The IUPAC InChI Chemical Structure Standard – Today and the Future

Stephen Heller

Project Director, InChI Trust

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

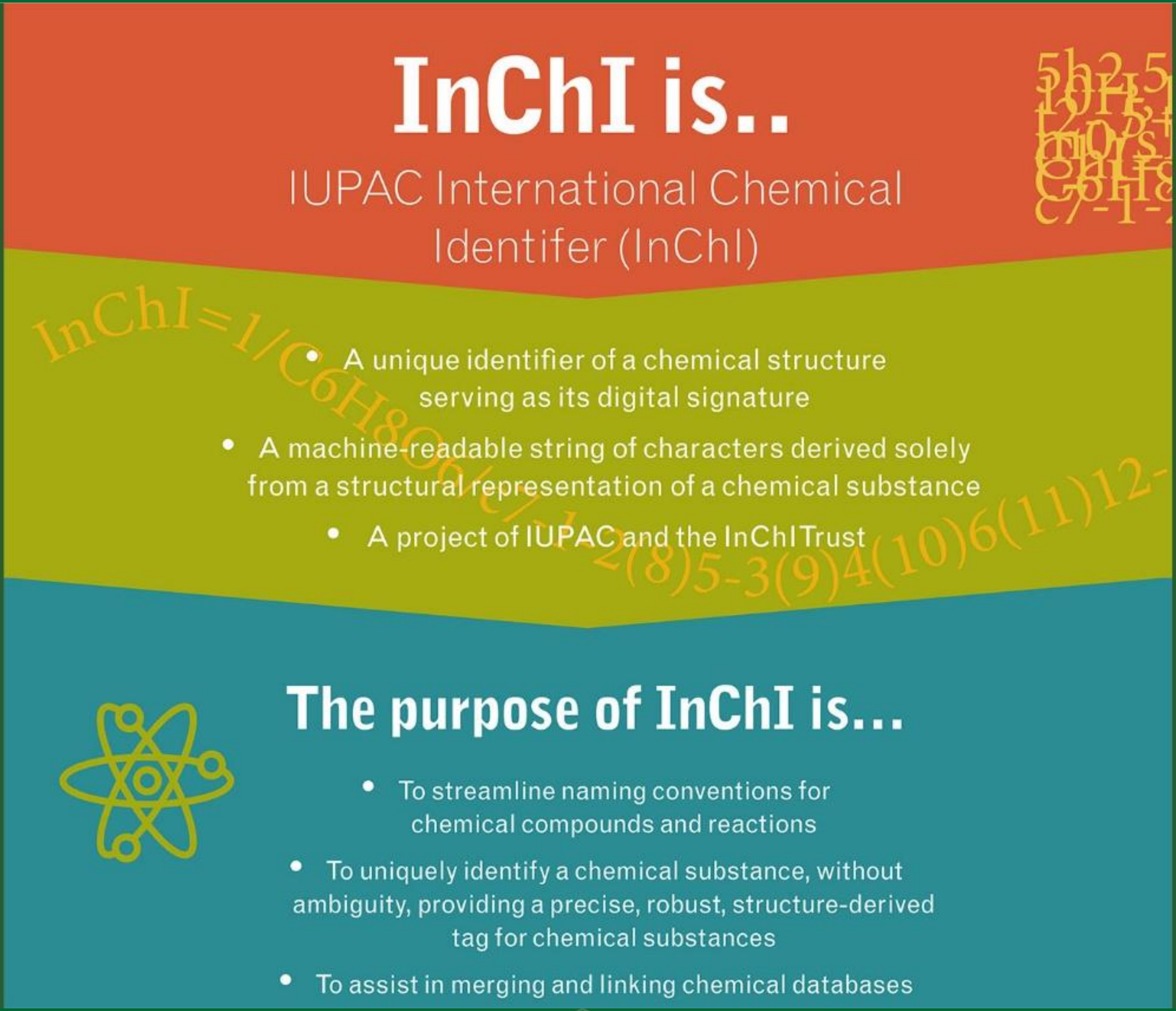
http://www.inchi-trust.org

China-5-6/2019

Slides are available at http://www.hellers.com/steve/china-2019.pdf

InChI Project Goal

To link everything about a chemical from many sources with the purpose of creating new information, and we hope new knowledge.



InChI is..

IUPAC International Chemical Identifer (InChI)

- A unique identifier of a chemical structure serving as its digital signature
- A machine-readable string of characters derived solely from a structural representation of a chemical substance
- A project of IUPAC and the InChI Trust

The purpose of InChI is...

- To streamline naming conventions for chemical compounds and reactions
- To uniquely identify a chemical substance, without ambiguity, providing a precise, robust, structure-derived tag for chemical substances
- To assist in merging and linking chemical databases

InChI is essential...

as the only structure representation standard in
the public domain, open-source and freely available
to the scientific community

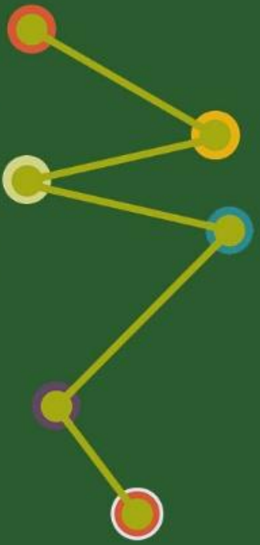
- **Anybody anywhere should be able to produce InChI from just the structural formula of a chemical substance**

- **Normalization to make structures of the same compound drawn under (reasonably) different styles and conventions close if not identical, which is essential for generating the same InChI**

- **Canonicalization of chemical structure upon generating InChI ensures strict uniqueness of the identifier.**

- **The layered structure allows targeting for specific applications (e.g., adding the ability to distinguish tautomers)**

The InChI Timeline



1999/2000
Steve Heller and Steve Stein initiate InChI project at the request of IUPAC

2005
Launch of InChI

2008
A fixed length (27 character) condensed digital representation of the InChI called the InChIKey was developed

2009
Formation of aTrust for a strategic vision and funding for the InChI project

2017
Latest version of InChI software was released along with InChI for reactions

And still more to come...

The InChI project is ongoing; not all of chemistry is yet covered by the software.

The vast majority of organic compounds can be encoded into InChIs, but many inorganic and organometallic compounds are still work in progress.

InChI TRUST

How did we get here?

1999: Steve Heller initiated a proposal for a freely available public domain structure representation standard
2000: Decided that InChI would be an IUPAC initiative
2001: The IUPAC Chemical Identifier project began
2005:version 1 was launched
2009: standard versions of InChI and the InChIKey were released, which took the original algorithm with its many variable parameters and fixed them so that interoperability between databases and resources with InChIs could be achieved
2009: the InChI Trust was formed
2011: version 1.04 released
2017: version 1.05 of the InChI, along with version 1.00 of Reaction InChI (RInChI)

Today publishers have both scientific/chemical journals and chemical databases. Before InChI 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

Why has it worked?

Need
Definition/Specification
Timing/Infrastructure
Acceptance/Use

And a fifth requirement for a standard. First rate staff to create, define, program, and deliver

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,
Former Associate Director for Data Science (Big Data), NIH

Definition/Specification

A computer algorithm to ensure consistency and reproducibility and to be able to call it a real 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. 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 local 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.

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)

"Better a diamond with a flaw than a pebble without."

— Confucius

Timing &Infrastructure

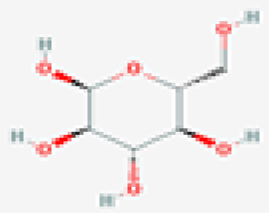
InChI has become a standard only because of the world has changed in the last 20 years.

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 of a dozen people at this talk in China.

What about SMILES as a standard?

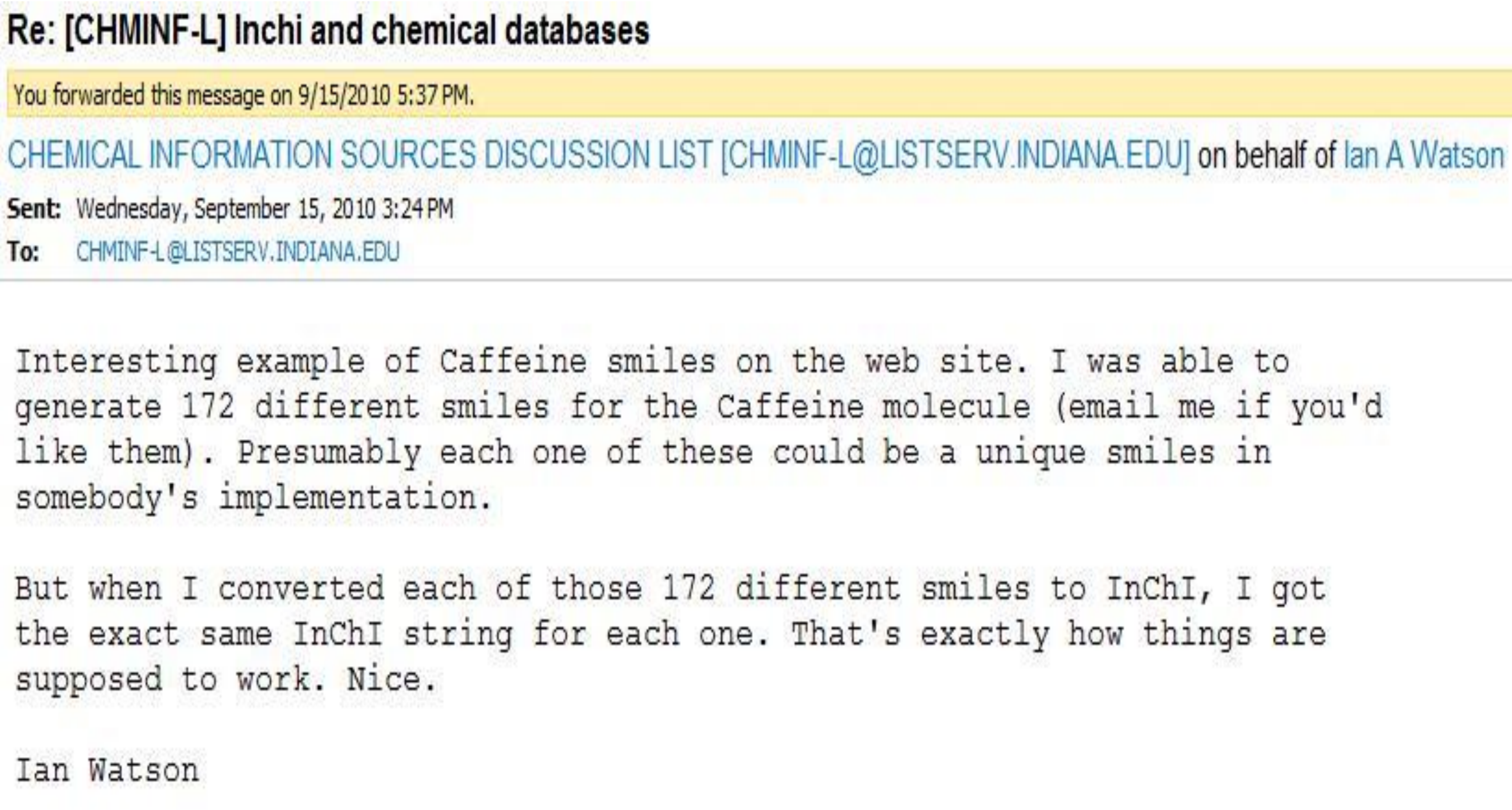
C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O)O)O)O)O



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



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

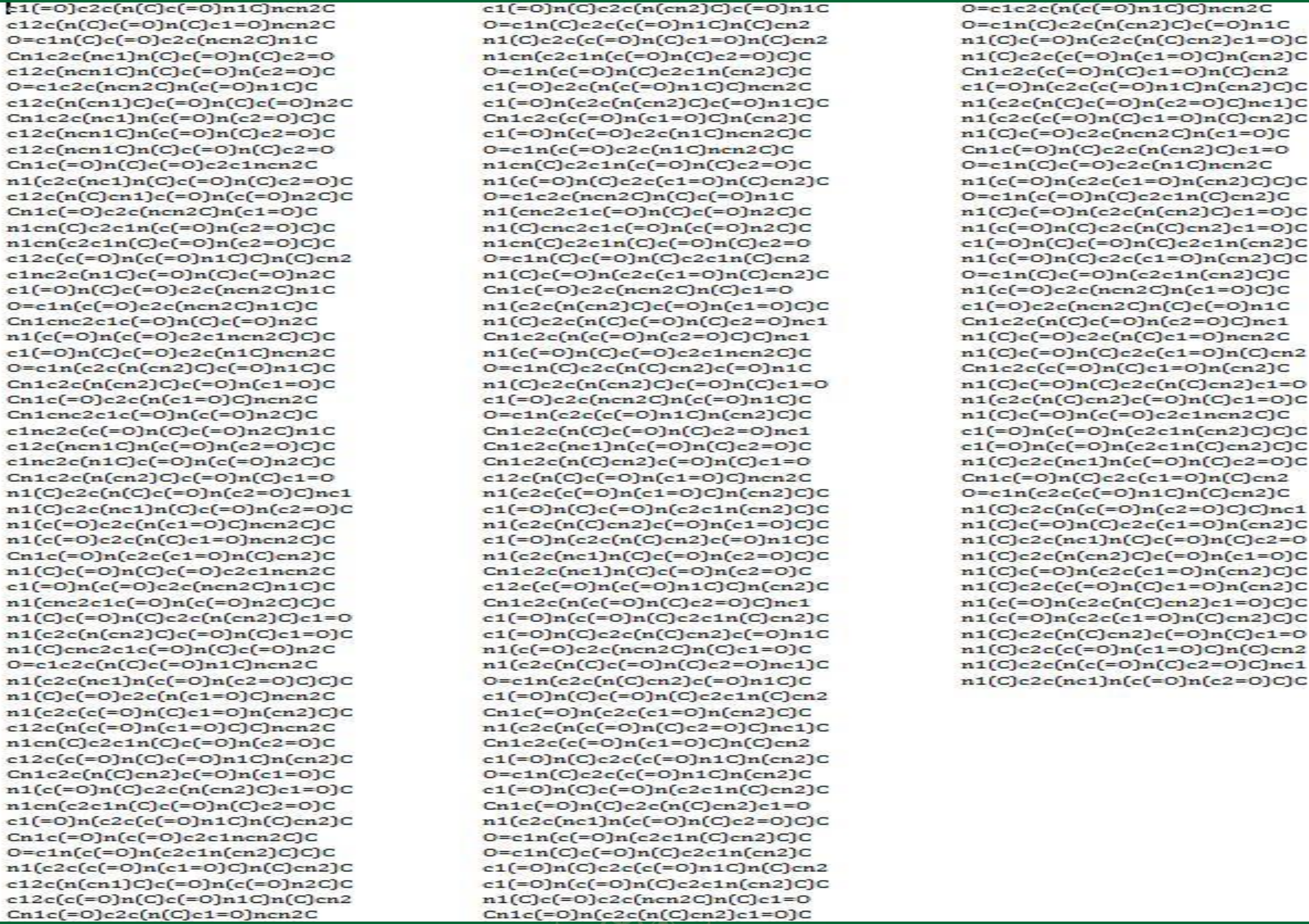
Sent: Wednesday, September 15, 2010 3:24 PM

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
```

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 cover enough of the every day needs of chemists and information specialists then the usage of InChI would not be as great as it is.

Large Real Databases with InChIs/InChIKeys

EBI UniChem –157 million
NIH/NCI/CADD/iRL – 133 million
NIH/PubChem - 97 million
RSC/ChemSpider – 71 million
Elsevier/Reaxys – 31 million
IUPAC – 0 million

Virtual Databases with InChIs/InChIKeys

GDB17 – 166 Billion
NIH/NCI/Argonne/SAVI – 4 Billion

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 chemists is in FINDING and LINKING information

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.

Project Director

The project Director oversees all aspects of the project. The IUPAC InChI subcommittee working groups defining the standards, the programming of these standards, lecturing on InChI, organizing meetings and workings on InChI.

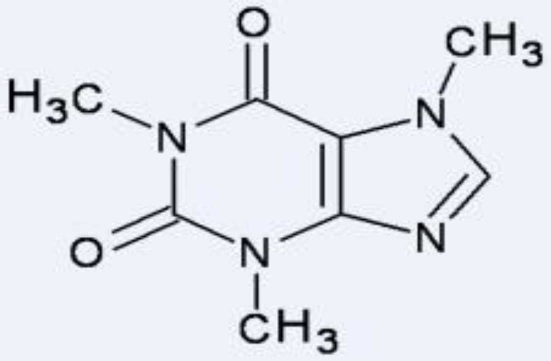
InChI layered structure design

The current InChI layers are:

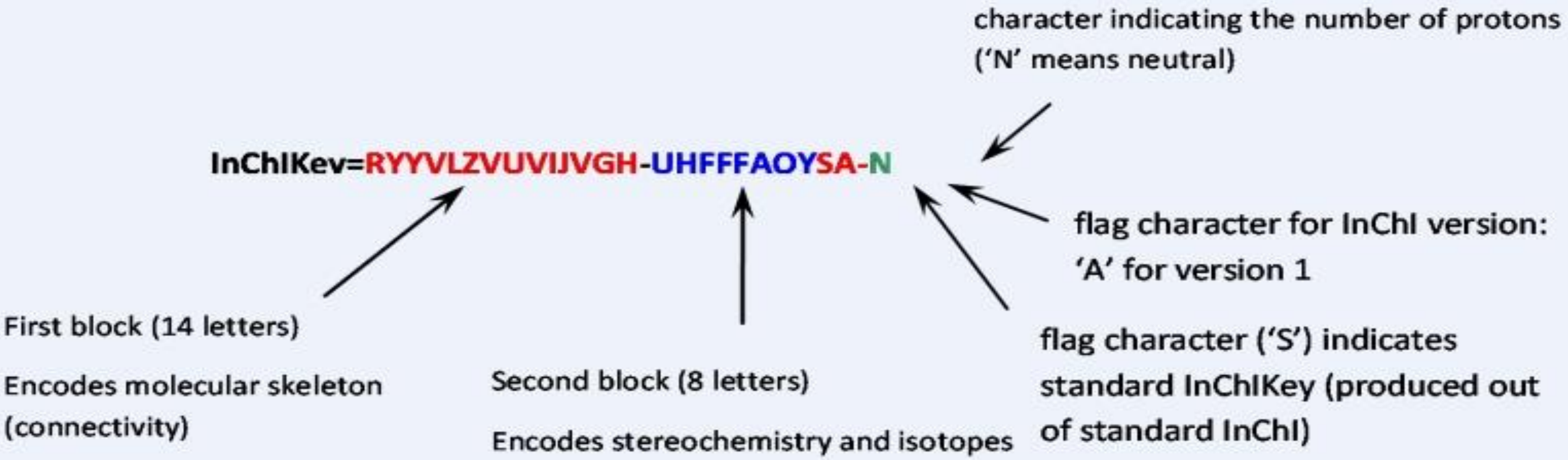
1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond (*Z/E)*
    b. tetrahedral (sp3)
5. Tautomers (on or off)

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



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

InChIKev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

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

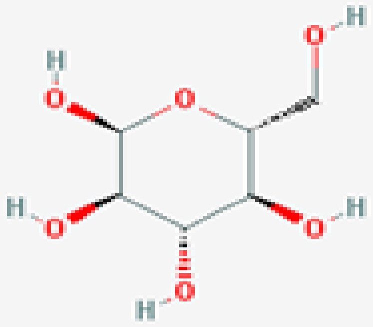
flag character ('S') indicates standard InChIKey (produced out of standard InChI)

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

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

“layered” line notation



alpha-D-Glucose

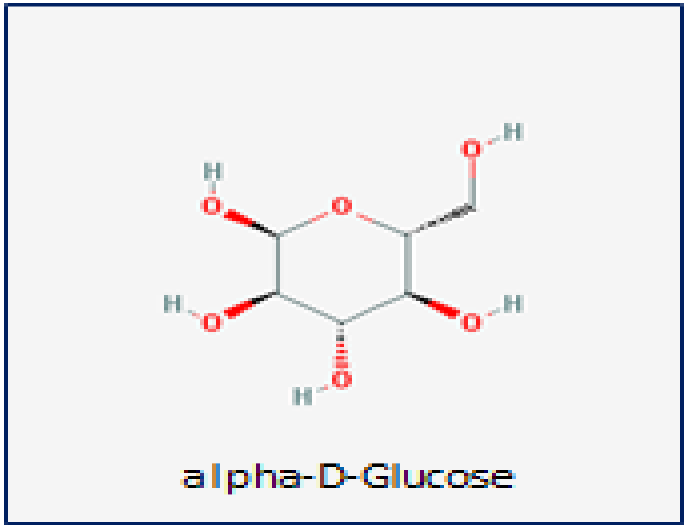
InChIKey is a "hashed" InChI

- 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

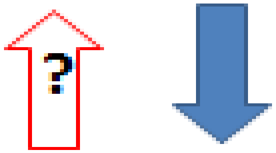
"layered" line notation



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



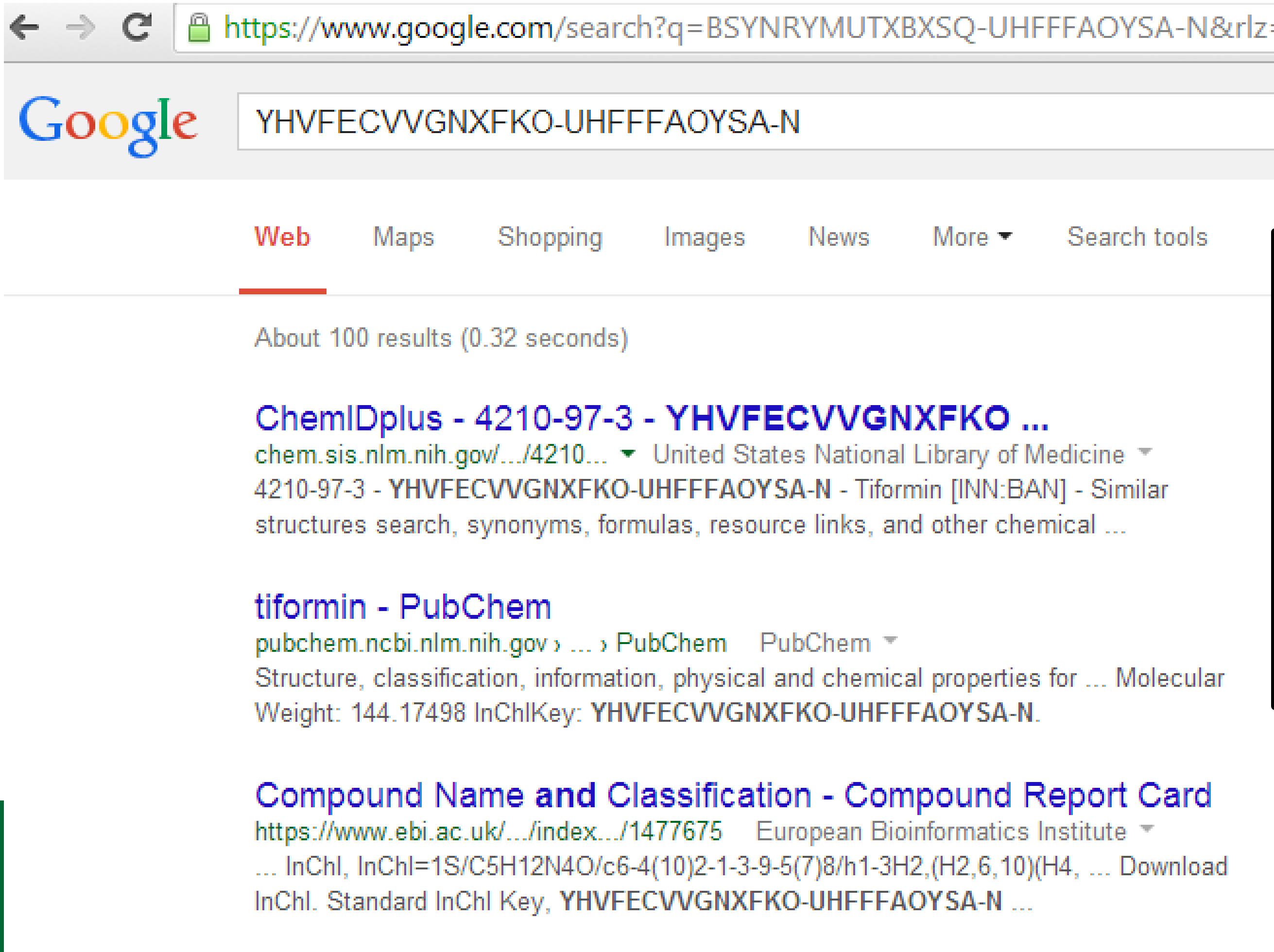
WQZGKKKJIJFFOK-DVKNGEFBSA-N

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

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

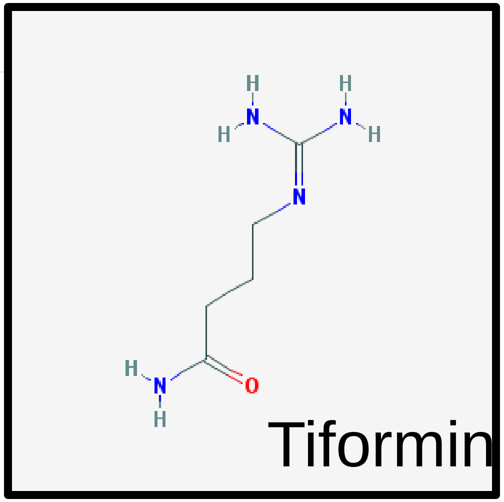
Google YHVFECVVGNXFKO-UHFFFAOYSA-N

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

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

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 formed May 2009

Mission

To deliver and support the implementation of the internationally agreed and widely adopted standard machine-readable chemical identifier, the IUPAC InChI, that enables global connections in chemistry for the advancement of science for the public benefit

Vision

We will have a strong community of InChI advocates and users

We will provide a sustainable organizational framework and the required financial support for the future of the InChI standard

Three strategic pillars

Global adoption and use

Increasing engagement with the chemistry community for the benefit of science and business

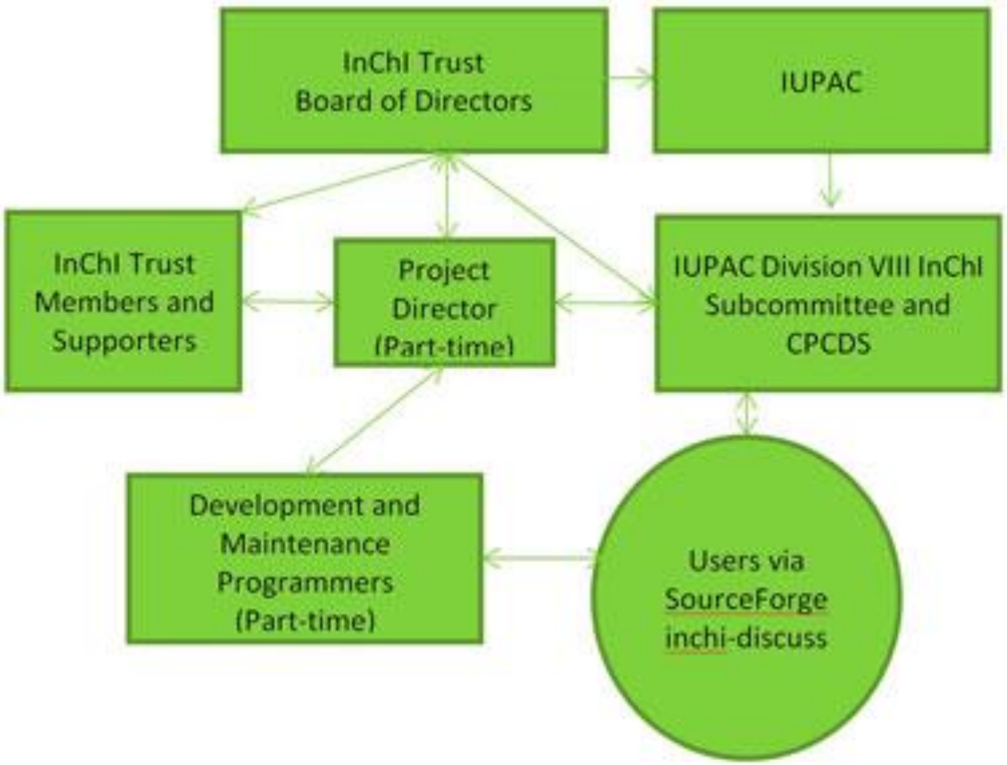
Maintenance & extension of the InChI and applications

To facilitate rapid and effective research discovery and business innovation

Governance

To provide an organizational framework that ensures the sustainability of the standard

Governance


InChI Trust
Board of Directors
IUPAC
InChI Trust
Members and
Supporters
Project
Director
(Part-time)
IUPAC Division VIII InChI
Subcommittee and
CPCDS
Development and
Maintenance
Programmers
(Part-time)
Users via
SourceForge
inchi-discuss

ACD/Labs

ALDRICH Chemistry

BIO-RAD

The Cambridge Crystallographic Data Centre

ChemAxon

CAS
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

IUPAC

FDA

NIH
National Institutes of Health

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

OpenEye

RELX Group

ROYAL SOCIETY OF CHEMISTRY

SPRINGER NATURE

Taylor & Francis Group
an informa business

WILEY
Publishers Since 1807

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
Version 1.05 released – 2017
Polymers
RInChI – InChI for Reactions
New API

Started/To be started

MInChI – InChI for Mixtures
InChI Resolver
QR codes for InChI
InChI teaching/educational materials
Large Molecules/Biopolymers/Macromolecules
Inorganics
Positional Isomers/Variability/Markush
Redesign of Handling of Tautomerism

Reactions - RInChI

First release implemented in Biovia software packages Draw, Direct and Pipeline Pilot. Used in Beilstein supplementary data

Planned enhancements

Additional input & output formats (currently restricted to RXN/RD file format)

Address failing reactions

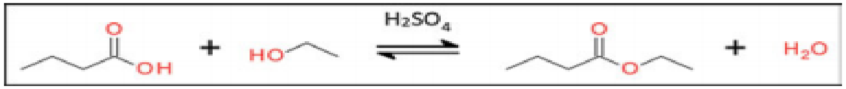
Workarounds for stereochemistry and tautomer restrictions

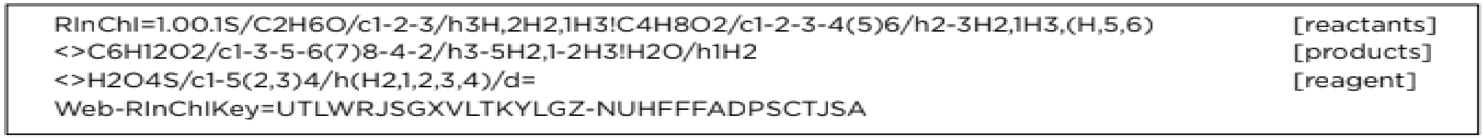
Reaction mapping (MapAuxInfo)

Address needs for big data analysis methods

Reaction properties (ProcAuxInfo)

Class code layer for reaction similarity clustering and pathway optimization, Transform layer for pathway optimization, Reaxys tool





```
RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)      [reactants]
<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3!H2O/h1H2                                    [products]
<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=                                                 [reagent]
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA
```

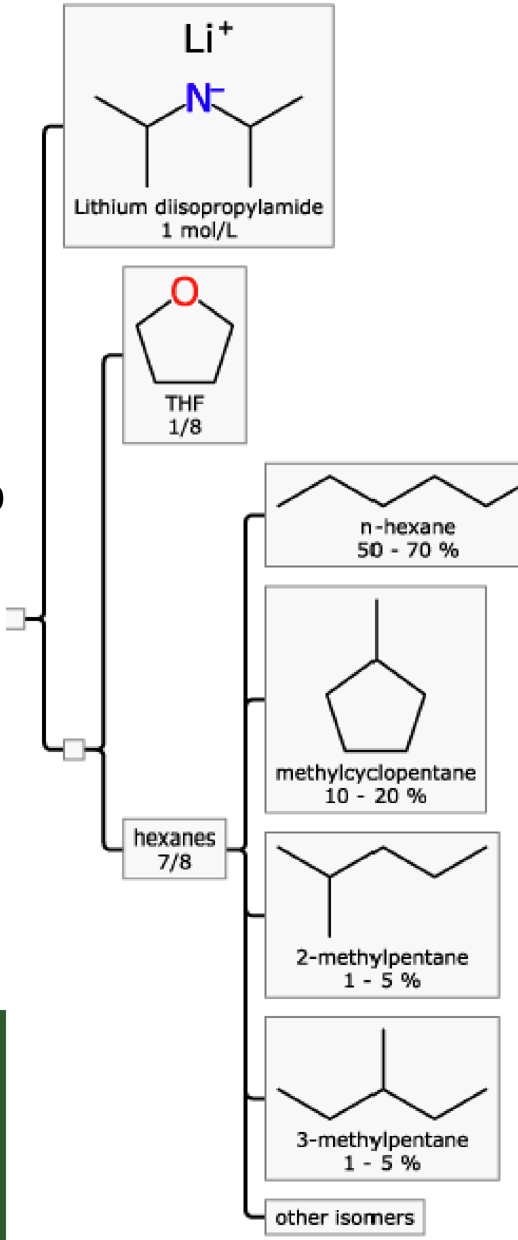
Figure 2. Reaction InChI (RInChI) string for the above reaction. (Image by G. Blanke, "Reaction InChI." InChI Workshop at NIH; Bethesda, MD; 16-18 August 2017.)

Mixtures - MInChI

Phase 1 includes a provisional spec of the notation, a pilot implementation and test case, and description of several target use cases.

The MInChI specification has been implemented as a proof of concept in a Mixture Editor tool currently in development by Collaborative Drug Discovery. This system is designed to parse composition descriptions of mixed substances in a number of forms, including information about components, concentration and mixture hierarchy. https://github.com/cdd/mixtures

The next phase of the project will be to transliterate the codebase to C++ and incorporate it into the RInChI project. This will become the reference implementation version 1.0.



Multi-component system notation

1.7M t-Butyllithium in Pentane:

MInChI=0.00.0S/
C4H9.Li/c1-4(2)3;/h1-3H3;/q-1;+1
&
C5H12/c1-3-5-4-2/h3-5H2,1-2H3&
/n{1&2}
/g{17mr-1&}

37% wt. Formaldehyde in Water with 10-15% Methanol:

MInChI=0.00.0S/
CH2O/c1-2/h1H2&
CH4O/c1-2/h2H,1H3&
H2O/h1H2
/n{{1&3}&2}
/g{{37wf-2&}&10:15vf-2}

- alphabetical order of components
- "**&**" separates components
- "**{}**" mixture groups (e.g., nested)
- "**/n**" indexes components (e.g., order)
- "**/g**" concentration (symbols detailed separately)

InChI for polymers

- Since v. 1.05, InChI supports regular single-strand polymers.
- Both structure-based and source-based representation and encoding of polymers are supported
- Support of polymers is an experimental feature. To emphasize this, InChI/InChIKey for a polymer uses the ‘B’ flag character (for “Beta”), instead of ‘S’ or ‘N’ for standard/non-standard InChI.
- This flag will be replaced by common standard/non-standard conventions if and when InChI for polymers is finally adopted.
- Polymer (‘/z’) layer is a modification layer which is optionally built “above” the other layers and does not affect their content.
- Source-based representation of polymers is based on the chemical structures of the starting material(s) with a special indication that the structure represents a polymer.

Examples

InChI for styrene-butadiene block copolymer, source-based representation:

InChI=1B/C8H8.C4H6/c1-2-8-6-4-3-5-7-8;1-3-4-2/h2-7H,1H2;3-4H,1-2H2/z200-9-12;200-1-8;330-1-12

InChIKey=MTAZNLWOLGHBHU-ZNVYRHKRBA-N 54

InChI for polycaprolactam, structure-based representation:

InChI=1B/C6H10O2/c7-6-4-2-1-3-5-8-6/h1-5H2/z101-1-8(1,2,1,3,2,4,3,5,4,6,5,8,6,8)

InChIKey=PAPBSGBWRJIAAV-CMRMDLKMBA-N

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.

Learn more here

Videos by the InChI Trust:
inchi-trust.org

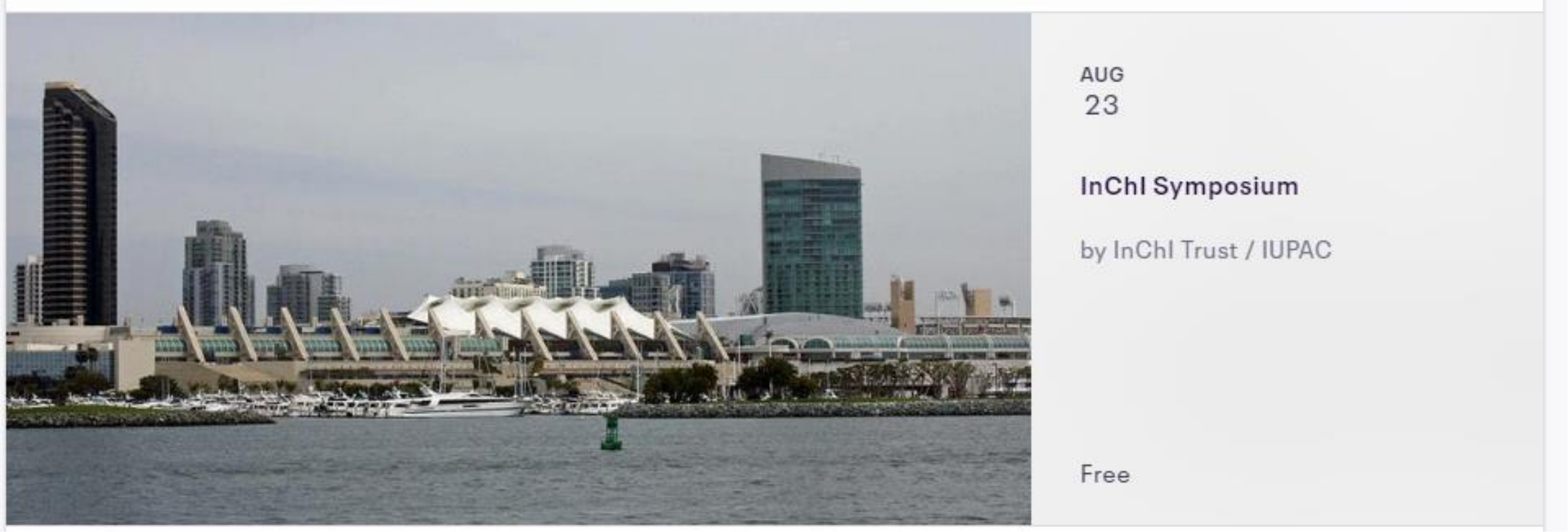
InChI Collection in J Cheminf:
biomedcentral.com/collections/InChI

Many InChIs and quite some feat
by Wendy A. Warr:
link.springer.com/article/10.1007%2Fs10822-015-9854-3
(or https://rdcu.be/M0kk)

Google tech:
youtube.com/watch?v=mpZj4b9elYE

IUPAC page on InChI:
iupac.org/who-we-are/divisions/division-details/inch

InChI TRUST



AUG
23

InChI Symposium

by InChI Trust / IUPAC

Free

State and Future of the IUPAC InChI

23-24 August 2019

San Diego / ACS National Meeting

Signup: www.inchi-trust.org

Summary

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

steve@inchi-trust.org
steve@hellers.com

Acknowledgements

(Current and past members for the IUPAC InChI subcommittee and associated InChI working groups – 5/19)

Steve Bachrach, Colin Batchelor, John Barnard, Bob Belford, Evan Bolton, Ray Boucher, Steve Boyer, Ian Bruno, Steve Bryant, Alex Clark, Szabolcs Csepregi, Rene Deplanque, Josef Eiblmaier, Vincent Scalfani, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Gary Mallard, Leah McEwen, Bill Milne, Hunter Moseley, Moss, Peter Murray-Rust, Heike Nau, Marc Nicklaus, Carmen Nitsche, Matthias Nolte, Steffen Pauly, 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.