The Development of the IUPAC InChI Chemical Structure Standard

Stephen Heller

NIST

&

InChI-Trust Project Director

steve@inchi-trust.org

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

http://www.inchi-trust.org **5/16/2016**

Slides are available at http://www.hellers.com/steve/wiley-5-16.pdf

InChI TRUST

This is a green talk –

These slides were made from 100% recycled electrons

InChI Project Goal

To link everything about a chemical from many sources with the purpose of creating new information.

Wiley has both scientific/chemical journals and chemical databases. Before InChI Wiley (as well as other publishers of both forms of information and data) were unable to connect and link the chemicals found in all these resources

What is InChI?

The IUPAC International Chemical Identifier, or InChI, is a non-proprietary, machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner.

InChIs are produced by computer from structures drawn on-screen with existing structure drawing software, and the original structure can be regenerated from an InChI with existing structure drawing software.

InChI is really just a synonym.

http://en.wikipedia.org/wiki/International_Chemical_Identifier

Unique InChI Features

Only IUPAC International structure standard

Only Open Source structure standard

Only structure standard support by a wide majority of publishers, database producers, and chemistry software companies

InChI Videos

1. What on Earth is InChI?

http://www.youtube.com/watch?v=rAnJ5toz26c

2. The Birth of the InChI

http://www.youtube.com/watch?v=X9c0PHXPfso

3. The Googlable InChIKey

http://www.youtube.com/watch?v=UxSNOtv8Rjw

4. InChI and the Islands

http://www.youtube.com/watch?v=qrCqJ0o4jGs

The InChI Team

(alphabetical order)

Stephen R. Heller
Alan McNaught
Igor Pletnev
Stephen E. Stein
Dmitrii Tchekhovskoi

Four Requirements for a Computer Representation Standard

Need

Definition/Specification

Timing/Infrastructure

Acceptance/Use

Need

There was no open source (freely available) standard method to “name” a chemical structure. That is, a method to give a structure an electronic signature – an identifier.

Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be found and LINKED to and combined with other content on the Internet. InChI provides an excellent ROI (return on investment). InChI increases productivity!

Date: Mon, 15 Nov 1999 18:48:30 -0500 (EST)
From: Stephen R. Heller<srheller@cliff.nal.usda.gov>
To: stein <sstein@enh.nist.gov>
Subject: Re: A strawman proposal

Steve-

First rough draft. Let's talk tomorrow about it.

Steve

11/15/99

An IUPAC Chemical Registry System

In response to the upcoming March 2000 IUPAC meeting - Representations of Molecular Structure: Nomenclature and its Alternatives - I would like to propose the creation of an IUPAC public domain chemical registry system.
…

Why InChI? - Too Many Good and Excellent Identifiers ("Standards")

Structure diagrams
- various conventions
- contain 'too much' information

Connection Tables/Notations
- MolFiles, SDF, SMILES, SLN, ROSDAL, …

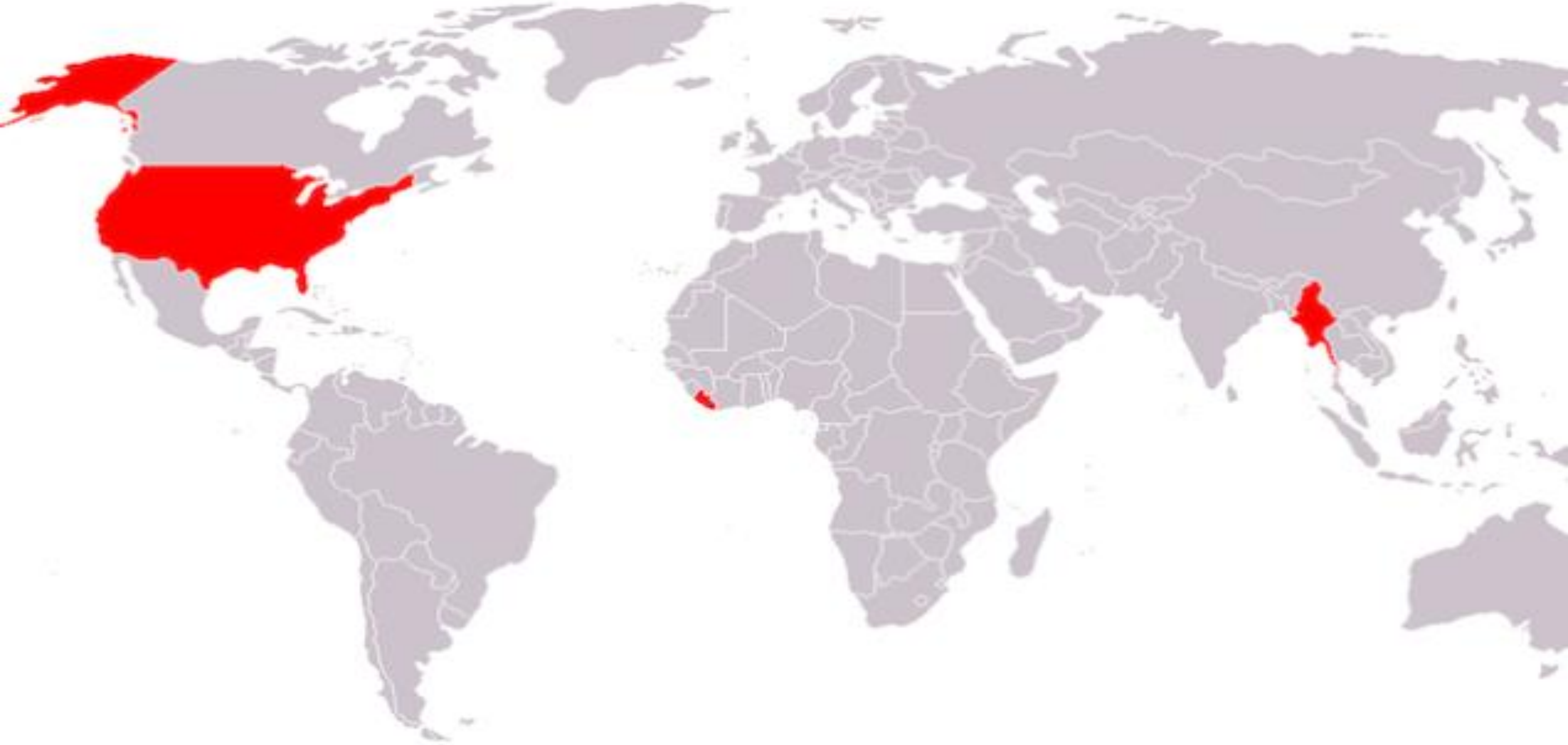
Pronounceable names (and mostly unpronounceable) and mostly complex names
- IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN, ASK, ISO

(Dumb) Index Numbers
EINECS, ELINCS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN, EC, ChemSpider ID, REACH, PubChem CID, BAN, NSC, ASK, KEGG, BP, IND, MARTINDALE, MESH, IT IS, RX-CUI, NDF-RT, ATC, AHPA, USP/NF, UNII, MFCD#, and so on

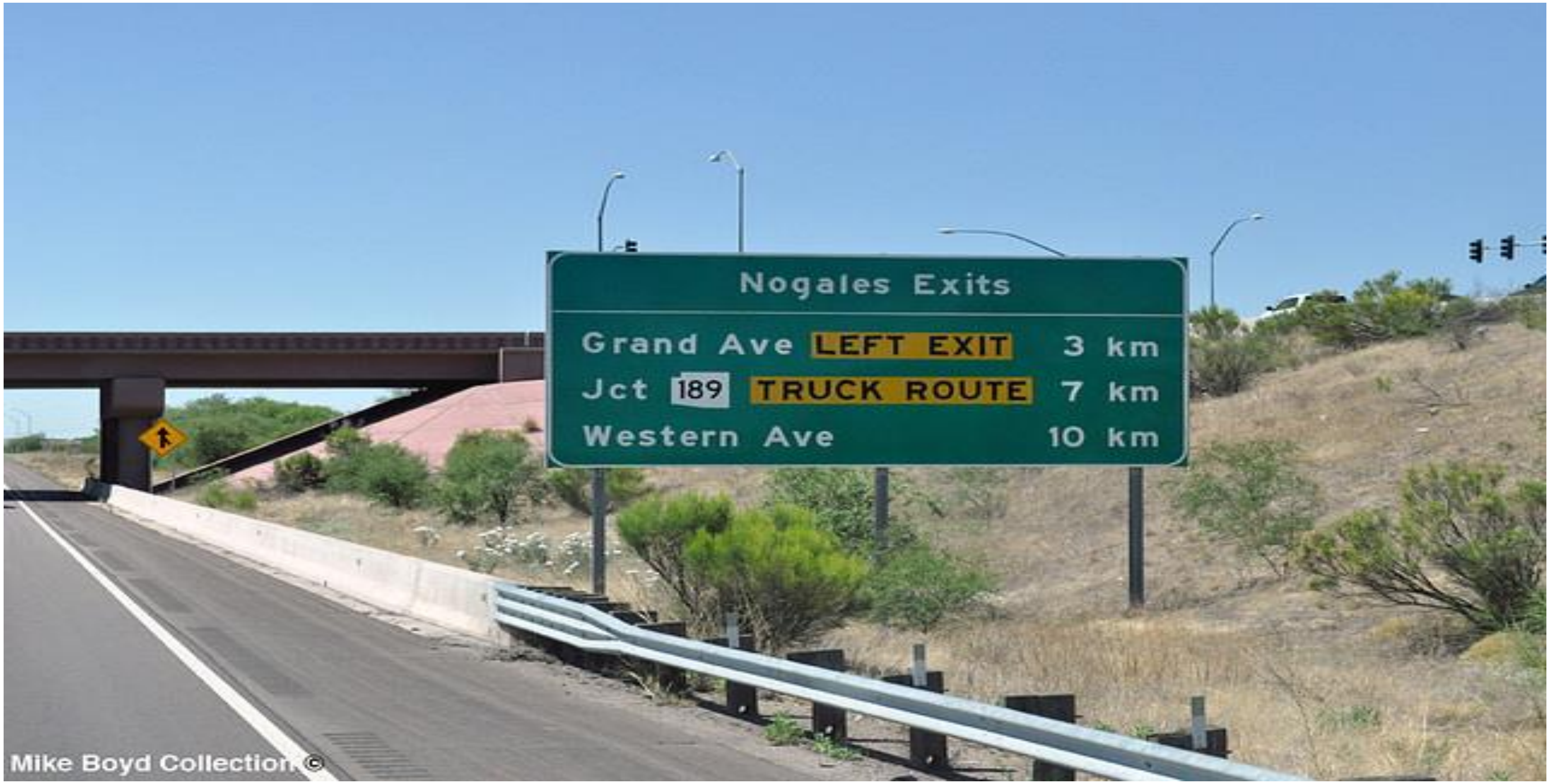
“Standards are like toothbrushes – everyone has one but no one wants to use someone else's.”

Phil Bourne,
Associate Director for Data Science, NIH

3 countries not using the Metric System



Only One US Highway uses the Metric System



Mike Boyd Collection ©

Definition/Specification

A computer algorithm to insure consistency and reproducibility.

What "*is*" the InChI standard?

The InChI standard programmed into the algorithm is an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane). In cases of more complex molecules where there is not agreement among chemists, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the SAME result using it – which is what a standard is!

InChI Characteristics

1. Easy to generate

2. Expressive (it will contain structural information)

3. Unambiguous/Unique

4. Does not require a centralized operation (it can be generated anywhere – can use crowdsourcing/free labor)

5. Easy to search for structure via Internet search engines (Google, Yahoo, Bing, etc.) using the InChI (hash) Key.

InChI is for computers

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, and InChI QR codes - InChIs are not designed to be read by humans.

Or, put another way – never send a human to do a machine's job!

Technology is at its best when it is invisible.

How difficult is it to create an InChI?

Today, all the major structure drawing programs (ChemDraw, MDL/Symyx /Accelrys/BIOVIA Draw, ISIS Draw, ChemAxon Marvin Sketch, ACD Labs ChemSketch, CLiDE, Jmol, and so on) have incorporated the InChI algorithm in their products, with usually an “InChI” button for generating the InChI.

An Open Source system keeps us on our toes. If things don't work or we don't respond as needed InChI won't remain a standard.

InChI is the worst computer readable structure representation except for all those other forms that have been tried from time to time.

With apologies to Sir Winston Churchill (House of Commons speech on November 11, 1947)

Timing &Infrastructure

InChI has become a standard only because of the world has changed in the last 20 years.

Without the Internet, without vast amounts of data and information becoming available in computer readable form for the first time, without Google (and other search engines), without structure drawing programs, and with most chemistry publishers now needing chemical structures in their products, InChI would be yet another interesting graph theory project that died like so many before it.

Without this perfect good storm that created a foundation for InChI, at best, I would be talking to a group a 5-7 people at ACS meeting talk.

Timing &Infrastructure

We got lucky. We had the perfect “good” storm of things happening.

And we had the right people who knew IUAPC and how it worked (or didn’t work).

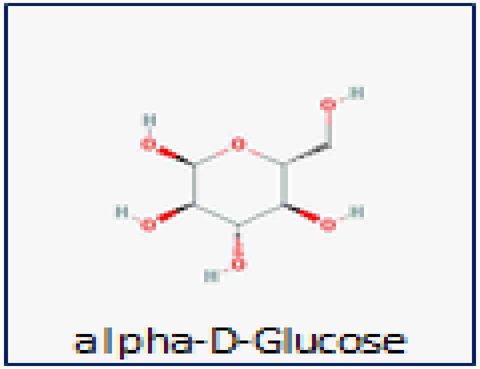
How did InChI succeed?

This project was the perfect “good” storm. The project came about in 1999 when Steve Heller retired and his wife threatened him with divorce unless he found some to do. (Yes, behind every successful project is a woman.) IUPAC discovered that nomenclature was for 20th, not 21st century. NIST, the US standards agency, needed a way to represent and link the structures from its standard property databases. The Internet (web 2.0) was taking off enabling silos and islands of information to be linked and searched if only there was a linking element. Publishers and database producers realized their information would be more valuable (i.e., they could sell more to more people) if only there was a way to link chemical structures from all the diverse resources on the Internet. With no funds to support the project, IUPAC needed the private sector to pay for the short and long term project needs. Lastly, the decentralized structure and hands-off management of the project enabled all the expert egos to be satisfied by putting everyone in charge of what they do best and giving them the final say - allowing for proper, scientific, bottom-up decisions.

Acceptance/Use

Easier said than done

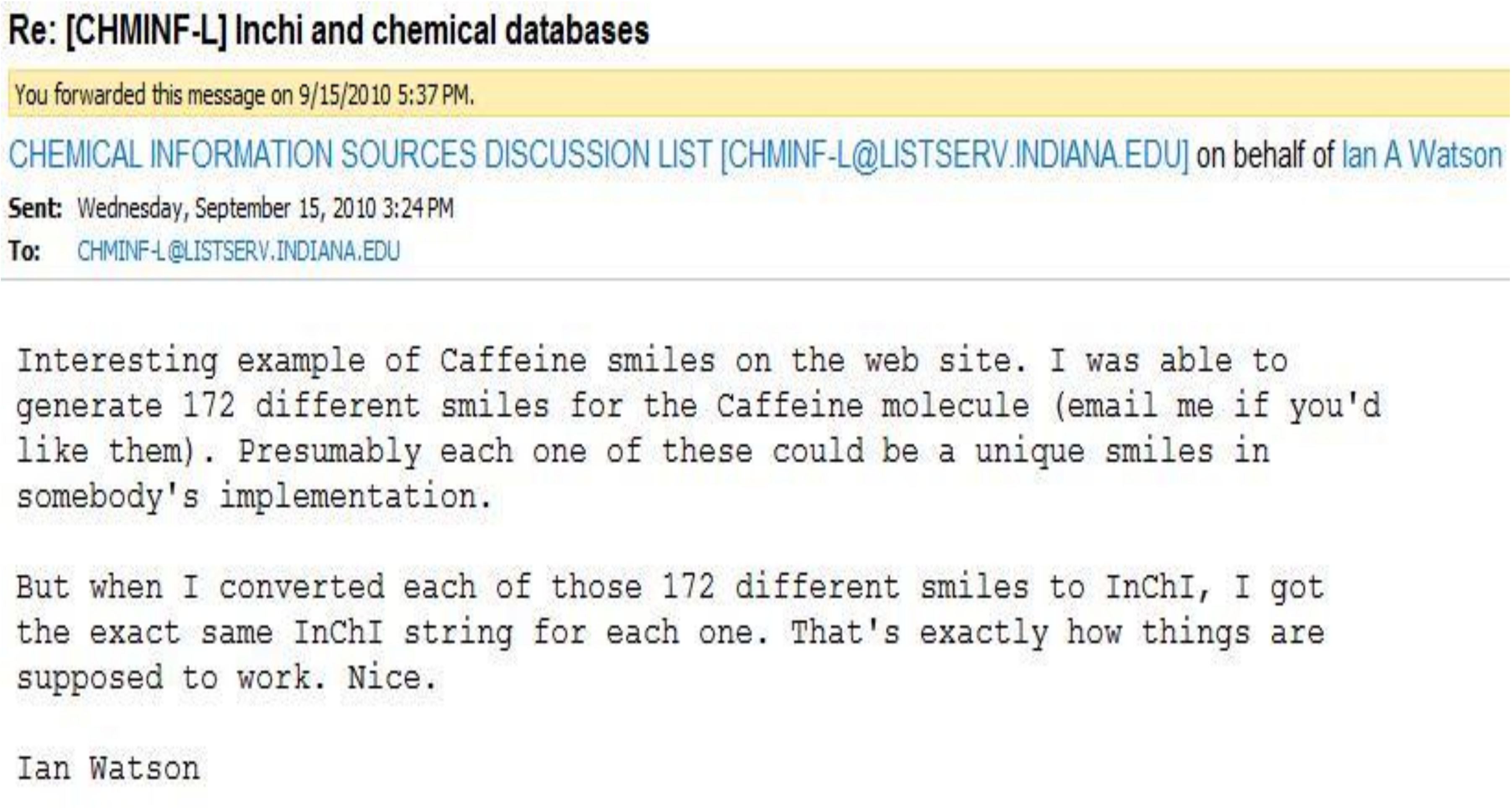
What about SMILES as a standard?


alpha-D-Glucose

C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O)O)O)O)O

- SMILES is a popular line notation
  - But not a published standard
- Every vendor has its own implementation
  - Differences in aromaticity models can lead to structure corruption
- Cannot reliably compare strings
  - Different software packages can make different strings for same structure
- No structure normalization
  - Different structural representations can yield different strings

Slide from Evan Bolton – NIH/PubChem



Re: [CHMINF-L] Inchi and chemical databases

You forwarded this message on 9/15/2010 5:37 PM.

CHEMICAL INFORMATION SOURCES DISCUSSION LIST [CHMINF-L@LISTSERV.INDIANA.EDU] on behalf of Ian A Watson

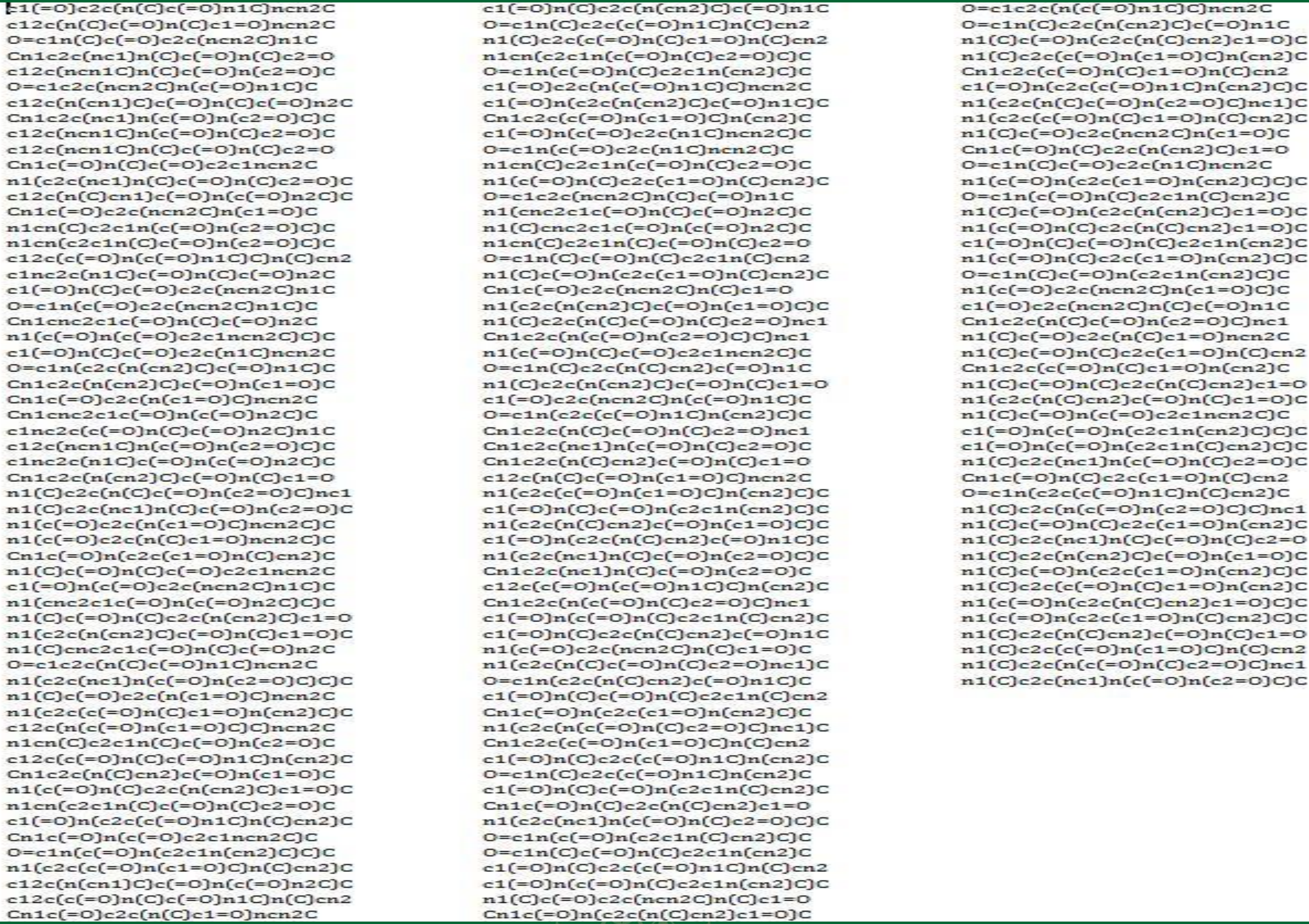
Sent: Wednesday, September 15, 2010 3:24 PM

To: CHMINF-L@LISTSERV.INDIANA.EDU

```
Interesting example of Caffeine smiles on the web site. I was able to
generate 172 different smiles for the Caffeine molecule (email me if you'd
like them). Presumably each one of these could be a unique smiles in
somebody's implementation.

But when I converted each of those 172 different smiles to InChI, I got
the exact same InChI string for each one. That's exactly how things are
supposed to work. Nice.

Ian Watson
```



```
c1(=O)c2c(n(C)c(=O)n1C)ncn2C
c12c(n(C)c(=O)n(C)c1=O)ncn2C
O=c1n(C)c(=O)c2c(ncn2C)n1C
Cn1c2c(nc1)n(C)c(=O)n(C)c2=O
c12c(ncn1C)n(C)c(=O)n(c2=O)C
O=c1c2c(ncn2C)n(c(=O)n1C)C
c12c(n(cn1)C)c(=O)n(C)c(=O)n2C
Cn1c2c(nc1)n(c(=O)n(c2=O)C)C
c12c(ncn1C)n(c(=O)n(C)c2=O)C
c12c(ncn1C)n(C)c(=O)n(C)c2=O
Cn1c(=O)n(C)c(=O)c2c1ncn2C
n1(c2c(nc1)n(C)c(=O)n(C)c2=O)C
c12c(n(C)cn1)c(=O)n(c(=O)n2C)C
Cn1c(=O)c2c(ncn2C)n(c1=O)C
n1cn(C)c2c1n(c(=O)n(c2=O)C)C
n1cn(c2c1n(C)c(=O)n(c2=O)C)C
c12c(c(=O)n(c(=O)n1C)C)n(C)cn2
c1nc2c(n1C)c(=O)n(C)c(=O)n2C
c1(=O)n(C)c(=O)c2c(ncn2C)n1C
O=c1n(c(=O)c2c(ncn2C)n1C)C
Cn1cnc2c1c(=O)n(C)c(=O)n2C
n1(c(=O)n(c(=O)c2c1ncn2C)C)C
c1(=O)n(C)c(=O)c2c(n1C)ncn2C
O=c1n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(n(cn2)C)c(=O)n(c1=O)C
Cn1c(=O)c2c(n(c1=O)C)ncn2C
Cn1cnc2c1c(=O)n(c(=O)n2C)C
c1nc2c(c(=O)n(C)c(=O)n2C)n1C
c12c(ncn1C)n(c(=O)n(c2=O)C)C
c1nc2c(n1C)c(=O)n(c(=O)n2C)C
Cn1c2c(n(cn2)C)c(=O)n(C)c1=O
n1(C)c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c2c(nc1)n(C)c(=O)n(c2=O)C
n1(c(=O)c2c(n(c1=O)C)ncn2C)C
n1(c(=O)c2c(n(C)c1=O)ncn2C)C
Cn1c(=O)n(c2c(c1=O)n(C)cn2)C
n1(C)c(=O)n(C)c(=O)c2c1ncn2C
c1(=O)n(c(=O)c2c(ncn2C)n1C)C
n1(cnc2c1c(=O)n(c(=O)n2C)C)C
n1(C)c(=O)n(C)c2c(n(cn2)C)c1=O
n1(c2c(n(cn2)C)c(=O)n(C)c1=O)C
n1(C)cnc2c1c(=O)n(C)c(=O)n2C
O=c1c2c(n(C)c(=O)n1C)ncn2C
n1(c2c(nc1)n(c(=O)n(c2=O)C)C)C
n1(C)c(=O)c2c(n(c1=O)C)ncn2C
n1(c2c(c(=O)n(C)c1=O)n(cn2)C)C
c12c(n(c(=O)n(c1=O)C)C)ncn2C
n1cn(C)c2c1n(C)c(=O)n(c2=O)C
c12c(c(=O)n(C)c(=O)n1C)n(cn2)C
Cn1c2c(n(C)cn2)c(=O)n(c1=O)C
n1(c(=O)n(C)c2c(n(cn2)C)c1=O)C
n1cn(c2c1n(C)c(=O)n(C)c2=O)C
c1(=O)n(c2c(c(=O)n1C)n(C)cn2)C
Cn1c(=O)n(c(=O)c2c1ncn2C)C
O=c1n(c(=O)n(c2c1n(cn2)C)C)C
n1(c2c(c(=O)n(c1=O)C)n(C)cn2)C
c12c(n(cn1)C)c(=O)n(c(=O)n2C)C
c12c(c(=O)n(C)c(=O)n1C)n(C)cn2
Cn1c(=O)c2c(n(C)c1=O)ncn2C
c1(=O)n(C)c2c(n(cn2)C)c(=O)n1C
O=c1n(C)c2c(c(=O)n1C)n(C)cn2
n1(C)c2c(c(=O)n(C)c1=O)n(C)cn2
n1cn(c2c1n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(C)c2c1n(cn2)C)C
c1(=O)c2c(n(c(=O)n1C)C)ncn2C
c1(=O)n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(c(=O)n(c1=O)C)n(cn2)C
c1(=O)n(c(=O)c2c(n1C)ncn2C)C
O=c1n(c(=O)c2c(n1C)ncn2C)C
n1cn(C)c2c1n(c(=O)n(C)c2=O)C
n1(c(=O)n(C)c2c(c1=O)n(C)cn2)C
O=c1c2c(ncn2C)n(C)c(=O)n1C
n1(cnc2c1c(=O)n(C)c(=O)n2C)C
n1(C)cnc2c1c(=O)n(c(=O)n2C)C
n1cn(C)c2c1n(C)c(=O)n(C)c2=O
O=c1n(C)c(=O)n(C)c2c1n(C)cn2
n1(C)c(=O)n(c2c(c1=O)n(C)cn2)C
Cn1c(=O)c2c(ncn2C)n(C)c1=O
n1(c2c(n(cn2)C)c(=O)n(c1=O)C)C
n1(C)c2c(n(C)c(=O)n(C)c2=O)nc1
Cn1c2c(n(c(=O)n(c2=O)C)C)nc1
n1(c(=O)n(C)c(=O)c2c1ncn2C)C
O=c1n(C)c2c(n(C)cn2)c(=O)n1C
n1(C)c2c(n(cn2)C)c(=O)n(C)c1=O
c1(=O)c2c(ncn2C)n(c(=O)n1C)C
O=c1n(c2c(c(=O)n1C)n(cn2)C)C
Cn1c2c(n(C)c(=O)n(C)c2=O)nc1
Cn1c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c2c(n(C)cn2)c(=O)n(C)c1=O
c12c(n(C)c(=O)n(c1=O)C)ncn2C
n1(c2c(c(=O)n(c1=O)C)n(cn2)C)C
c1(=O)n(C)c(=O)n(c2c1n(cn2)C)C
n1(c2c(n(C)cn2)c(=O)n(c1=O)C)C
c1(=O)n(c2c(n(C)cn2)c(=O)n1C)C
n1(c2c(nc1)n(C)c(=O)n(c2=O)C)C
Cn1c2c(nc1)n(C)c(=O)n(c2=O)C
c12c(c(=O)n(c(=O)n1C)C)n(cn2)C
Cn1c2c(n(c(=O)n(C)c2=O)C)nc1
c1(=O)n(c(=O)n(C)c2c1n(C)cn2)C
c1(=O)n(C)c2c(n(C)cn2)c(=O)n1C
n1(c(=O)c2c(ncn2C)n(C)c1=O)C
n1(c2c(n(C)c(=O)n(C)c2=O)nc1)C
O=c1n(c2c(n(C)cn2)c(=O)n1C)C
c1(=O)n(C)c(=O)n(C)c2c1n(C)cn2
Cn1c(=O)n(c2c(c1=O)n(cn2)C)C
n1(c2c(n(c(=O)n(C)c2=O)C)nc1)C
Cn1c2c(c(=O)n(c1=O)C)n(C)cn2
c1(=O)n(C)c2c(c(=O)n1C)n(cn2)C
O=c1n(C)c2c(c(=O)n1C)n(cn2)C
c1(=O)n(C)c(=O)n(c2c1n(C)cn2)C
Cn1c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(nc1)n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(c2c1n(C)cn2)C)C
O=c1n(C)c(=O)n(C)c2c1n(cn2)C
c1(=O)n(C)c2c(c(=O)n1C)n(C)cn2
c1(=O)n(c(=O)n(C)c2c1n(cn2)C)C
n1(C)c(=O)c2c(ncn2C)n(C)c1=O
Cn1c(=O)n(c2c(n(C)cn2)c1=O)C
O=c1c2c(n(c(=O)n1C)C)ncn2C
O=c1n(C)c2c(n(cn2)C)c(=O)n1C
n1(C)c(=O)n(c2c(n(C)cn2)c1=O)C
n1(C)c2c(c(=O)n(c1=O)C)n(cn2)C
Cn1c2c(c(=O)n(C)c1=O)n(C)cn2
c1(=O)n(c2c(c(=O)n1C)n(cn2)C)C
n1(c2c(n(C)c(=O)n(c2=O)C)nc1)C
n1(c2c(c(=O)n(C)c1=O)n(C)cn2)C
n1(C)c(=O)c2c(ncn2C)n(c1=O)C
Cn1c(=O)n(C)c2c(n(cn2)C)c1=O
O=c1n(C)c(=O)c2c(n1C)ncn2C
n1(c(=O)n(c2c(c1=O)n(cn2)C)C)C
O=c1n(c(=O)n(C)c2c1n(C)cn2)C
n1(C)c(=O)n(c2c(n(cn2)C)c1=O)C
n1(c(=O)n(C)c2c(n(C)cn2)c1=O)C
c1(=O)n(C)c(=O)n(C)c2c1n(cn2)C
n1(c(=O)n(C)c2c(c1=O)n(cn2)C)C
O=c1n(C)c(=O)n(c2c1n(cn2)C)C
n1(c(=O)c2c(ncn2C)n(c1=O)C)C
c1(=O)c2c(ncn2C)n(C)c(=O)n1C
Cn1c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c(=O)c2c(n(C)c1=O)ncn2C
n1(C)c(=O)n(C)c2c(c1=O)n(C)cn2
Cn1c2c(c(=O)n(C)c1=O)n(cn2)C
n1(C)c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(n(C)cn2)c(=O)n(C)c1=O)C
n1(C)c(=O)n(c(=O)c2c1ncn2C)C
c1(=O)n(c(=O)n(c2c1n(cn2)C)C)C
c1(=O)n(c(=O)n(c2c1n(C)cn2)C)C
n1(C)c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c(=O)n(C)c2c(c1=O)n(C)cn2
O=c1n(c2c(c(=O)n1C)n(C)cn2)C
n1(C)c2c(n(c(=O)n(c2=O)C)C)nc1
n1(C)c(=O)n(C)c2c(c1=O)n(cn2)C
n1(C)c2c(nc1)n(C)c(=O)n(C)c2=O
n1(C)c2c(n(cn2)C)c(=O)n(c1=O)C
n1(C)c(=O)n(c2c(c1=O)n(cn2)C)C
n1(C)c2c(c(=O)n(C)c1=O)n(cn2)C
n1(c(=O)n(c2c(n(C)cn2)c1=O)C)C
n1(c(=O)n(c2c(c1=O)n(C)cn2)C)C
n1(C)c2c(n(C)cn2)c(=O)n(C)c1=O
n1(C)c2c(c(=O)n(c1=O)C)n(C)cn2
n1(C)c2c(n(c(=O)n(C)c2=O)C)nc1
n1(C)c2c(nc1)n(c(=O)n(c2=O)C)C
```

canonicalization (sometimes called **standardization** or **normalization**) is a process for converting data that has more than one possible representation into a "standard", "normal", or canonical form.

In chemistry it means no matter which atom in a structure you start with you get the same unique numbering for each atom.

InChI does this, SMILES does not uniformly (i.e., each version of SMILES can and often is different) nor does HEML (and exchange notation for large molecules developed by Pfizer)

HELM was first conceived at Pfizer in the summer of 2008 to support the Pfizer oligonucleotide therapeutic unit and molecules were first registered into the Pfizer corporate database using HELM in December 2008.
HELM was designed to create a single notation that can encode the structure of all biomolecules.

SMILES reference: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

InChI is a descriptor for a chemical

HELM is a file format for a chemical

Whatever the controversies and different opinions, InChI has now been more widely adopted than SMILES. In addition three US Government agencies - FDA, NIH, NIST - now have become paying members of the InChI Trust which would seem to indicate more official and institutional support leading to further widespread usage.

Current InChI Status

At present, practically speaking, InChI can handle simple organic molecules, which turns out to cover 99%+ of what people deal with every day. If it did not the every day needs of chemists and information specialists then the usage of InChI would not be as great as it is.

Large Databases with InChIs/InChIKeys

NIH/NCI – 110 million
NIH/PubChem - 91 million (68 million online)
EBI UniChem – 91 million
RSC/ChemSpider – 34 million
Elsevier/Reaxys – 30 million
IUPAC – 0 million

Why is InChI a Success

InChI is able to put things together in a new way. We took IUPAC, the Internet, Open Source software, crowdsourcing (SourceForge), Graph theory, existing representation algorithms, digitized data available on the web, and search engines, combines them, and created a very valuable tool.

InChI only works because of new technology. Without these factors above, for all practical purposes, no one would even know InChI existed.

Success is uncoerced adoption

InChI is not a replacement for any existing internal structure representations. InChI is in ADDITION to what one uses internally. Its value to the USPTO is in FINDING and LINKING information

Internal

Your representation (e.g. WLN, SMILES)

Your format(s)

External

Same representation (Standard InChI/InChIKey)

Same one format

InChI Staff and Collaborators

The InChI project has had the unusual perfect “good storm” of cooperation and support. It is a truly international project with programming in Moscow, computers in the cloud, incorporated in the UK, and a project director in the USA. Collaborators from over a dozen countries, from academia, Pharma, publishers, and the chemical information industry, have all offered, and continue to offer, senior scientific staff to develop the InChI standard.

Critical words/phrases for InChI

Link

Addition; not replacement

Algorithm

Synonym

No bureaucracy/Almost no staff

Decentralized

A Bottoms Up Project

InChI layered structure design

The current InChI layers are:

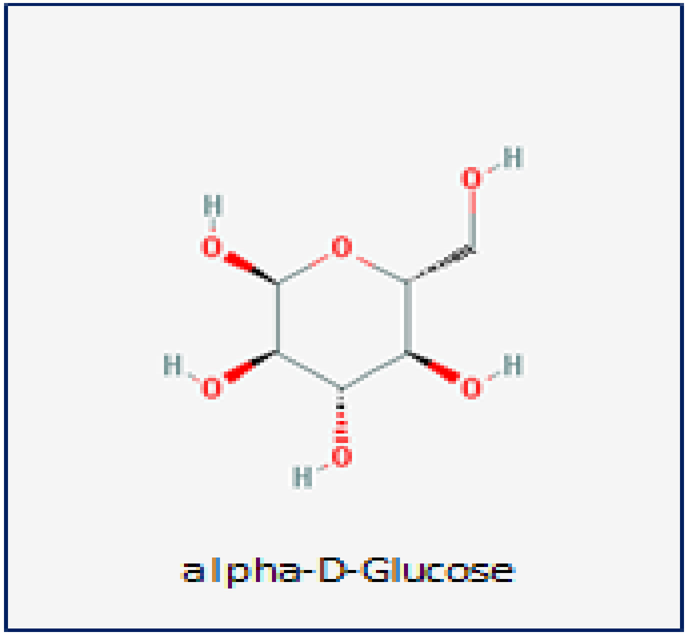
1. Formula
2. Connectivity (no formal bond orders)
   a. disconnected metals
   b. connected metals
3. Isotopes
4. Stereochemistry
   a. double bond (*Z/E)*
   b. tetrahedral (sp3)
5. Tautomers (on or off)

Charges are added to end of the string

The InChI Algorithm normalizes chemical representation and includes a "standardized" InChI, and the 'hashed' form called the InChIKey

InChI is a string

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



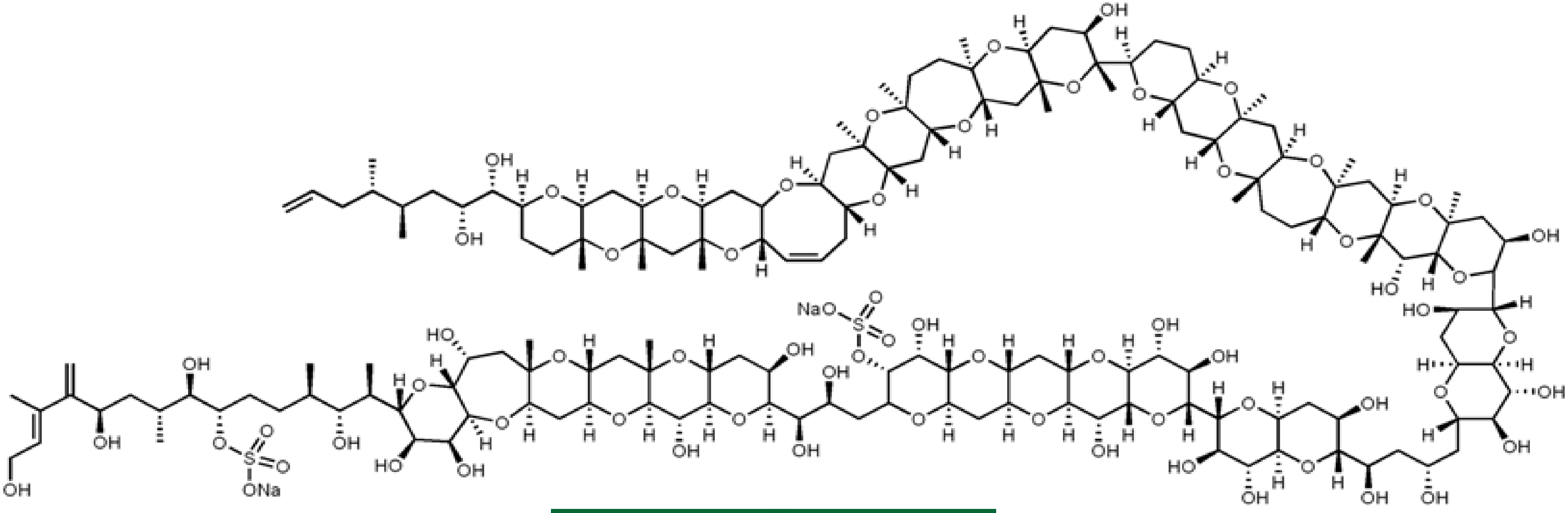
alpha-D-Glucose

Version/Type
Chemical formula
Connectivity
Charge/Proton
Stereochemical
Other (e.g., Isotopic)

“layered” line notation

InChI for Maitotoxin (from Nextmove Software, UK)

```
InChI=1S/C164H258O68S2.2Na/c1-24-26-65(2)68(5)41-74(168)117(179)85-33-36-152(11)106(203-85)55-109-162(21,231-152)64-161(20)105(210-109)51-89-83(220-161)28-
25-27-82-99(199-89)59-157(16)108(202-82)56-107-153(12,230-157)39-38-151(10)112(211-107)61-158(17)111(224-151)54-101(176)163(22,232-158)103-32-31-84-90(204-
103)53-110-156(15,219-84)62-113-150(9,223-110)37-34-102-155(14,225-113)63-114-164(23,227-102)147(192)149-159(18,226-114)58-81(175)134(218-149)133-79(173)47-93-
136(216-133)120(182)119(181)92(200-93)44-72(166)43-76(170)131-77(171)46-94-137(214-131)122(184)124(186)143(207-94)145-126(188)125(187)144-146(217-
145)128(190)139-97(208-144)50-88-87(206-139)49-96-138(205-88)127(189)141(229-234(196,197)198)95(201-96)45-75(169)118(180)132-78(172)48-98-140(215-
132)129(191)148-160(19,221-98)60-100-91(209-148)52-104-154(13,222-100)57-80(174)135-142(212-104)123(185)121(183)130(213-135)71(8)115(177)67(4)29-30-86(228-
233(193,194)195)116(178)69(6)42-73(167)70(7)66(3)35-40-165;;/h24-25,28,35,65,67-6 9,71-149,165-192H,1,7,26-27,29-34,36-64H2,2-6,8-
23H3,(H,193,194,195)(H,196,197,198);;/q;2*+1/p-2/b28-25-,66-35+;;/t65-,67+,68+,69+,71+,72+,73+,74+,75-,76+,77+,78+,79+,80+,81+,82+,83-,84+,85-,86-,87-,88+,89+,90-,91-
,92-,93-,94-,95-,96+,97-,98+,99-,100+,101+,102+,103-,104+,105-,106-,107+,108-,109+,110+,111-,112-,113+,114+,115+,116+,117-,118+,119-,120+,121+,122+,123-
,124+,125+,126+,127+,128+,129-,130-,131-,132-,133+,134+,135+,136+,137+,138+,139-,140+,141-,142-,143+,144-,145+,146+,147-,148+,149+,150-,151+,152+,153-,154-,155-
,156-,157+,158+,159-,160-,161+,162-,163+,164+;;/m0../s1
```



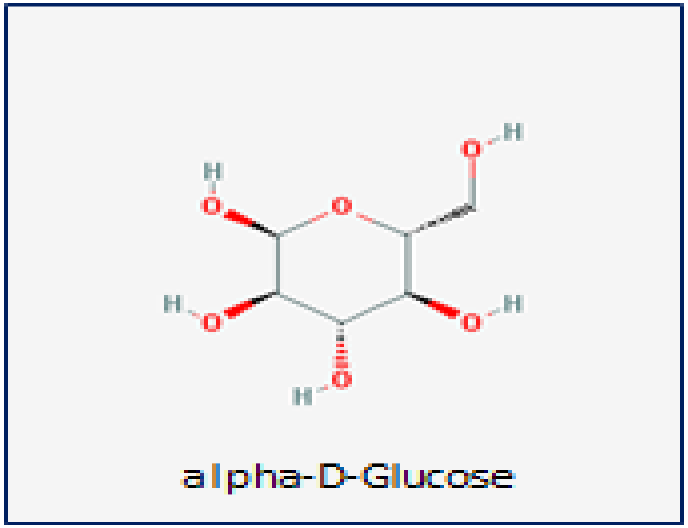
InChIKey is a "hashed" InChI

- Search engine friendly InChI
- May allow for 'secure' lookup of a chemical

WQZGKKKJIJFFOK-DVKNGEFBSA-N

Chemical formula
Connectivity
Stereochemical
Other (e.g., Isotopic)
Type
Version
Charge/Proton

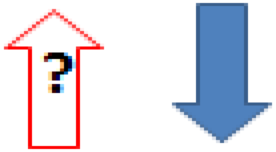
"layered" line notation



alpha-D-Glucose

InChIKey can be a 'secret'

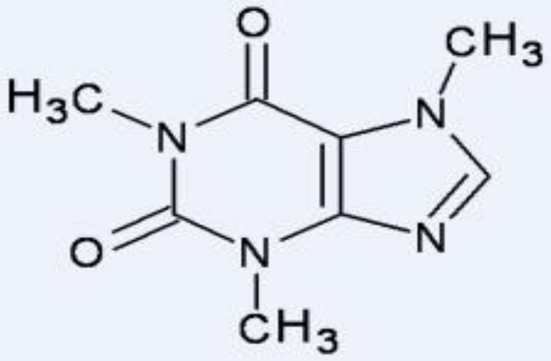
InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



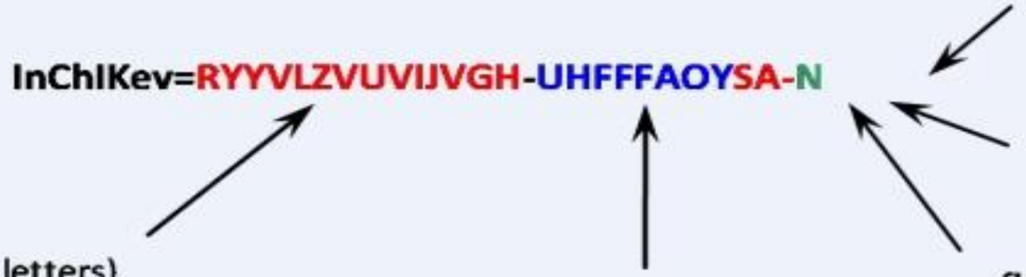
WQZGKKKJIJFFOK-DVKNGEFBSA-N

There is no chemical information in an InChIKey … if you do not know the InChI, you cannot convert the InChIKey back into a chemical structure

Slide from Evan Bolton/NIH/PubChem



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H.1-3H3 (caffeine)



InChIKev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

character indicating the number of protons ('N' means neutral)

flag character for InChI version: 'A' for version 1

First block (14 letters)

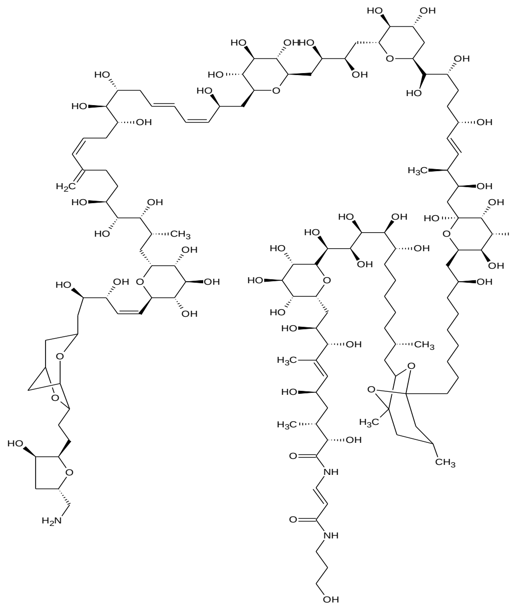
Encodes molecular skeleton (connectivity)

Second block (8 letters)

Encodes stereochemistry and isotopes

flag character ('S') indicates standard InChIKey (produced out of standard InChI)

Really long InChI (Palytoxin)



Palytoxin

Isolated from Hawaiian soft coral

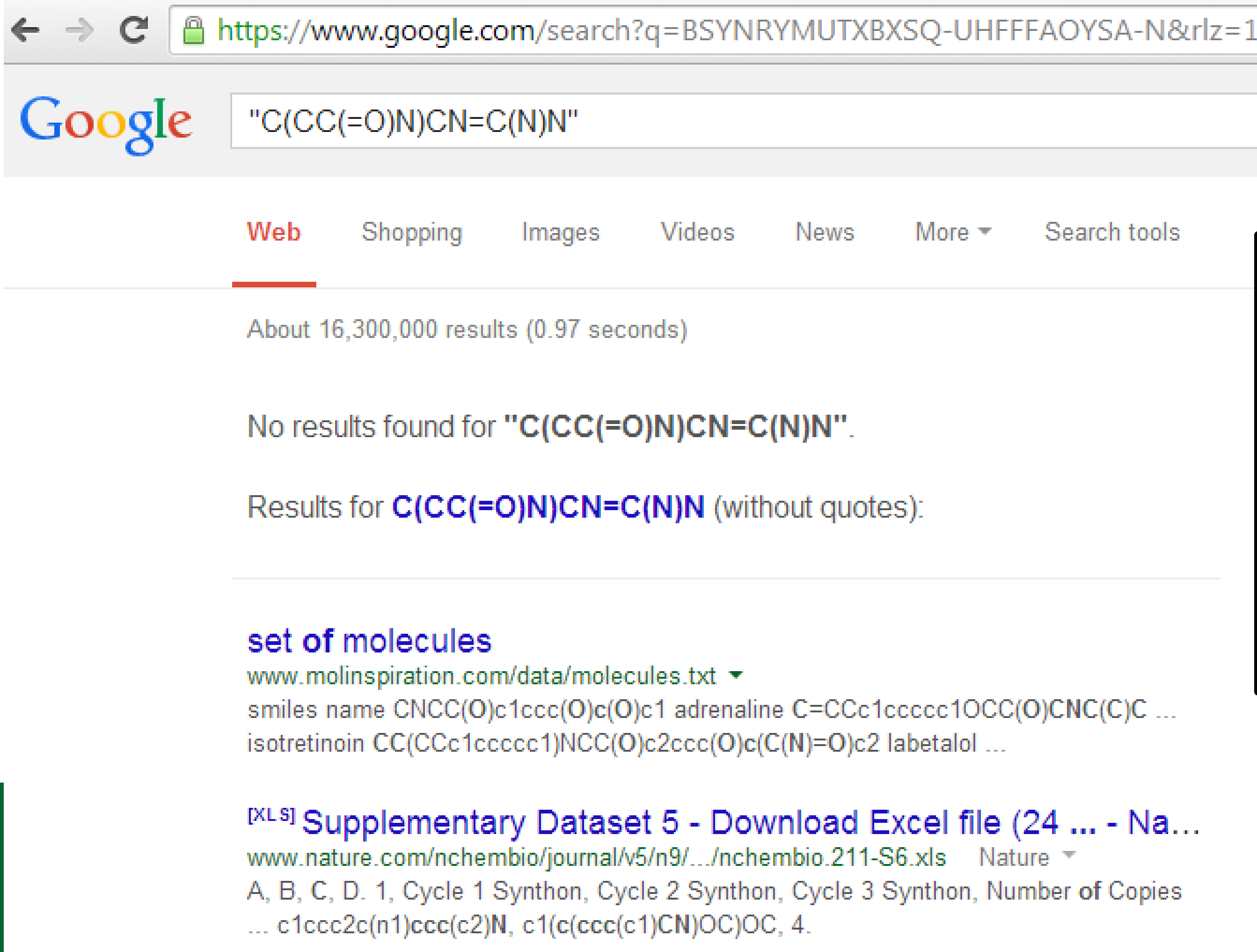
One of the most toxic non-peptide substances

Contains >70 stereochemical elements

InChI=1S/C129H223N3O54/c1-62(29-33-81(143)108(158)103(153)68(7)47-93-111(161)117(167)110(160)91(180-93)36-35-76(138)82(144)51-73-50-74-53-92(178-73)90(177-74)38-37-89-85(147)52-75(61-130)179-89)23-20-28-78(140)105(155)77(139)26-18-13-16-25-70(135)48-94-112(162)118(168)113(163)97(181-94)55-84(146)83(145)54-95-107(157)87(149)57-96(182-95)106(156)80(142)34-32-69(134)31-30-65(4)88(150)60-129(176)125(174)123(173)115(165)99(184-129)49-71(136)24-15-10-9-11-19-40-128-59-64(3)58-127(8,186-128)100(185-128)44-63(2)22-14-12-17-27-79(141)109(159)116(166)120(170)122(172)124-121(171)119(169)114(164)98(183-124)56-86(148)102(152)66(5)45-72(137)46-67(6)104(154)126(175)132-42-39-101(151)131-41-21-43-133/h13,16,18,20,23,25,30-31,35-36,39,42,45,63-65,67-100,102-125,133-150,152-174,176H,1,9-12,14-15,17,19,21-22,24,26-29,32-34,37-38,40-41,43-44,46-61,130H2,2-8H3,(H,131,151)(H,132,175)/b18-13+,23-20-,25-16-,31-30+,36-35-,42-39+,66-45+/t63-,64?,65-,67+,68+,69+,70+,71-,72-,73?,74?,75-,76+,77+,78+,79+,80+,81-,82+,83+,84+,85+,86-,87+,88-,89+,90?,91+,92?,93+,94-,95+,96-,97+,98+,99+,100?,102+,103+,104-,105-,106?,107-,108+,109-,110+,111-,112-,113+,114-,115-,116-,117-,118+,119+,120+,121-,122-,123+,124?,125+,127?,128?,129-/m0/s1

InChIKey=CWODDUGJZSCNGB-DCBUCRFRSA-N

Search engines cannot use SMILES


https://www.google.com/search?q=BSYNRYMUTXBXSQ-UHFFFAOYSA-N&rlz=1C1CHFX_enUS4(

Google "C(CC(=O)N)CN=C(N)N"

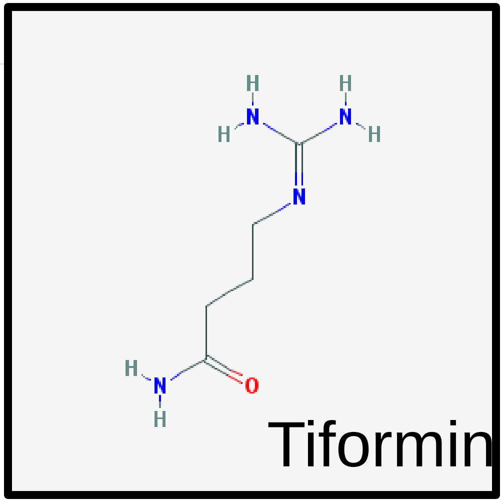
About 16,300,000 results (0.97 seconds)

No results found for **"C(CC(=O)N)CN=C(N)N"**.

Results for **C(CC(=O)N)CN=C(N)N** (without quotes):

set of molecules
www.molinspiration.com/data/molecules.txt
smiles name CNCC(O)c1ccc(O)c(O)c1 adrenaline C=CCc1ccccc1OCC(O)CNC(C)C ... isotretinoin CC(CCc1ccccc1)NCC(O)c2ccc(O)c(C(N)=O)c2 labetalol ...

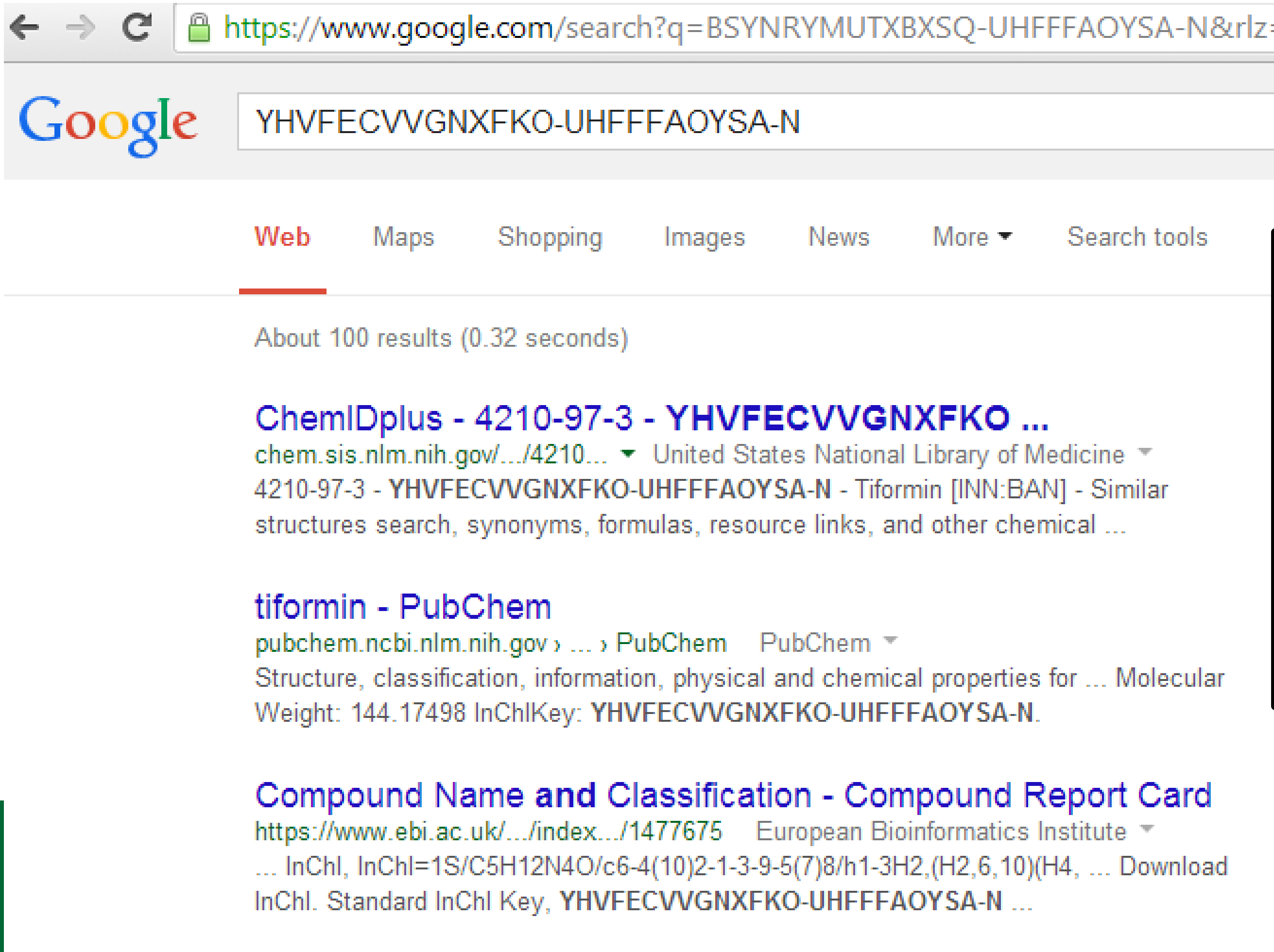
[XLS] Supplementary Dataset 5 - Download Excel file (24 ... - Na...
www.nature.com/nchembio/journal/v5/n9/.../nchembio.211-S6.xls Nature
A, B, C, D. 1, Cycle 1 Synthon, Cycle 2 Synthon, Cycle 3 Synthon, Number of Copies ... c1ccc2c(n1)ccc(c2)N, c1(c(ccc(c1)CN)OC)OC, 4.



Tiformin

Search Engines can use InChIKey

They can use InChI too! ... but your mileage may vary



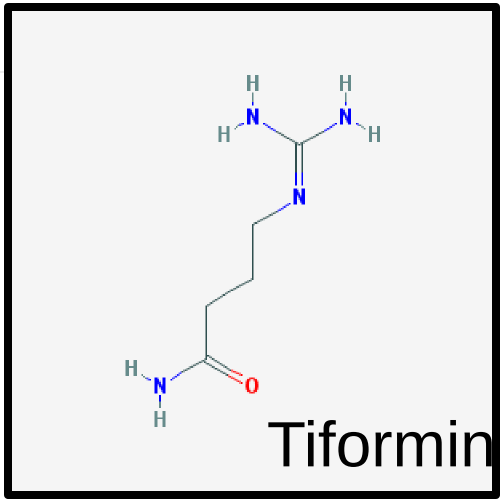
https://www.google.com/search?q=BSYNRYMUTXBXSQ-UHFFFAOYSA-N&rlz=1C1CHFX_enUS4

About 100 results (0.32 seconds)

ChemIDplus - 4210-97-3 - YHVFECVVGNXFKO ...
chem.sis.nlm.nih.gov/.../4210... United States National Library of Medicine
4210-97-3 - **YHVFECVVGNXFKO-UHFFFAOYSA-N** - Tiformin [INN:BAN] - Similar structures search, synonyms, formulas, resource links, and other chemical ...

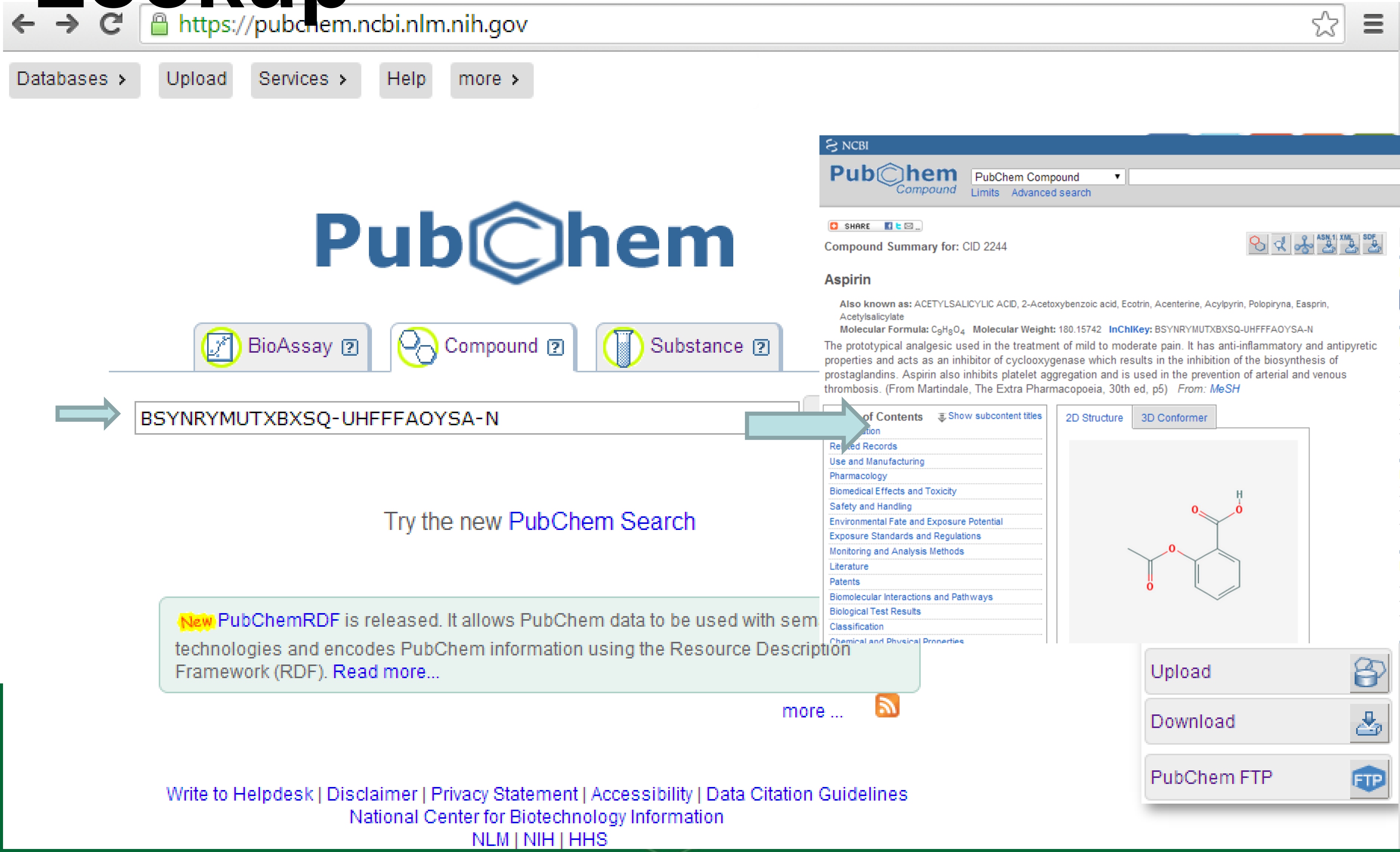
tiformin - PubChem
pubchem.ncbi.nlm.nih.gov › ... › PubChem PubChem
Structure, classification, information, physical and chemical properties for ... Molecular Weight: 144.17498 InChIKey: **YHVFECVVGNXFKO-UHFFFAOYSA-N**.

Compound Name and Classification - Compound Report Card
https://www.ebi.ac.uk/.../index.../1477675 European Bioinformatics Institute
... InChI, InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4, ... Download InChI. Standard InChI Key, **YHVFECVVGNXFKO-UHFFFAOYSA-N** ...



Tiformin

InChIKey Compound-based Lookup



https://pubchem.ncbi.nlm.nih.gov

BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Try the new PubChem Search

New PubChemRDF is released. It allows PubChem data to be used with sem technologies and encodes PubChem information using the Resource Description Framework (RDF). Read more...

more ...

Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines
National Center for Biotechnology Information
NLM | NIH | HHS

Compound Summary for: CID 2244

Aspirin

Also known as: ACETYLSALICYLIC ACID, 2-Acetoxybenzoic acid, Ecotrin, Acenterine, Acylpyrin, Polopiryna, Easprin, Acetylsalicylate

Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742 **InChIKey:** BSYNRYMUTXBXSQ-UHFFFAOYSA-N

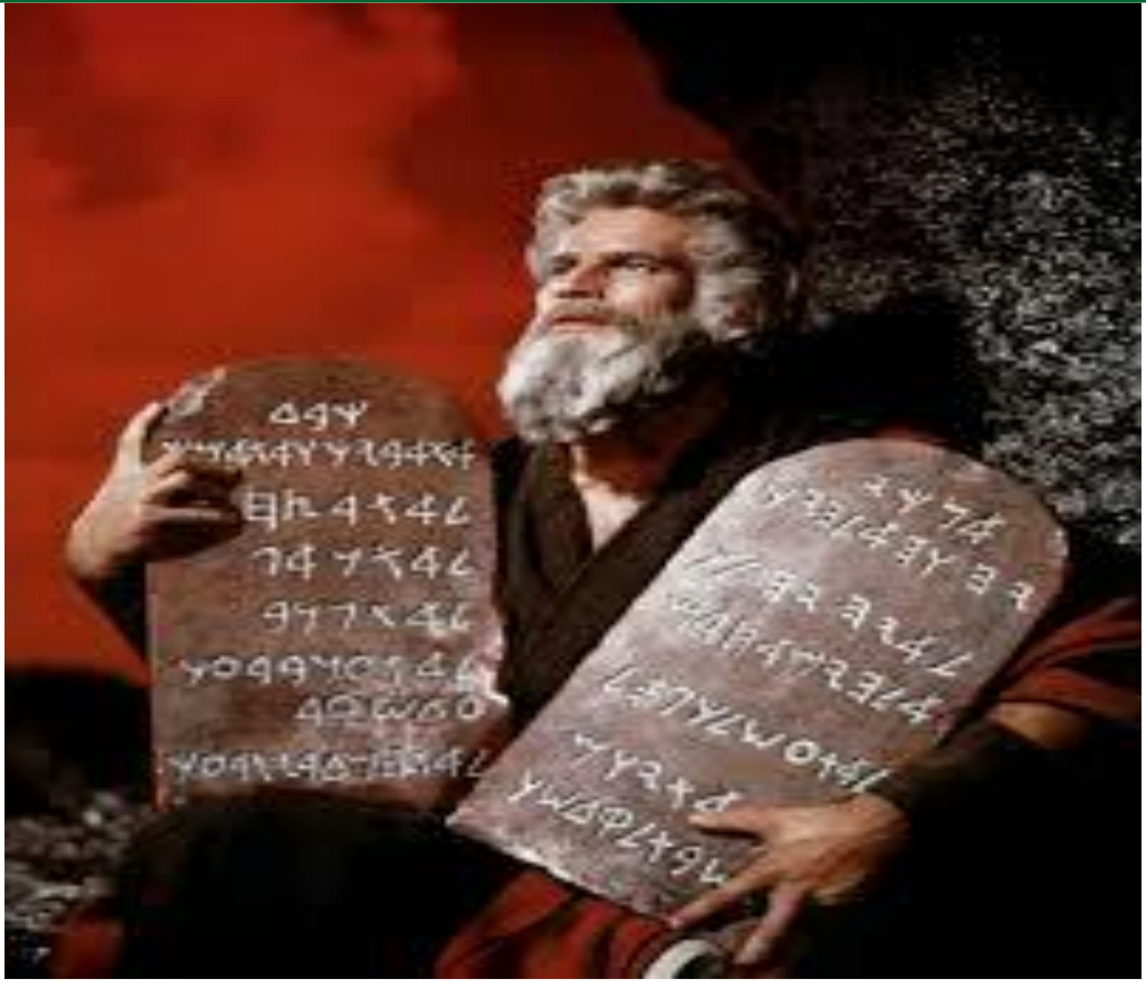
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5) *From: MeSH*

of Contents Show subcontent titles

- tion
- Re ed Records
- Use and Manufacturing
- Pharmacology
- Biomedical Effects and Toxicity
- Safety and Handling
- Environmental Fate and Exposure Potential
- Exposure Standards and Regulations
- Monitoring and Analysis Methods
- Literature
- Patents
- Biomolecular Interactions and Pathways
- Biological Test Results
- Classification
- Chemical and Physical Properties

InChI/InChIKey Use and Utility

- InChI
  - Enabler of data exchange
  - Provides chemical structure normalization
- InChIKey
  - Compact form for structure lookup
  - Allows "secret" chemical information exchange



What about funding ?

Don’t give up - Moses was once a basket case

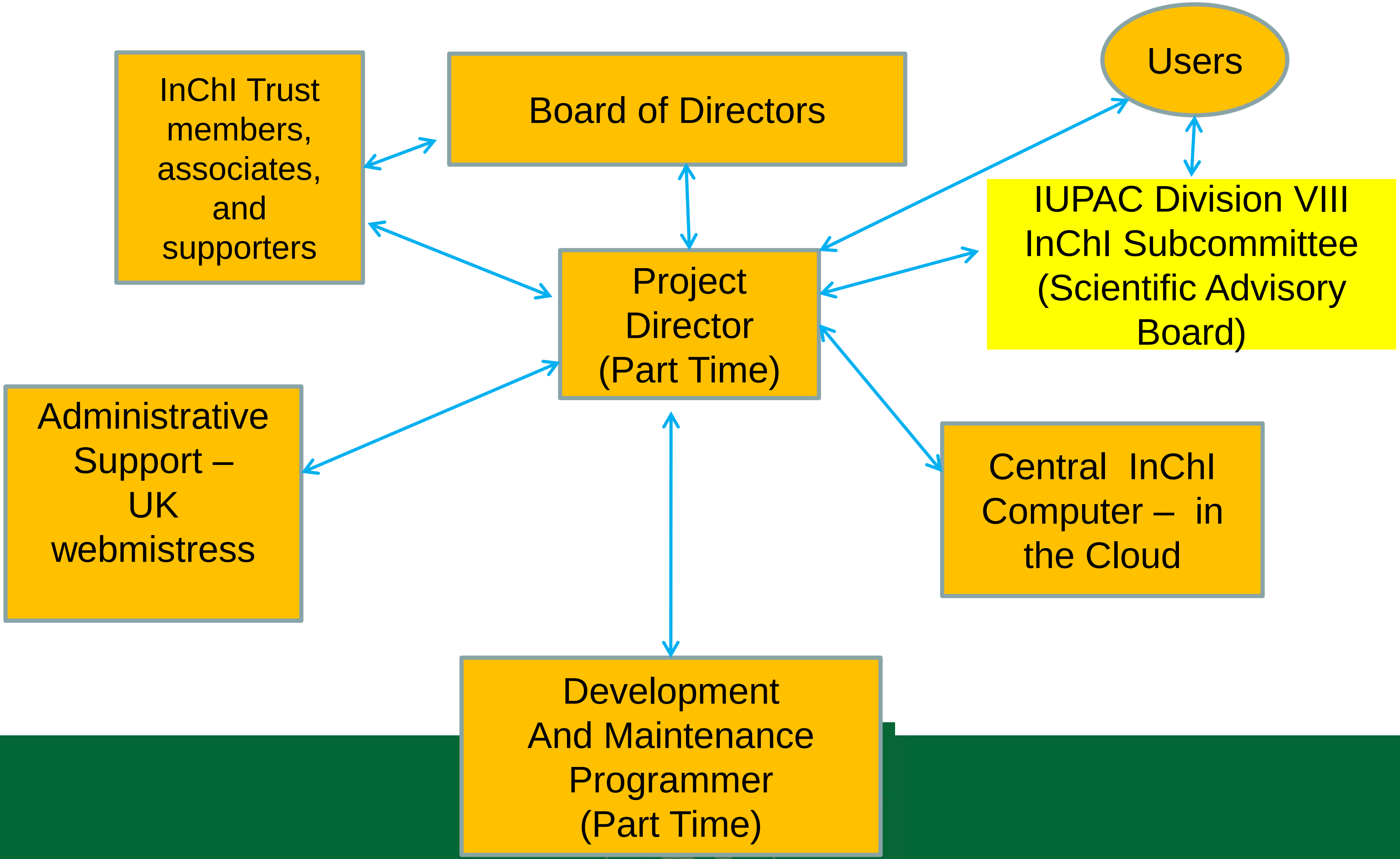
While InChI did not make the top 10, it is #14

(Thou shall use InChI for structure representation.)

The InChI Trust

To function and succeed, InChI had to become personality independent. InChI had to be "institutionalized". If the work of this project was to be enduring it needed to turned over to an entity that would ensure its ongoing activities and be acceptable to the community. It was concluded that a not-for-profit organization would best fit the ongoing and future project needs. Thus the decision to create and incorporate the "InChI Trust“ as a UK charity.

InChI Trust Organization



Users

InChI Trust members, associates, and supporters

Board of Directors

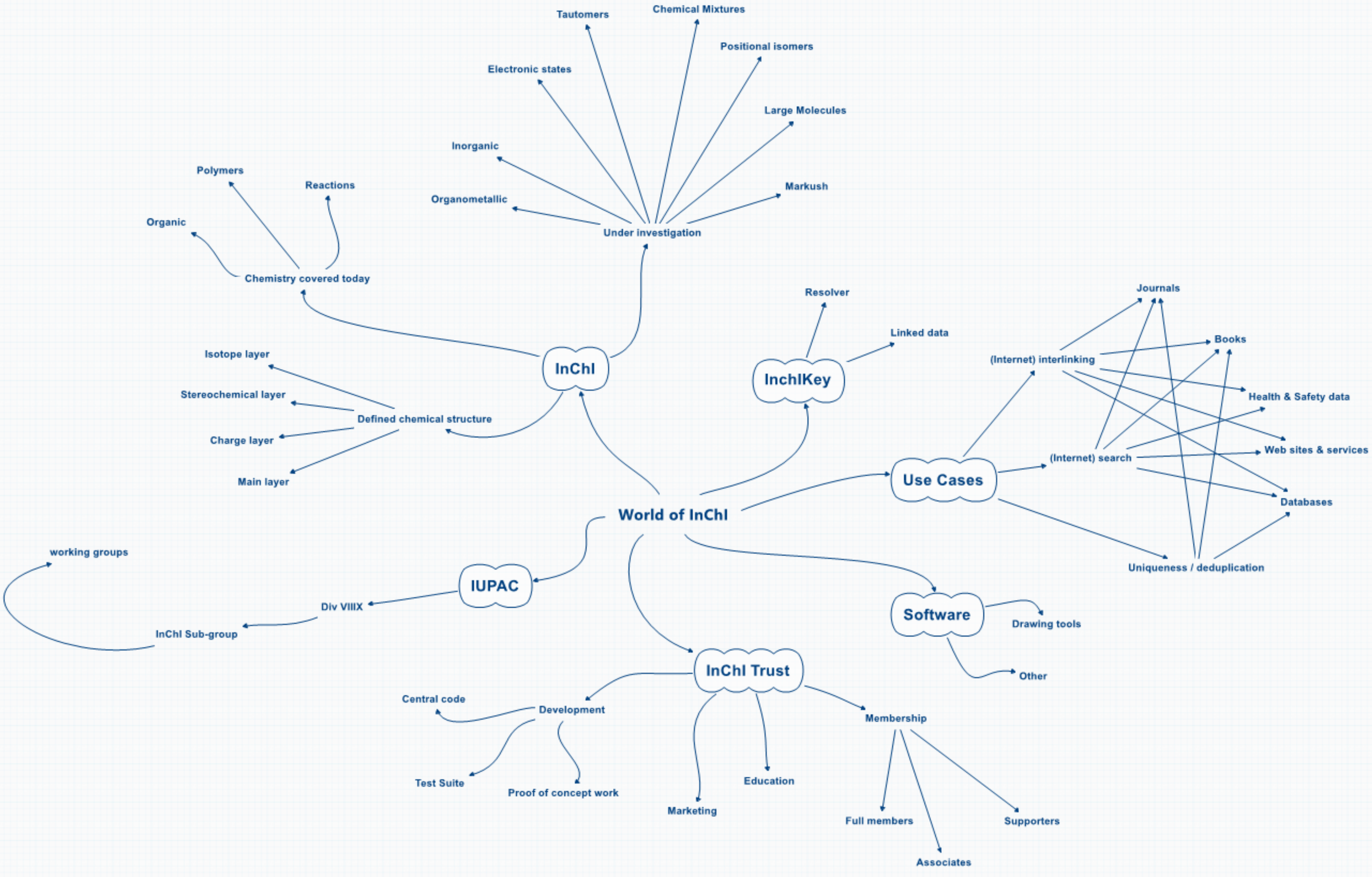
IUPAC Division VIII InChI Subcommittee (Scientific Advisory Board)

Project Director (Part Time)

Administrative Support – UK webmistress

Central InChI Computer – in the Cloud

Development And Maintenance Programmer (Part Time)



World of InChI

InChI

Chemistry covered today

Organic

Polymers

Reactions

Under investigation

Organometallic

Inorganic

Electronic states

Tautomers

Chemical Mixtures

Positional isomers

Large Molecules

Markush

Defined chemical structure

Isotope layer

Stereochemical layer

Charge layer

Main layer

InchIKey

Resolver

Linked data

Use Cases

(Internet) interlinking

(Internet) search

Journals

Books

Health & Safety data

Web sites & services

Databases

Uniqueness / deduplication

Software

Drawing tools

Other

IUPAC

Div VIIIX

InChI Sub-group

working groups

InChI Trust

Development

Central code

Test Suite

Proof of concept work

Marketing

Education

Membership

Full members

Associates

Supporters

InChI characteristics

Consensus
Technical competence
Political and technical cooperation
Precompetitive collaboration – publishers, databases, software
No competition with commercial products
No mission creep
IUPAC blessing/endorsement & rapid IUPAC acceptance
Excellent understanding of what the Internet and how it can be effectively used in Chemical Information

Vision of the future

Current IUPAC Working Groups & Projects

Completed:

Revised FAQ’s from Cambridge- Nick Day/Peter Murray-Rust
InChI Certification Suite
Version 1.05 released – 5/16
Markush (contract to be signed when funded)
Polymers
RInChI – InChI for Reactions
New API

Started/To be started

Electronic/Excited States
Mixtures
InChI Resolver
QR codes for InChI
InChI teaching/educational materials
Large Molecules/Biopolymers/Macromolecules
Material Science (MGI – Materials Genome Initiative)
Inorganics
Crystal/3D structures
Redesign of Handling of Tautomerism

The Future

InChI has become mainstream for publishers, databases providers, and software developers. Over the next 5-10 years, publishers will use data mining to create both better abstracts, useful indexing, and concept terms. Search engines will be able to search for appropriate text and structures and direct users to the original (fee or free/Open Access/Open Data) sources.

Keep Calm and Use InChI

Summary

If you are not part of the solution; you are part of the precipitate

Acknowledgements

(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Colin Batchelor, John Barnard , Evan Bolton, Steve Boyer, Steve Bryant, Szabolcs Csepregi , Rene Deplanque, Gary Mallard, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau , Marc Nicklaus, Carmen Nitsche, Matthias Nolte , Igor Pletnev, Josep Prous, Peter Murray-Rust, Hinnerk Rey, Ulrich Roessler, Roger Schenck , Martin Schmidt, Steve Stein, Peter Shepherd, Markus Sitzmann , Chris Steinbeck, Keith Taylor, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, Tony Williams, Andrey Yerin.

Special Acknowledgement: Ted Becker& Alan McNaught for their vision and leadership of the future of IUPAC nomenclature.

Have any questions?

If you think of a question later, email me:

steve@inchi-trust.org