

Chemoinformatics in Europe: Research and Teaching



Chemical informatics: where are we headed?

Can Chemoinformatics move to a new paradigm ?

- Previously
 - Data.....modelling.....prediction
- Future
 - Information.....knowledge.....decision

What is changing ?

■ Previously

– Data.....modelling.....prediction

- Data : numbers, pictures, descriptions without context, disconnected and incomplete – single use databases, no metadata
- Modelling : localised in analysis, unconnected, often artifactual, not phenomenological (most of QSAR)
- Prediction : limited to one application, data dependant, not transferable – or often not reproducible

■ Future

– Information.....knowledge.....decision

- Information is contextual, marked up, transferrable, can be integrated with other data, adheres to standards

– Knowledge

- Abstractions from relationships between data, deeper understanding of underlying rules, more robust prediction, amenable to natural language queries

– Decision

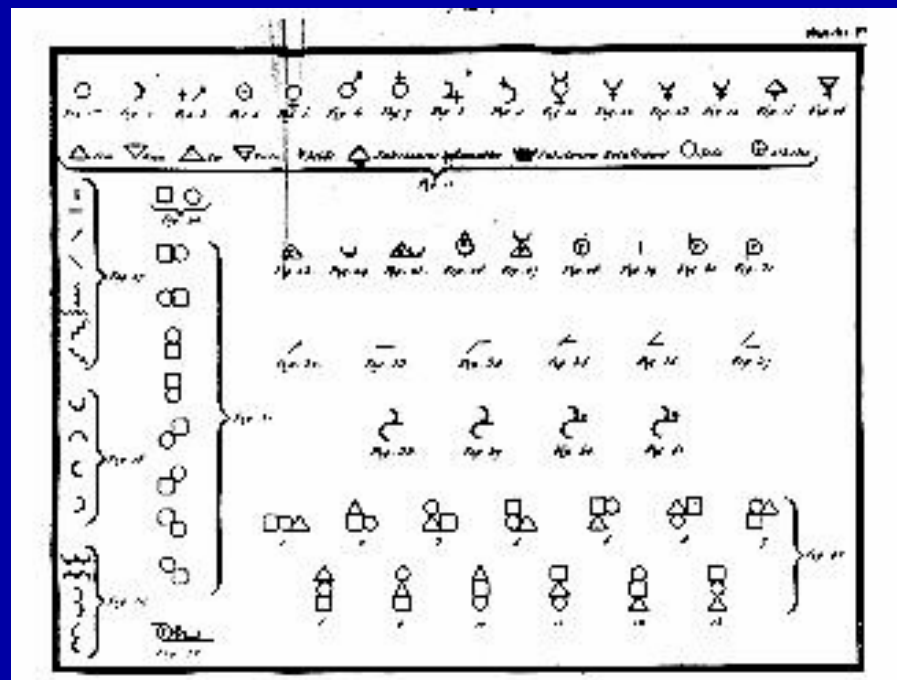
- Moving from single predictions e.g. a compound is predicted active in an assay to
- What factors influence activity, is a compound fit for development, what experiments need to be done, is there a market, has it been made before?

A bit of history on how molecules are described and how this affects what we can do.....

- How molecules have been represented
- Early 'Informatics'

Representing compounds as pictures

- Early chemical symbols and those created by Jean Henri Hassenfratz and Pierre Auguste Adet to complement the *Methode de Nomenclature Chimique* (1787).





Berzelius : introduced our familiar alphabetic symbols, isolated cerium, selenium, and thorium, and invented words such as allotropy, isomerism, and protein catalyst.

He proposed that names for chemicals should be ‘definitions of the composition of the substances’ – gave rise to our atomic symbols

The ‘connections’ between symbols of Scott-Couper gave rise to structural diagrams

An early application of infomatics

On pourrait écrire beaucoup d'autres corps, par exemple l'acide tartarique :

$$\begin{array}{c} \text{O} - \text{OH} \\ | \\ \text{O} \\ | \\ \text{H} \\ | \\ \text{O} - \text{OH} \\ | \\ \text{H} \\ | \\ \text{O} \\ | \\ \text{O} - \text{OH} \end{array}$$

et l'acide biborique dérivé de l'acide tartarique par l'action de la chaleur sans perdre :

$$\begin{array}{c} \text{O} - \text{OH} \\ | \\ \text{O} \\ | \\ \text{H} \\ | \\ \text{O} - \text{OH} \\ | \\ \text{H} \\ | \\ \text{O} \\ | \\ \text{O} - \text{OH} \end{array}$$

Il résulte de là que le carbone et l'azote combinés, de manière à constituer tous deux les limites de leur pouvoir de combinaison, forment un corps dans l'affinité libre s'exercera en formant un équivalent d'hydrogène ou d'un autre élément.

Ainsi la formule de l'acide cyanhydrique sera

$$\begin{array}{c} \text{H} \\ | \\ \text{C} \end{array} \text{Az}$$

L'acide cyanique sera

$$\text{HO} = \begin{array}{c} \text{O} \\ | \\ \text{C} \end{array} \text{Az}$$

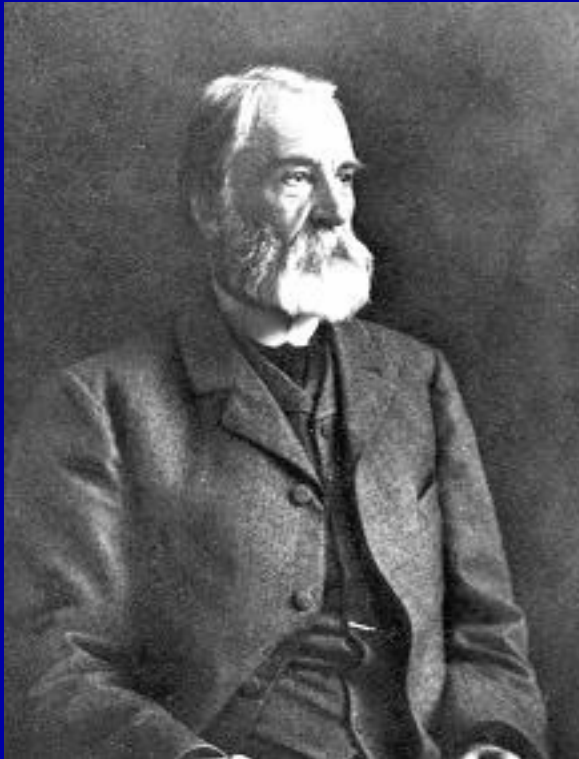
L'acide cyanurique

$$\text{HO} = \begin{array}{c} \text{O} - \text{C} - \text{Az} - \text{C} - \text{Az} - \text{O} - \text{OH} \\ | \\ \text{C} \\ | \\ \text{O} \end{array} \text{Az}$$

Dans cette dernière formule, les atomes de carbone et d'azote sont liés par 3 unités d'affinité et non par 4, comme dans les deux premiers exemples.

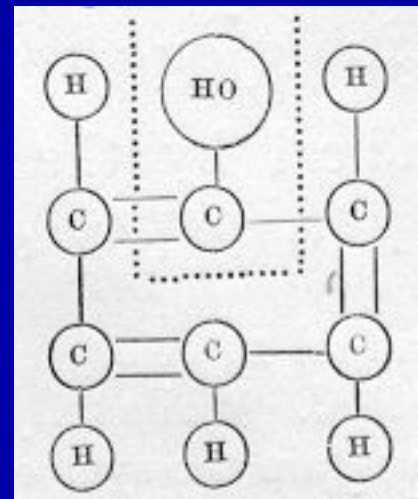
Archibald Scott Couper's bond lines.

Frederick Beilstein



In the Handbuch the naming of compounds was an integral part of their storage and retrieval from indexes.

This drive for efficient indexing dominated chemical information for the next half century (still does today)



Machine readable structure representation.

■ Line Notations

– WLN

- L66J BMR& DSWQ IN1&1

– ROSDAL

- 1=-5=-10=5,10-1,1-11N-12-
=17=12,3-18S-
19O,18=20O,18=21O,8-22N-
23,22-24

– SMILES

- c1ccccc1Nc2cc(S(=O)(=O)O)c3c2cc(N(C)C)cc3



Example of an SD file

GABA analog

28 27

2.8660 0.0000 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0

. Atom block: X Y Z SYMBOL MASS CHARGE STEREO

. Bond block: ATOM1 ATOM2 TYPE BOND ST NX TOPOL

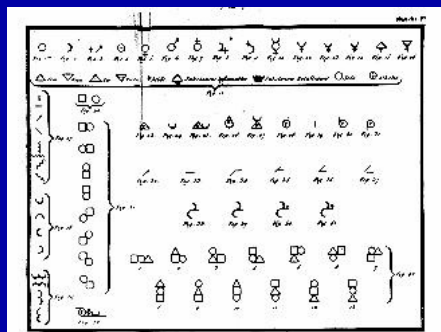
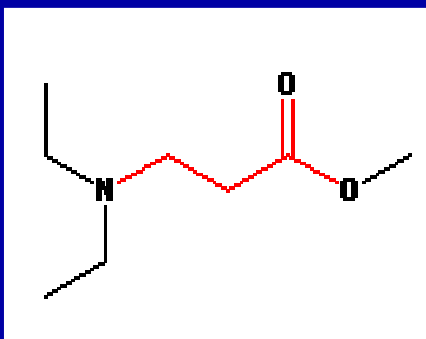
1 2 1 0 0 0 0

1 3 1 0 0 0 0

> <E_CAS>

5351-01-9

\$\$\$\$



I could ask the question –

How much have we moved on ?

Until recently, encoding chemistry has got stuck
(the ‘format’ I used in 1970 (.mol) is still used.....–
but that’s changing

Encoding chemistry

- Cheminformatics has concentrated on:
 - Re-writing chemical structures and data (from the literature) for computer storage and retrieval – usually based solely on structure
 - Construction of ‘single use’ databases – structure and searching capability are fixed
 - Centralised database resources updated infrequently
 - Annotation is fixed and not transferable
 - We deal with abstract connection tables – not properties of substances
- This approach has served us well, but with the capabilities of the internet, markup of data, metadata, pipelining of procedures (the ‘GRID’), multiple databases, high throughput data production, a huge increase in the literature...new opportunities emerge
- Appropriate data can be aggregated from multiple sources, checked, annotated further and analysed with more complex queries

XML and molecules

- XML is a computer language that allows 'metadata' to be stored along with the primary data
- Chemical Markup Language (CML) is being developed specifically for chemistry
 - Structure
 - Reactions
 - Calculations
 - Measurement
- In the future, much more information will be stored with molecules allowing greater re-use of data
 - E.g. not just melting point, but its range, units, method etc...
- see : Chemical Markup, XML and the World-Wide Web. Part I. Basic principles. P. Murray-Rust and H. S. Rzepa, J. Chem. Inf. Comp. Sci., 1999, 39, 928.

Some questions...

- Where do we find chemical data (data discovery)
 - literature, trusted sites, computation, in-house
 - Can we automate the process
- Is the data appropriate
 - are the units compatible, consistent among multiple sources, relevant to the questions asked
 - Can the data be audited
- Is it validated
 - Misprints, mistyped, missing, wrong, inconsistent
 - Can the data be checked
- Can we share it
 - No access, copyright, incomplete, hidden
 - Is data deposition open to a new paradigm ?
 - Will publishers share data ?

So, the first step, putting data into the computer
How useful is the data ?

- Can be wrong – so we can use -
 - computational methods to check it (best before it appears in a database – authoring tools)
- However, a more serious problem is that the data can simply be inappropriate to the problem
 - E.g. modelling and the real world don't match
- Start with some examples of data checking (authoring) and move on to data correction and data abstraction

Using information from Legacy data

- 100 years of chemistry – in books
- 20 years of data destruction – PDF
- How can it be read and checked? Increasingly can validate using ‘robots’
- However
 - Need data to be abstracted to a computer readable form
 - Need data to be standardised
 - Need data to be available to ‘robots’ for processing, checking, analysis – and to ‘talk’ to other robots

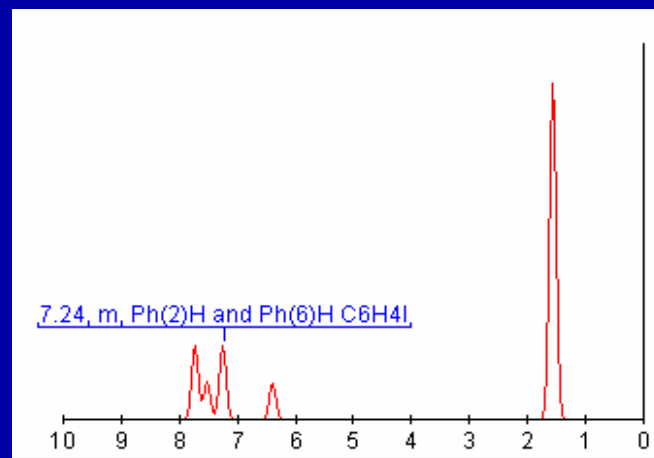
RSC/UCC markup project

- Uses CML, natural language processing, knowledge of chemistry
- An authoring tool
- A data checker
- Data abstraction from chemistry papers -> computer readable database

Experimental data checker: better information for organic chemists S. E. Adams, J. M. Goodman, R. J. Kidd, A. D. McNaught, P. Murray-Rust, F. R. Norton, J. A. Townsend and C. A. Waudby *Org. Biomol. Chem.* 2004, **2**, 3067-3070.

Highlighted in *Chemical Science* 2004, **1**, C33.

Chemical documents: machine understanding and automated information extraction J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman and P. Murray-Rust *Org. Biomol. Chem.* 2004, **2**, 3294-3200



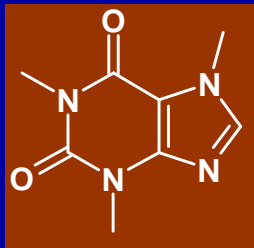
6a: 6,7b-Dihydro-3(2H)-oxepinone. Eluent for flash chromatography: pentane-Et₂O 9 : 1; colourless oil; yield 70%; HRMS calcd. for C₆H₈O₂: 112.0524. Found: 112.0524; max/cm⁻¹ (neat) 2932, 1655; H (250 MHz, CDCl₃) 6.59 (1 H, td, J 12.4, 4.4), 6.07 (1 H, td, J 12.4, 1.8), 4.33 (2 H, s), 3.96 (2 H, t, J 5.4), 2.71 (2 H, dq, J 5.4, 1.8); C (63 MHz, CDCl₃) 204.4, 145.2, 130.3, 78.9, 69.6, 35.2; m/z (EI) 112 (43, M⁺), 84 (83), 83 (43), 81 (54), 55 (20), 54 (100), 53 (41).

6b: 2,3,7,8-Tetrahydro-4H-oxocin-4-one. Eluent for flash chromatography: pentane-EtOAc 1 : 1; colourless oil; yield 73%; HRMS calcd. for C₇H₁₀O₂: 126.0680. Found: 126.0681; max/cm⁻¹ (neat) 2950, 1659; H (250 MHz, CDCl₃) 6.49 (1 H, td, J 12.1, 8.0), 6.23 (1 H, d, J 12.1), 3.91 (2 H, t, J 6.8), 3.69 (2 H, t, J 6.6), 2.90 (2 H, t, J 6.8), 2.76 (2H, q, J 6.6); C (63 MHz, CDCl₃) 201.0, 139.0, 136.0, 65.03, 64.98, 44.2, 29.4; m/z (EI) 126 (21, M⁺), 99 (29), 96 (14), 81 (11), 68 (100), 54 (50).

Once the data is in, automated calculation can aid database quality

Reliability of data ?

Caffein solubility



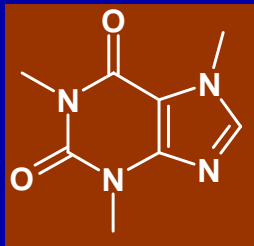
Temperature	Solubility g/l	Reference
25	2.132	[1]
25	896.2	[2]

[1] Oliveri-Mandala, E. (1926), *Gazzetta Chimica Italiana* 56, 896-901

[2] Ochsner, A. B., Belloto, R. J., and Sokoloski, T. D. (1985), *Journal of Pharmaceutical Sciences* 74, 132-135

Reliability of data ?

Caffeine solubility



Temperature	Solubility g/l	Reference
25	2.132	[1]
25	896.2	[2]

[1] Oliveri-Mandala, E. (1926), *Gazzetta Chimica Italiana* 56, 896-901

[2] Ochsner, A. B., Belloto, R. J., and Sokoloski, T. D. (1985), *Journal of Pharmaceutical Sciences* 74, 132-135

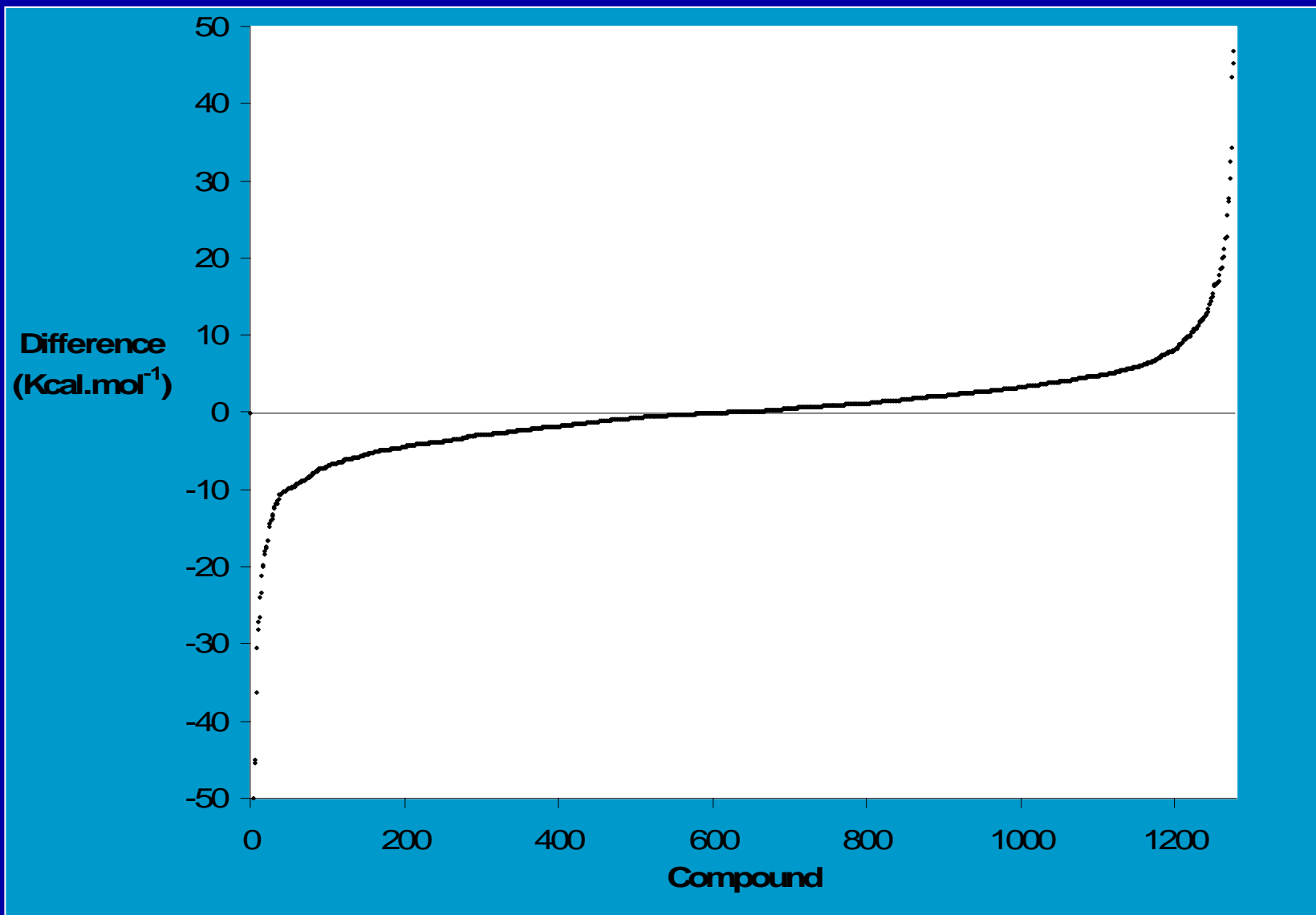
Conclusion : from 1926 to 1985, the solubility of caffeine increased at a rate of ca. 15 g/l/year

More seriously, there are ten different definitions of solubility, the value is the solubility of what form, under what conditions, with what solid form ? This could all be marked up.

Heats of formation Errors Detected by Calculation on the NIST Databook (from James JP Stewart)

Compound	Formula	Expt. ΔH_f	This Work		B88-LYP	
			Calc. ΔH_f	Diff	Calc. ΔH_f	Diff
Pentafluoriodobenzene	C ₆ F ₅ I	-131.1± 3.0	-176.6	-45.5	-178.1	-47.0
Perfluorobutadiene	C ₄ F ₆	-225.2	-253.4	-28.2	-241.6	-16.4
Bis-(n-perfluoropropyl ether)	C ₆ OF ₁₄	-742.1±0.8	-769.3	-27.2	-757.3	-15.2
Dodecafluorocyclohexane	C ₆ F ₁₂	-566.5± 2.0	-590.5	-24.0	-587.4	-20.9
Bromopentafluorobenzene	C ₆ F ₅ Br	-170.2±1.3	-191.3	-21.1	-191.8	-21.6
Perfluoroacetone	C ₃ OF ₆	-325.2	-342.6	-17.4	-337.7	-12.5
Hexafluorobenzene	C ₆ F ₆	-228.5±0.29	-242.5	-14.0	-241.6	-13.1
Thietane	C ₃ H ₆ S	14.6± 0.3	4.4	-10.2	8.0	-6.6
DL-3,4-di-1-cyclohexen-1-yl-2,2,5,5-tetramethyl hexane	C ₂₂ H ₃₈	-62.1±1.5	-50.4	11.7	-17.5	44.6
Dioxybismethanol	C ₂ H ₆ O ₄	-136.6±1.6	-124.5	12.1	-119.2	17.4
2,5,8-Trioxanonane	C ₆ H ₁₄ O ₃	-138.9±0.25	-124.6	14.3	-124.2	14.7
2,4,6-Trimethylphenyl isocyanide	C ₁₀ H ₁₁ N	40.0	56.6	16.6	48.8	8.8
6-(1,1-dimethylethyl)-2,3-dihydro-1,1-dimethyl-1H-Indene	C ₁₅ H ₂₂	-41.7	-24.9	16.7	-19.2	22.4
n-Perfluorobutane	C ₄ F ₁₀	-533.9	-515.3	18.6	-515.2	18.7
5,6-Dibutyl-5,6-bis(4-tert-butylphenyl)decane	C ₃₈ H ₆₂	-83.8±0.8	-58.6	25.2	-40.1	43.7

Distribution of Differences



Conclusions from this work:

- Most errors in original work is caused by arithmetic errors, typographic errors, transcription errors, and not by the experimental work.
 - Semiempirical methods can be used for predicting heats of formation of well-behaved systems.
 - About 4-5% of all reference thermochemical data are inaccurate by at least 2 kcal/mol.
 - Computational methods can, *and should*, be used for checking entries in thermochemical reference data compendia.
- 1 Stewart J. J. P., "Comparison of the Accuracy of Semiempirical and some DFT Methods for predicting Heats of Formation", *J. Mol. Modelling* 10, 6-12 (2004).

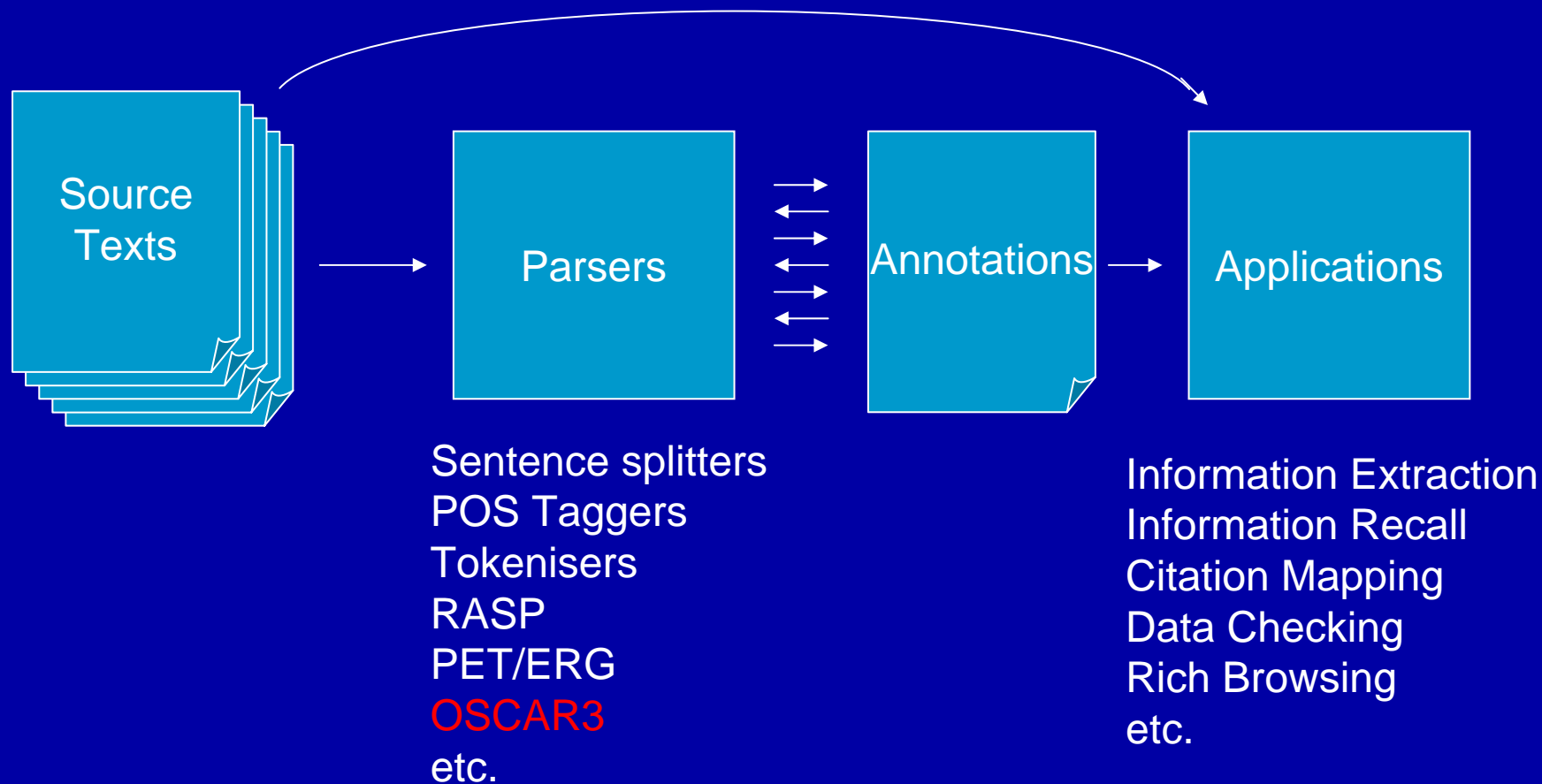
Given that Chemistry is 'all over the place' – how can we explore chemistry

- Cheminformatics discovery
 - Our current paradigm is to construct a database, extract relevant information when required, create relationships (e.g. SAR, trends in properties) and rebuild the database periodically
 - Another option is to dynamically update or acquire data interactively – data discovery – since much of the data is in the literature, this requires 'journal eating robots' to aggregate the data we require.
 - In paper-based publication the medium and message are inextricably linked and can normally only be processed by humans – this needs separated and marked up
 - For documents to be reliably machine-processable the paper image is not sufficient. It is necessary to identify the various components of a document, both in their intrinsic nature and their role in the document structure. This process is termed *markup* and has been adopted by many publishers through the standard generalised markup language (SGML)
 - -a lot of what we require is available internally to publishers

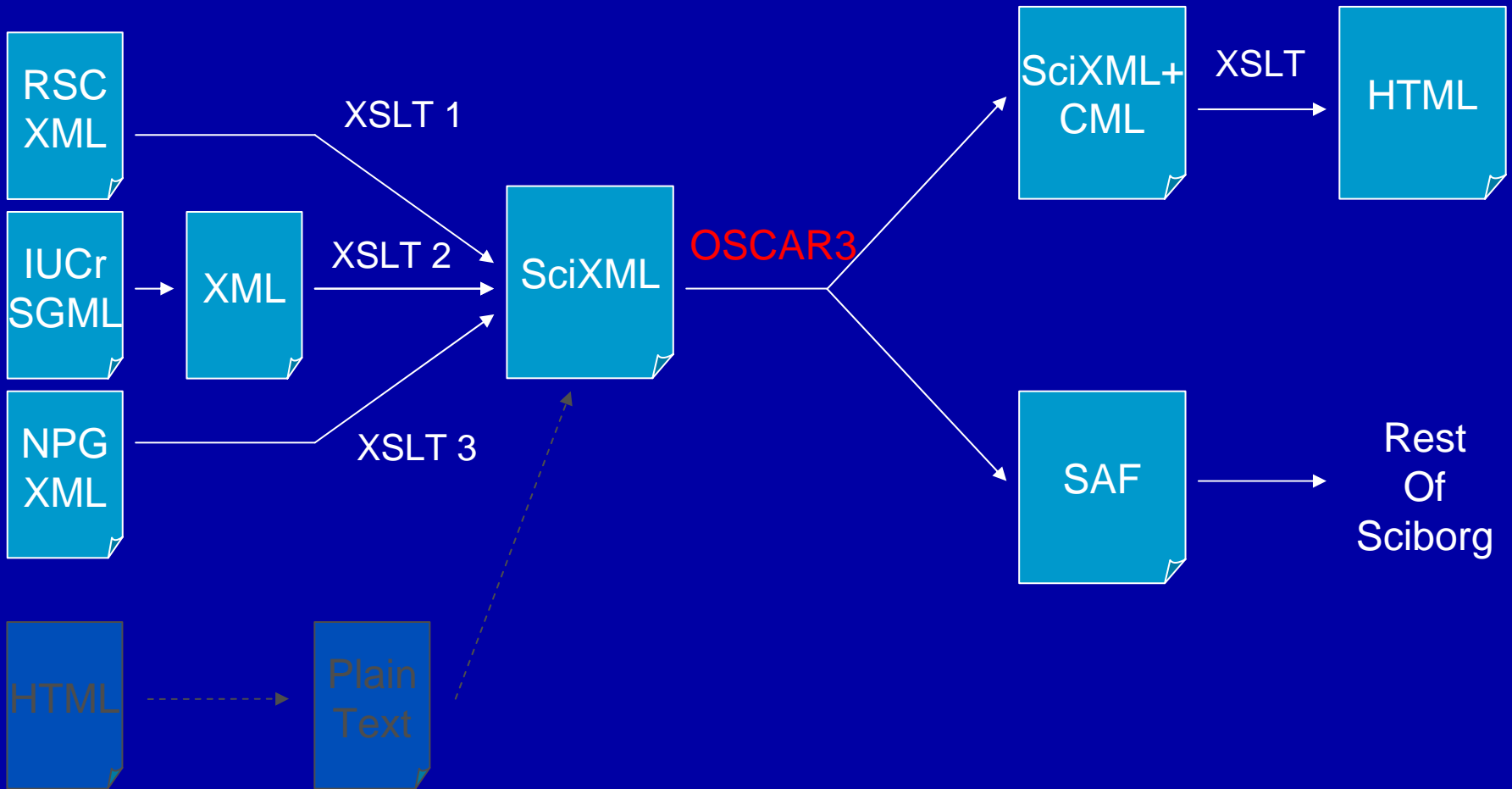
Encoding chemistry

- The representation of information in electronic form usually involves several layers: encoding, syntax, semantics, and ontology.
 - Creating a *generally agreed tag set*.
 - Adding *programmable functionality through (modular) code*.
 - Linking to *glossaries/ontologies/metadata*.
 - **Ontologies**. These represent formal descriptions of terms, concepts, behavior, etc.
 - Information loss between software, and ontological loss (or corruption e.g. bond order, consistently. In some systems bond orders are related to measured lengths, in others to calculated properties, and in others to electron counts.)
- Chemists are trained to recognize features in molecules and frequently to make mental translations between them. Common examples might be tautomers, delocalized systems, conformations, and representations of bond types (double, dative, bipolar). This is not possible completely in a computer program, but some limited progress is being made

Reading chemistry – development of enhanced annotation, recognition and natural language processing



OSCAR3 In Context



Name Recognition

- From lists
- Bayesian analysis:
 - Divide word into 4-letter fragments
 - “^ace” “acet” “ceto” “eton” “tone” “one\$”
 - Compare against smoothed fragment frequencies in English and Chemical training data
 - Factor in features from context, suffix etc.
 - Calculate a score

Name Recognition

- From lists
- Bayesian analysis
- Regular Expressions
 - E.g. Formulae
 - like: $((H|He|Li|B|Be|etc)\backslash d^*)^+$ but more complicated

Name Recognition

- From lists
- Bayesian analysis
- Regular Expressions

- Tokenisation + Chunking Rules
 - Hexane-ethyl acetate vs. tert-butyl peroxide

Name Resolution

- Attachment of structural information to names
- Output in SMILES, InChI, CML
- Vision: data to CML

The problem of language and name recognition...

- It was a cold January day in St. Louis (MO), USA
- I went into a shop (Famous Barr)
- I asked where the 'jumpers' were
- The assistant asked who it was for
- Me, I said

- My picture of a 'jumper'
(and Marks and Spencers
A large store in Great Britain)



**Cotton-Rich
Jumper**
from £7.00
to £14.00

- The assistants picture of a 'jumper'

The problem of language and name recognition...

- It was a cold January day in St. Louis (MO), USA
- I went into a shop (Famous Barr)
- I asked where the 'jumpers' were
- The assistant asked who it was for
- Me, I said

- My picture of a 'jumper'
(and Marks and Spencers)



Cotton-Rich
Jumper
from £7.00
to £14.00

- The assistants picture of a 'jumper'



Girls' @ Class Khaki Twill
Jumper with Side Pleats

Serious problem with a 'similar' language

Names for cocaine.....

Applications Actions Fri 24 Mar, 16:23

PubChem To HandBag - Mozilla Firefox

File Edit View Go Bookmarks Tools Help del.icio.us

http://localhost:8181/test/HandBagForm?url=http%3A%2F%2Fpubchem.ncbi.nlm.nih.gov%2Fsummary%2Fsummary Go

del.icio.us Connotea Bookmarklets OSCAR3 up JS Shell partial source PubChem to HandBag Search PubChem for... Search PubChem for...

PubChem To HandBag

InChI: InChI=1/C17H21NO4/c1-18-12-8-9-13(18)15(17(20)21-2)14(10-12)22-16(19)11-6-4-3-5-7-11/h3-7,12-15H,8-10H2,1-2H3/t12-,13+,14-,15+/m0/s1
SMILES: CN1C2CCC1C(C(C2)OC(=O)C3=CC=CC=C3)C(=O)OC

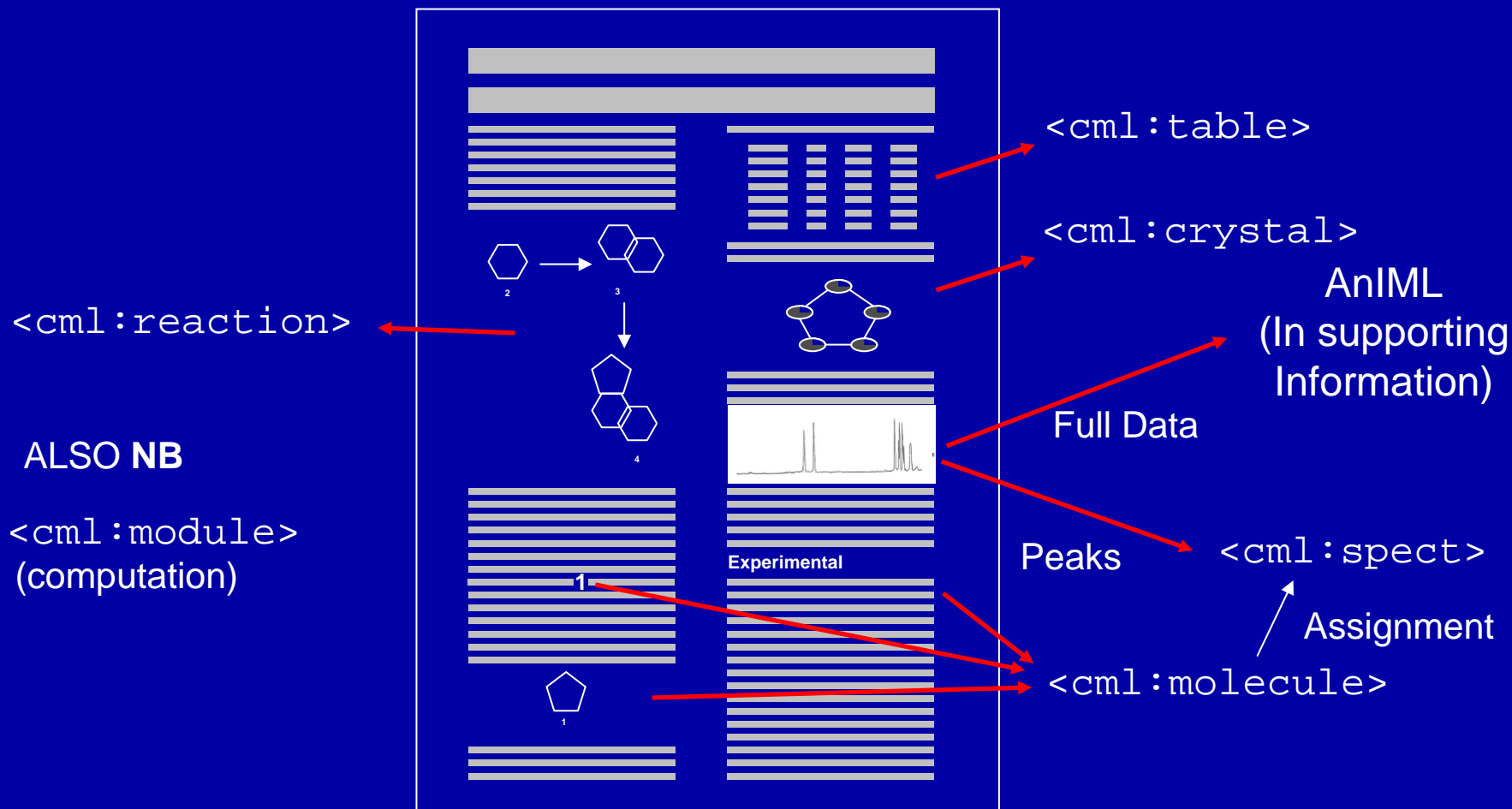
- Freeze
- Toot
- Star-spangled powder
- Dust
- Happy dust
- 1-alpha-H,5-alpha-H-Tropane-2-beta-carboxylic acid, 3-beta-hydroxy-, methyl ester, benzoate
- Line
- Lady
- Star dust
- Rock
- Bump
- Cecil
- Foo Foo
- Cocaina
- Benzoylmethylecgonine
- Methyl Benzoylecgonine
- Bouncing Powder
- Girl
- Charlie
- methyl (1R,2R,3R,5R)-8-methyl-3-(phenylcarbonyloxy)-8-azabicyclo[3.2.1]octane-2-carboxylate
- Ecgonine, methyl ester, benzoate (ester)
- 1-Cocaine
- Trails
- G-Rock
- 3-(Benzoyloxy)-8-methyl-8-azabicyclo-(3.2.1)octane-2-carboxylic acid methyl ether
- Green gold
- Kokain
- Moonrocks
- Pimp's drug
- Gold dust
- Kibbles n' Bits
- Cocaine
- Jam
- Blizzard
- Hell
- Kokayeen
- Coke
- 1-alpha-H,5-alpha-H-Tropane-2-beta-carboxylic acid, 3-beta-hydroxy-, methyl ester, benzoate (ester) (8CI)
- Candy

Done

A new entry to 9-azabicyclo[3.3.1]nonanes using radical PubChem To HandBag - Mozilla Firefox XMMS - 4. The Cure - From The Edge Of The Deep Gree ()

The result

CML



XSD Schema + further validation by JUMBO

Applications Actions Fri 24 Mar, 16:21
A new entry to 9-azabicyclo[3.3.1]nonanes using radical translocation/cyclisation reactions of 2-(but-3-ynyl)-1-(o-iodobenzoyl)piperidines - Mozilla Firefox
File Edit View Go Bookmarks Tools Help del.icio.us
file:///home/ptc24/output/oscar3/reduced_corpus/rsc/b203243k.html
del.icio.us Connotea Bookmarklets OSCAR3 up JS Shell partial source PubChem to HandBag Search PubChem for... Search PubChem for...
National Expre... BBC NEWS | M... Friends - The Chemistr... The PubChem ... file:///...577d.xml xml works in I.E... Using XML Dat... A new entry ...

was subjected to **desilylation**, ozonolysis, and subsequent 1,2-transposition of the resulting **carbonyl** group to give **9-benzoyl-1-methyl-9-azabicyclo[3.3.1]nonan-3-one**, a potential precursor for the synthesis of **(±)-euphococcinine**.

Introduction

Bridged **azabicyclic** rings are widely found as the basic structural elements in biologically active alkaloids such as **cocaine**, **atropine**, and **epibatidine**. Recently we have developed a new synthetic method for the **7-azabicyclo[2.2.1]heptane** and **8-azabicyclo[3.2.1]octane** ring systems **4** which involves treatment of **1-(o-iodobenzoyl)-2-(prop-2-ynyl)-pyrrolidines** and **-piperidines 1** with **tributyltin hydride (Bu₃SnH)** in the presence of **azoisobutyronitrile (AIBN)** in boiling **toluene**.¹ The formation of **4** can be formulated as proceeding *via* the **α-acylamino** radicals **3** which are generated by a **1,5-hydrogen** transfer (a radical translocation)² of the initially formed **aryl** radicals **2**. The radicals **3** then undergo a **5-exo-dig** cyclisation to **lead** to **4** (Scheme 1). **As** a further extension of this reaction, we have now investigated the Bu₃SnH-mediated radical reaction of the **2-(but-3-ynyl)piperidines 16a-c** and found that the translocation and **6-exo-dig** cyclisation reactions proceed in a regioselective manner to afford the expected **9-azabicyclo[3.3.1]nonane** ring system.³ **In** this paper we also describe a transformation of the cyclised product **17c** into **9-benzoyl-1-methyl-9-azabicyclo[3.3.1]nonan-3-one 28**, a protected form of **(±)-euphococcinine 29**.⁴

Results and discussion

The radical precursor **16a** was prepared starting from **methyl N-Boc-pipecolate 5**⁵ according to the procedures previously described for the synthesis of the **pyrrolidine** congener^{1d} (Scheme 2). Thus, **5** was subjected to **allylation**, and

Currently viewing

Search PubChem for cocaine

- **Compound(s) with structure**
 - **Compound(s), without structure**
 - **Group**
 - **Element**
 - **Reaction**

Applet viewer started

A new entry to 9-azabicyclo[3.3.1]nonanes using radical translocation/cyclisati XMMS - 4. The Cure - From The Edge Of The Deep Gree (7:44)

Document marked up, structure recognised and displayed
(like a spell checker, if the structure is not recognised, it can be added and in
future, recognised and displayed)

Applications Actions Fri 24 Mar, 16:27

Mozilla Firefox

File Edit View Go Bookmarks Tools Help del.icio.us

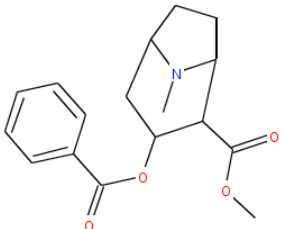
http://localhost:8181/test/OSCAR3.xml

del.icio.us Connotea Bookmarkslets OSCAR3 up JS Shell partial source PubChem to HandBag Search PubChem for... Search PubChem for...

National Ex... BBC NEWS ... Friends The Chem... The PubCh... file://...7d.xml xml works i... Using XML ... A new entr... http:...xml Index of file...

Bridged azabicyclic rings are widely found as the basic structural elements in biologically active alkaloids such as cocaine, atropine, and epibatidine. Recently we have developed a new synthetic method for the 7-azabicyclo[2.2.1]heptane and 8-azabicyclo[3.2.1]octane ring systems 4 which involves treatment of 1-(o-iodobenzoyl)-2-(prop-2-ynyl)-pyrrolidines and piperidines 1 with tributyltin hydride (Bu₃SnH) in the presence of azoisobutyronitrile (AIBN) in boiling toluene. The formation of 4 can be formulated as proceeding via the 945-acylamino radicals 3 which are generated by a 1,5-hydrogen transfer (a radical translocation) of the initially formed aryl radicals 2. The radicals 3 then undergo a 5-exo-dig cyclisation to lead to 4 (Scheme 1). As a further extension of this reaction, we have now investigated the Bu₃SnH-mediated radical reaction of the 2-(but-3-ynyl)piperidines 16a₃c and found that the translocation and 6-exo-dig cyclisation reactions proceed in a regioselective manner to afford the expected 9-azabicyclo[3.3.1]nonane ring system. In this paper we also describe a transformation of the cyclised product 17c into 9-benzoyl-1-methyl-9-azabicyclo[3.3.1]nonan-3-one 28, a protected form of (±)-euphoccinine 29.4

Currently viewing



Molecule weight: 303

Search PubChem by InChI

- Compound(s) with structure
- Compound(s), without structure
 - Group
 - Element
 - Reaction

Applet viewer started

Mozilla Firefox PubChem Substance - Mozilla Firefox XMMS - 4. The Cure - From The Edge Of Th

Example – extracting the latest information from Acta. Cryst

- Regularly updated journal with small molecule X-ray structures
- Articles contain associated structure files in supplemental data. These can be extracted.
- Data is annotated/marked up and associated with the article.
- Gives - Instant data access, RSS feeds, document summary, data checking, enhanced data searching (e.g. Google via InChI), additional computed properties.

Example structure from Acta Cryst. E.

(IUCr) Structure Reports Online Contents - Windows Internet Explorer

http://journals.iucr.org/e/issues/2006/06/00/issconts.html

Structure Reports Online

Journals Online
search
help
subscribe

back issues
previous issue
next issue
electronic archive

Acta Crystallographica Section E

Structure Reports Online

For publication in Volume 62, Part 6 (June 2006)

author index volume author index

- metal-organic compounds
- organic compounds

show schemes

metal-organic compounds

html pdf abstract cif 3d view structure factors checkCIF buy

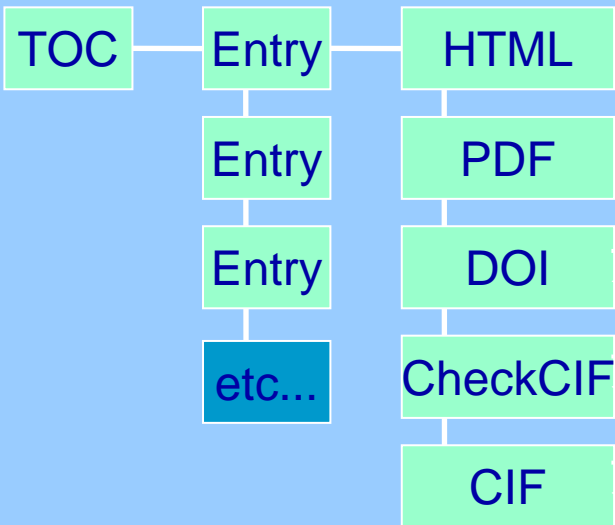
Acta Cryst. (2006). E62, m1181-m1182 [doi:10.1107/S1600536806014681]

catena-Poly[[[tetraaquacopper(II)]- μ -sulfato- κ^2 O:O']-[bis(malonamide- κ^2 O,O')copper(II)]- μ -sulfato- κ^2 O:O'] dihydrate]

start | obernai | Inbox - Outlook Express | Microsoft PowerPoint ... | (IUCr) Structure Rep... | Preview and Enhance... | 11:09

IUCr Website

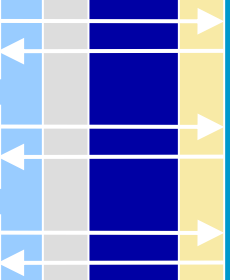
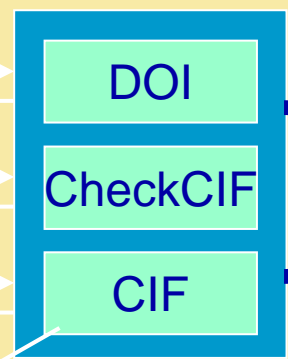
Acta Crystallographica



ORTEP

Cambridge - WWMM

Robot



CML

2D image

Summary

InChI Web Service

Example data capture from the latest literature

Fetch Documents

Access new edition
Of Acta Cryst-E

Download HTML from
Table of Contents

Parse into XML

Xpath query to get each
Separate entry

Each entry- extract each
Link to CIF and Check-CIF

Send out Robot to
Get text (CIF (in ASCII)
And HTML (CheckCif)

Process documents

Convert CIF to CML

Recalculate CheckCIF

Makes XML document from checkCIF

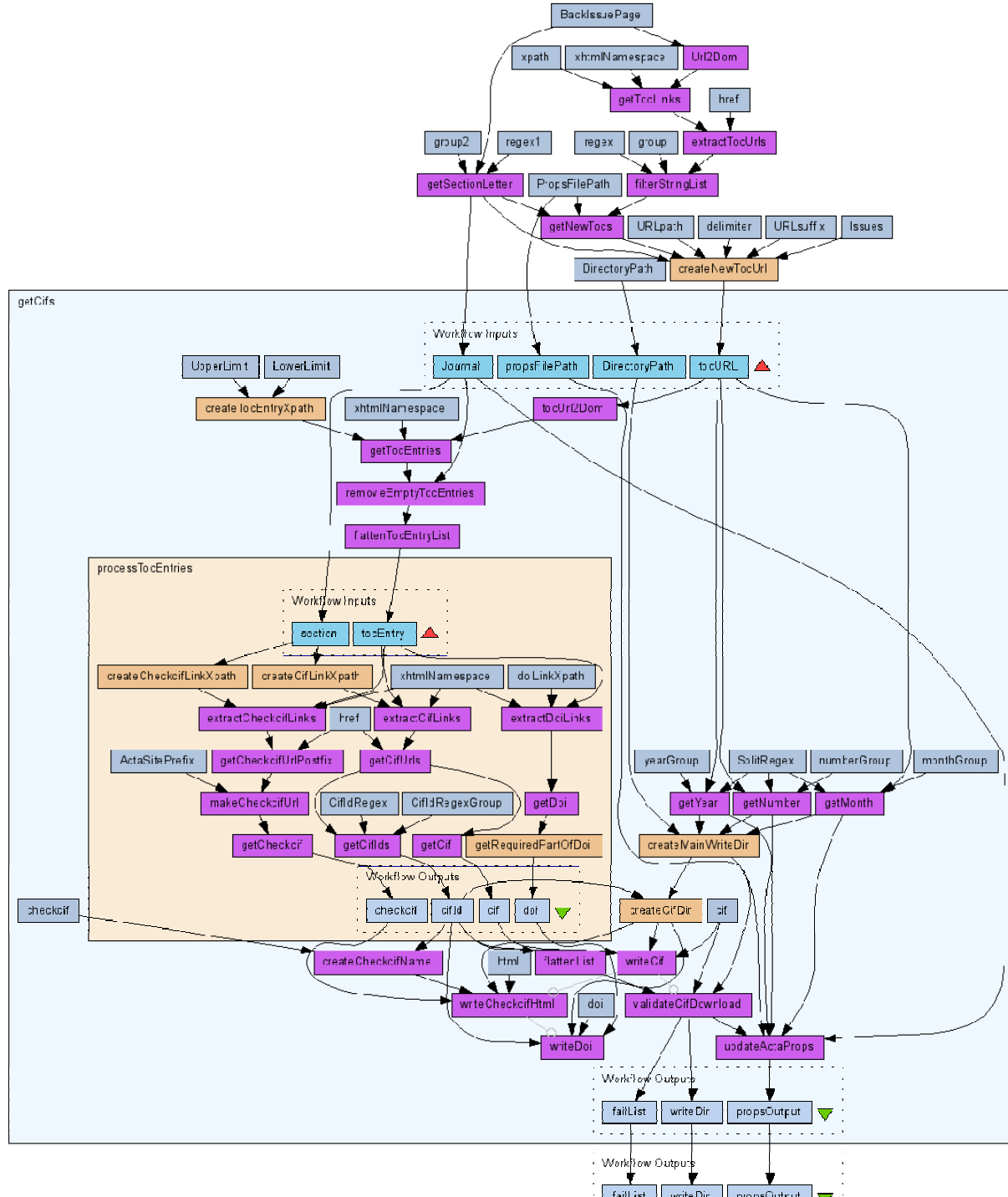
Calculate InChI

Add InChI/DOI/CheckCIF to CML

Get a CIF summary from CML

Create summary page

Using a workflow approach
- Using open source
components
e.g. Taverna, JMOLE,
checkCIF etc. and online
access to Acta Cryst E.



Workflow for discovering and extracting the latest structures

Scufl Workbench v1.3.1, built Fri Jan 27 17:00:56 GMT 2006

Tools and Workflow Invocation

Taverna Scufl Workbench v1.3

<http://taverna.sf.net>

Advanced model explorer

Workflow | Object properties

Load | Load from web | Save | New subworkflow | Offline | Reset

Workflow object	Retries	Delay	Back...	Thre...	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
writeDir					
Processors					
xpath : //x:a[@target="*_parer	0	0	1	1	<input type="checkbox"/>
regex1 : C:/Documents and Set	0	0	1	1	<input type="checkbox"/>
URLsuffix : isscontsbdy.html	0	0	1	1	<input type="checkbox"/>
DirectoryPath : C:/Documents :	0	0	1	1	<input type="checkbox"/>
createNewToaUrl	0	0	1	1	<input type="checkbox"/>
group2 : 1	0	0	1	1	<input type="checkbox"/>
group : 1	0	0	1	1	<input type="checkbox"/>

Workflow diagram

Save as | Configure diagram

Rendering done.

Available services

Search list | Watch loads

- Available Processors
 - Local Services
 - String Constant
 - Local Java widgets
 - conditional
 - Fail if true
 - Fail if false
 - list
 - Merge string list to string
 - Flatten l(l) to l()
 - Remove duplicate strings
 - Echo list
 - net
 - Get image URLs from HTTP docume
 - Send an email
 - Get web page from URL
 - Check a URL exists
 - Get image from URL
 - text
 - Byte[] to String
 - String list union
 - Concatenate two strings
 - String list difference
 - Filter list of strings by regex
 - Split string into string list by regula
 - Pad numeral with leading 0s
 - Filter list of strings extracting matc
 - String list intersection
 - metadata
 - Get internal LSID of input
 - base64
 - Encode byte[] to base64
 - Decode base64 to byte[]
 - XML
 - HTML2DOM
 - Get elements from DOM by XPath

start | 2 Outlook E... | 3 Windows ... | SCiPharm 20... | Microsoft Po... | C:\WINDOW... | Scufl Workbe... | Preview and ... | 14:49

Workflow for annotation and storing latest structures (running)

The screenshot displays the Taverna Scufil Workbench v1.3 interface. The main window shows a workflow graph with the following steps: CIF2CML, CML2JmolCML, CML2PNG, and createSummaryPage. The 'Enactor invocation' window is open, showing the workflow status as 'Running'. Below the status, there is a table of processor states.

Name	Last event	Event timestamp	Event detail	Breakpoint
CmlMime	ProcessComplete	10-May-2006 14:51:50		-
WriteDir	ProcessComplete	10-May-2006 14:51:50		-
CIF2CML	ProcessComplete	10-May-2006 14:51:53		-
SummaryDir	ProcessComplete	10-May-2006 14:51:50		-
CML2JmolCML	ProcessComplete	10-May-2006 14:51:54		-
CML2PNG	Invoking	10-May-2006 14:51:54		-
createSummaryPage	ProcessScheduled	10-May-2006 14:51:50		-

The workflow graph shows a vertical sequence of steps: CIF2CML, CML2JmolCML, CML2PNG, and createSummaryPage. The 'Enactor invocation' window also includes a 'Graph' tab with 'Intermediate inputs' and 'Intermediate outputs' sections, which are currently empty.

Results from the latest journal issue

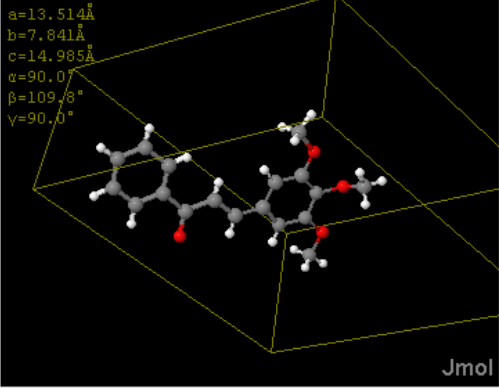
Untitled Document - Mozilla Firefox

file:///C:/Documents%20and%20Settings/rcg28/Desktop/actademo/summary/index.html

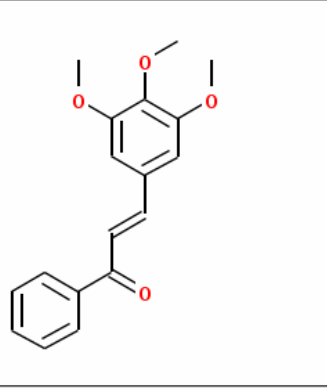
Acta Crystallographica - Summary

Published Formula	Article Link	Summary
<chem>C19H9N5O8</chem>	yes	yes
<chem>C16H10N2OS</chem>	yes	yes
<chem>C18H18O4</chem>	yes	yes
<chem>(C14H14NO+)(C7H7O3S-)</chem>	yes	yes
<chem>C15H11BrO</chem>	yes	yes
<chem>C11H12O3</chem>	yes	yes
<chem>C13H19NO3</chem>	yes	yes
<chem>C28H24N6O4</chem>	yes	yes

a=13.514Å
b=7.841Å
c=14.985Å
α=90.0°
β=109.8°
γ=90.0°



Jmol



Jmol script completed

start | 2 Outlook E... | 3 Windows ... | Microsoft Po... | C:\WINDOW... | Scuff Workbe... | Preview and ... | Untitled Doc... | 14:54

Summary for each molecule, including automatic CIF checking, adding InCHI

1-Phenyl-3-(3,4,5-trimethoxyphenyl)prop-2-en-1-one

Contact Author: Prof. Fun Hoeng Kun
e-mail: hfun@usm.my
DOI: 10.1107/S15205360060203461
Compound Class: organic
Date Recorded: 2006-01-19

Data collection parameters

Chemical formula sum	C ₁₈ H ₁₆ O ₄
Chemical formula moiety	C ₁₈ H ₁₆ O ₄
Crystal system	Monoclinic
Space group I-M	P 2 ₁ /c
Space group Hall	-P 2 ₁ /bc
Cell length a	13.6136
Cell length b	7.8414
Cell length c	14.9854
Cell angle alpha	90.0
Cell angle beta	109.823
Cell angle gamma	90.0
Data collection temperature	120.00(10)

Refinement results

R Factor (Obs)	0.0367
R Factor (All)	0.0403
Weighted R Factor (Obs)	0.0879
Weighted R Factor (All)	0.0907

Available Files

- ci2002_L_ellipsoid.jpeg
- ci2002_1_ID.png

Final Result

- ci2002.cif
- ci2002_1.cml.xml

Validation

- ci2002 checkcif.html

InChI=1C19HZ7O2K1-14(16-9-6-5-7-9-16)10-11-15-12-17(20-21)9(21-3)10(1-3-15)22-4h5-17,19H,1-4H3q+15



Can also exchange Information with PubChem

SID 841206 - PubChem Substance Summary - Mozilla Firefox

http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=841206

NCBI PubChem National Library of Medicine NLM

Substance Summary:

Compound Displayed: PubChem

SID: 841206
CID: 446220
PubMed: 4 Links
Related Substances: 0 Links
Same, Connectivity: 26 Links
Same, Isotopes: 32 Links
Similar Substances: 134 Links
Structure Search

Source: BIND (349)

Medical Subject Annotations: (Total 2) Display: Next 1 | All

Cocaine
An alkaloid ester extracted from the leaves of plants including coca. It is a local anesthetic and vasoconstrictor and is clinically used for that purpose, particularly in the eye, nose, and throat. It also has powerful central nervous system effects similar to the amphetamines and is a drug of abuse. Cocaine, like amphetamines, acts by multiple mechanisms on brain catecholaminergic neurons; the mechanism of its reinforcing effects is thought to involve inhibition of dopamine uptake.

The kind of problem that could be addressed with this approach...

- A good example is newsfeeds – news is highly time dependant
 - What are the latest molecules active in my area of science ?
 - What patents have been filed recently ?
 - Is there information not yet in the abstracts ?
 - Have different groups obtained the same answer (e.g. solubility of a compound)
 - Provide a summary that is up to the minute
 - Also, Robots can do data checking/annotation/add value to information before we receive it
 - Natural language processing can refine or expand our search

Conclusion

Chemistry has a language that is in some ways quite formal and systematic. However, like most languages, there is a degree of 'artistic licence' and converting this into a computer representation is challenging, but allows a new way of doing science, knowledge discovery. In earlier years, we have been to some extent limited in that we have applied the standard scientific method of hypothesis and experiment followed by acceptance or revision of the hypothesis (QSAR is an excellent example). There are of course many examples of successes in drug discovery or property prediction using this approach. However, with large amounts of data, the internet and cheminformatics tools, science can be performed in a new way in which hypotheses are generated from relationships between the data, sometimes revealing unexpected results. Indeed, serendipity could potentially be systematised. Although we are not there yet, I believe that this is where we are headed. Recent applications are designed to both improve the quality of data, and also to discover new relationships between cheminformatics data.

Acknowledgements

- Previous OSCAR Authors
 - Chris Waudby, Sam Adams, Joe Townsend, Jonathan Goodman, Peter Murray-Rust, Peter Corbett
 - Support from DTI, RSC & NPG
- James Stewart