Royal Society of Chemistry: Community Driven Open Science

Brian O'Connor
UK & Northern Europe Journal Sales Executive

9th Munin Conference 2014

27-11-2014

Royal Society of Chemistry

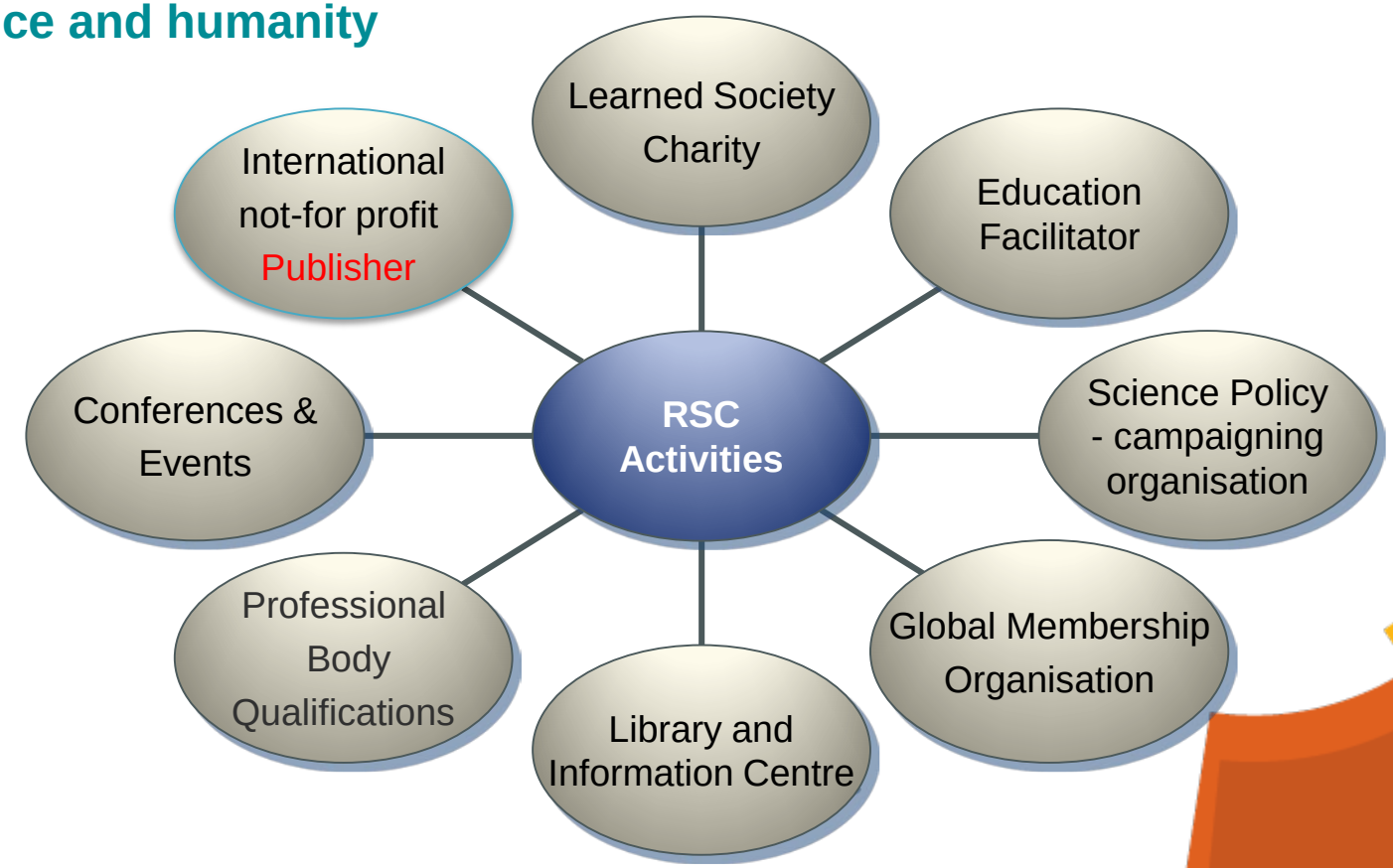
The world's leading chemistry community

- Over 51,000 members
- Leading international not-for-profit publisher
- Scientific policy and education
- Conferences and events

The Royal Society of Chemistry

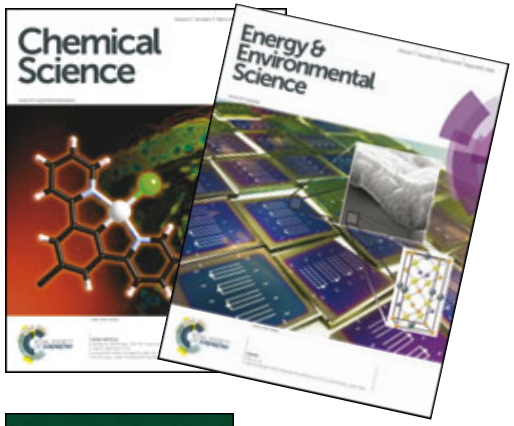
Advancing excellence in the chemical sciences

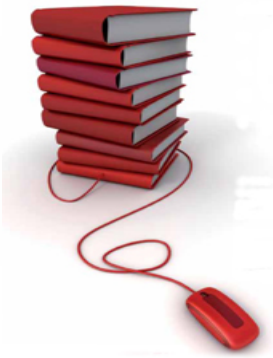
Shaping the future of the chemical sciences for the benefit of science and humanity

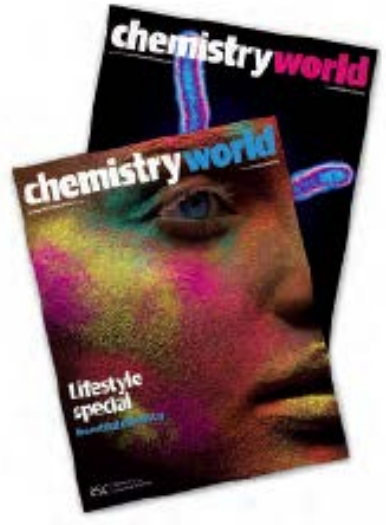


Leading international publisher

- Our publishing activities span books, e-books, journals, databases and magazines







THE MERCK INDEX* Online

ChemSpider
The free chemical database

**The name THE MERCK INDEX is owned by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Whitehouse Station, N.J., U.S.A., and is licensed to The Royal Society of Chemistry for use in the U.S.A. and Canada.*

Summary: Current situation

OA is fast moving, we are in a transition period

Individual funding agency mandates - varying policies which can be confusing

Publisher experimentation continues

Rapid growth of OA journals and institutional repositories

Librarians as 'Guardians' of Open Access

RSC wants to work with the community to support them, and be involved in the on-going discussions

Our Stance

*"RSC supports Open Access models which seek to ensure that scholarly publishing activities operate in a **long-term sustainable way**"*

- Maximise availability and accessibility of research
- Maintain standards and ethics of publishing
- Provide a quality service

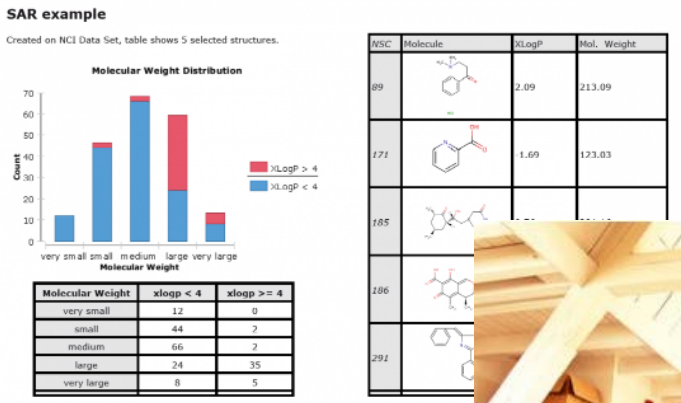
Gold OA is the best option – sustainable and provides immediate access

Green OA is also supported with our Chemical Sciences Article Repository: http://www.rsc.org/Chemical-Sciences-Repository/articles/

Gold for Gold

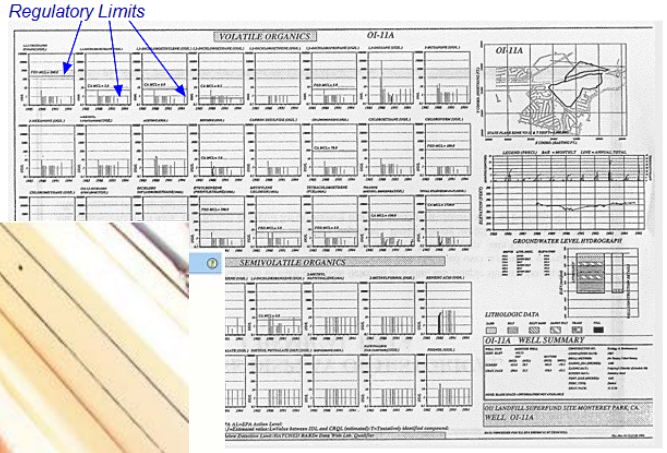
- Number of voucher codes received is what the institution pays for RSC Gold divided by £1,600
- We recognise researchers are being asked to publish OA, but may not have the funding
- A **reward** for all RSC Gold subscribing institutions
- Institutions get **voucher codes to publish Gold OA free of charge**
- **> £9Million** invested in free Gold OA voucher codes for RSC Gold members

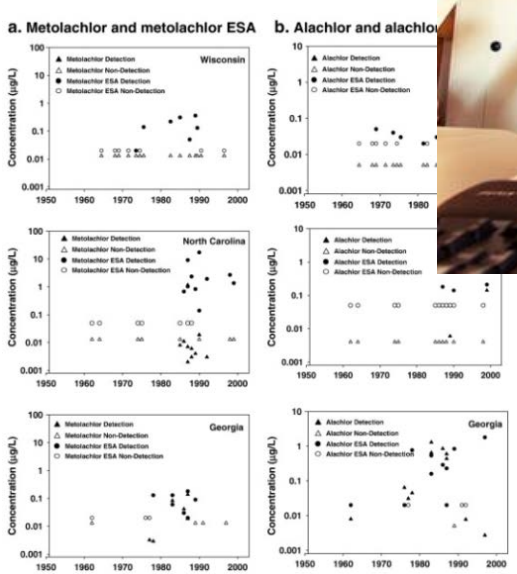
Overwhelmed with data…

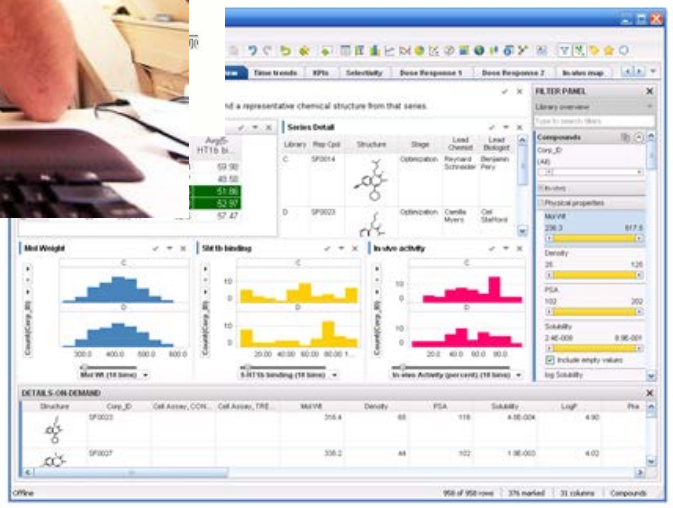


Molecular Weight	xlogp < 4	xlogp >= 4
very small	12	0
small	44	2
medium	66	2
large	24	35
very large	8	5

NSC	Molecule	XLogP	Mol. Weight
89		2.09	213.09
171		1.69	123.03
185			
186			
291			



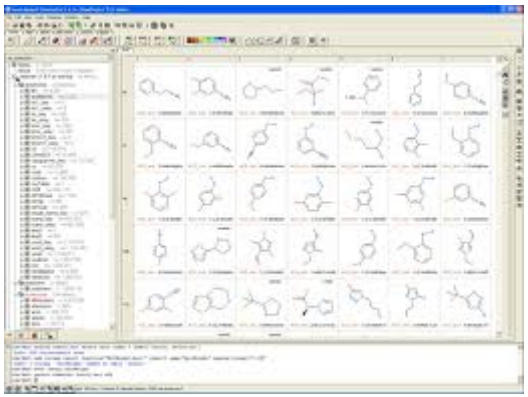




So much online data…

ChemSpider
The free chemical database

A free database for virtual screening
ZINC
is not commercial



NATIONAL CANCER INSTITUTE

ChEMBL

un1Chem

ctd™
Illuminating how chemicals affect human health.
Comparative Toxicogenomics Database

Chemical Property Data Search

eChemPortal

ACToR | ToxRefDB | ToxCastDB | ExpoCastDB | DSSTox

Organizations releasing data

For Immediate Release: Tuesday, March 13, 2012

NIH and Lilly to generate public resource of approved and investigational medicines

Collaboration may make drug development pipelines more productive

GSK gives update on plans to share detailed clinical trial data as part of its commitment to transparency

News

GlaxoSmithKline goes public with malaria data

Company to place structures and properties of drug leads in the public domain.

Funders encourage openness

We model data – then lose it

What if we could share models and the underlying data via a central repository?

This is MOSTLY not a technology issue!!!

CDD MODELS

GUSAR ONLINE

RELIABLE QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIPS FOR YOUR CHEMICAL COMPOUNDS

CHEM BENCH

HOME | MY BENCH | DATASET

ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

Online chemical database with modeling environment

Home | Database | Models

eADMET

THE REFERENCE IN CHEMOINFORMATICS

PERSPECTIVE

www.rsc.org/loc | Lab on a Chip

Precompetitive preclinical ADME/Tox data: set it free on the web to facilitate computational model building and assist drug development

Sean Ekins*abc and Antony J. Williams*d

Pharma Companies Repeat Work

Pre-competitive Informatics:

Pharma are all accessing, processing, storing & re-processing external research data

Literature Genbank
Patents PubChem

Databases

Downloads

X

Repeat at each company

Data Integration

Data Analysis

Firewalled Databases

Publications lock up data

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

Phedias Diamandis[1–4], Jan Wildenhain[4], Ian D Clarke[1,2], Adrian G Sacher[1,2], Jeremy Graham[1,2], David S Bellows[3], 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]

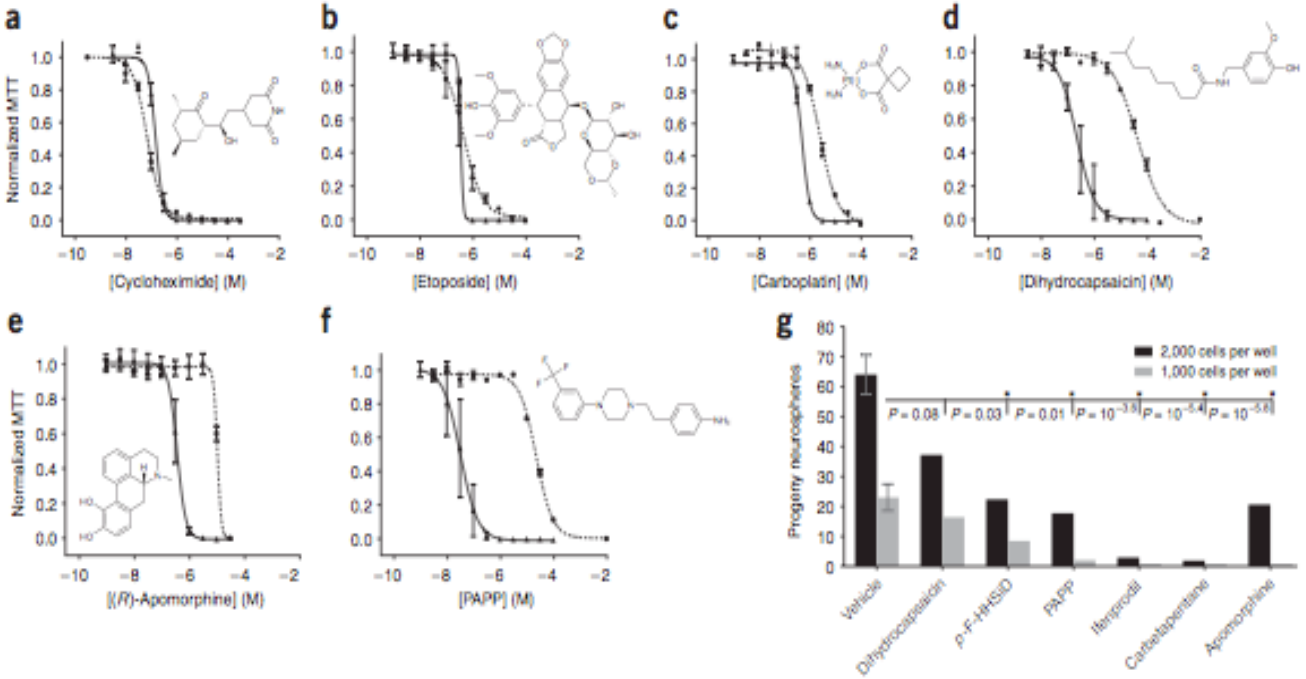


Figure 2 Identification of potent NPC-specific compounds. (**a**–**f**) Dose-response curves and chemical structures of controls: cycloheximide (**a**), etoposide (**b**) and carboplatin (**c**), and of selected newly identified compounds: dihydrocapsaicin (**d**), apomorphine (**e**) and PAPP (**f**).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

But what if we could navigate?

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

Phedias Diamandis[1–4], Jan Wildenhain[4], Ian D Clarke[1,2], Adrian G Sacher[1,2], Jeremy Graham[1,2], David S Bellows[3], 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]

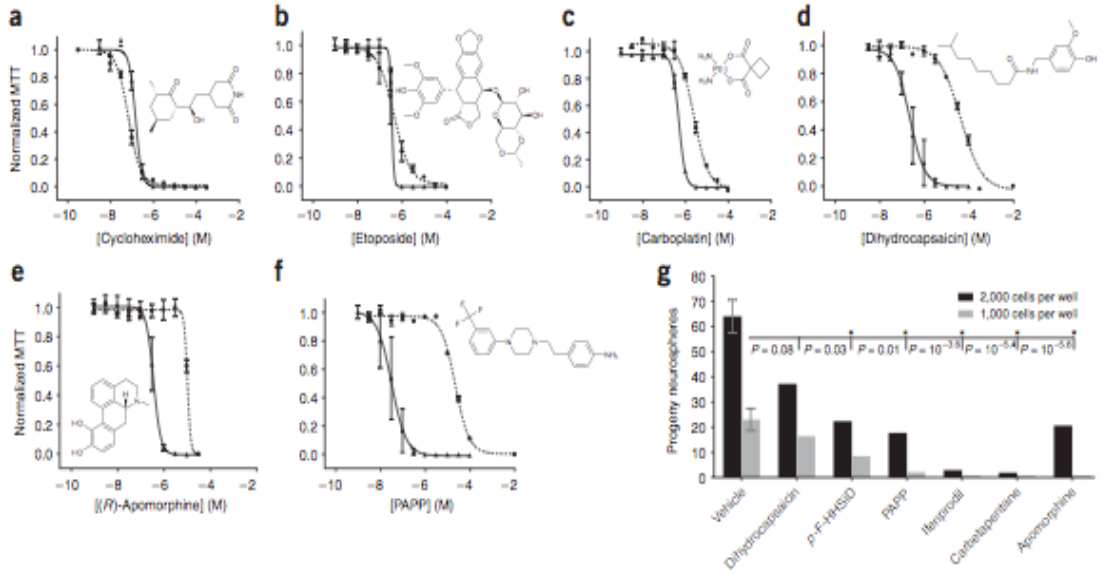


Figure 2 Identification of potent NPC-specific compounds. (**a**–**f**) Dose-response curves and chemical structures of controls: cycloheximide (**a**), etoposide (**b**) and carboplatin (**c**), and of selected newly identified compounds: dihydrocapsaicin (**d**), apomorphine (**e**) and PAPP (**f**).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


KEEP
CALM
AND
DON'T
LOSE HOPE

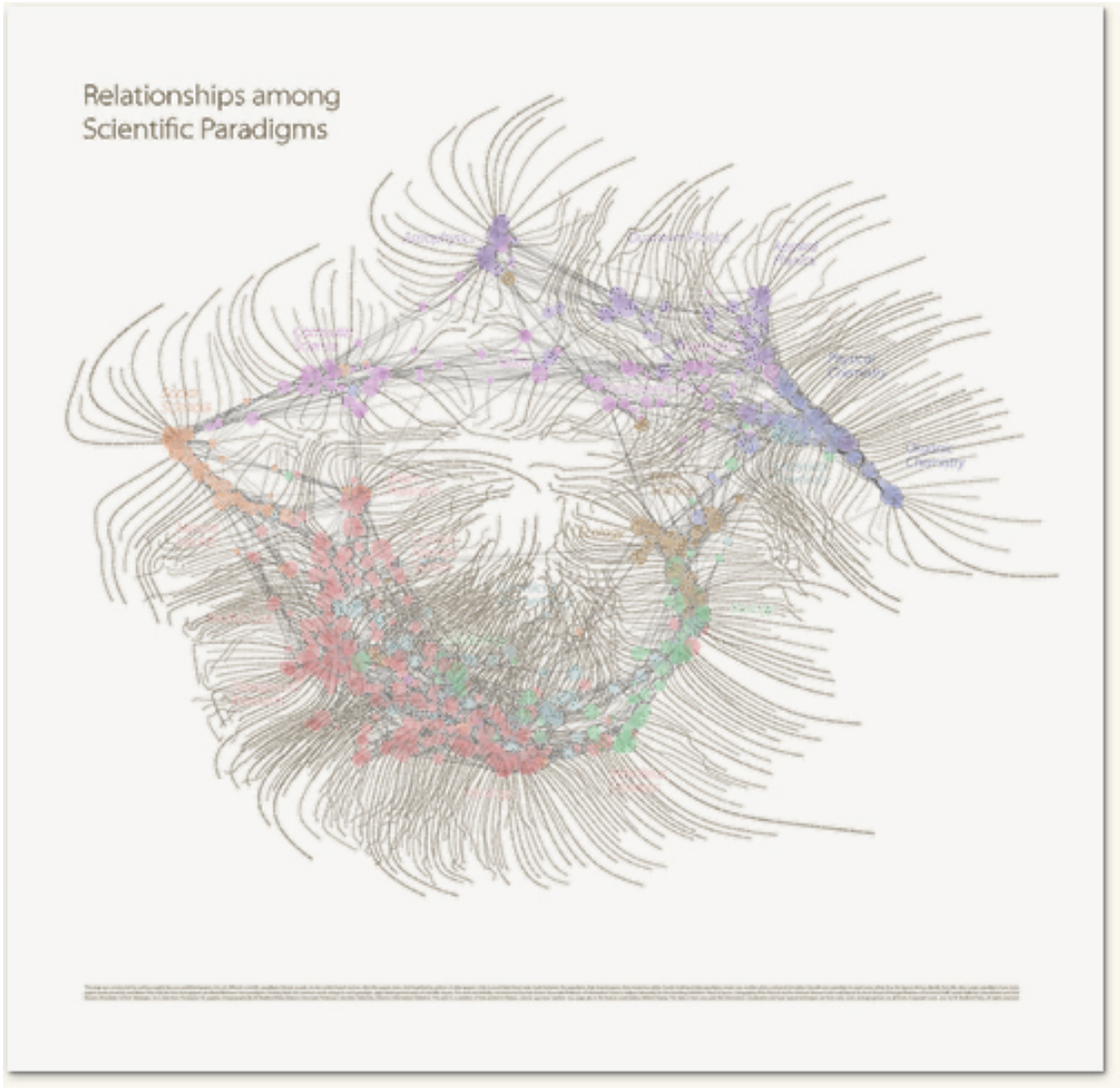
What about this….

- We’re going to map the world
- We’re going to take photos of as many places as we can and link them together
- We’ll let people annotate and curate the map
- Then let’s make it available free on the web
- We’ll make it available for decision making
- Put it on Mobile Devices, give it away…

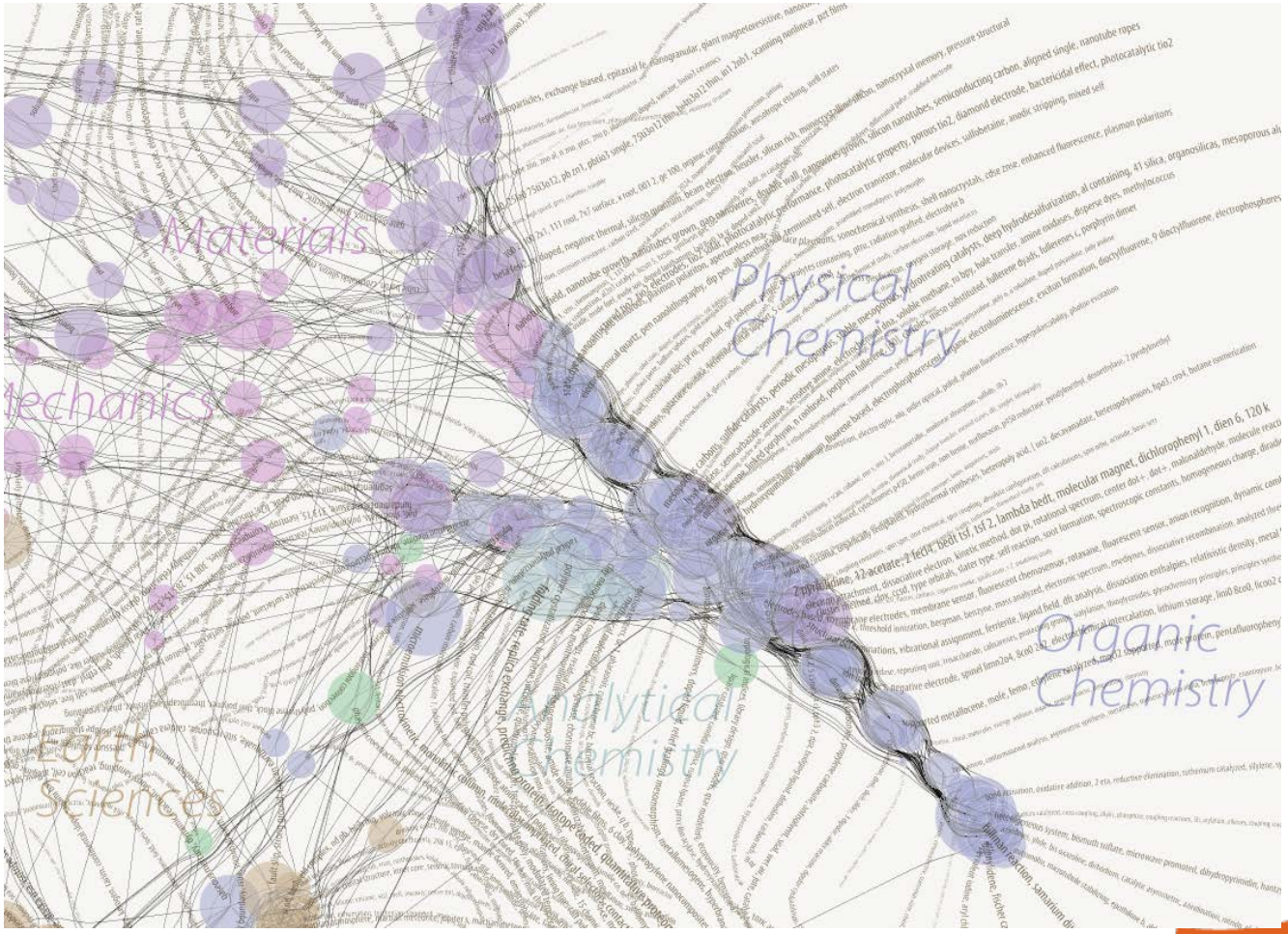
Whoa…

- So the world can be mapped…
- We can enter a 3D world within the map
- We can add annotations
- We can use the data, reference it, we can extract it, we can make decisions with it
- And we can do it on our lap, in our hands
- **Let’s do this for chemistry…**

Science map


Relationships among
Scientific Paradigms
Physics
Chemistry
Engineering
Biology
Infectious Disease
Medical Research
Brain Research
Health Sciences
Social Sciences
Computer Science

Chemical sciences map



Chemistry data is of value?

- Reference databases generate hundreds of millions of dollars/euros per year
- So much data generated that could go public
- **Maybe 5%** of all data generated is **published**
- There is **no** “Journal of Failed Experiments”
- Funding agencies start to demand **Open Data**
- Scientists want funding but also **recognition**

Taking on a big challenge…

- Let’s map together all historical chemistry data and build systems to integrate
- Heck, let’s integrate chemistry and biology data and add in disease data too
- Let’s **model** the data and see if we can extract new relationships – quantitative and qualitative
- Let’s make it all available on the web

ChemSpider
Search and share chemistry

- ~30 million chemicals and growing
- Data sourced from hundreds of sources
- Crowd sourced curation and annotation
- Ongoing deposition of data from our journals and our collaborators
- Structure centric hub for web-searching
- …and a really big dictionary!!!

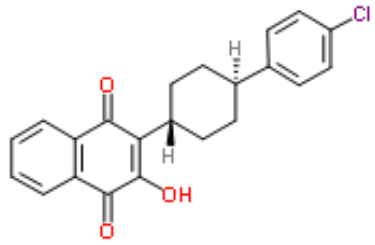
A Chemistry Data Repository to Serve Them All

Chemistry for the Community

- The Royal Society of Chemistry as a provider of chemistry for the community:
  - As a charity
  - As a scientific publisher
  - As a host of commercial databases
  - As a partner in grant-based projects
  - As the host of ChemSpider
  - And now in development : the RSC Data Repository for Chemistry

ChemSpider

Search term: **atovaquone** (Found by approved synonym)



2D 3D Save Edit Zoom

- 2 of 2 defined stereocentres

Atovaquone

ChemSpider ID: **10482034**

Molecular Formula: $C_{22}H_{19}ClO_3$

Average mass: 366.837494 Da

Monoisotopic mass: 366.102264 Da

▼ Systematic name

2-[trans-4-(4-Chlorophenyl)cyclohexyl]-3-hydroxy-1,4-naphthoquinone

▶ SMILES and InChIs

▶ Cite this record

Wikibox

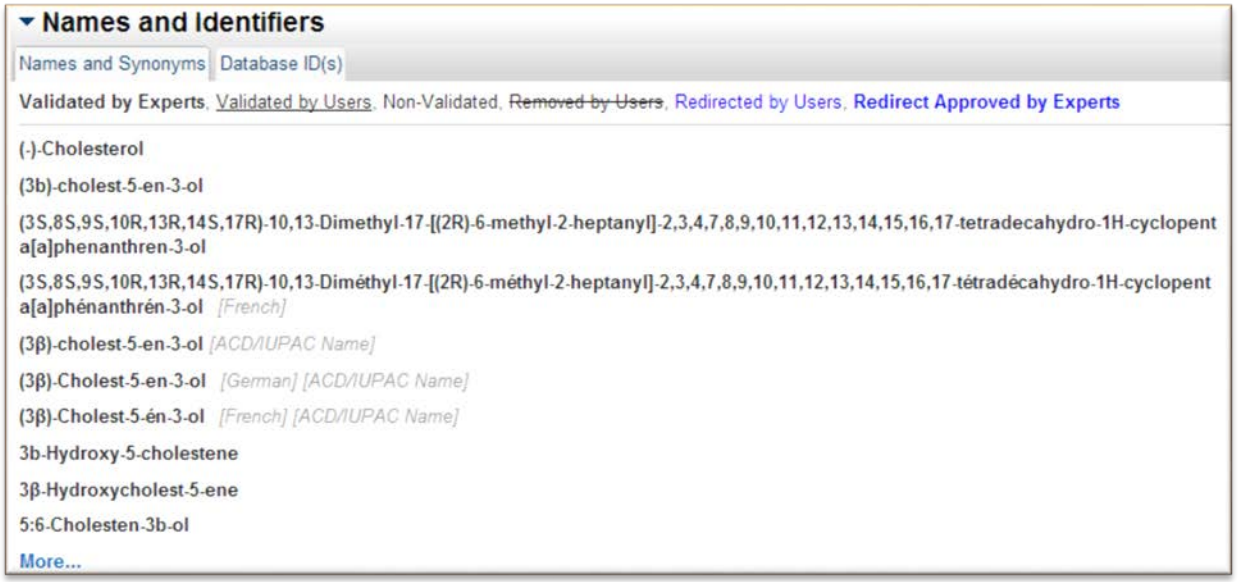
Embed

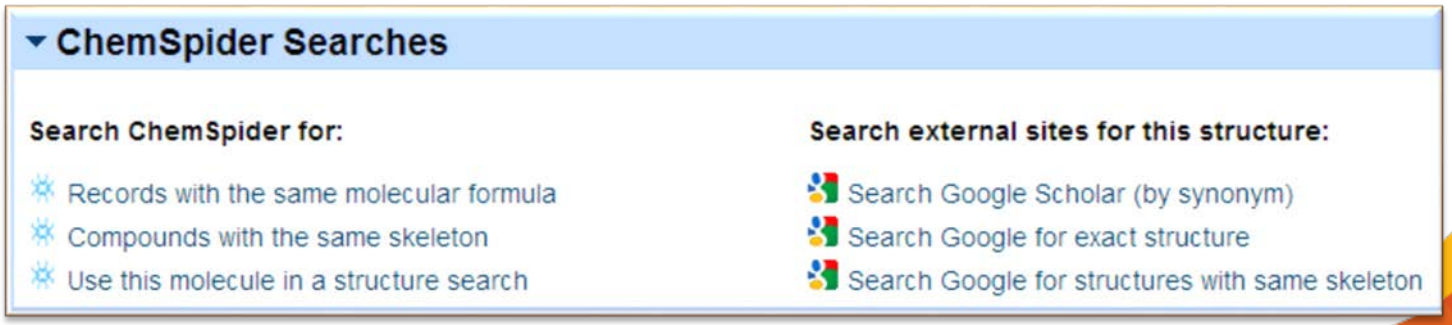
Deprecate

Watch this record

Manage data slice

ChemSpider


Names and Identifiers
Names and Synonyms
Database ID(s)
Validated by Experts, Validated by Users, Non-Validated, Removed by Users, Redirected by Users, Redirect Approved by Experts
(-)-Cholesterol
(3b)-cholest-5-en-3-ol
(3S,8S,9S,10R,13R,14S,17R)-10,13-Dimethyl-17-[(2R)-6-methyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[a]phenanthren-3-ol
(3S,8S,9S,10R,13R,14S,17R)-10,13-Diméthyl-17-[(2R)-6-méthyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tétradécahydro-1H-cyclopenta[a]phénanthrén-3-ol [French]
(3β)-cholest-5-en-3-ol [ACD/IUPAC Name]
(3β)-Cholest-5-en-3-ol [German] [ACD/IUPAC Name]
(3β)-Cholest-5-én-3-ol [French] [ACD/IUPAC Name]
3b-Hydroxy-5-cholestene
3β-Hydroxycholest-5-ene
5:6-Cholesten-3b-ol
More...


ChemSpider Searches
Search ChemSpider for:
Records with the same molecular formula
Compounds with the same skeleton
Use this molecule in a structure search
Search external sites for this structure:
Search Google Scholar (by synonym)
Search Google for exact structure
Search Google for structures with same skeleton

Experimental/Predicted Properties

Properties

Experimental data | **Predicted** - ACD/Labs | **Predicted** - EPISuite | **Predicted** - ChemAxon

Predicted data is generated using the ACD/Labs' ACD/PhysChem Suite, for more information see their website.

ACD/LogP:	9.85±0.28	# of Rule of 5 Violations:	1
ACD/LogD (pH 5.5):	9.85	ACD/LogD (pH 7.4):	9.85
ACD/BCF (pH 5.5):	1000000.00	ACD/BCF (pH 7.4):	1000000.00
ACD/KOC (pH 5.5):	5438366.00	ACD/KOC (pH 7.4):	5438366.00
#H bond acceptors:	1	#H bond donors:	1
#Freely Rotating Bonds:	6	Polar Surface Area:	20.23 Å^2
Index of Refraction:	1.525	Molar Refractivity:	120.0±0.4 cm^3
Molar Volume:	391.4±5.0 cm^3	Polarizability:	47.6±0.5 10^{-24}cm^3
Surface Tension:	38.2±5.0 dyne/cm	Density:	1.0±0.1 g/cm^3
Flash Point:	209.3±12.4 °C	Enthalpy of Vaporization:	85.9±6.0 kJ/mol
Boiling Point:	480.6±14.0 °C at 760 mmHg	Vapour Pressure:	0.0±2.7 mmHg at 25°C

Literature references

Articles

Links & Reference | RSC Journals | RSC Books | PubMed | MeSH Literature | Google Books

- H.-S. Shieh and C. E. Nordman. Cholesterol hemimethanol solvate, *Acta Cryst. (2002). E58, o79-o80Â Â Â Â*
  [DOI: 10.1107/S1600536801020487]
- Bosco et al.. Elevated levels of oxidized cholesterol metabolites in Lewy body disease brains accelerate alpha-synuclein fibrilization, *Nature Chemical Bi*
  [DOI: 10.1038/nchembio782]
- Polozov et al.. Progressive Ordering with Decreasing Temperature of the Phospholipids of Influenza Virus, *Nature Chemical Biology, 2008*
  [DOI: 10.1038/nchembio.77]
- Meloni et al.. Metal swap between Zn7-metallothionein-3 and amyloid-beta Cu protects against amyloid-beta toxicity, *Nature Chemical Biology, 2008*
  [DOI: 10.1038/nchembio.89]
- Gerlach et al.. HIV-1 Nef membrane association depends on charge, curvature, composition and sequence, *Nature Chemical Biology, 2009*
  [DOI: 10.1038/nchembio.268]
- Tobias Kind, Martin Scholz, Oliver Fiehn. How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry, *PLoS ONE 4(5):*
  [DOI: 10.1371/journal.pone.0005440]
  Calculating the metabolome size of species by genome-guided reconstruction of metabolic pathways misses all products from orphan genes and from en
  annotated genes. Hence, metabolomes need to be determined experimentally. Annotations by mass spectrometry would greatly benefit if peer-reviewed
  could be queried to compile target lists of structures that already have been reported for a given species. We detail current obstacles to compile such a k
  metabolites.
- Mével Mathieu. Novel neutral imidazole-lipophosphoramides for transfection assays, *Chemical Communications, 2008*
  [DOI: 10.1039/b805226c]
- Aparicio Jesús F.. Microbial cholesterol oxidases: bioconversion enzymes or signal proteins?, *Molecular BioSystems, 2008*
  [DOI: 10.1039/b717500k]

Patents references

Google Patents

Cholesterol efflux assay probe formulations, methods of making and using
US Pat. WO2013155374A2 - Apr 12, 2013 - The Ohio State University
A **cholesterol** efflux assay probe formulation having a core comprised of a biocompatible hydrophobic material at least partially coated with a ...

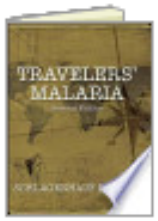
Sperm protective agent, namely cholesterol sulfate, for frozen semen diluent
US Pat. CN103238584A - Feb 10, 2012 - 郑云胜, 倪利平
The invention discloses a freezing protective agent, namely **cholesterol** sulfate, added in a frozen animal semen diluent. According to a using ...

METHOD FOR REMOVING CHOLESTEROL FROM FOOD USING ß-CYCLODEXTRIN CROSSLINKED ...
US Pat. EP1893652B1 - Dec 02, 2005 - Kwak, Hae-soo
A method for removing **cholesterol** from food, comprising treating the **cholesterol**- containing food with a cross-linked beta-cyclodextrin to trap ...

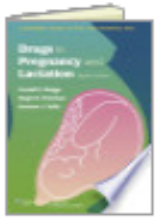
Methods for quantitating high-density lipoprotein cholesterol
US Pat. WO1999010526A1 - Aug 25, 1998 - Daiichi Pure Chemicals Co Ltd, Koichi Hino, Mitsuhisa Manabe, Mitsuhiro Nakamura, Kazuo Nakanishi
Methods for quantitating HDL **cholesterol**, comprising adding to serum a surfactant selected from among polyoxyethylene alkylene phenyl ethers and ...

Books

Links & Reference | RSC Journals | RSC Books | PubMed | MeSH Literature | Google Books



Travelers' Malaria
by unknown
2007 - 408 pages
books.google.com



Drugs in Pregnancy and Lactation for PDA: A Reference Guide to Fetal and ...
by Gerald G. Briggs, Roger K. Freeman, Sumner Jason Yaffe
2008 - 2117 pages
books.google.com

Emerging Protozoan Pathogens
by unknown
2008 - 384 pages
books.google.com

Vendors and data sources

Chemical Vendors

Data Source	External ID(s)
ABI Chemicals	AC1Q1PAZ
Alfa Aesar	A11470
Alfa Chemistry	ACM57885, ACM80356145
Angene	AG-G-04640
Aronis	BBC/344
Chembo Pharma	KB-48897, 91320
Extrasynthese	3551, 3506
Finetech Industry	FT-57-88-5
Fluorochem	226906, 121530
Glentham Life Sciences	GE1295
InterBioScreen	BB_NC-0135, STOCK1N-54186, BB_NC-00135
King Scientific	KSC497S9L
Labseeker	86-35931, SC-18458
Matrix Scientific	058463
Molport	MolPort-002-506-911
Paragos	490008
Pharmten	S-PTN26633
R&D Chemicals	3551
Research Organics	RES1387C-A102X, RES1387C-A103X
Santa Cruz Biotechnology	sc-202539
Sigma-Aldrich	S5442, SIGMA, C3292, SIGMA, C1231, SIGMA, C3137, SIGMA, C3045, SIGMA, C8503, SIGMA

Aspirin on ChemSpider

2-(Acetyloxy)benzoic acid

2-(Acetyloxy)benzolcarbonsäure *[German] [ACD/IUPAC Name]*

200-064-1 *[EINECS]*

2-Acetoxybenzenecarboxylic acid

2-Acetoxybenzoesäure *[German] [ACD/IUPAC Name]*

2-Acetoxybenzoic acid *[ACD/IUPAC Name]*

2-Acetyloxybenzoic acid

50-78-2 *[RN]*

A.S.A.

Acesan

Acetard

Acetoxybenzoic acid

acetyl salicylic acid

Acetyl-SAL

ACETYLSALICYLIC ACID

Acetyonyl

Acetysalicylic acid

Acide 2-(acétyloxy)benzoïque *[French]*

Acide 2-acétoxybenzoïque *[French] [ACD/IUPAC Name]*

Ácido acetilsalicílico *[Portuguese]* *[Wiki]*

acidum acetylsalicylicum *[Latin]* *[INN]*

Asatard

Asetilsalisilik asit *[Turkish]* *[Wiki]*

Aspirin *[Wiki]* *[USP]* *[BAN]* *[JAN]* *[JP15]*

aspirina *[Basque]*

Aspropharm

Benzoic acid, 2-(acetyloxy)- *[ACD/Index Name]*

ECM

Kyselina 2-acetoxybenzoova *[Czech]*

Kyselina acetylsalicylova *[Czech]*

Melhoral

Miniasal

o-(Acetyloxy)benzoic Acid

o-Acetylsalicylic acid

QVR BOV1 *[WLN]*

Rhodine NC RP

Salicylic acid, acetyl-

Salospir

Tasprin

Toldex

Triaminicin

Ασπιρίνη *[Modern Greek (1453-)]* *[Wiki]*

Ацетилсалициловая кислота *[Russian]*

Ацетилсаліцилова кислота *[Ukrainian]*

アセチルサリチル酸 *[Japanese]* *[Wiki]*

אספירין *[Hebrew]* *[Wiki]*

एस्पिरिन *[Hindi]* *[Wiki]*

阿司匹林 *[Chinese]*

2-(acetoyloxy)benzoic acid

2-(acetyloxy)-benzoic acid

2-Carboxyphenyl acetate *[Spanish]*

2-O-Acetylsalicylic acid

4-10-00-00138 *[Beilstein]*

779271 *[Beilstein]*

Acenterine

Acesal

Aceticyl

Acetilsalicilico

Acenterine

Acesal

Aceticyl

Acetilsalicilico

Acetilum acidulatum

Acetisal

acetol

Acetophen

Acetosal

Acetosalic acid

Acetosalin

ACETYL SALYCYLIC ACID

Acetylin

Acetylsal

acetylsalicylicacid

Acetylsalicylsaeure

Acetylsalicylsaure *[German]*

acetyl-salicylsyra

acetylsalicylzuur

Acetylsaliycilic acid

Acetylsalycilic acid

Acetysal

acide 2-(acetyloxy)benzoique

acido acetilsalicilico *[Italian]*

Acido O-acetil-benzoico *[Italian]*

acidum acetylsalicylicum

Acimetten

Acisal

Acylpyrin

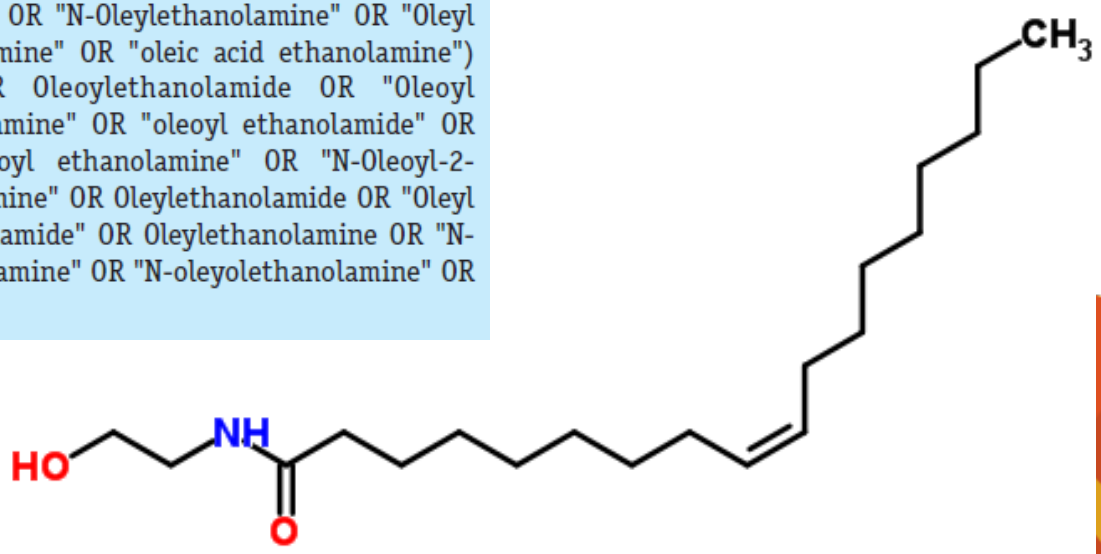
Adiro

AIN

Many Names, One Structure

The final strategy for FreePatentsOnline:

TTL/(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine") OR **ABST/**(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine") OR **ACLM/**(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine")



The ultimate “dictionary”

- Search all forms of structure IDs

- Systematic name(s)
- Trivial Name(s)
- SMILES
- InChI Strings
- InChIKeys
- Database IDs
- Registry Number

Crowdsourced “Annotations”

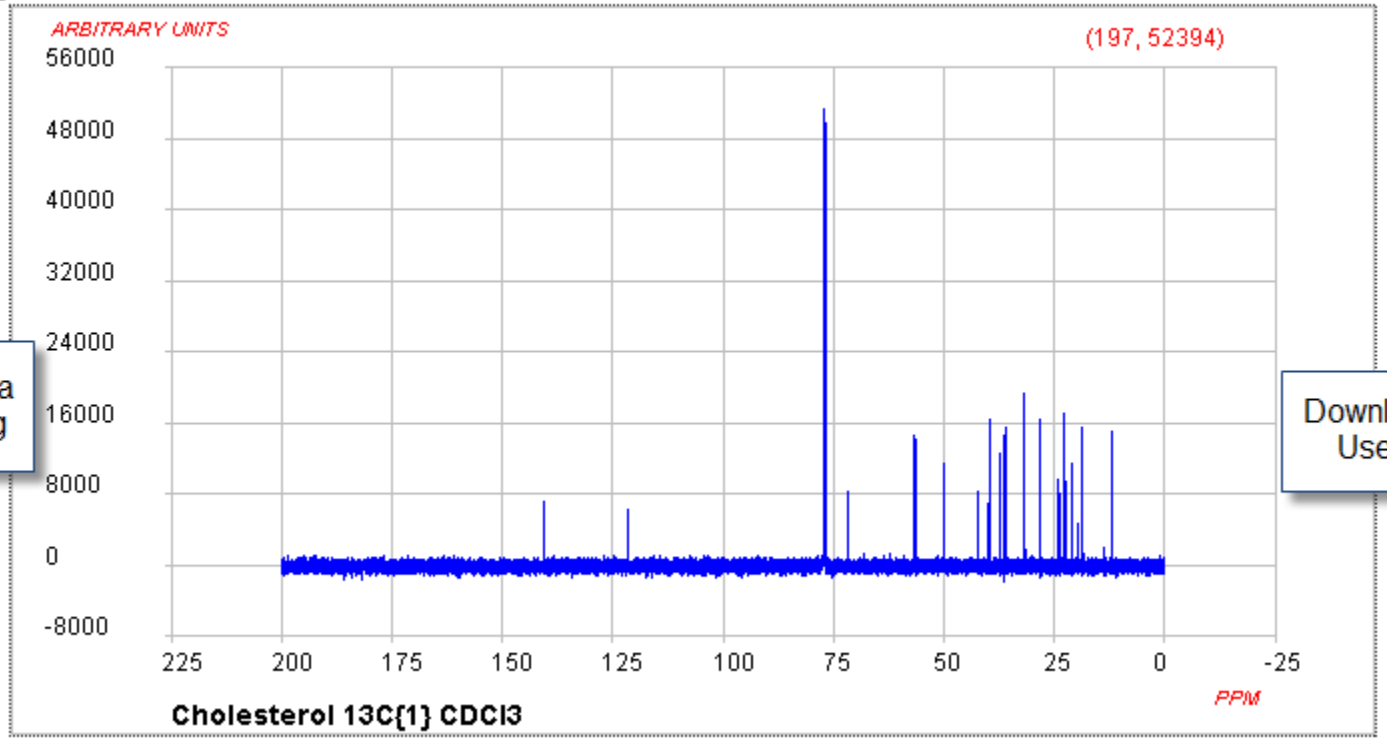
- Users can add
  - Descriptions, Syntheses and Commentaries
  - Links to PubMed articles
  - Links to articles via DOIs
  - Add spectral data
  - Add Crystallographic Information Files
  - Add photos
  - Add MP3 files
  - Add Videos

ChemSpider

- ChemSpider allowed the community to participate in linking the internet of chemistry & crowdsourcing of data
- Successful experiment in terms of building a central hub for integrated web search
- More people are “users” than “contributors”
- Yet basic feedback and game-play helps

ChemSpider Spectra

- **Type:** CNMR
  Approved: Yes



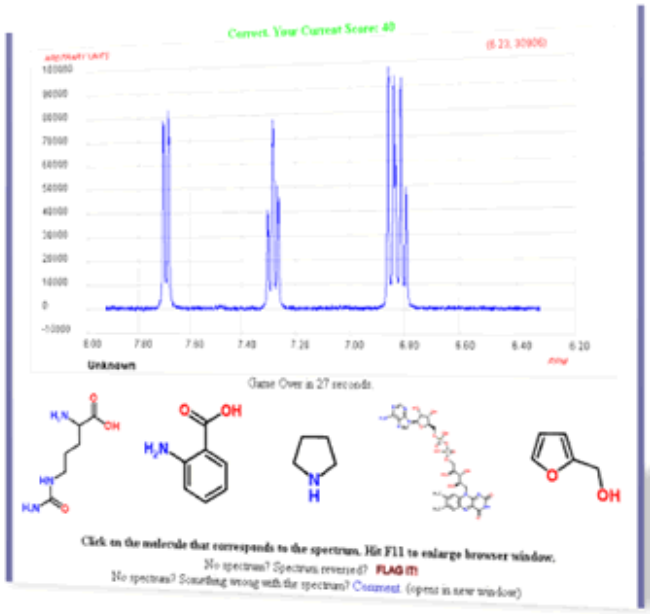
Open Data Licensing

Download and Use Data

Download

www.SpectralGame.com

http://www.jcheminf.com/content/1/1/9



Powered by ChemSpider
Building community for chemists

Powered by ChemDoodle WEB COMPONENTS

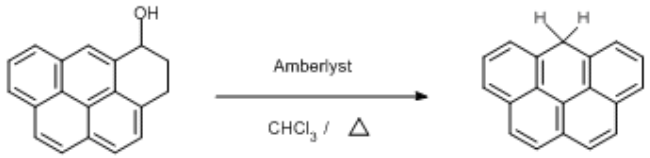
How to play

Each round you'll be presented with a new spectrum. You have to select the molecule that matches the spectrum. For each molecule you identify correctly you'll get one point. The game continues until you get one wrong. Enter your name, choose a group (optional), select the type of spectrum you want and click play.

Micropublishing with Peer Review (a chemical synthesis blog?)

Dehydration of 3,4-dihydro-5H-Benzo[cd]pyren-5-ol; 6H-Benzo[cd]pyrene

SyntheticPage 542
DOI: 10.1039/SP542
Submitted Mar 15, 2012, published May 31, 2012
Anish Mistry (a.mistry@warwick.ac.uk)
A contribution from Fox Group, Warwick University



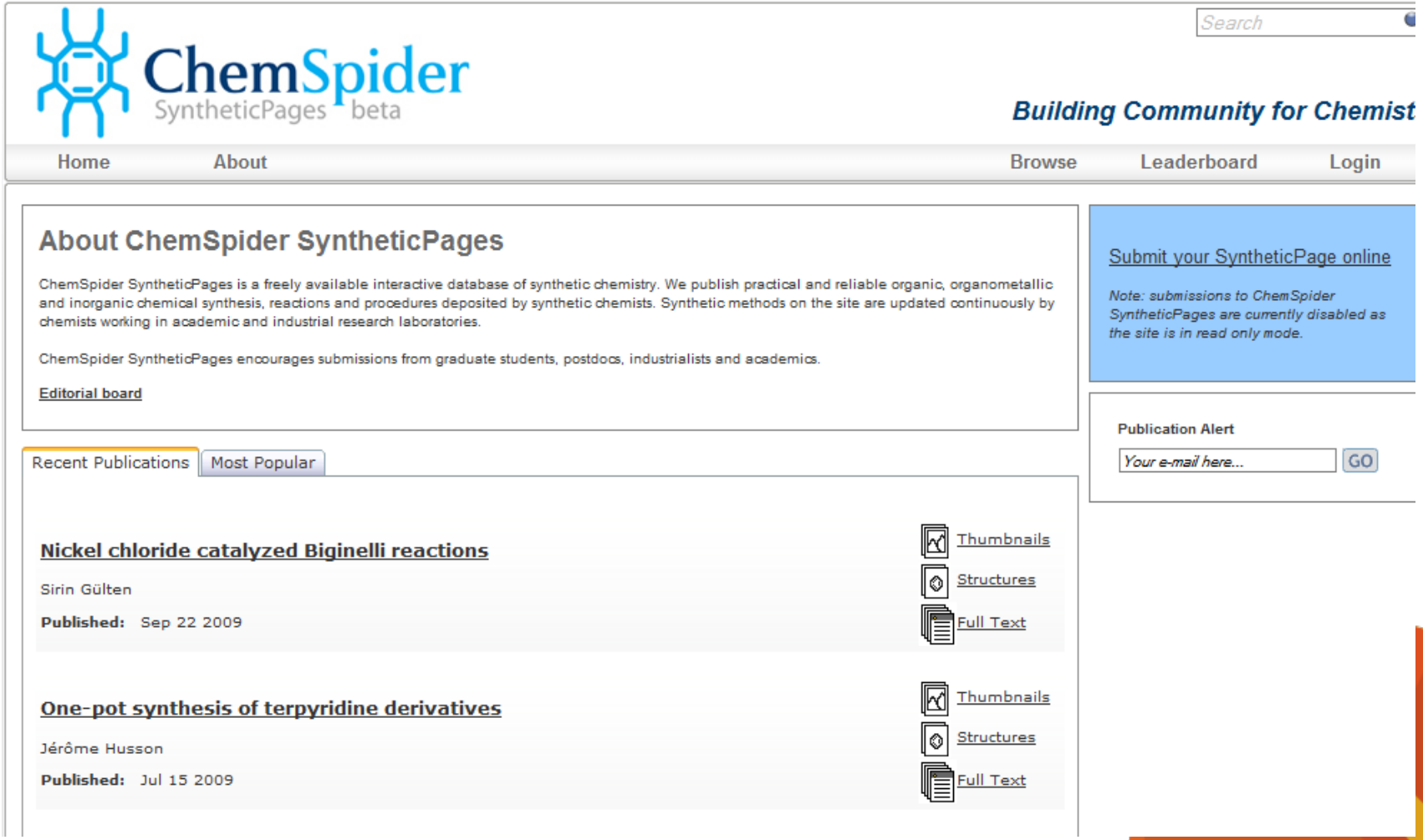
Chemicals Used

3,4-dihydro-5H-Benzo[cd]pyren-5-ol (prepared)
Amberlyst 15 (Sigma-Aldrich)
Chloroform

Procedure

3,4-dihydro-5H-Benzo[cd]pyren-5-ol (0.1 g, 0.39 mmol) was dissolved in chloroform (30 ml) and Aberlyst 15 (0.1 g) added under a dinitrogen atmosphere. The reaction was heated to 30°C and left overnight under the inert atmosphere. The solution was then filtered to seperate the Amberlyst and washed with chloroform. The combined solvents were removed under vacuum using a Rotary evaporator. The crude product was column chromatographed under a dinitrogen atmosphere eluting with 1:1 chloroform:petroleum ether 40-60°C. A white solid was obtained using this method (50 mg, 54%).

ChemSpider SyntheticPages

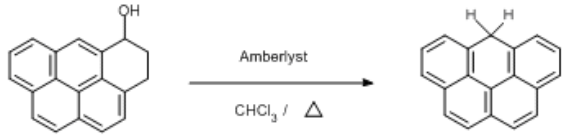


Multi-Step Synthesis

Dehydration of 3,4-dihydro-5H-Benzo[cd]pyren-5-ol

Anish Mistry

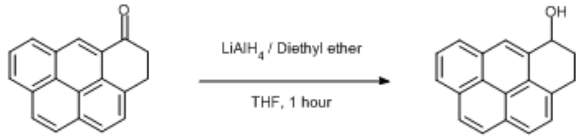
Published: May 31 2012



Reduction of 3,4-dihydro-5H-benzo[cd]pyren-5-one.

Anish Mistry

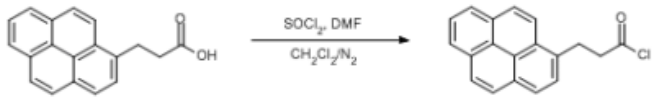
Published: Mar 12 2012



Chlorination of a carboxylic acid

Anish Mistry

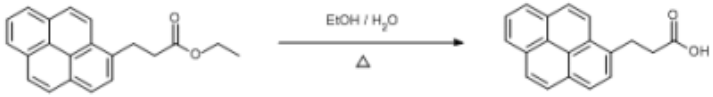
Published: Nov 27 2011



Hydrolysis of Ethyl 3-(1-pyrenyl)propanoate

Anish Mistry

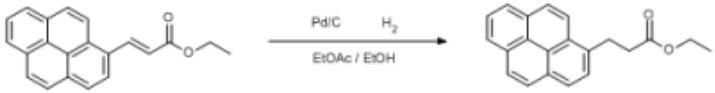
Published: Oct 11 2011



Hydrogenation of Ethyl 3-(1-pyrenyl)acrylate

Anish Mistry

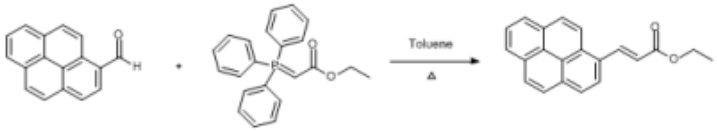
Published: Oct 06 2011



Wittig Reaction

Anish Mistry

Published: Sep 14 2011

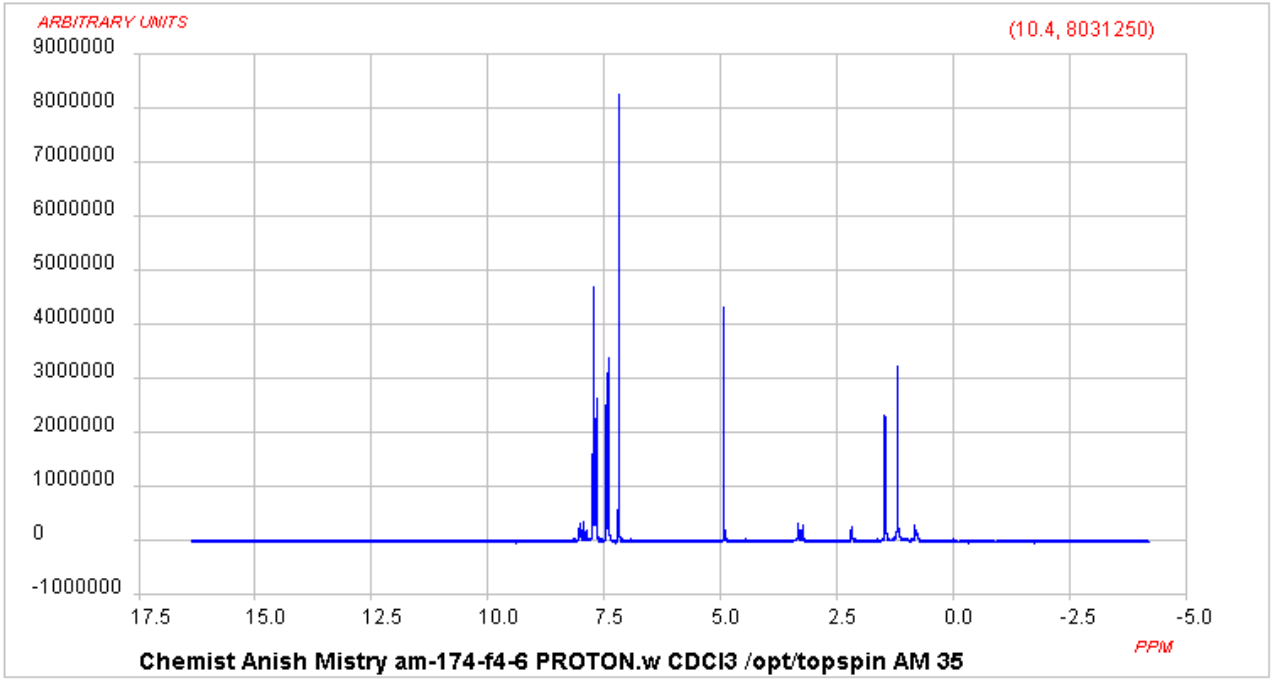


Interactive Data

Supplementary Information

Proton NMR shows the symmetric dihydrogen species as the major product, the small peaks which can be observed in the spectrum (around 2, 3.5 ppm and aromatic region) are presumably other isomers of the compound.

1H NMR spectrum of Olympicene - Click to view



1H NMR (Jan11-2012.dx)

Get structure file (.cdx, .sk2, .mol)

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?

Deposition of Research Data

- If we manage compounds, syntheses and analytical data…
- If we have security and provenance of data…
- If we deliver user interfaces to satisfy the various use cases…
- Then we have delivered electronic lab notebooks for chemistry laboratories. ELNs are research data repositories

What did we learn???

Data Quality is an enormous challenge

Crowd sourced annotation can help!

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation

Antony J. Williams[1], Sean Ekins[2] and Valery Tkachenko[1]

[1] Royal Society of Chemistry, US Office, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[2] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay Varina, NC 27526, USA

In recent years there has been a dramatic increase in the number of freely accessible online databases serving the chemistry community. The internet provides chemistry data that can be used for data-mining, for computer models, and integration into systems to aid drug discovery. There is however a responsibility to ensure that the data are high quality to ensure that time is not wasted in erroneous searches, that models are underpinned by accurate data and that improved discoverability of online resources is not marred by incorrect data. In this article we provide an overview of some of the experiences of the authors using online chemical compound databases, critique the approaches taken to assemble data and we suggest approaches to deliver definitive reference data sources.

Antony J. Williams graduated with a Ph.D. in chemistry as an NMR spectroscopist. Dr Williams is currently VP, Strategic development for ChemSpider at the Royal Society of Chemistry. Dr Williams has written chapters for many books and authored or ≥120 peer reviewed papers and book chapters on NMR, predictive ADME methods, internet-based tools, crowdsourcing and database curation. He is an active blogger and participant in the internet chemistry network.

Sean Ekins graduated from the University of Aberdeen; receiving his M.Sc., Ph.D. and D.Sc. He is Principal Consultant for Collaborations in Chemistry and Collaborations Director at Collaborative Drug Discovery Inc. He has written more than 180 papers and book chapters on topics, including drug drug

Crowdsourced Enhancement

- The community can clean and enhance the database by providing Feedback and direct curation

- Tens of thousands of edits made

ChemSpider is a building block

Methods

SimpleSearch

Run a simple search which tries to interpret a query string as anything it can search by (Synonym, SMILES, InChI, ChemSpider ID etc.)

Parameters

Name	Type	Description
searchOptions	SimpleSearchOptions	Simple search options
commonOptions	CommonSearchOptions	Common search options like HasSpectra or HasPatents
scopeOptions	SearchScopeOptions	Scope options that specify the set where the results should be found. Not supported at the full manner at the moment.
resultOptions	SearchResultOptions	Some search results properties like limit, pagination etc.
limit	Int32	Search limit. Specify how many results return back during the search. Obsolete and will be removed soon. Please use property from SearchResultOptions object

Return Value

Type: String
Request ID string that uniquely identify search request on the server side and that should be used to get information related to search like search status or search results

Example

Run the SimpleSearch operation and search for Aspirin

http://parts.chemspider.com/JSON.ashx?op=SimpleSearch&searchOptions.QueryText=Aspirin

Methods

SimpleSearch
ExactStructureSearch
SubstructureSearch
SimilaritySearch
IntrinsicPropertiesSearch
DataSourceSearch
ElementsSearch
PredictedPropertiesSearch
AdvancedSearch
GetSearchStatus
GetSearchResult
GetSearchResultWithRelevance
GetSearchResultAsCompounds
GetSearchResultAsSdf
GetRecordsAsCompounds
GetRecordsAsSdf
ChemSpiderPrefix
ExtIdSearch
ConvertToStructure
ConvertTo

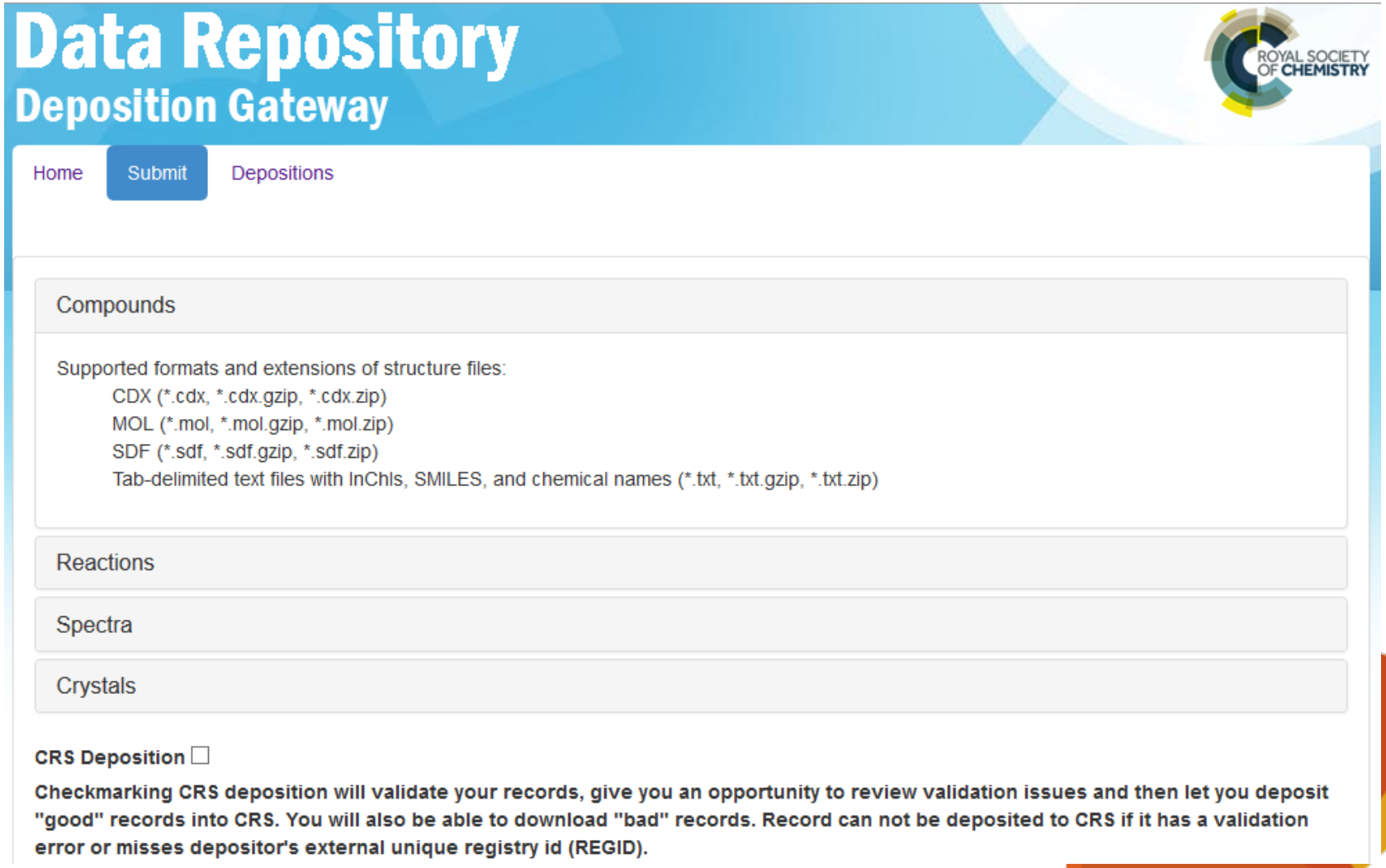
Types

SimpleSearchOptions
CommonSearchOptions
SearchScopeOptions
SearchResultOptions
ExactStructureSearchOptions
SubstructureSearchOptions
SimilaritySearchOptions
IntrinsicPropertiesSearchOptions
DataSourceSearchOptions
ElementsSearchOptions
PredictedPropertiesSearchOptions
AdvancedSearchOptions
StructureSearchOptions
KeywordSearchOptions
LassoSearchOptions
SuppInfoSearchOptions
TextPropertySearchOptions
NumericPropertySearchOptions
AnnotationSearchOptions
CmpIdListSearchOptions
RequestStatus
ResultRecord
Compound
Identifier
Synonym
Reference
Blob
DatasourceType
Datasource
N2SResult
ConvertOptions

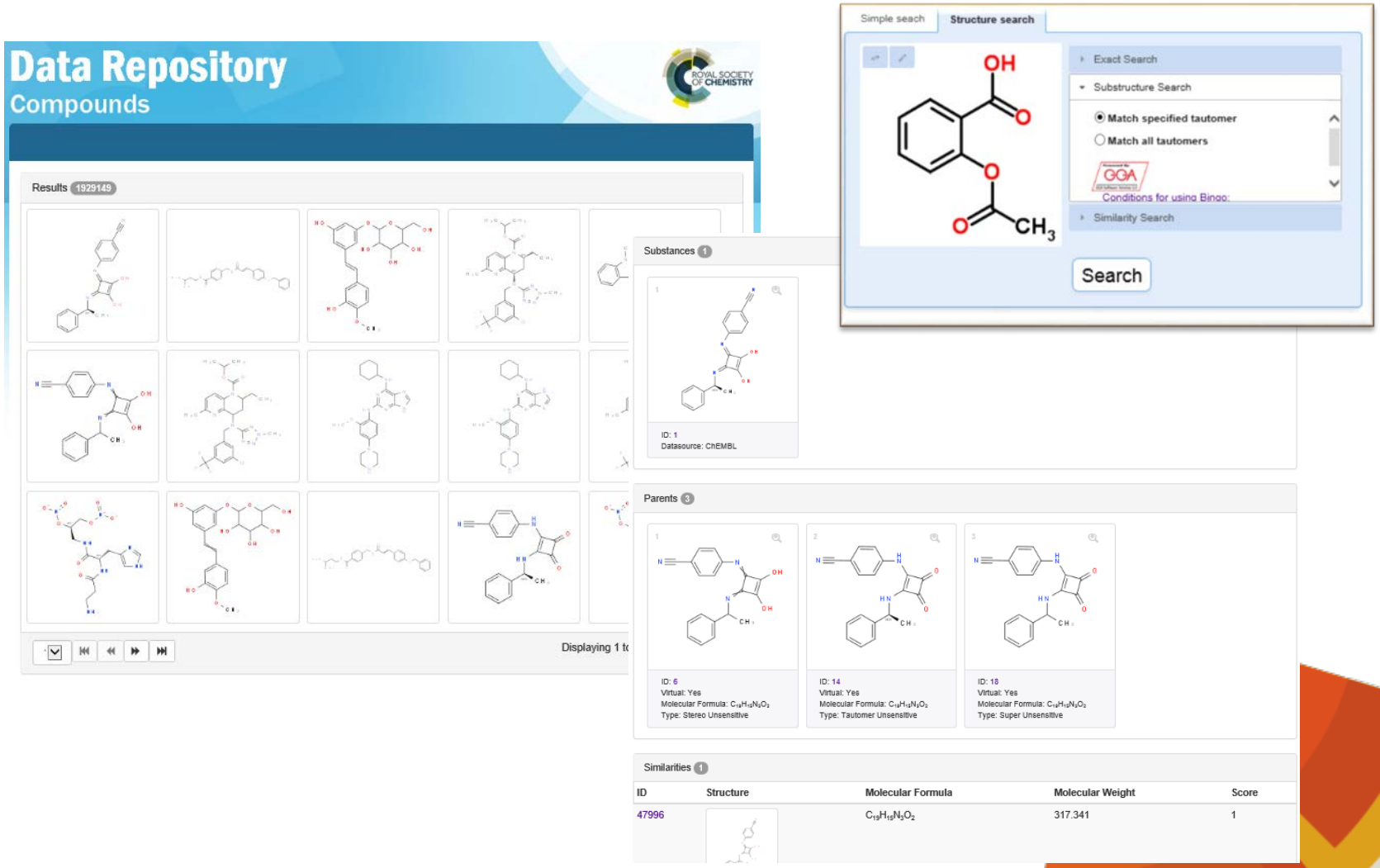
What are we building?

- We are building the “RSC Data Repository”
- Containers for compounds, reactions, analytical data, tabular data
- Algorithms for data validation and standardization
- Flexible indexing and search technologies
- A platform for modeling data and hosting existing models and predictive algorithms

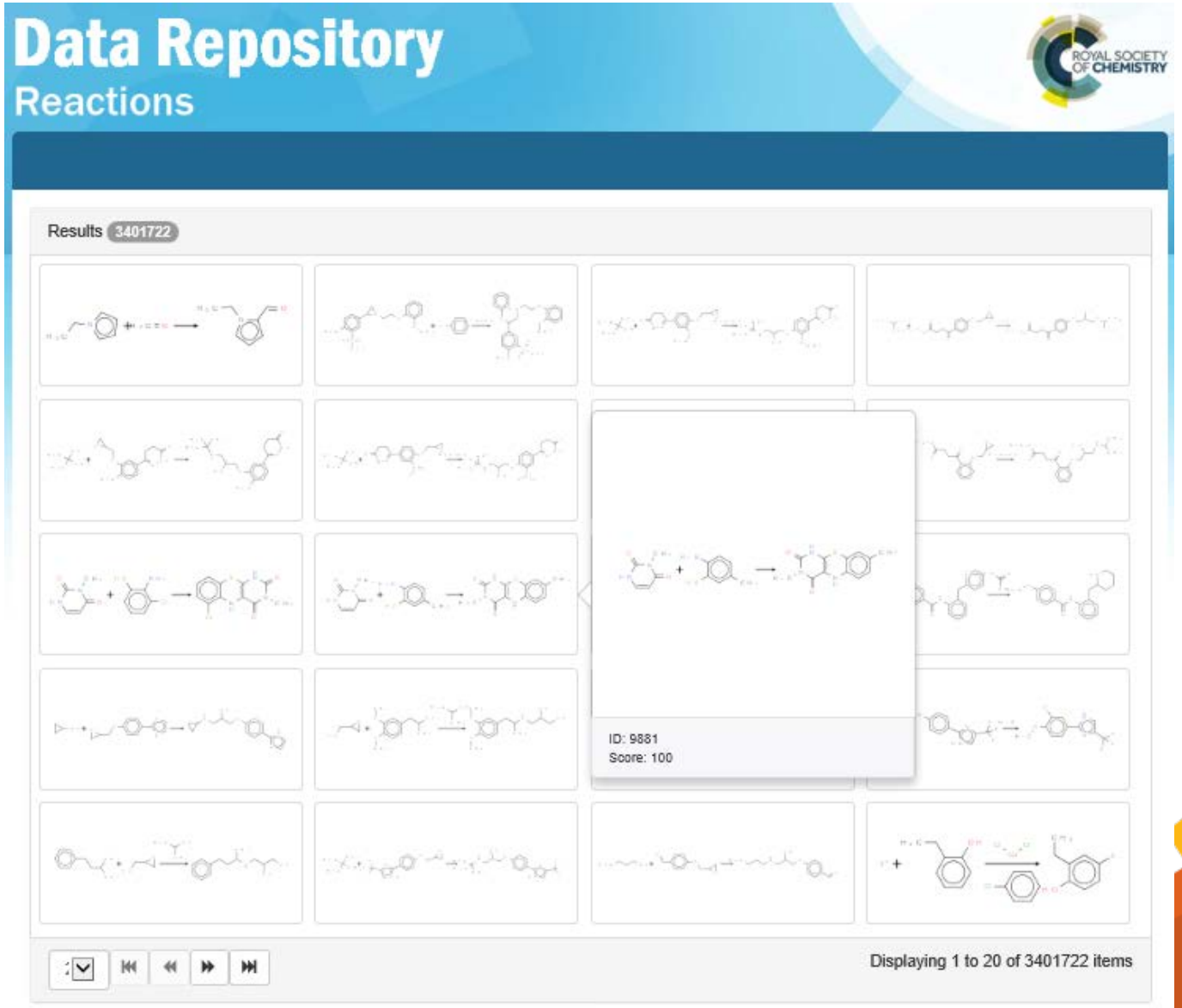
Deposition of Data



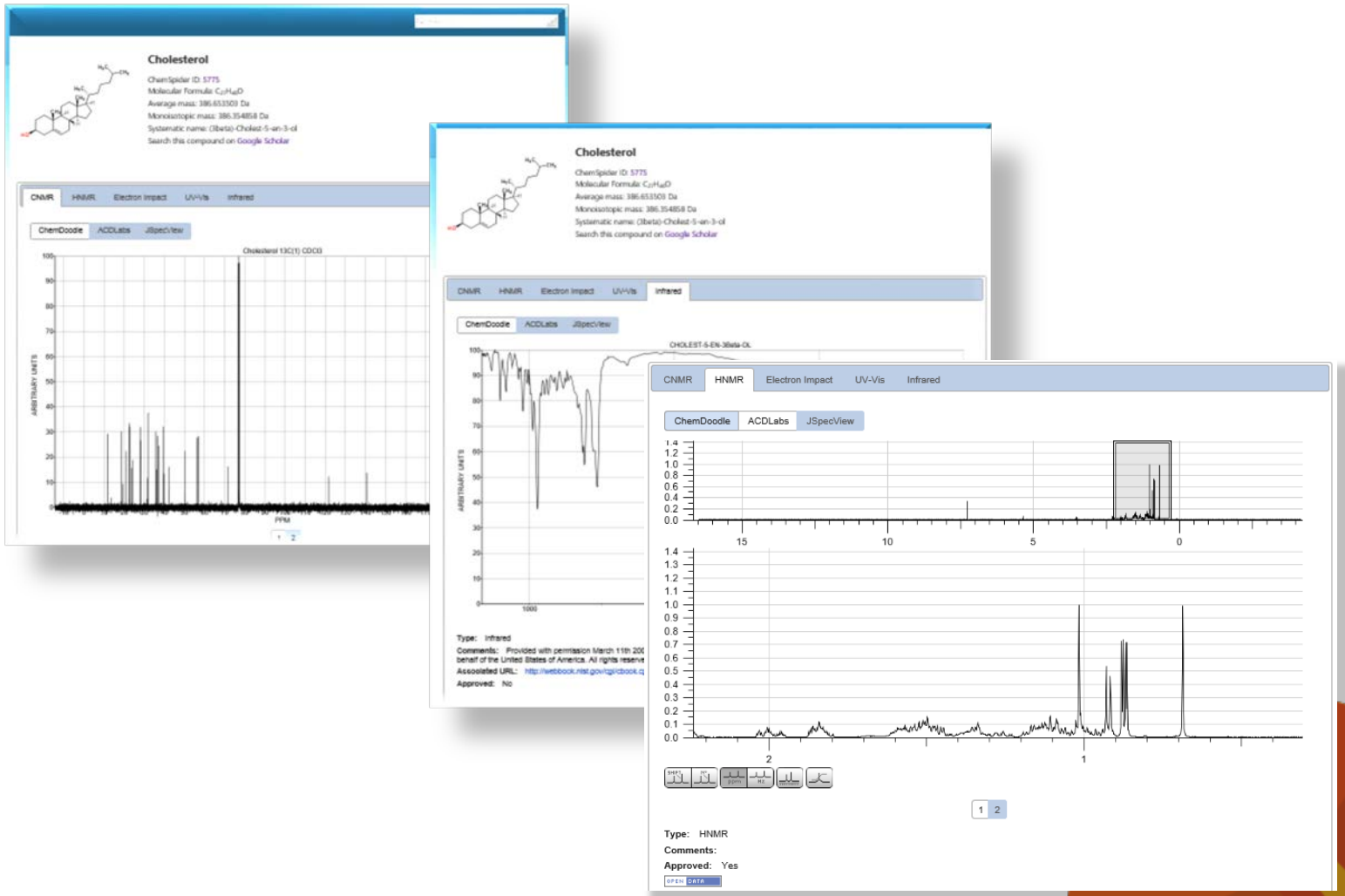
Compounds


Data Repository
Compounds
ROYAL SOCIETY OF CHEMISTRY
Results 1929149
Displaying 1 to
Simple seach
Structure search
Exact Search
Substructure Search
Match specified tautomer
Match all tautomers
Conditions for using Bingo:
Similarity Search
Search
Substances 1
ID: 1
Datasource: ChEMBL
Parents 3
ID: 6
Virtual: Yes
Molecular Formula: C19H19N3O2
Type: Stereo Unsensitive
ID: 14
Virtual: Yes
Molecular Formula: C19H19N3O2
Type: Tautomer Unsensitive
ID: 18
Virtual: Yes
Molecular Formula: C19H19N3O2
Type: Super Unsensitive
Similarities 1
ID
Structure
Molecular Formula
Molecular Weight
Score
47996
C19H19N3O2
317.341
1

Reactions


Data Repository
Reactions
ROYAL SOCIETY OF CHEMISTRY
Results 3401722
ID: 9881
Score: 100
Displaying 1 to 20 of 3401722 items

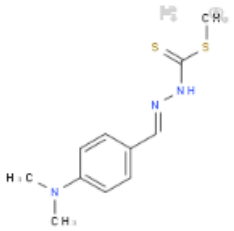
Analytical data



Crystallography data

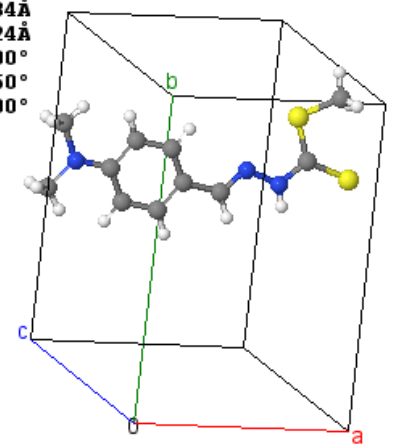
Data Repository

Crystals



ID	529
Molecular Formula	$C_{11}H_{15}N_3S_2$
Monoisotopic Mass	253.070740 Da
Molecular Weight	253.386902 Da
SMILES	CSC(=S)N/N=C/C1C=CC(=CC=1)N(C)C
Std. InChI	InChI=1S/C11H15N3S2/c1-14(2)10-6-4-9(5-7-10)8-12-13-11(15)16-3/h4-8H,1-3H3,(H,13,15)/b12-8+
Std. InChIKey	NKOZYBZPSHOLOD-XYOKQWHBSA-N
ChemSpider ID	5279366

HM:P 1 1 21/b
a=8.364Å
b=12.684Å
c=12.224Å
α=90.000°
β=92.360°
γ=90.000°



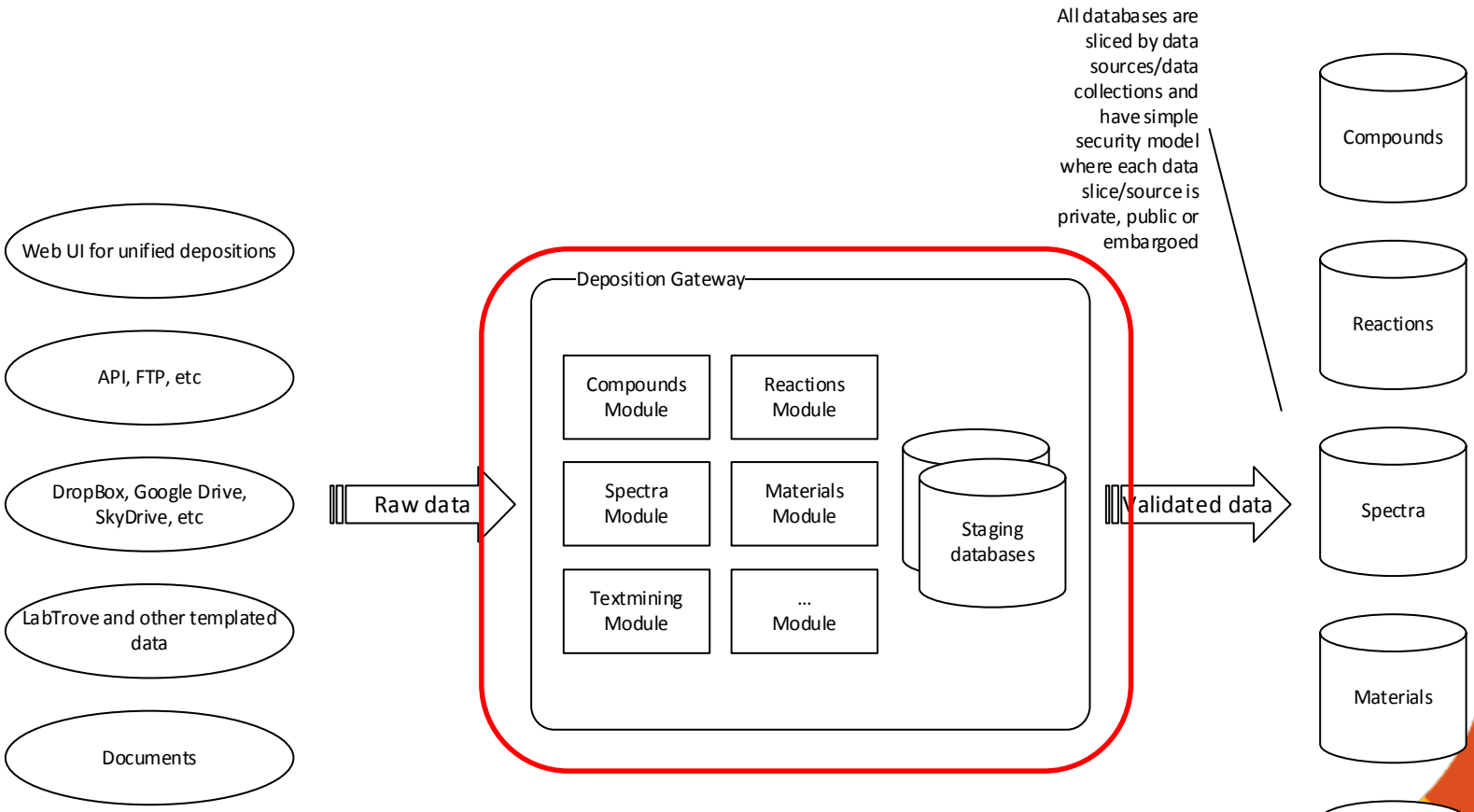
Jmol

Deposition of Data

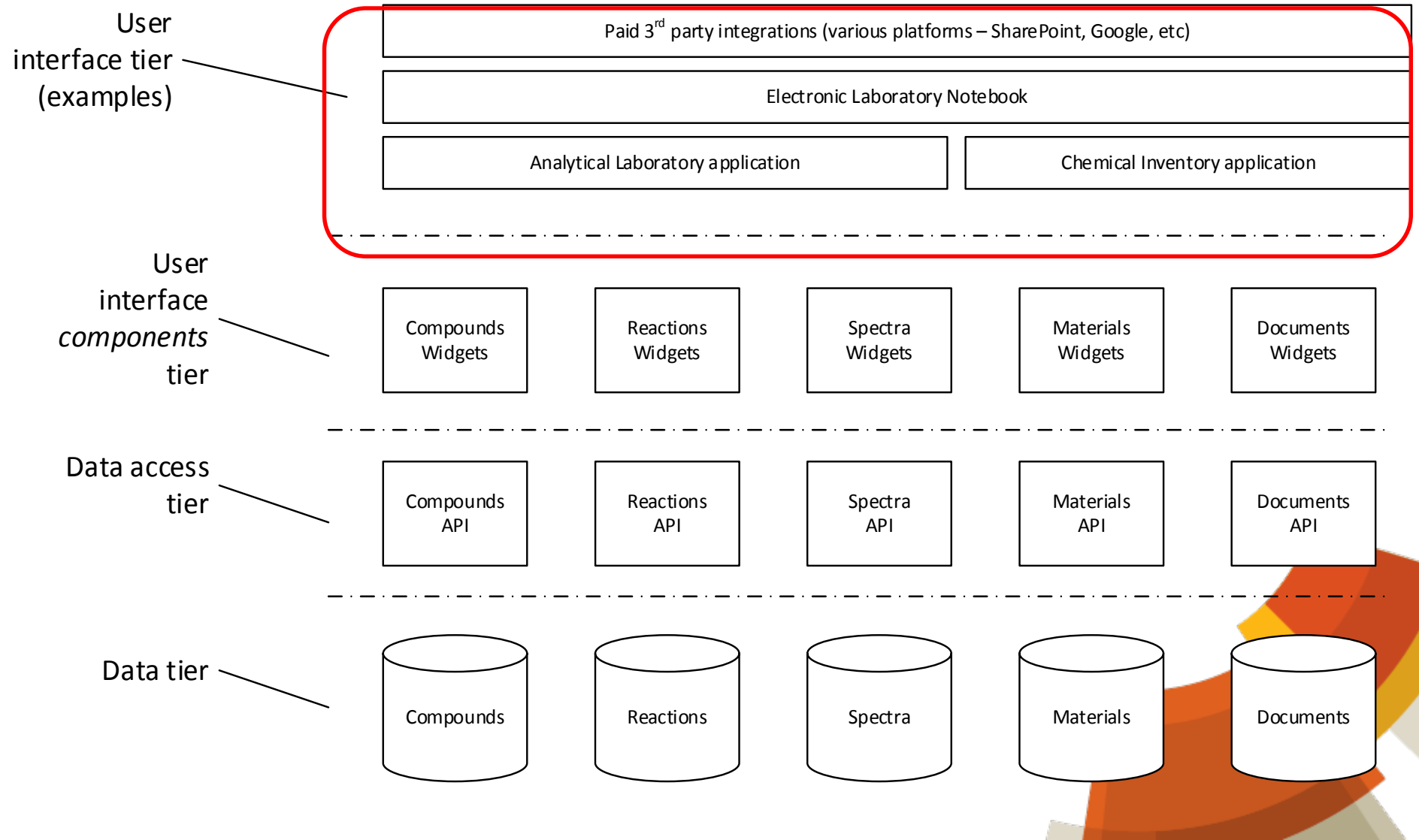
- Developing systems that provides feedback to users regarding data quality
  - Validate/standardize chemical compounds
  - Check for balanced reactions
  - Checks spectral data

- EXAMPLE Future work
  - Properties – compare experimental to pred.
  - Automated structure verification - NMR

Input data pipeline



User Interface Approach



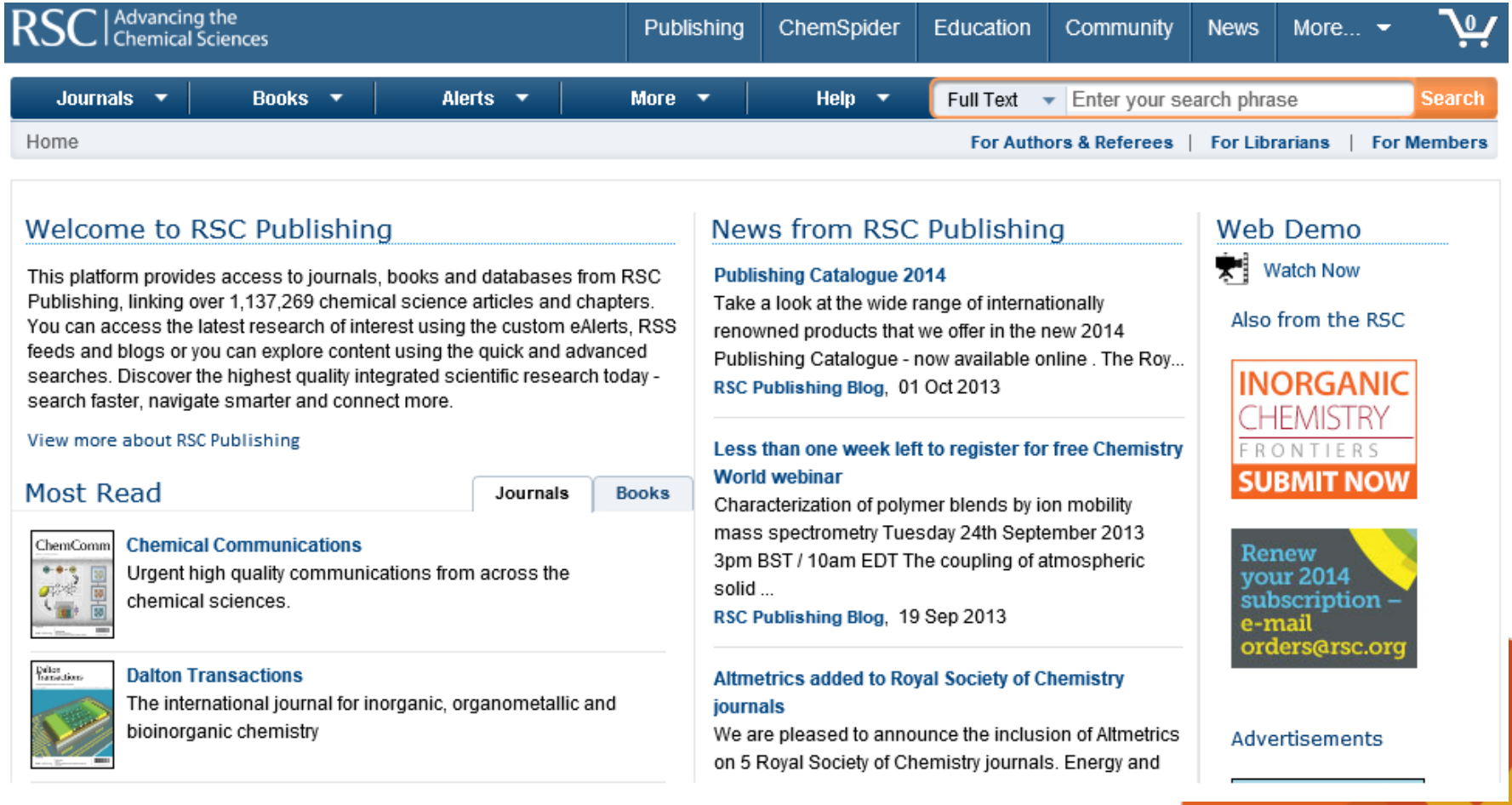
Can we get historical data?

- Text and data can be mined
- Spectra can be extracted and converted
- SO MUCH Open Source Code available

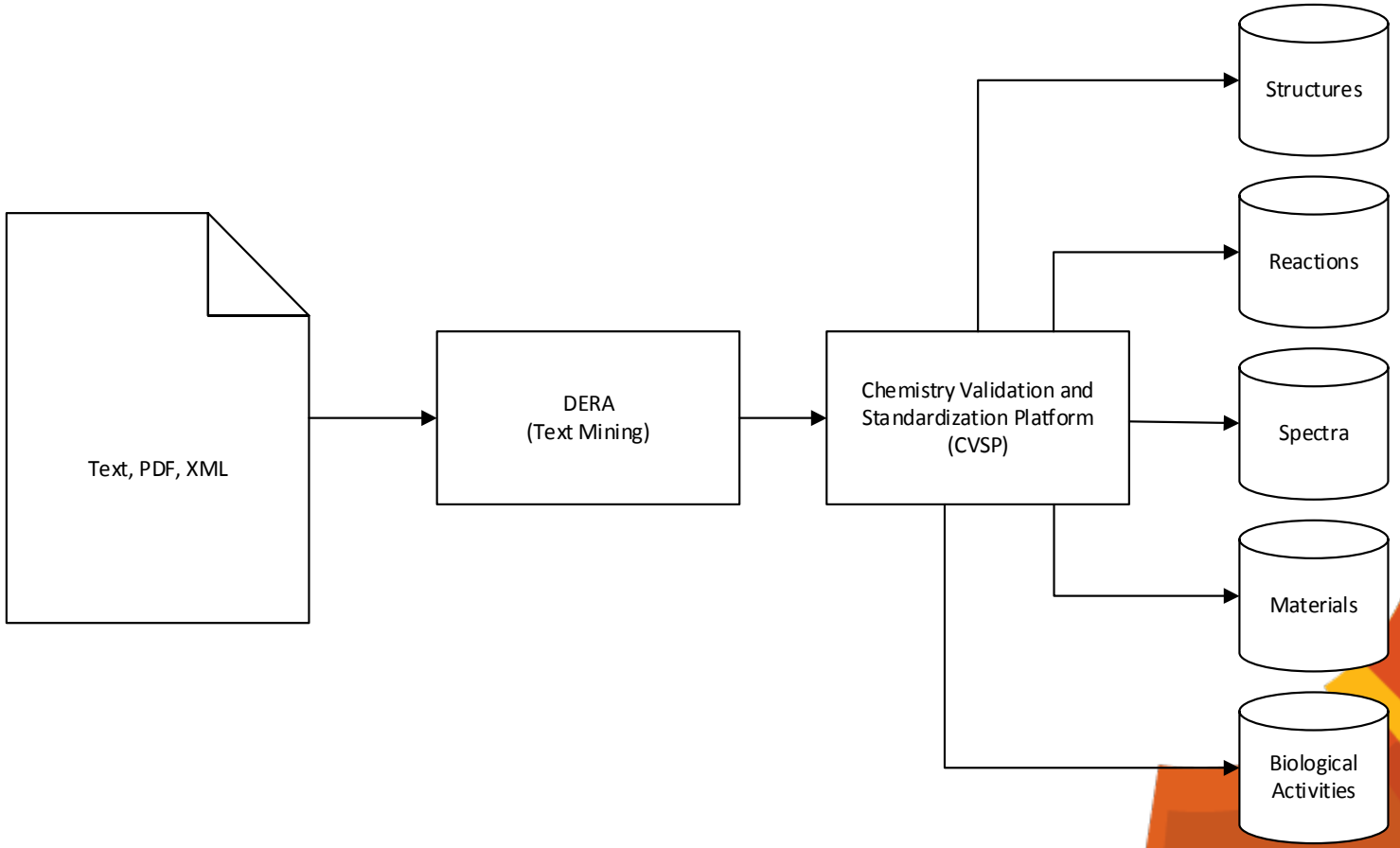
Extracting our Archive

- What could we get from our archive?
  - Find chemical names and generate structures
  - Find chemical images and generate structures
  - Find reactions
  - Find data (MP, BP, LogP) and deposit
  - Find figures and database them
  - Find spectra (and link to structures)

RSC Archive – since 1841

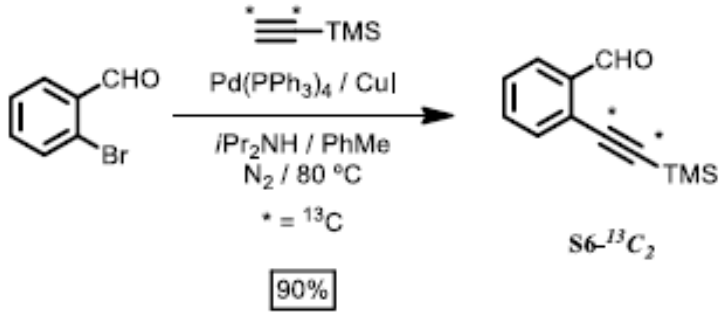


Digitally Enabling RSC Archive



SO MANY reactions!

Scheme S13. Synthesis of **S6-$^{13}C_2$**



Synthesis of S6-$^{13}C_2$. 2-bromobenzaldehyde (0.526 g 2.84 mmol), CuI (5 mg, 0.028 mmol), and $Pd(PPh_3)_4$ (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 N_2 three times. Anhydrous PhMe (3 mL) and anhydrous *i*Pr$_2$NH (1 mL) were added via cannula under N_2. The mixture was bubbled with N_2 for 20 min and trimethylsilylacetylene-$^{13}C_2$ (99% atom ^{13}C, 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 NH_4Cl (*aq*), and extracted with CH_2Cl_2 (3 x 10 mL). The combined organic extracts were rinsed with saturated NH_4Cl (*aq*), water, and brine. The solution was dried over anhydrous $MgSO_4$, filtered over celite, and concentrated to dryness. The obtained dark residue was purified by column chromatography (SiO_2, 5% *v/v* THF/hexanes) to provide **S6-$^{13}C_2$** (0.554 g, 90% yield) as a yellow oil. ^1H NMR (400 MHz, $CDCl_3$) 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). ^{13}C NMR (100 MHz, $CDCl_3$) δ 102.74 (d, J_{CC} = 136.2 Hz), 100.02 (d, J_{CC} = 136.3 Hz). EI-MS: calcd for $[C_{10}{}^{13}C_2H_{14}OSi]^+$ 204.09, found 203.15.

Text Mining

The N-(β-hydroxyethyl)-N-methyl-N'-(2-trifluoromethyl-1,3,4-thiadiazol-5-yl)urea prepared in Example 6 , thionyl chloride (5 ml) and benzene (50 ml) were charged into a glass reaction vessel equipped with a mechanical stirrer , thermometer and reflux condenser .

The reaction mixture was heated at reflux with stirring , for a period of about one-half hour .

After this time the benzene and unreacted thionyl chloride were stripped from the reaction mixture under reduced pressure to yield the desired product N-(β-chloroethyl)-N-methyl-N'-(2-trifluoromethyl-1,3,4-thiaidazol-5-yl)urea as a solid residue

Text Mining

The N-(β-hydroxyethyl)-N-methyl-N'-(2-trifluoromethyl-1,3,4-thiadiazol-5-yl)urea prepared in Example 6 , thionyl chloride (5 ml) and benzene (50 ml) were charged into a glass reaction vessel equipped with a mechanical stirrer , thermometer and reflux condenser .

The reaction mixture was heated at reflux with stirring , for a period of about one-half hour .

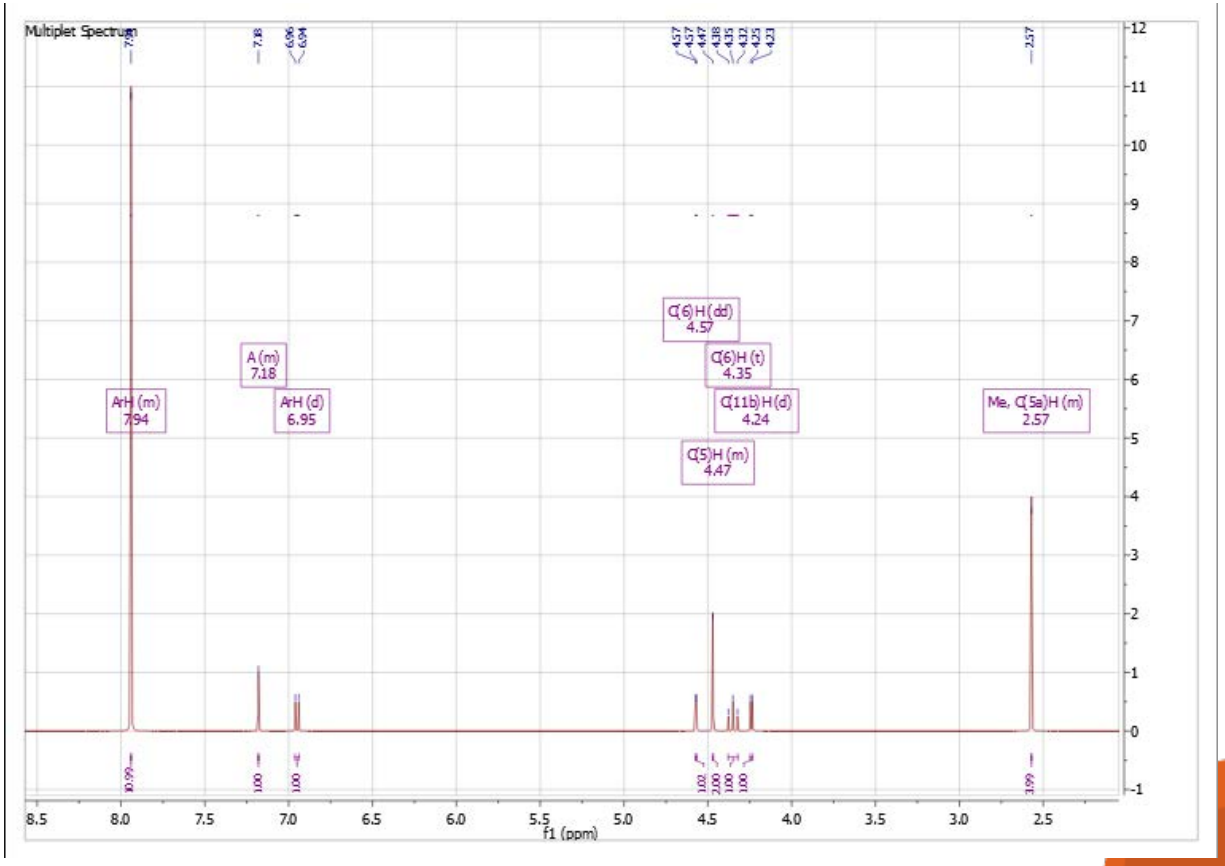
After this time the benzene and unreacted thionyl chloride were stripped from the reaction mixture under reduced pressure to yield the desired product N-(β-chloroethyl)-N-methyl-N'-(2-trifluoromethyl-1,3,4-thiaidazol-5-yl)urea as a solid residue

Text spectra?

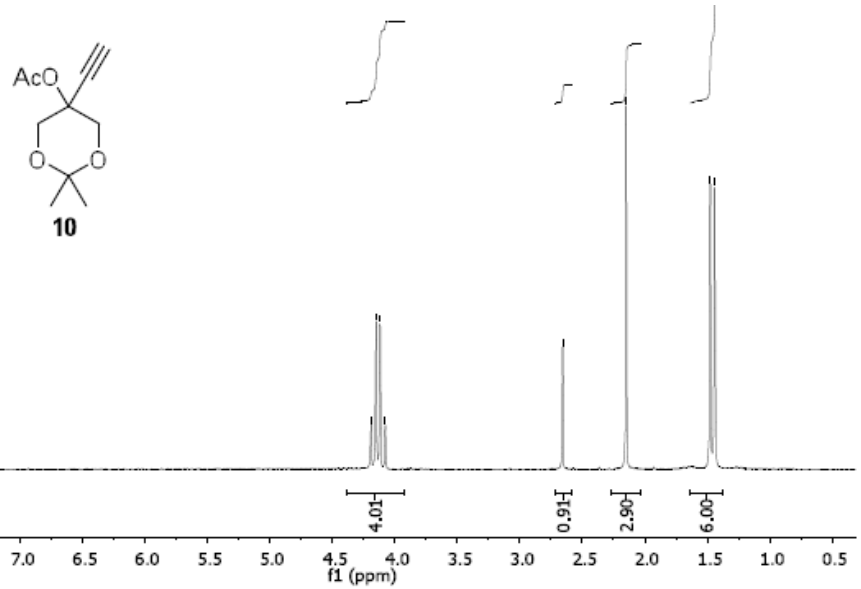
13C NMR (CDCl3, 100 MHz): δ = 14.12 (CH3), 30.11 (CH, benzylic methane), 30.77 (CH, benzylic methane), 66.12 (CH2), 68.49 (CH2), 117.72, 118.19, 120.29, 122.67, 123.37, 125.69, 125.84, 129.03, 130.00, 130.53 (ArCH), 99.42, 123.60, 134.69, 139.23, 147.21, 147.61, 149.41, 152.62, 154.88 (ArC)

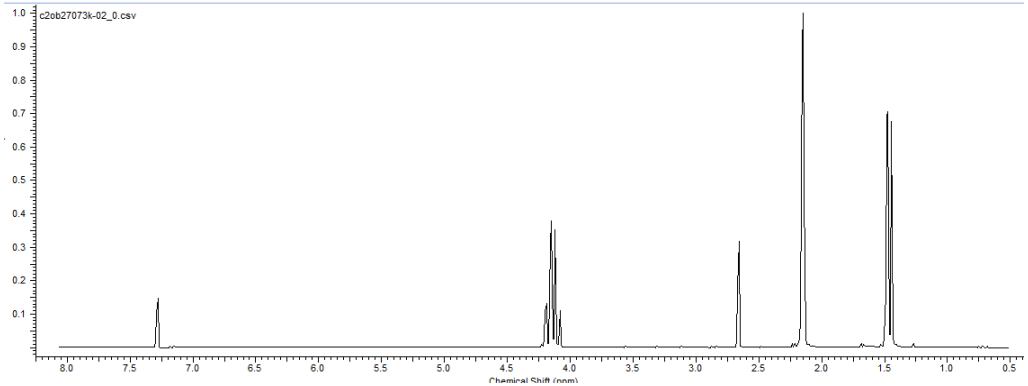
1H NMR (CDCl3, 400 MHz):

δ = 2.57 (m, 4H, Me, C(5a)H), 4.24 (d, 1H, J = 4.8 Hz, C(11b)H), 4.35 (t, 1H, Jb = 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)

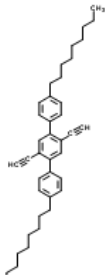


Turn “Figures” Into Data





Make it interactive



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

ChemSpider ID: **29211602**

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

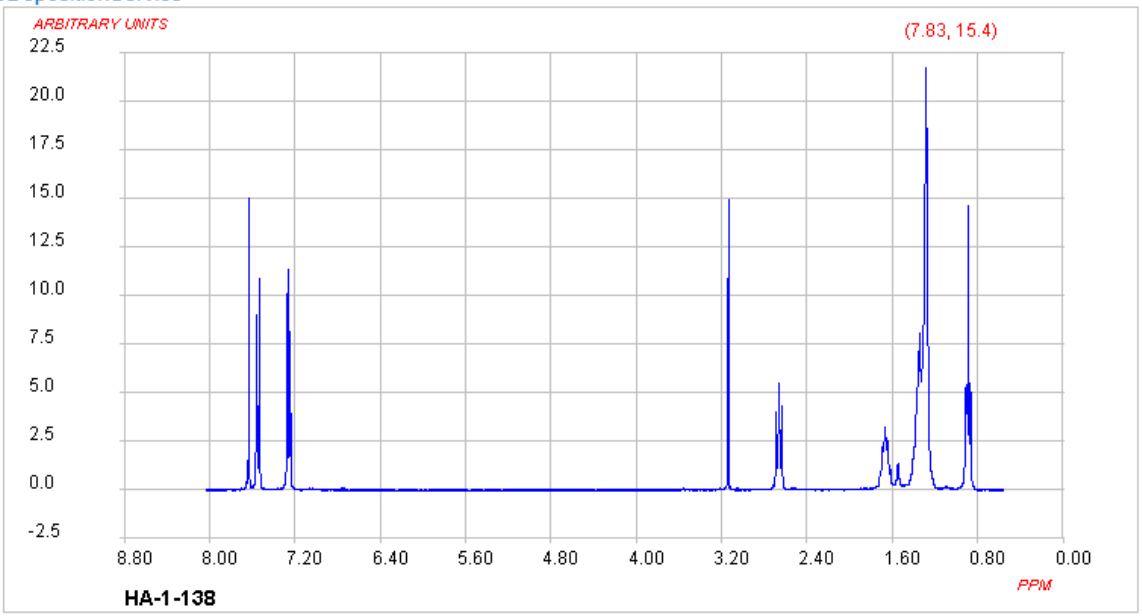
Average mass: 530.825012 Da

Monoisotopic mass: 530.391235 Da

▼ Systematic name

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

- **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



OPEN DATA

Download

How is DERA going?

- We have text-mined all 21st century articles… >**100k** articles from 2000-2013
- Marked up with XML and published onto the HTML forms of the articles
- Required multiple **iterations** based on dictionaries, markup, text mining iterations
- New visualization tools in development – not just chemical names. Add chemical and biomedical terms markup also!

Is It Easy?

ChEBI
UniProt
the Gene Ontology
RSC ontologies (methods, reactions)

Curated dictionaries for known names

Dictionary (ontologies)
Dictionary (chemistry)

ChEMBL
DrugBank
Open Data Drug & Drug Target Database
ChEBI

ChemSpider
Search and share chemistry

Unknown names: automated name to structure conversion

Production processes

Text-mining

OPSIN

ACD N2S

XML ready for publication

Chemical structures SD file

CDX integration (coming soon)

Marked-up XML

Support grant-based services

- Multiple European consortium-based grants
  - **PharmaSea** (FP7 funded)
  - **Open PHACTS** (IMI funded)
- UK National **Chemical Database Service** (http://cds.rsc.org) – developing data repository for lab data, integrate Electronic Lab Notebooks

PHARMASEA

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.

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

Related Stories

http://www.openphacts.org

Open PHACTS is an Innovative Medicines Initiative (IMI) project, aiming to reduce the barriers to drug discovery in industry, academia and for small businesses.

Semantic web is one of the corner stones

ROYAL SOCIETY OF CHEMISTRY

- 3-year Innovative Medicines Initiative project
- Integrating chemistry and biology data using semantic web technologies
- **Open code, open data, open standards**
- Academics, Pharmas, Publishers…
- To put medicines in the pipeline…

The Open PHACTS community ecosystem

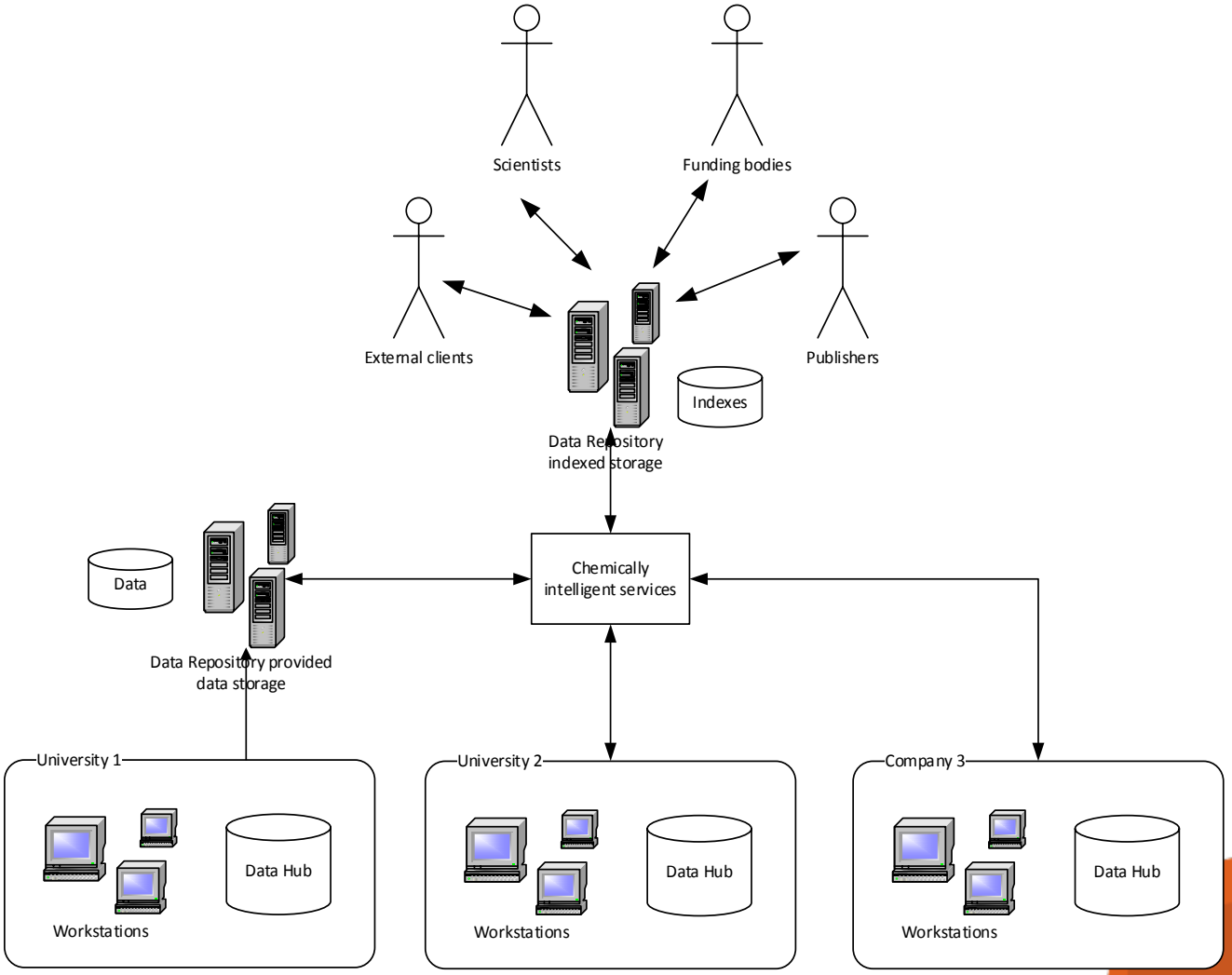
UK Chemical Database Service

- The National Chemical Database Service is for **UK** academics

EPSRC

Engineering and Physical Sciences Research Council

National Data Repository



National Chemistry Database

National Chemical Database Service

About NCDS | Contact us

The National Chemical Database Service offers access to a suite of commercial databases and resources, with additional development to create a chemistry data repository to take place. All UK academic institutions are eligible for access to the Service, access will initially be authenticated on institutional IP address - a username and password can be obtained if this is not possible.

ACD/I-Lab2

Physicochemical, ADME and toxicity property prediction (ACD/Labs Inc.).

Further information

CSD

Organic and organometallic crystal structures (CCDC).

Further information

DETHERM

DECHEMA
Gesellschaft für Chemische Technik und Biotechnologie e.V.

Database of thermophysical data for pure substances and mixtures.

Further information

ICSD

FIZ Karlsruhe ICSD

>160,000 inorganic and related crystal structures (FIZ Karlsruhe GmbH).

Further information

Available Chemicals Directory

accelrys

Provides supplier information for building block molecules.

Further information

ARChem

ARChem
SimBioSys, Inc.

Retrosynthetic tool for chemical analysis of target organic molecules.

Further information

Chemicalize

ChemAxon

Physicochemical property prediction tools with Lipinski-like filters.

Further information

ChemSpider

ChemSpider
The free chemical database

An online database of molecules from >400 datasources (RSC).

Further information

SPRESIweb

SPRESIweb

In partnership with the EPSRC

EPSRC
Engineering and Physical Sciences

Vision for the Service PART 1

- Provide access to databases and services of interest to the academic community to serve their needs. Access to services to include:
  - Crystallography data – Organic and inorganic materials
  - Thermophysical data
  - Reactions Data including retrosynthetic analysis
  - Prediction technologies – name generation, physicochemical parameters, NMR prediction

Vision for the Service PART 2

- Response to the call for proposals included our vision for a 21st Century data repository
- At a time of Open Access, Open Data and funding agency requirement to make data public – build a data repository
- Funding is split for licensing content and services (VAST MAJORITY) and some funding for research and development

An Initial “Vague” Vision Set

- Manage “all” of the chemistry data associated with chemical substances
- Data to be downloadable, reusable, interactive
- Build a platform that enables the scientist
  - Data storage, validation, standardization and curation
  - Collaborative data sharing
- **Provide data platform that can enable and enhance publishing of scientific papers**

10 years ago…

- There was no iPad, no iPhone, no Android
- Facebook had only just been released

Powered by RSC Data

- What will it be like when we are hosting chemistry data that doesn’t get published???
- Or hosting all data UNTIL it gets published??
- What will it be like when computer models are being rebuilt every time there is a new dataset – validating the data, flagging data
- What will it be like when publications are not only peer-reviewed but also computer reviewed?

The Future

Individual Scientists

Internet Data

Published Chemistry

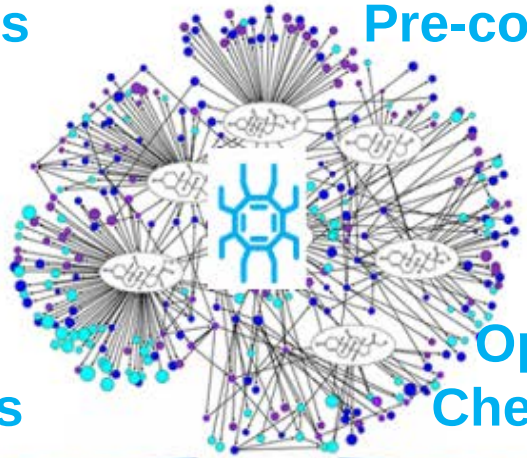
RSC Advancing the Chemical Sciences RSC Data

Electronic Lab Notebooks

Enterprise Laboratory Platform lims eln

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

A Global Chemistry Network

- The Global Chemistry Network is much bigger than just data - scientific networking, micro/publishing, integration hub.
- The data repository as a handler for data, GCN as a submission interface, GCN as a profile handler, rewards and recognition platform etc.
- Data repository architecture designed to deliver the underpinning data containers and visualization widgets etc.

Thank you!

Come visit us at our stand to find out more

More information:
http://rsc.li/open-access
http://rsc.li/goldforgold

E-mail: OConnorB@RSC.ORG
Twitter: @btdoconnor