peakAnnotation>
  </peak>
  <peak>
    <peakValue>7.58</peakValue>
    <peakAnnotation>d, 4H; J=8.5 Hz, Ar-H2,6</peakAnnotation>
  </peak>
  <peak>
    <peakValue>10.53</peakValue>
    <peakAnnotation>s, 2H, OH</peakAnnotation>
  </peak>
</parse>
```

1H-NMR (DMSO-d6, 400 MHz): 1.04 (t, 6H; J=7.9 Hz, -CH3), 1.38 (q, 4H; J=7.9 Hz, Ge-CH2-), 6.88 (d, 4H; J=8.5 Hz, Ar-H3,5), 7.58 (d, 4H; J=8.5 Hz, Ar-H2,6), 10.53 (s, 2H, OH)

NMR extracted as $f(\text{year})$

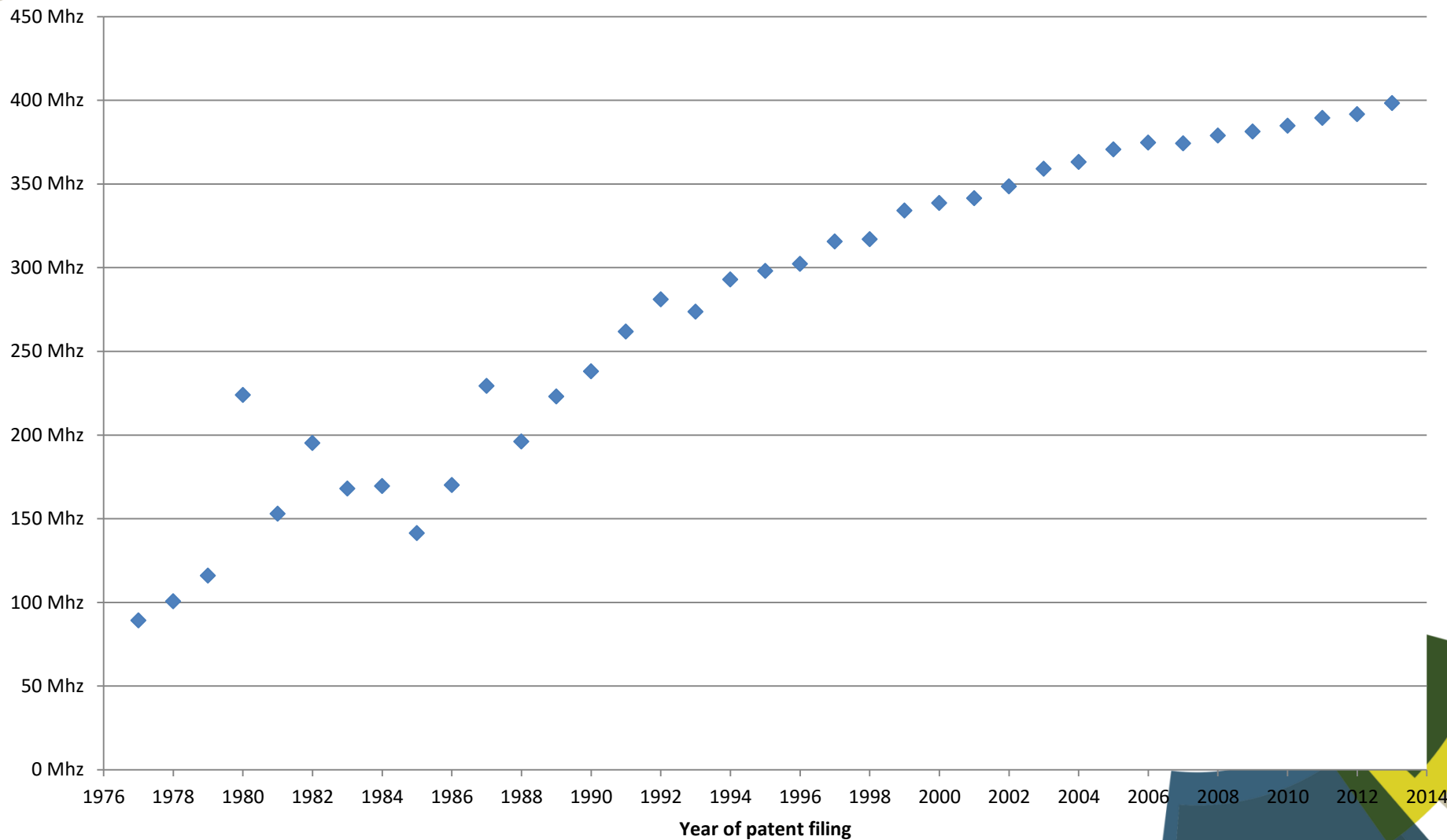


NMR solvents



Others: CD₂Cl₂, CD₃CN-d₃, C₆D₆, Pyridine-d₅, THF-d₈, CD₃Cl, dimethylformamide-d₇, d₁-trifluoroacetic acid, methanol-d₃, acetic acid-d₄, toluene-d₈, sulfuric acid-d₂, 1,1,2,2-tetrachloroethane-d₂, CD₃OCD₃, dioxane-d₈, 1,2-dichloroethane-d₄,

1H-NMR frequency over time





Sounds easy right?

- Potential for errors with names
 - No name extracted for structure
 - Incomplete names extracted
 - Misassociation of names with structures
 - Incorrect conversion of names to structures





BIGGEST problem - BRACKETS

- Brackets in names is a big problem- either an additional bracket or a missing bracket



Cannot be converted

- <https://www.google.co.uk/patents/US20050187390A1>
- 2-[2-(4'-carbamoyl-4-methoxy-**biphen-2-yl**)-quinolin-6-yl]-1-cyclohexyl-1H-benzoimidazole-5-carboxylic Acid
- OPSIN expects **biphenyl-2-yl**

OCR error Correction



- <https://www.google.co.uk/patents/WO2012150220A1>
- di-ter~~f~~-butyl (4S)-~~/V~~-(~~f~~ert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

CaffeineFix corrected to:

- di-ter~~t~~-butyl (4S)-~~N~~-(ter~~t~~-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

Corrections made: f--> t , / V --> N, f --> t



Sounds easy right?

- Textual Spectrum descriptions have issues
 - Transcription errors (rare)
 - Subjective interpretation (very common)
 - Incomplete listing of shifts
 - No/incomplete couplings/multiplicities listed
 - Overlap of multiplets (very common)
 - Labile protons – included/excluded/partial
- 

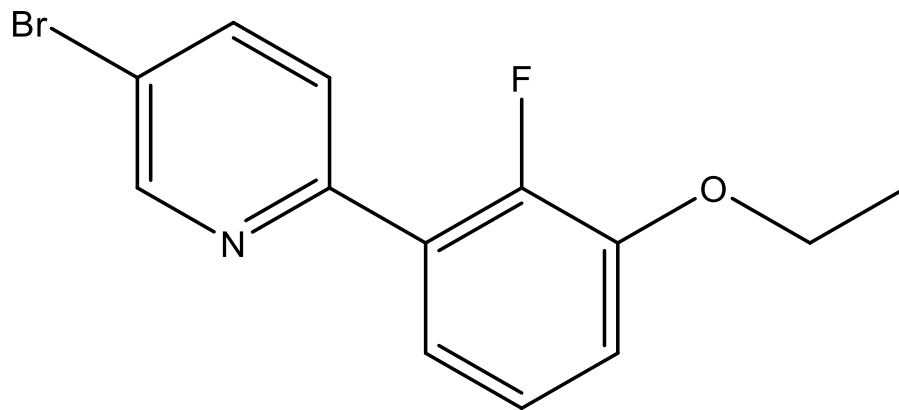


Sounds easy right?

- Textual Spectrum descriptions have issues
 - No peak width indications – especially labiles
 - No peak shape indications – dynamic exchange
 - Presence of rotamers
 - Impurities included or misidentified
 - Solvent peak belonging to the compound
 - Wrong number of nuclei
- 

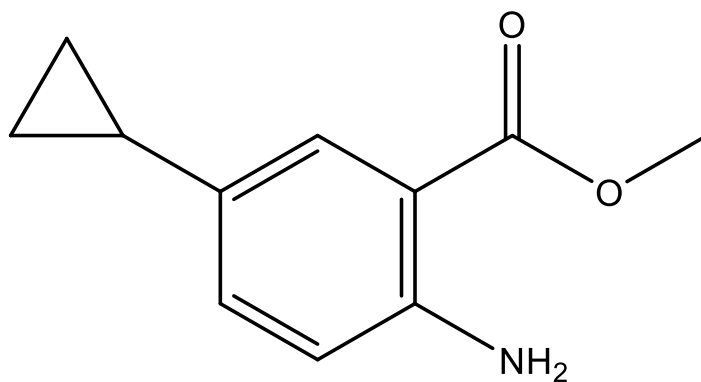
Problems Generating Spectra

- Multiplicities no coupling constants
- δ ^1H NMR (300 MHz, CDCl_3): 1.48 (t, 3H), 4.15 (q, 2H), 7.03 (td, 1H), 7.16 (td, 1H), 7.49 (m, 1H), 7.70 (dd, 1H), 7.88 (dd, 1H), 8.77 (d, 1H)



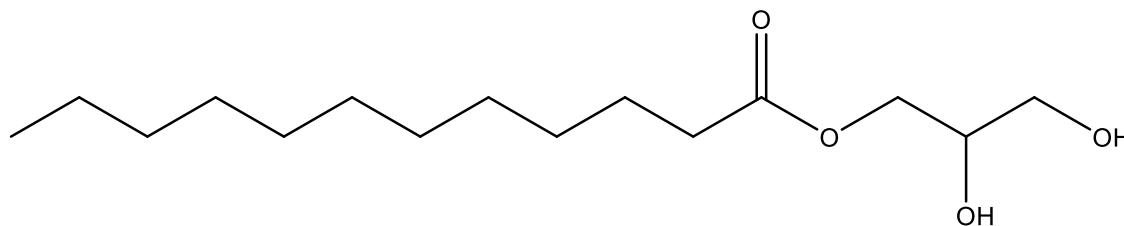
Problems Generating Spectra

- PARTIAL couplings only for ca. 90% of spectra!
- δ ^1H NMR (300 MHz, CDCl_3): 0.48-0.66 (m, 2H) 0.75-0.95 (m, 2H), 1.80 (s, 1H), 3.86 (s, 3H), 5.56 (s, 2H), 6.59 (d, **$J=8.50$ Hz**, 1H), 7.03 (dd, **$J=8.50$, 2.15 Hz**, 1H), 7.60 (s, 1H)



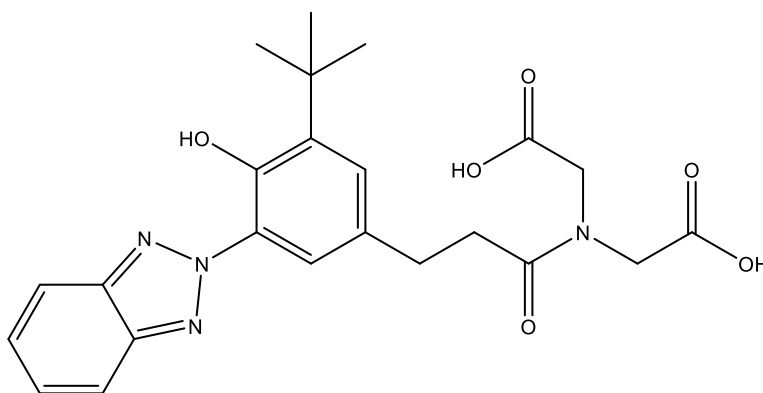
Error Detection

^1H NMR (400 MHz, CDCl_3) δ ppm 11.47-12.05 (1H), 7.97-8.24 (1H), 7.61-7.97 (2H), 7.28-7.61 (2H), 7.21 (1H), 5.27 (1H), 3.70-4.74 (8H), 2.80-3.16 (2H), 2.46-2.80 (2H), 1.87-2.45 (2H), 1.35-1.77 (11H), 1.24 (18H), 0.87 (3H) associated with Glycerol Monolaurate



Error Detection

- 54 hydrogens counted in the reported spectrum. Glyceryl Monolaurate has only 30 hydrogens.
- Title was: “Polymerization of **Monomer 4** with Glyceryl Monolaurate”
- Text-mining title missed compound: **Monomer 4** is the compound below





Text-mined spectra

- In the process of converting spectra into visual depictions many challenges identified
 - Validation approaches include:
 - NMR prediction and validation
 - Hosting “extracted text spectra” plus depictions
 - full provenance to source
 - Application to RSC archive will come later
- 

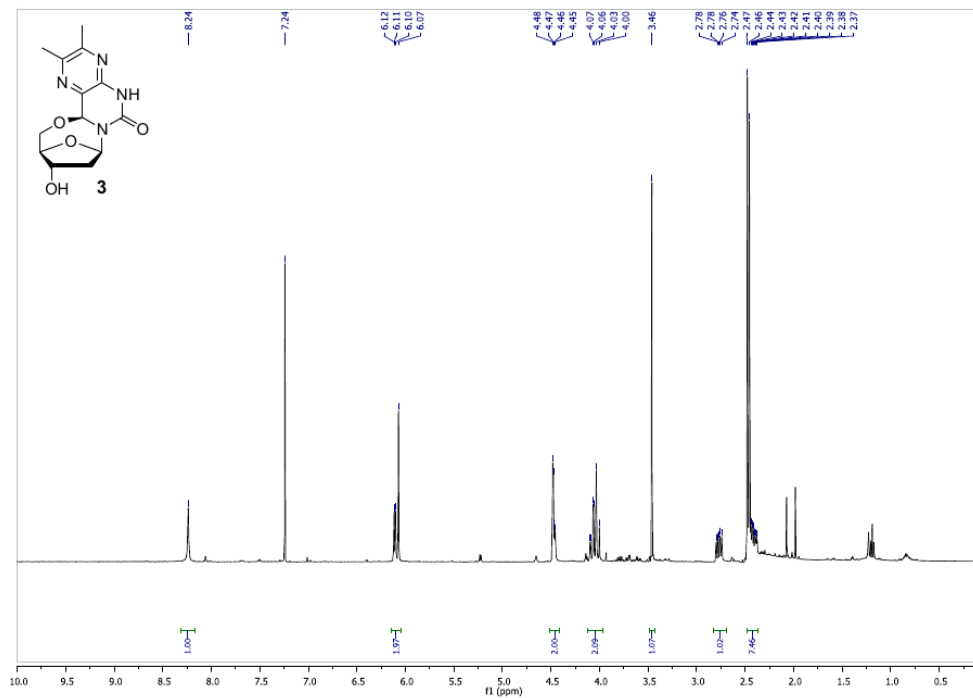
ESI Data also contains figures

Electronic Supplementary Material (ESI) for Organic & Biomolecular Chemistry
This journal is © The Royal Society of Chemistry 2013

Supporting information for Gislason, Gophane and Sigurdsson, *Org. Biomol. Chem.*

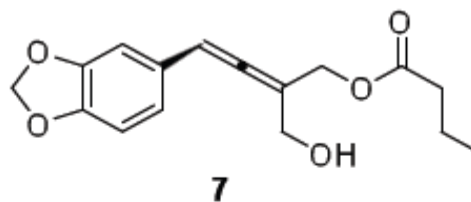
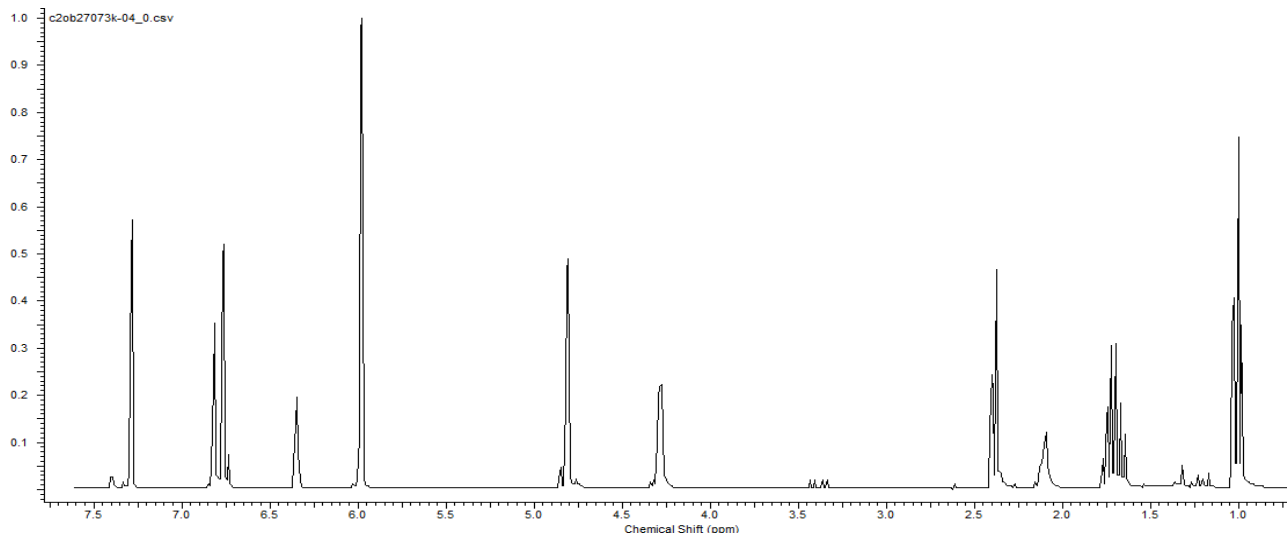
Page S7

NMR-spectra

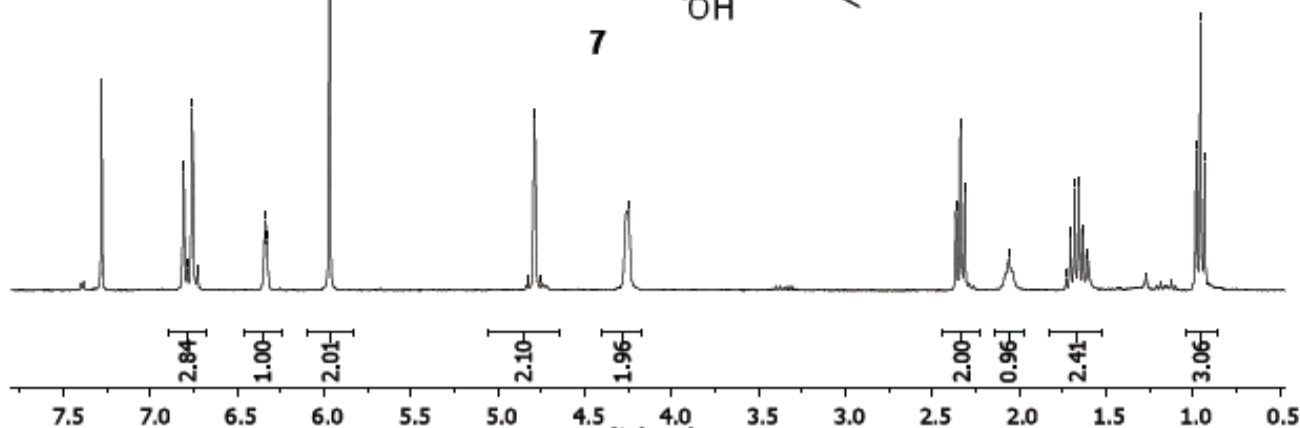


“Where is the real data please?”

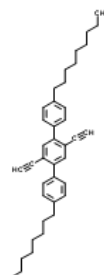
DATA



FIGURE



Data added to ChemSpider



2',5'-Diethynyl-4,4''-dinonyl-1,1':4',1''-terphenyl

ChemSpider ID: **29211602**

Molecular Formula: $C_{40}H_{50}$

Average mass: 530.825012 Da

Monoisotopic mass: 530.391235 Da

▼ Systematic name

2',5'-Diethynyl-4,4''-dinonyl-1,1':4',1''-terphenyl

1,1':4,4'':2,5'-terphenyl

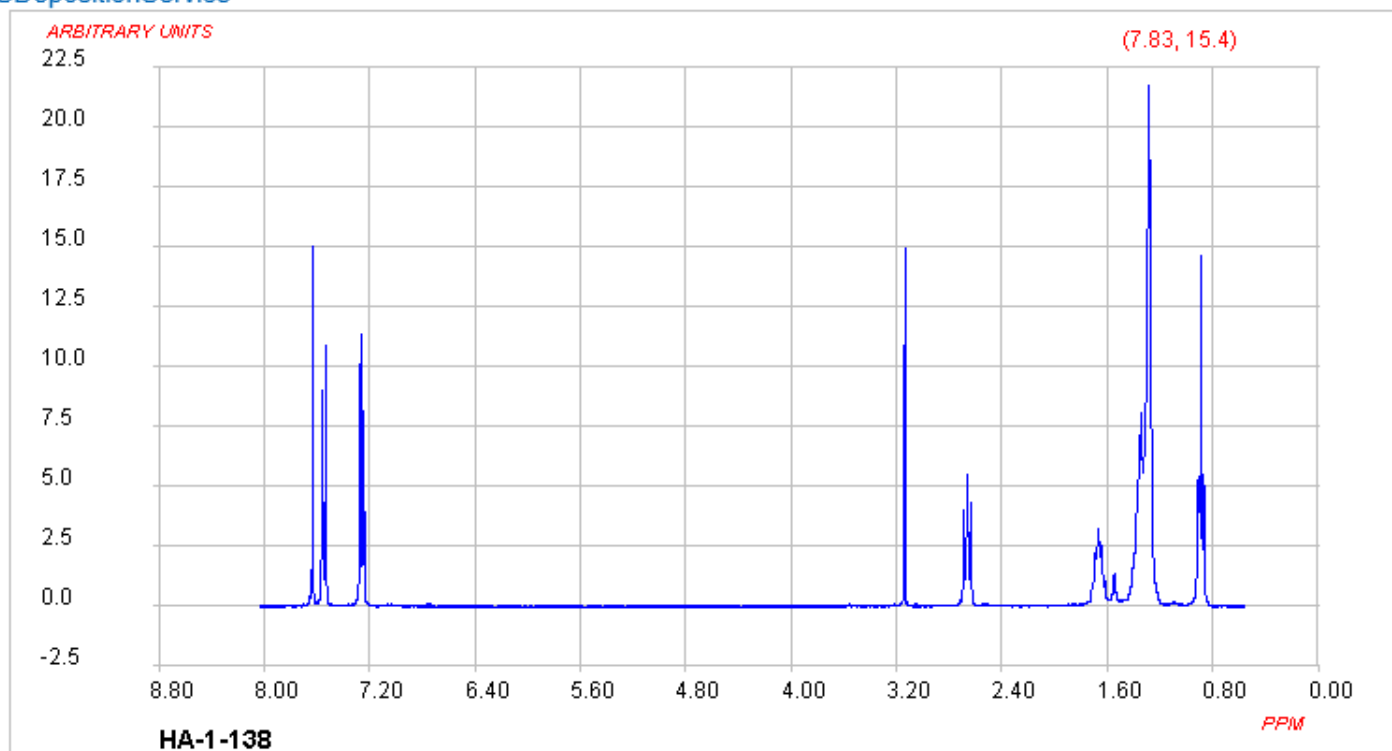
• Type: HNMR

Associated Hyperlink: <http://dx.doi.org/10.1039/C3SC51212F>

Comments: Spectral data kindly provided by the research group of Prof. William Dichtel at Cornell University- <http://www.williamdichtel.com/>

Approved: No

Submitted by: [CSDepositionService](#)





Manual Curation Layer

- ChemSpider has had a manual curation layer for >8 years
 - Users can annotate data on ChemSpider
 - We do receive useful feedback from the community on the data and are optimistic!
- 



Extraction is the **WRONG WAY**

- We should NOT mine data out – digital form!
 - Structures should be submitted “correctly”
 - Spectra should be digital spectral formats, not images
 - ESI should be RICH and interactive
 - Data should be open, available, with meta data and provenance
 - Can we encourage depositions????
- 



An EPSRC Call

EPSRC NATIONAL CHEMICAL DATABASE SERVICE



Issue date: 06 Jan 2012

“...the identification of the need for a UK national service for the provision of a searchable, electronic chemical database for the UK academic research community.”



National Chemical Database Service

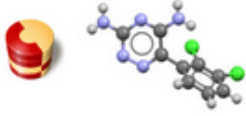
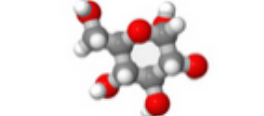
CDS National Chemical Database Service

In partnership with
EPSRC



Home About CDS Learning Resources External Resources Help Contact Us

The National Chemical Database Service brings together tools and resources for UK researchers in chemistry and related fields. All web-based services are freely accessible from any UK academic network. Further information about access to the NCDS can be found on the [About page](#).

ACD/I-Lab  Physchem and NMR prediction and database (ACD/Labs Inc.). Further information	CSD  Organic and organometallic crystal structures (CCDC). Further information	DETERM  Database of thermophysical data for pure substances and mixtures. Further information	ICSD  >160,000 inorganic and related crystal structures (FIZ Karlsruhe GmbH). Further information
Available Chemicals Directory  Provides supplier information for building block molecules. Further information	Chemicalize  Physicochemical property prediction tools with Lipinski-like filters. Further information	ChemSpider  An online database of molecules from >400 datasources (RSC). Further information	SPRESIweb  Online chemical structure and reaction database (InfoChem GmbH). Further information
CrystalWorks  Crystallographic data from the CSD, ICSD and CrystMet (STFC Daresbury). Further information	Introductory NCDS Video  A brief introductory video to the National Chemical Database Service. Watch the video here	In partnership with the EPSRC  Engineering and Physical Sciences Research Council	



Community Data Repository

- Automated depositions of data
 - Electronic Lab Notebooks as feeds
 - National services feeding the repository – crystallography, mass spectrometry
 - Accessing open data from other projects
- 

Antibiotics search to focus on sea bed

Researchers are embarking on an £8m project to discover new antibiotics at the bottom of the ocean.

A team, led by scientists at Aberdeen University, is hunting for undiscovered chemicals among life that has evolved in deep sea trenches.

Prof Marcel Jaspars said the team hoped to find "the next generation" of infection-fighting drugs.

England's chief medical officer has warned of an "antibiotic apocalypse" with too few new drugs in the pipeline.



KIRSTI HELLAND - UNIVERSITY OF TROMSØ

Scientists will test unique chemical compounds from marine samples found in deep sea trenches

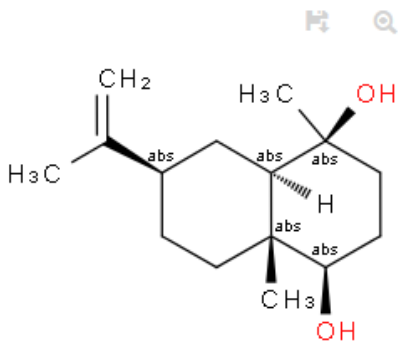
[Related Stories](#)

The PharmaSea Website

PHARMASEA



Compound Record

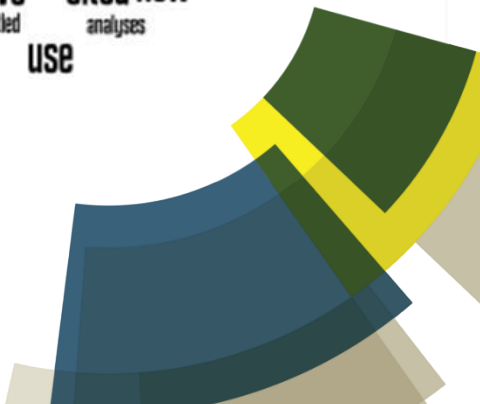


ID	18
Virtual	No
Molecular Formula	C ₁₅ H ₂₆ O ₂
Monoisotopic Mass	238.193283 Da
Molecular Weight	238.365738 Da
SMILES	<chem>C[C@]1(O)CC[C@@H](O)[C@]2(C)CC[C@H](C[C@@H]12)C(C)=C</chem>
Non Std. InChI	InChI=1/C15H26O2/c1-10(2)11-5-7-14(3)12(9-11)15(4,17)8-6-13(-14)16/h11-13,16-17H,1,5-9H2,2-4H3/t11-,12-,13-,14-,15+/m1/s1
Non Std. InChIKey	LGKGTMWCBFNQHP-RYPNDVFKNA-N
Std. InChI	InChI=1S/C15H26O2/c1-10(2)11-5-7-14(3)12(9-11)15(4,17)8-6-13(-14)16/h11-13,16-17H,1,5-9H2,2-4H3/t11-,12-,13-,14-,15+/m1/s1
Std. InChIKey	LGKGTMWCBFNQHP-RYPNDVFKSA-N
ChemSpider ID	28282620



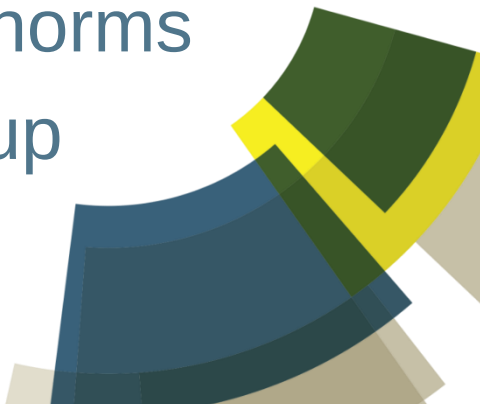
What can drive participation?

- What can drive scientists to participate and contribute?
 - Ensuring provenance of their data for reuse
 - Mandates from funding agencies
 - Improved systems to ease contribution
 - Additional contributions to science
 - Improved publishing processes
 - **Recognition** for contributions
- 





My opinions...

- Yes, platform development is critical
 - Yes, ease-of-use/efficiency is necessary
 - Yes, standards can be improved
 - The **greatest shifts** will come from:
 - An increased willingness to share
 - More training in chemical information
 - Working towards new community norms
 - The majority of change is bottom-up
- 

The Future



Individual Scientists

Internet Data

Published Chemistry



RSC | Advancing the
Chemical Sciences

RSC Data

Electronic Lab Notebooks



Small organic molecules

Undefined materials

Organometallics

Nanomaterials

Polymers

Minerals

Particle bound

Links to Biologicals

Commercial Software

Pre-competitive Data

Open Science

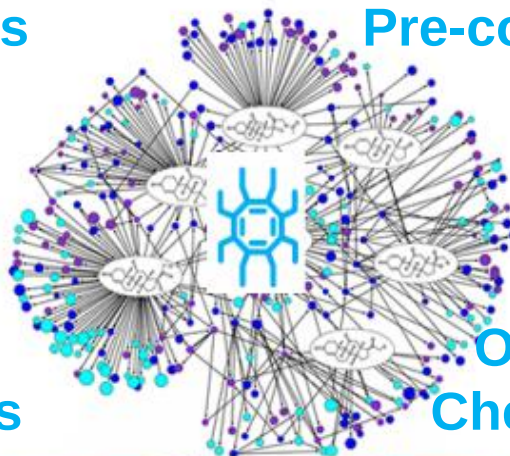
Open Data

Publishers

Educators

Open Databases

Chemical Vendors



Aggregate
Data

Search
Network

Generate
Models

Integrate
and Federate



Acknowledgments

- Data Repository Team and ChemSpider Team
 - Daniel Lowe (NextMove software)
 - Igor Tetko (HelmholtzZentrum München)
 - Carlos Coba (Mestrelab Research)
- 



Thank you

Email: tony27587@gmail.com

ORCID: [0000-0002-2668-4821](https://orcid.org/0000-0002-2668-4821)

Twitter: [@ChemConnector](https://twitter.com/ChemConnector)

Personal Blog: www.chemconnector.com

SLIDES: www.slideshare.net/AntonyWilliams

