Using the IUPAC International Chemical Identifier standard (InChI) to find information on chemicals

Stephen Heller

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

5/15&16/2014

Slides are available at http://www.hellers.com/steve/pub-talks/SIN-5-14.pdf

What is InChI ?

The IUPAC International Chemical Identifier structure representation standard, 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 appropriate software.

InChI is really just a synonym.

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

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

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

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

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

(Dumb) Index Numbers
- EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN, EC, ChemSpider ID, REACH, MFCD#, …

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

Phil Bourne, Associate Director for Data Science, NIH

“No, no, not another structure standard!!!”

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., CH_4 - 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 is plumbing. InChI is an (enabling) tool. It is a means to an end. InChI is a modern enabling technology.

For all but small group of chemists developing it, InChI is not something anyone should want to know about.

All you want to do is use InChI to find information on the web.

InChI is helping scientists to do better work and find/link to the latest information.

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

Without InChI, finding something on the Internet is like trying to find the bathroom in a house with 1,000,000 unmarked doors

The Internet is like drinking from a fire hydrant; InChI will cut it to a faucet drip.

The problem with too much information on the Internet: **Lack of integration**

multiple applications
multiple repositories
multiple interfaces and protocols

InChI does not replace any internal, local system pieces. Your language and format remain as is.

But even though we communicate around the world in English, there are still over 2,500 language versions of the Bible.

As you see on many drug labels “InChI for external use only.”(1) The InChI standard is like universal language -- like English.

(1) http://www.macmillandictionary.com/us/dictionary/american/external

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.

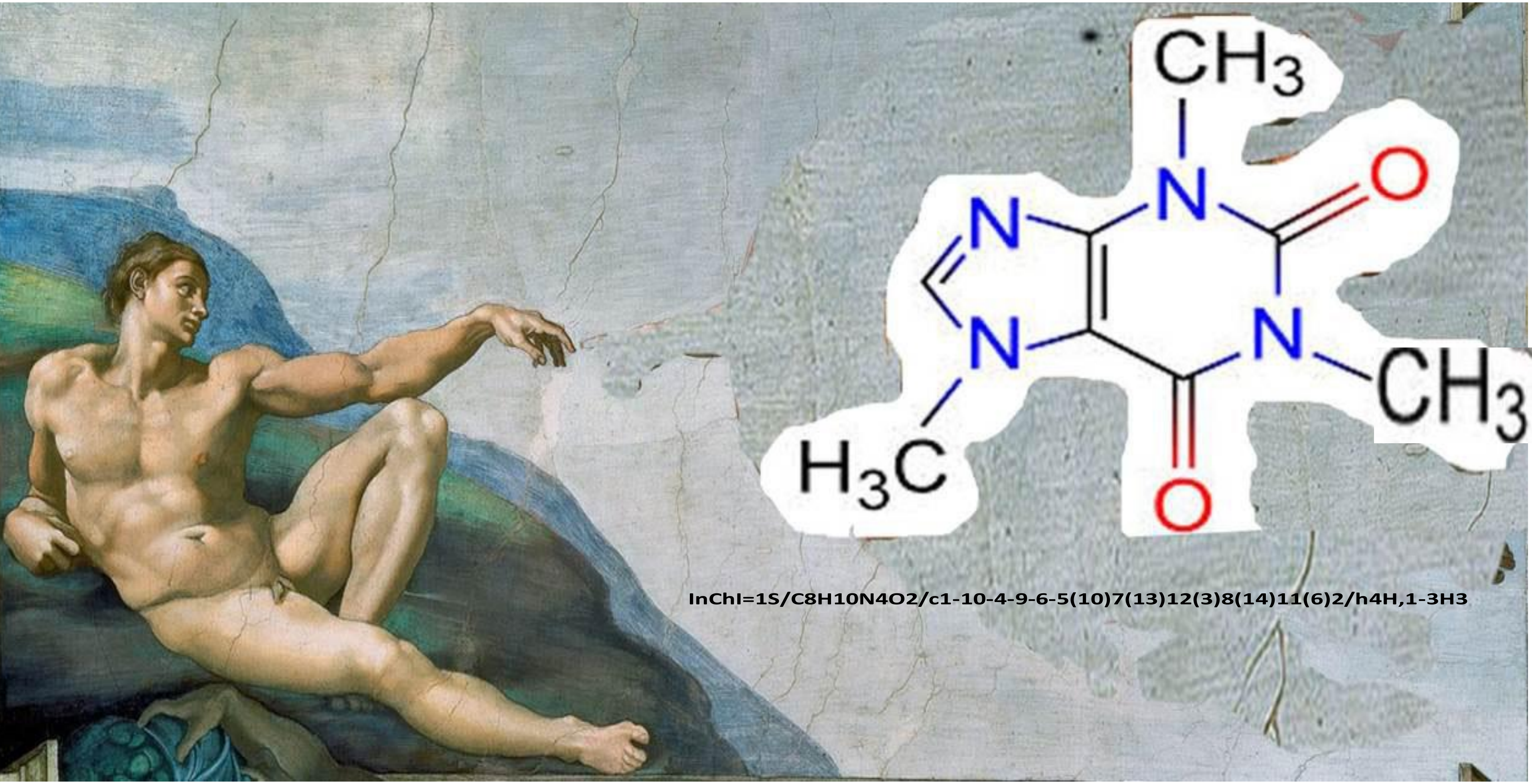
InChI YouTube Videos

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

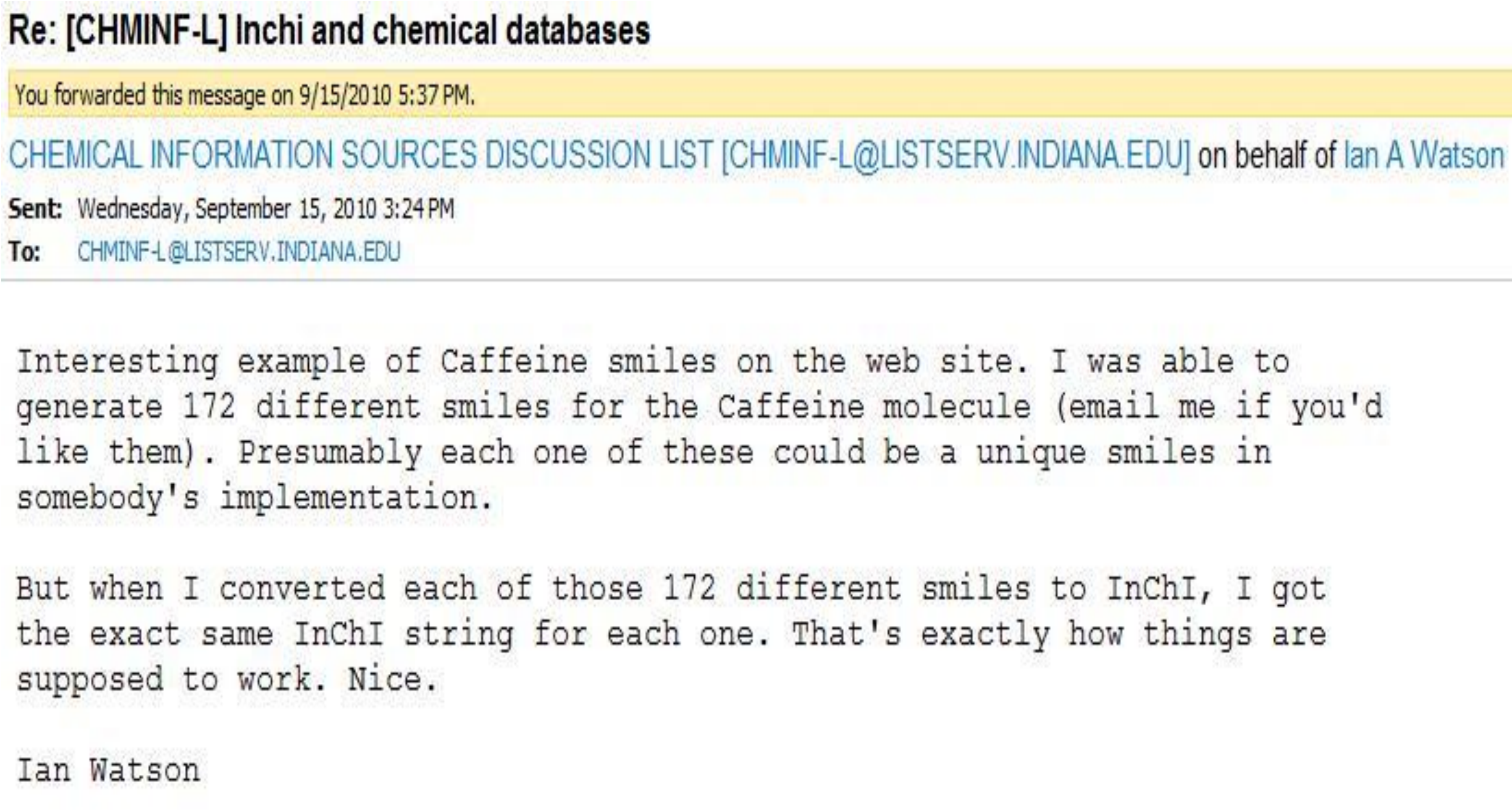


InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

With apologies to Michelangelo -- God created man & man created InChI

How do I create an InChI ?

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 appropriate software.



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

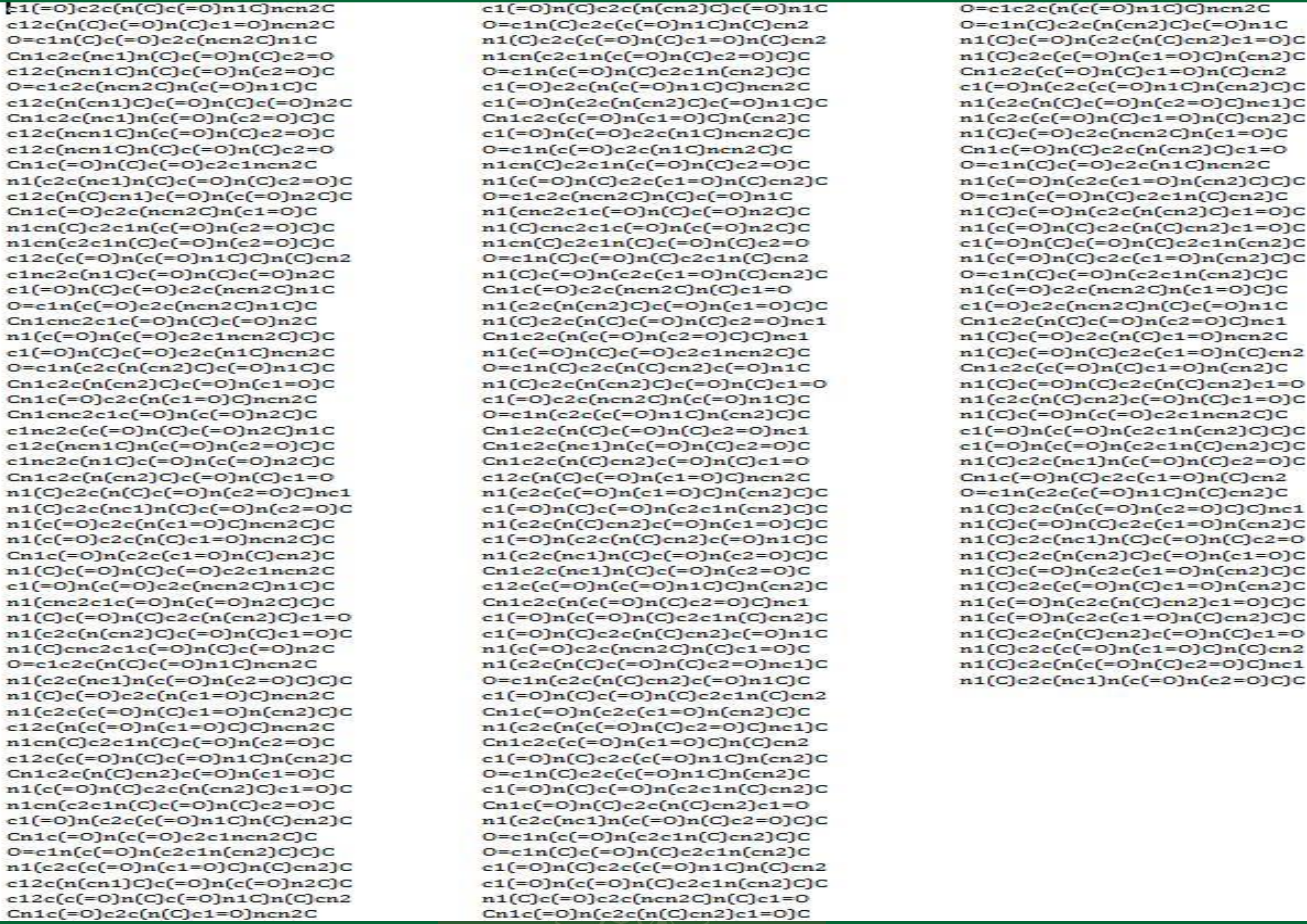
Sent: Wednesday, September 15, 2010 3:24 PM

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.

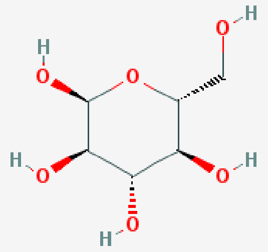
Ian Watson



```
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
```

What about SMILES as a standard?

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



alpha-D-Glucose

- 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

Too many “standards” actually slow things down and make getting to the information you want and need take a lot longer time and effort than it would take with InChI

E Pluribus Unum
Out of many, One

InChI

172 SMILES representations

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 Nov. 11, 1947)**

InChI Characteristics

1. Easy to generate (It will use existing software.)

2. Expressive (It will contain structural information.)

3. Unique/Unambiguous

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

InChI as a web index for molecules

“We have now discovered, serendipitously, that these InChIs have been comprehensively and accurately indexed by the Google search engine. From preliminary exploration it appears that every known document in which an InChI appears has been indexed and that all are retrievable by standard queries with virtually 100% precision. This means that standard Web-based indexers, without any alteration, are capable of acting as completely precise chemical search engines. Although we have many years of developing chemistry on the web, this was an unexpected and very welcome finding”

Murray-Rust et al. 2004 http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/

Where are InChIs?

PubChem ~ 50 million
ChemSpider ~ 30 million
Reaxys ~ 30 million
PubChem from patents (all sources) ~ 15 million
PubChem journal sources (PubMed + ChEMBL) ~ 1 million
SciFinder ~ 60 million (estimated as input for searches)
Web sources outside the above (no idea)

Chris Southan BioIT 2014 lecture

The Googleable InChIKey

InChI is an international computer readable standard not just for chemists, but rather has very wide technical and non-technical use for linking and connecting information in many areas of scientific and everyday activities - -

- abstracting services
- biochemistry
- biology/genomics databases
- bio-activity databases
- books
- chemical biology
- chemical spills
- chemistry databases
- clinical trials
- company annual reports
- drug discovery
- drug information
- drug overdoses
- electronic books
- environmental information
- food additives
- lawsuits
- magazines
- medicinal chemistry
- medical information
- medical records
- metabolomics
- newspapers
- patents
- packages/bottles/transportation labels/ everyday product labels
- pharmacology
- scientific journals
- toxicology
- toxicological information

Using InChI to link to information

InChI characteristics

Consensus
Technical competence
Political and technical cooperation
Precompetitive collaboration
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

While InChI is an Open Source, public domain, system for creating a unique computer-readable identifier (“name”) it is NOT a registry system. InChI’s are created only by those who choose to adopt and use the algorithm. Registry systems which index the literature are complementary to any InChI databases that anyone creates. Of course if one wants to create a chemical registration system, InChI along with other notations can be used.

Critical words/phrases for InChI

Link

Addition; not replacement

Algorithm

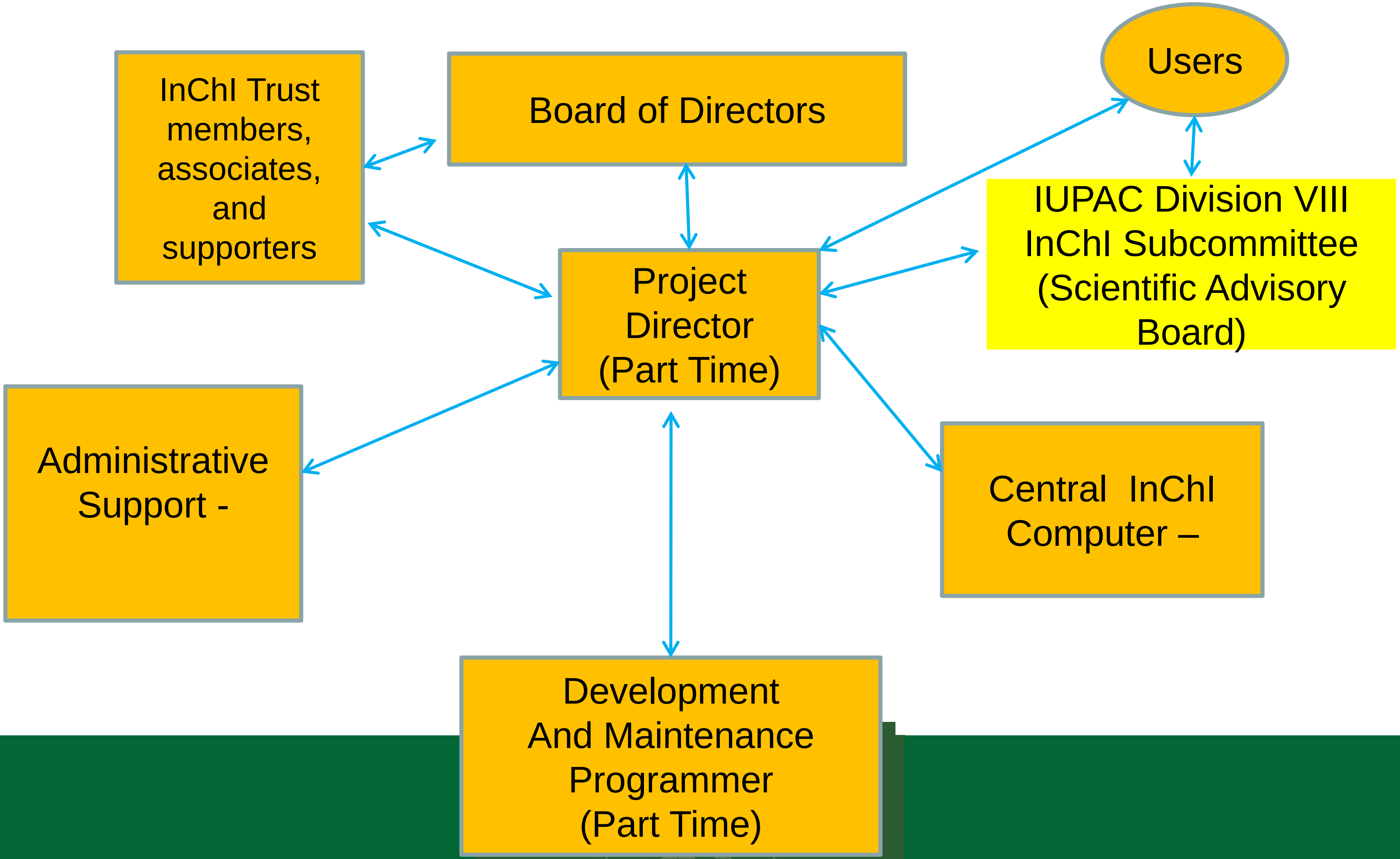
Synonym

No bureaucracy

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



InChI Trust members, associates, and supporters

Board of Directors

Users

IUPAC Division VIII InChI Subcommittee (Scientific Advisory Board)

Project Director (Part Time)

Administrative Support -

Central InChI Computer –

Development And Maintenance Programmer (Part Time)

Total number of Members, Associate Members, and (non paying) Supporters ~60

(Please consider joining !!)

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 senior scientific staff to develop the InChI standard.

Why InChI is a success

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

2. InChI is a public domain algorithm that anyone, anywhere can freely use. And they sure use it!

Success is uncoerced adoption

Bypassing IUPAC procedures

The usual very, lengthy IUPAC approval process was hijacked and sped up by sending the IUPAC bureaucracy, not a white paper with InChI rules, but rather unreadable and unintelligible C code.

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.

Introduction to InChI video

InChI layered structure design

The current InChI layers are:

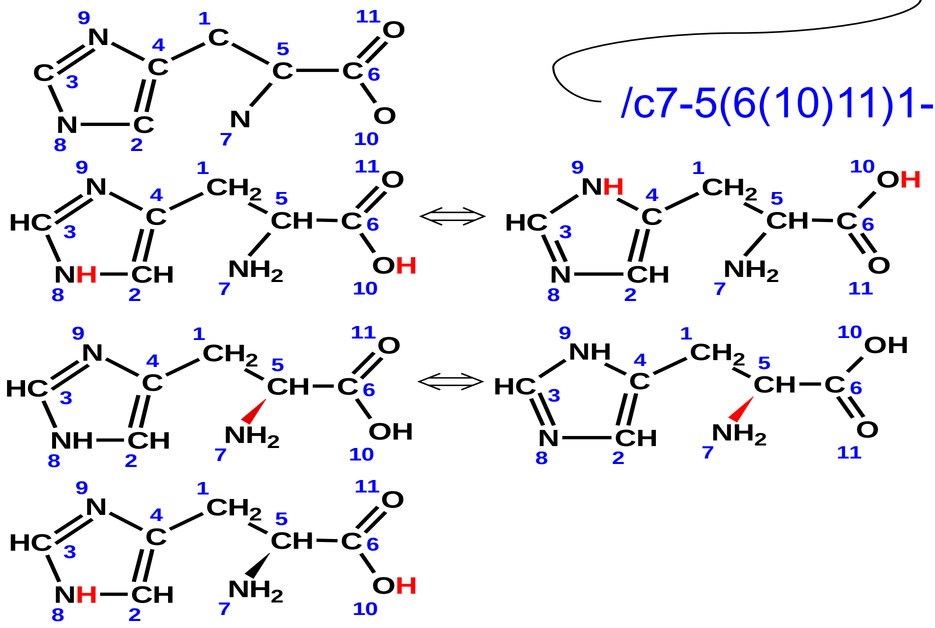
1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond (*Z/E)*
    b. tetrahedral (sp3)
5. Tautomers (on or off)

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 Layers: L-Histidine

$C_6H_9N_3O_2$



InChI=1/C6H9N3O2 — Formula

/c7-5(6(10)11)1-4-2-8-3-9-4 — Connections

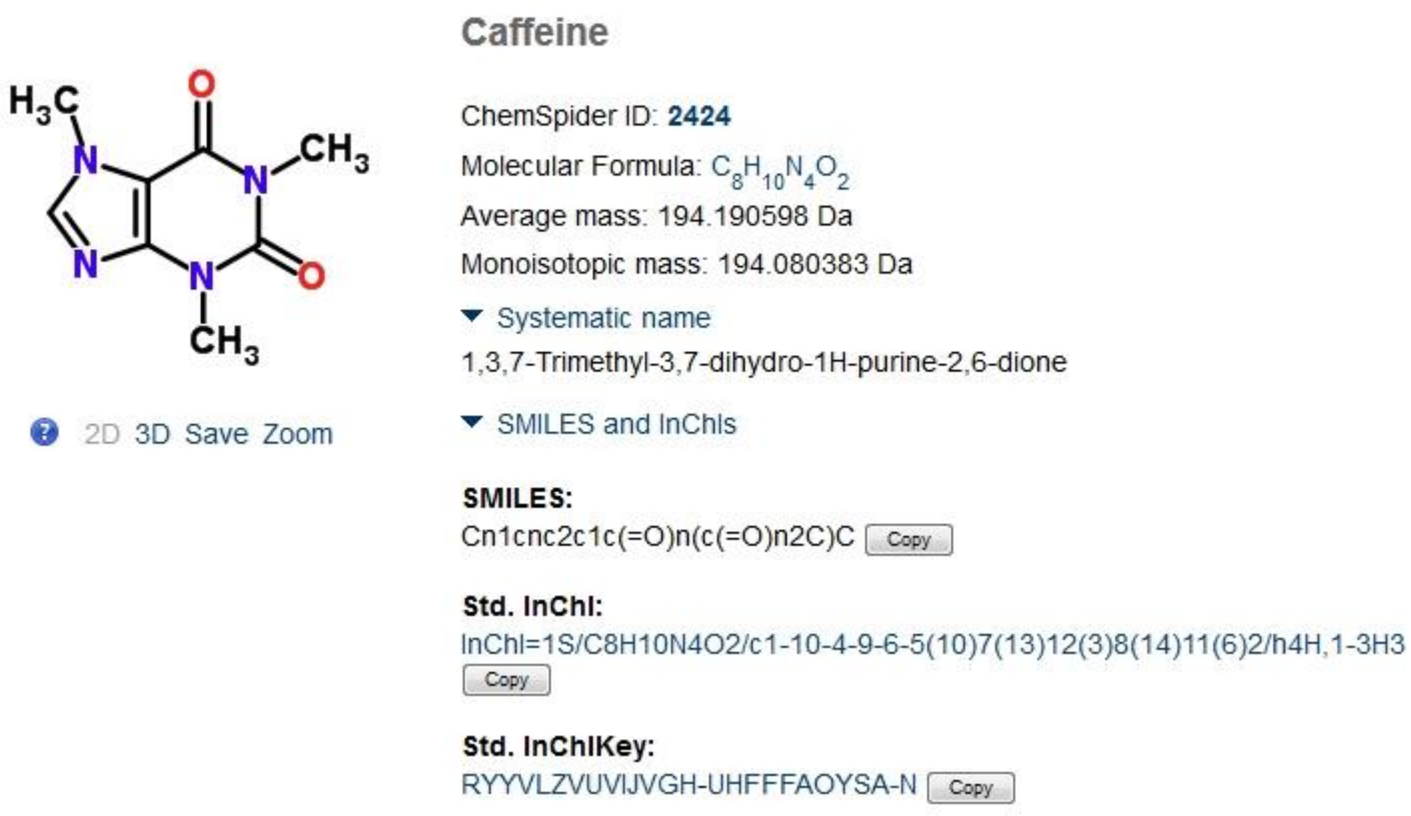
/h2-3,5H,1,7H2,(H,8,9)(H,10,11) — Hydrogens (mobile)

/t5-/m0/s1 — Stereo

/f/h8,10H — Hydrogens (fixed)

InChI=1/C6H9N3O2/c7-5(6(10)11)1-4-2-8-3-9-4/h2-3,5H,1,7H2,(H,8,9)(H,10,11)/t5-/m0/s1/f/h8,10H

InChIKey=HNDVDQJCIGZPNO-QLMCEAFFNA-N **InChIKey=HNDVDQJCIGZPNO-YFKPBYRVSA-N**



Caffeine

ChemSpider ID: **2424**

Molecular Formula: $C_8H_{10}N_4O_2$

Average mass: 194.190598 Da

Monoisotopic mass: 194.080383 Da

▼ Systematic name

1,3,7-Trimethyl-3,7-dihydro-1H-purine-2,6-dione

▼ SMILES and InChIs

SMILES:

Cn1cnc2c1c(=O)n(c(=O)n2C)C

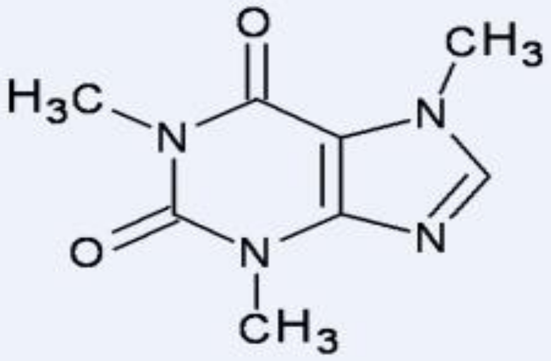
Std. InChI:

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

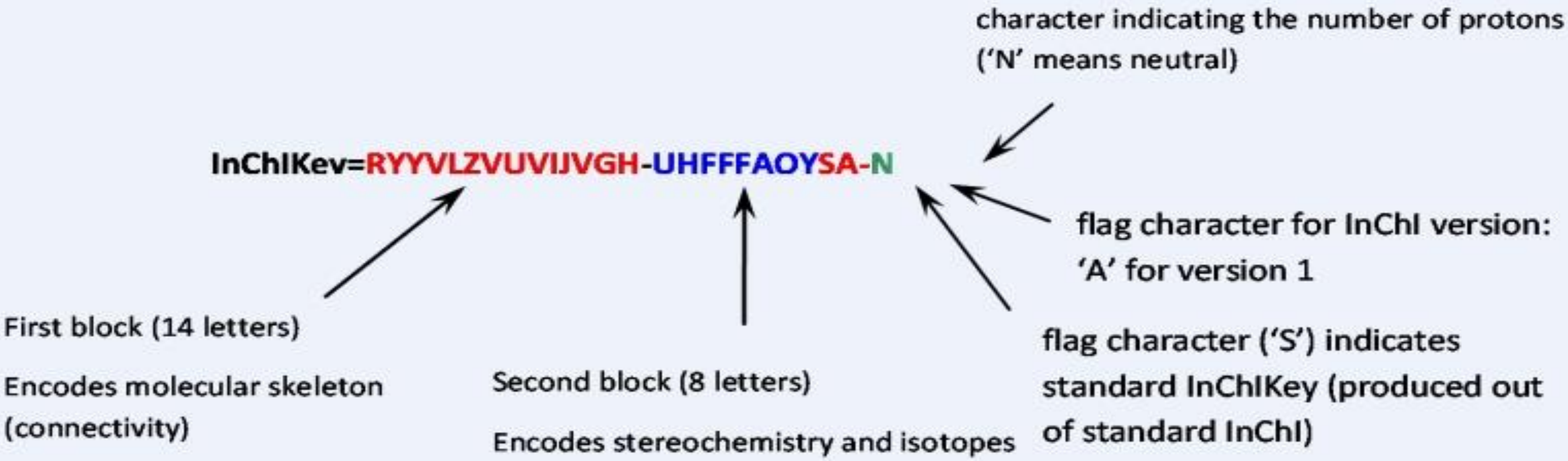
Std. InChIKey:

RYYVLZVUVIJVGH-UHFFFAOYSA-N

Caffeine from ChemSpider database



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)

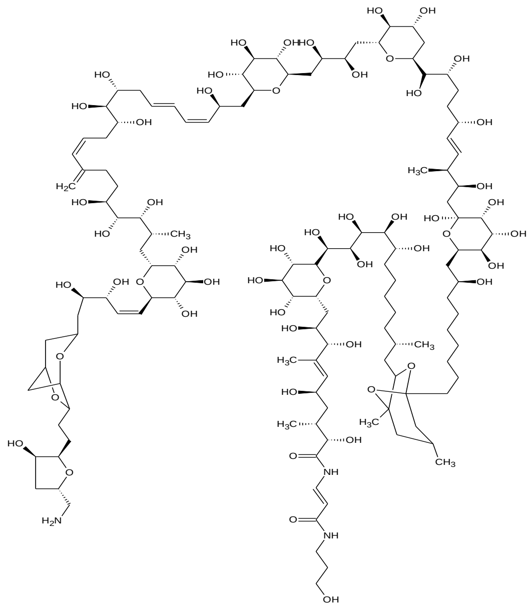
Encodes molecular skeleton (connectivity)

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

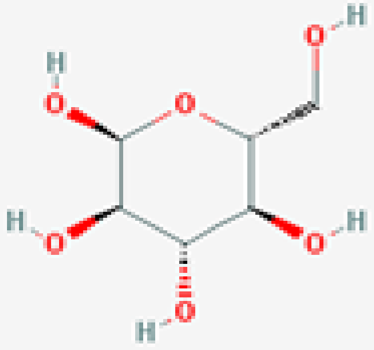
InChI=1S/C129H223N3O54/c1-62(29-33-81(143)108(158)103(153)68(7)47-93-111(161)117(167)110(160)91(180-93)36-35-76(138)82(144)51-73-50-74-53-92(178-73)90(177-74)38-37-89-85(147)52-75(61-130)179-89)23-20-28-78(140)105(155)77(139)26-18-13-16-25-70(135)48-94-112(162)118(168)113(163)97(181-94)55-84(146)83(145)54-95-107(157)87(149)57-96(182-95)106(156)80(142)34-32-69(134)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

InChI TRUST

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

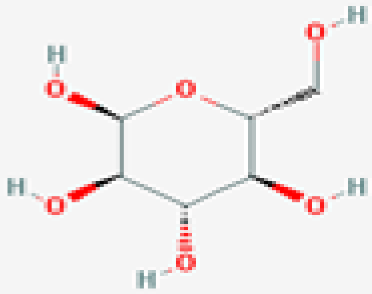
InChIKey is a “hashed” InChI

- 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

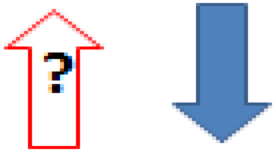
“layered” line notation



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

QA/QC - InChI Certification Suite

The InChI certification suite is a software package designed to check that your installation of the InChI program has been performed correctly. The programs test your installation against a broad set of structures (which are provided with the Suite) to assure the InChIs and InChIKeys are correct and valid. Only this way is it possible to know that the InChIs have been generated properly and consistently.

Unlike other Trust products (software and documentation) the Certification Suite is NOT free, except to members and supporters who use for non-commercial activities. It costs $5,000 per year.

Current IUPAC Working Groups & Projects

In Progress:
Organometallics
InChI Resolver

Completed:
Revised FAQ’s from Cambridge- Nick Day/Peter Murray-Rust
InChI Certification Suite
Version 1.04 released – 9/11
Markush (contract to be signed when funded)
Polymers/Mixtures
RInChI – InChI for Reactions (contract to be signed when funded)
New API

Started/To be started in 2013/2014:
Electronic/Excited States
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.

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, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Gary Mallard , 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