The IUPAC InChI Chemical Structure Standard – Today and the Future

Ray Boucher, Stephen Heller, and Richard Kidd

The main web sites for the IUPAC InChI project are:
http://www.iupac.org/inchi
and
http://www.inchi-trust.org

4/9/2019

Slides are available at http://www.hellers.com/steve/ICSDV-4-19.pdf

www.inchi-trust.org

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

www.inchi-trust.org

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 TRUST

InChI Project Goal

To create a free, nonproprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking and finding of data compilations

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

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)**

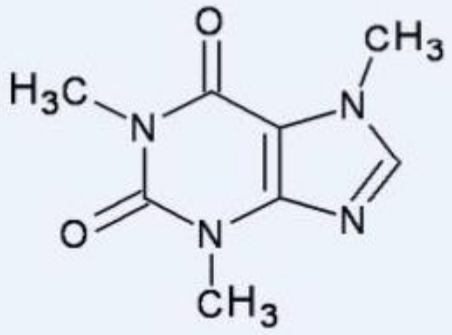
InChI layered structure design

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)

InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N

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

First block (14 letters)

Encodes molecular skeleton (connectivity)

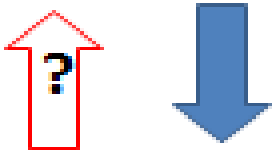
Second block (8 letters)

Encodes stereochemistry and isotopes

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

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

“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

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

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

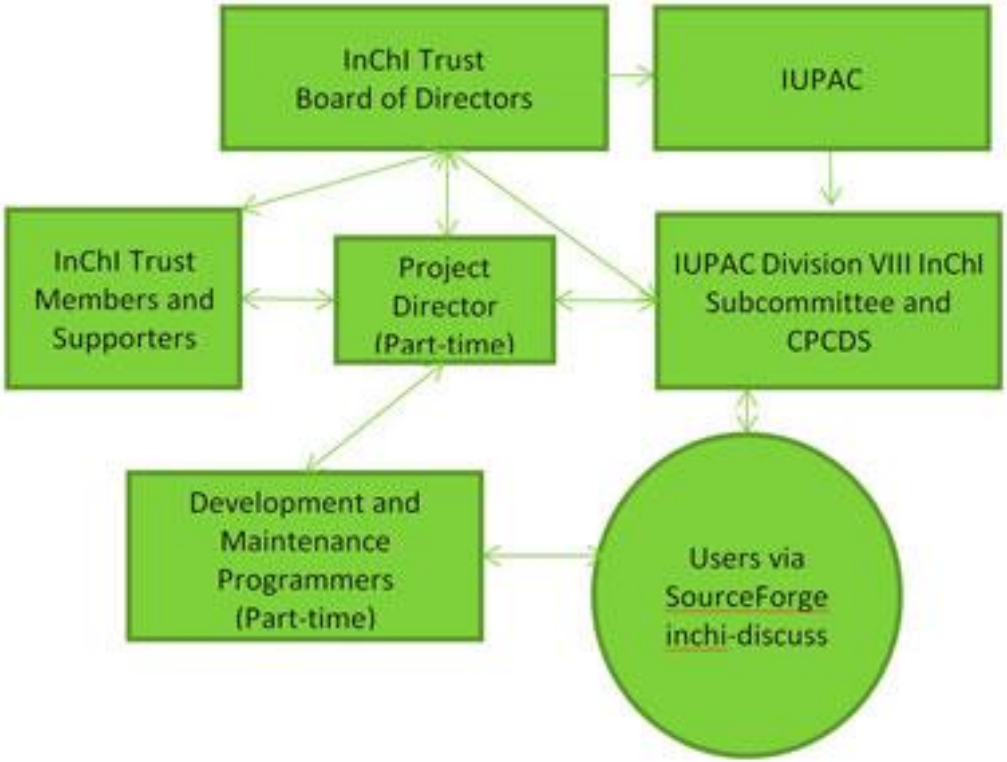
Maintenance & extension of the InChI and applications

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


And where are we now?
Where are we going?
What should we do?
What can we do?
How far? Extension *vs* application?
Do we widen the coverage, to new areas (and areas well outside traditional chemistry)?

Extension

Organometallics
More Complete Tautomerism
Inorganics
Positional isomers
Polymers
Large molecules
Markush

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

Application

Reactions
Mixtures
InChI Resolver
QR codes for InChI
InChI open educational resources

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

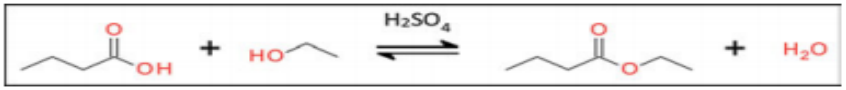
Reaction mapping (MapAuxInfo)

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	

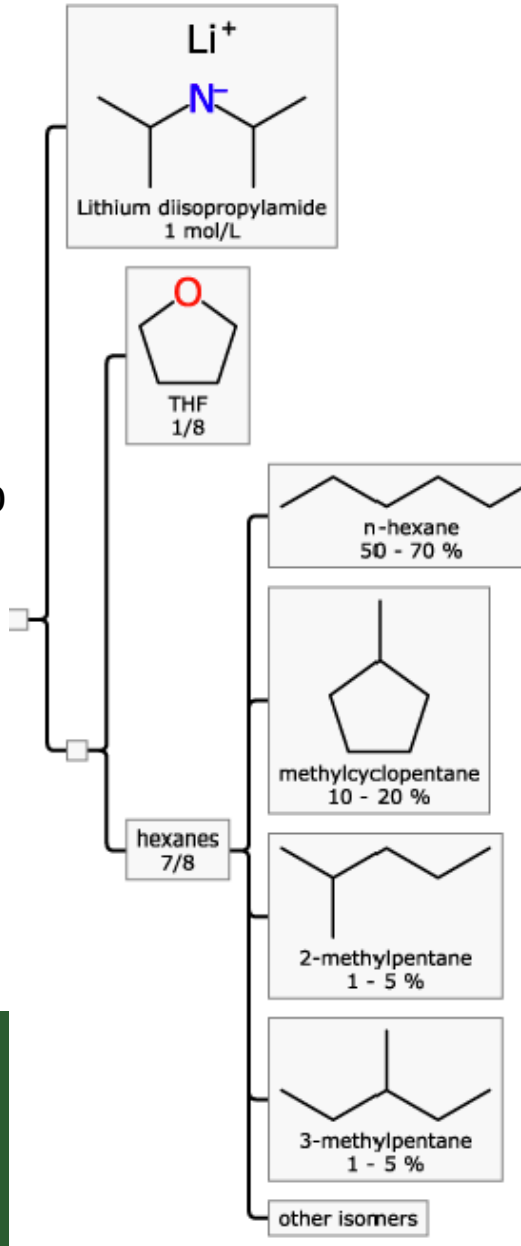
Figure 2. Reaction InChI (RInChI) string for the above reaction. (Image by G. Blanke, "Reaction InChI." InChI Workshop at NIH; Bethesda, MD; 16-18 August 2017.)

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.



QR Codes

InChI QR Code: Standard Form

IUPAC-InChI

[database]/[type]/[identifier]

[Source URL]/InChIKey/[Key]

1. Institution incorporates into local app
2. Vendor generated and linked to their database
3. Public version (QRInChI.org for re-direction e.g. to info.identifiers.org)

How standard?

We have the standard InChI & InChIKey

We have multiple experimental and optional flags

People use InChI in ways that were not intended, and to experiment. To what extent should we actively encourage this? Is it an identifier or a representation?

How to support a standard?

No written standard
Software is the implementation
Centralised expert resource

How should we encourage the best open source practices?
Community input and contributions while maintaining a standard

Governance and funding
Participation through time and money

Education and support

This OER initiative, IUPAC project 2018-012-3-024, is being created to provide a resource on InChI related resources to assist practicing scientists and educators in learning about and benefiting from the use of InChI.

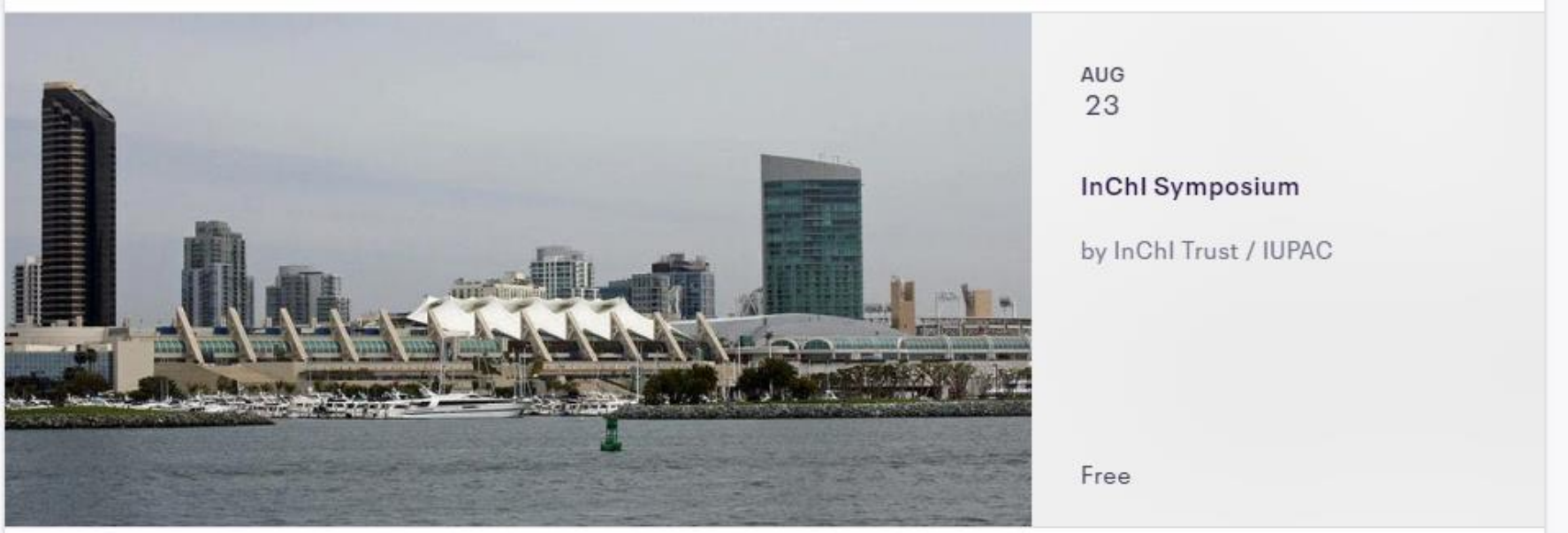
https://www.inchi-trust.org/oer/

Rolling workshops

Mar 2017 – EBI Hinxton
Aug 2017 – NIH Bethesda

Aug 2018 – Boston MA

Feb 2019 – Cambridge UK
Aug 2019 – San Diego CA



State and Future of the IUPAC InChI

23-24 August 2019

San Diego / ACS National Meeting

Signup: www.inchi-trust.org

InChI TRUST

www.inchi-trust.org

Be part of the InChI community...

Help build the chemical web and encourage others to do the same!

Explore more than **90 000 000** chemical compounds with an InChI here:
pubchem.ncbi.nlm.nih.gov

Explore more than **156 000 000** chemical structures here:
ebi.ac.uk/unichem

Join the discussion **HERE**

InChI TRUST

Learn more here

Videos by the InChITrust:
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

InChITRUST

Acknowledgements

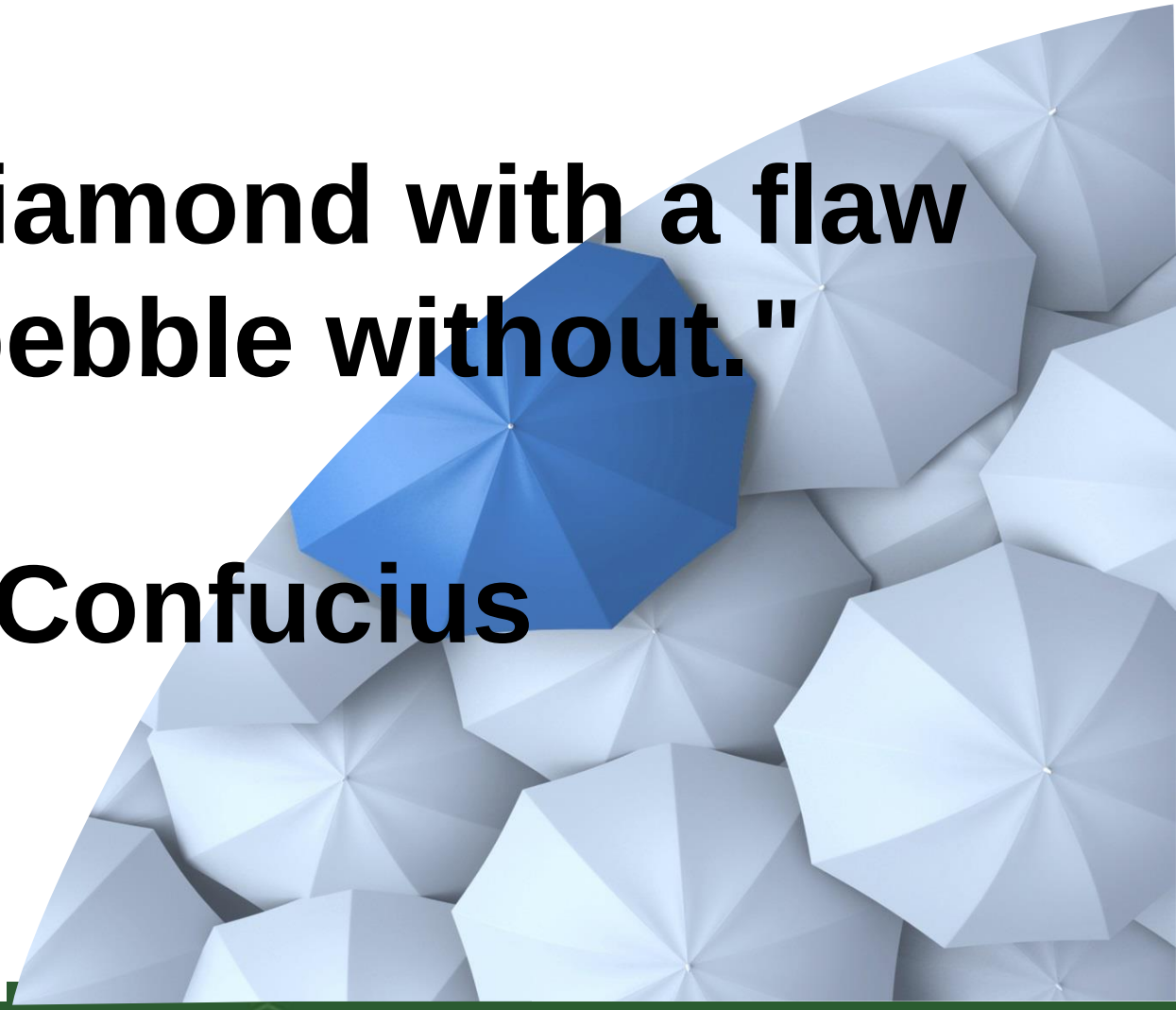
(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Greg Banik, John Barnard, Colin Batchelor, Bob Belford, Gerd Blanke, Evan Bolton, Ray Boucher, Steve Boyer, Ian Bruno, Steve Bryant, Alex Clark, Szabolcs Csepregi, Andrew Dalke, Rene Deplanque, Josef Eiblmaier, David Evans, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec, Richard Kidd, Hans Kraut, Alexander Lawson, Peter Linstrom, Gary Mallard, John Mayfield, Leah McEwen, Bill Milne, Hunter Moseley, 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 Sayle, Vincent Scalfani, 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 and others

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

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

— Confucius



Questions?

InChI TRUST