

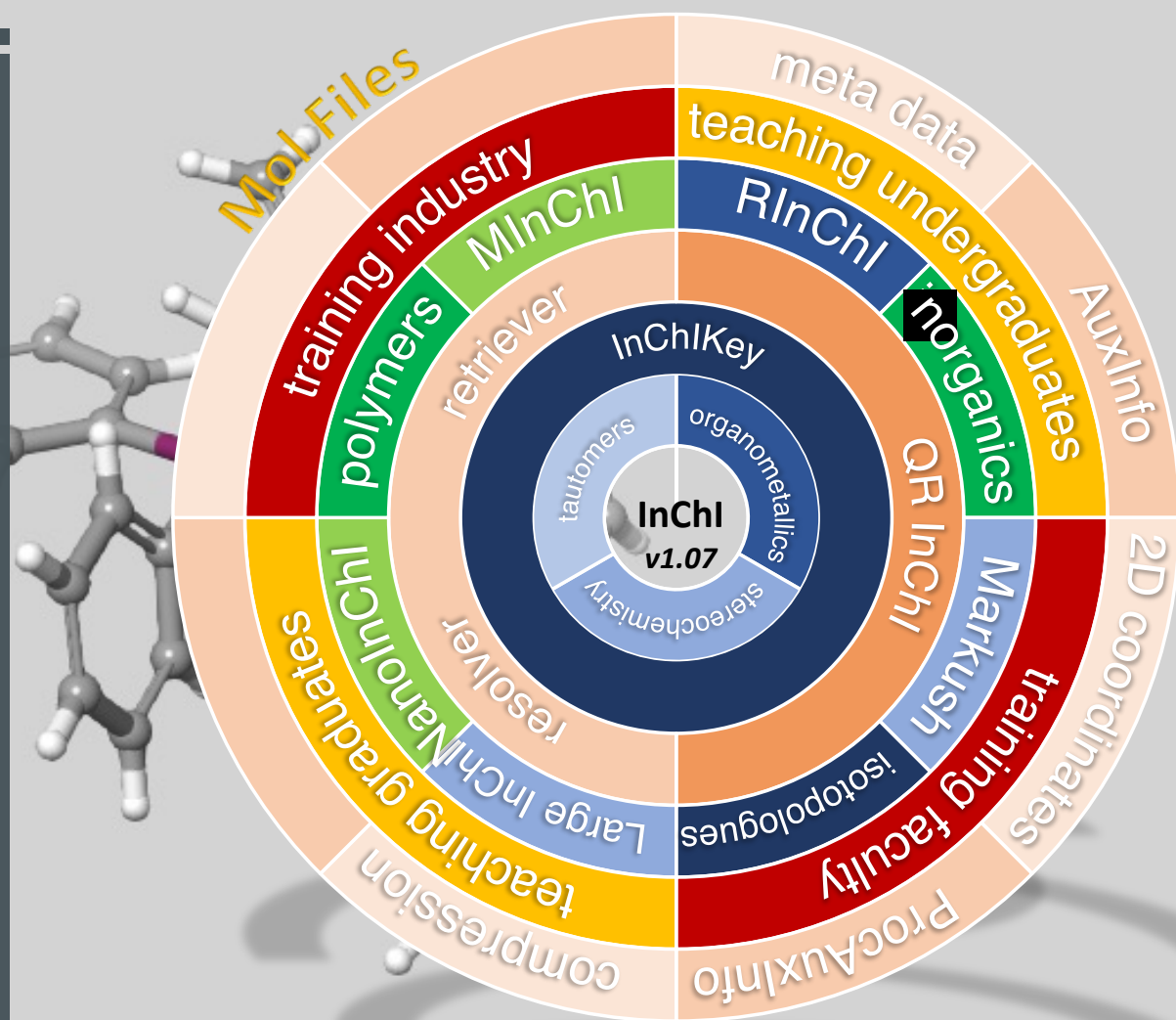
# IODIDE



INCHI OVERVIEW DISCUSSION  
IN DENVER FOR EVERYBODY

SATURDAY AUGUST 17<sup>TH</sup> 2024

DENVER



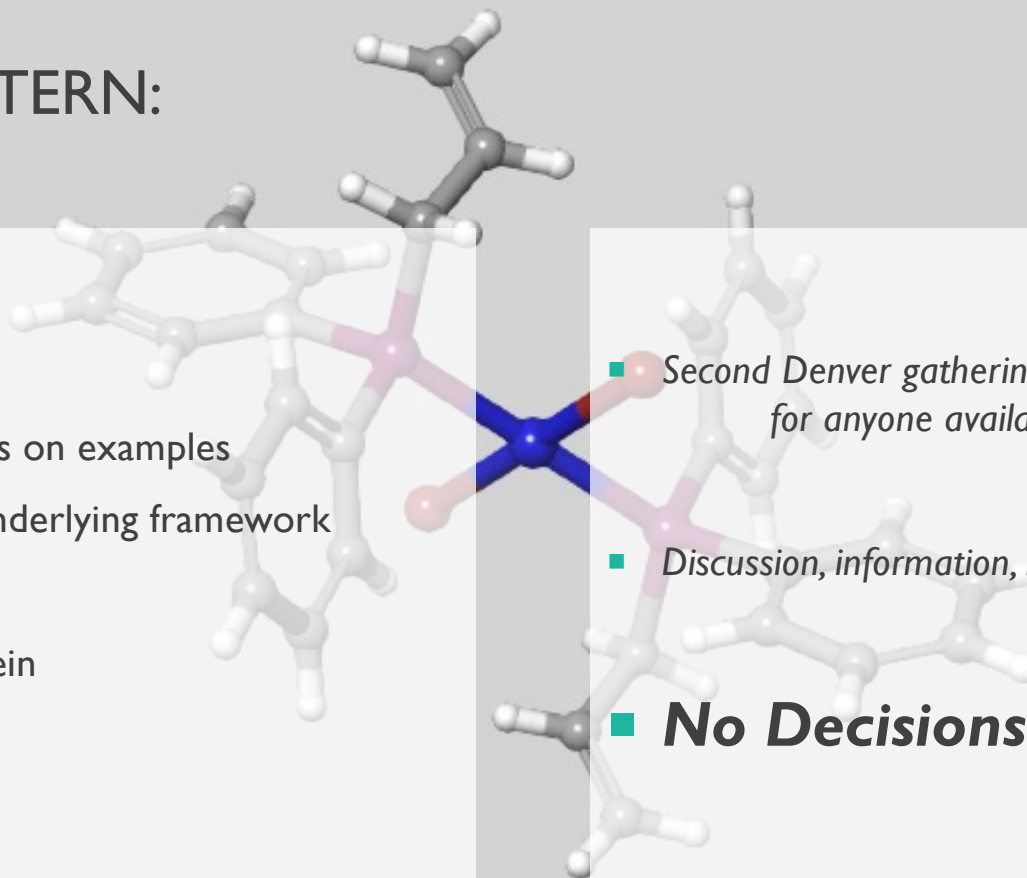
## WORKING PATTERN:

- 9:00 Overview - focus on examples
- 9:30 (RMN)InChI - underlying framework
- 10:30 break
- 10:45 InChI and Beilstein

■ *Second Denver gathering in next few days  
for anyone available*

■ *Discussion, information, issues*

■ **No Decisions**



## INCHI TECHNICAL PAPERS

*These actually exist!*

- <https://github.com/IUPAC-InChI>
- QRInChI:  
Frey, Jeremy G., Hartshorn, Richard M. and McEwen, Leah R..  
"Specification of International Chemical Identifier (InChI) QR codes for linking labels on containers of chemical samples to digital resources (IUPAC Recommendations 2021)" Pure and Applied Chemistry, vol. 94, no. 10, 2022, pp. 1195-1206. <https://doi.org/10.1515/pac-2021-0604>
- Isotopologues:  
Moseley, H.N.B., Rocca-Serra, P., Salek, R.M. et al. InChI isotopologue and isotopomer specifications. J Cheminform 16, 54 (2024). <https://doi.org/10.1186/s13321-024-00847-8>
- InChI / SMILES+ comparison appeared in Summer 2024 CICAG newsletter (Jonathan Goodman / Vin Scalfini) <http://www.rscicag.org/newsletters.htm>
- Herres-Pawlis S, Blanke G, Brammer J, Baljovic D, Khan N, Lange F, et al. Making the InChI FAIR and sustainable by moving to open-source on GitHub. ChemRxiv. 2024; doi:10.26434/chemrxiv-2024-w6kws

## INCHI TECHNICAL PAPERS

*Still under development...*

- “Think like an InChI” is now planned and partially written (Jonathan Goodman)
  - Outline of the InChI algorithm, emphasizing canonical numbering
  - Highlight how difficult this is and the edge cases where it is hard
  - Illustrate how the InChI can be used to curate a database of molecules
- The InChI Algorithm: desirable, but at an early stage
  - This cannot be written unless we know exactly what the algorithm does
- A paper on InChI Molecular Inorganics has been outlined and writing has begun (Sonja Herres-Pawlis / Gerd Blanke)
- Vin Scalfini, Jonathan Goodman and others?: PAC technical article on InChI vs. SMILES, use-cases, etc
- Wendy Warr and Jonathan Goodman considering a perspective
- Denver ACS presentations: next two slides



# InChI: what is it and what will it be?

Gerd Blanke,<sup>1</sup> Jonathan M. Goodman<sup>2</sup>

1. StructurePendium Technologies GmbH, Essen, Germany;  
2. Yusuf Hamied Department of Chemistry, University of Cambridge, CB2 1EW, UK

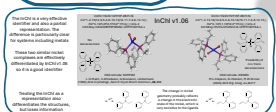
The InChI is the international chemical identifier, providing dependable, canonical names for molecular structures.

The latest release, InChI v1.07, [1] has security enhancements and refactored code, building on v1.06. [1]

The code is all available on GitHub. [1]

Work on the InChI continues

extended stereochemistry  
tautomers  
organometallics  
molecular inorganics  
polymers  
nanomaterials  
isotopologues  
macromolecules  
Applications are addressing new areas of InChI-based data-handling, Mixture-InChI, Markush-InChI, Nano-InChI, Reaction-InChI



Try it out with the new Web Demo!

InChI v1.07

Try it now!

<https://iupac-inchi.github.io/InChI-Web-Demo/>

Join the discussion on:

<https://github.com/IUPAC-InChI/InChI/issues>

InChI can also be installed in a Docker container. The files for this are included in the release: [https://github.com/IUPAC-InChI/InChI/releases/tag/InChI-1-TEST\\_210](https://github.com/IUPAC-InChI/InChI/releases/tag/InChI-1-TEST_210)



2. J. M Goodman, I. Pletnev, P Thiessen, E. Bolton and S. R. Heller. InChI version 1.06: now more than 99.99% reliable. *J. Cheminform.* 2021, **13**, 40.



The InChI is an

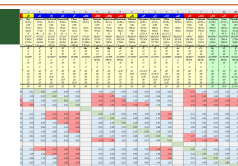
# InChI for inorganic chemists: organometallics and molecular inorganics

Gerd Blanke<sup>1</sup>, Sonja H. Herres-Pawlis<sup>2</sup>, Jonathan M. Goodman<sup>3</sup>, Ulrich Schatzschneider<sup>4</sup>, Andrey Yerin<sup>5</sup>, Richard Hartshorn<sup>6</sup>, Felix Bänsch<sup>7</sup>, Nauman Khan<sup>2</sup>, Djordje Baljovic<sup>2</sup>, Jan Brammer<sup>2</sup>

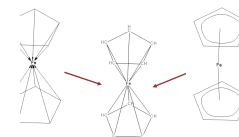
1. StructurePendium Technologies GmbH, Essen, Germany; 2. Institute of Inorganic Chemistry, RWTH Aachen University, Aachen, Germany; 3. Yusuf Hamied Department of Chemistry, University of Cambridge, CB2 1EW, UK; 4. Institute of Inorganic Chemistry, Julius Maximilians University Würzburg, Germany; 5. Advanced Chemistry Development Inc; 6. University of Canterbury, New Zealand; 7. Beilstein Institut, Frankfurt/Main, Germany

The standard InChI, v1.07, was developed with organic molecules as the focus. The algorithm generates InChI for organometallics

Disconnection Check Table



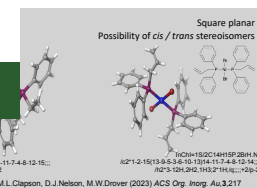
by one InChI but may be represented by sw InChI program will interpret diverse netallics. Structure depictions of the same he bond types used for the chemical le, and triple bonds, other bond types like input structure depictions must be processed of unique InChIs. Hydrogen atoms directly be drawn explicitly. Guidance on constructing



ChI=1/2C5H5.Fe/c2\*1-2-4-5-3-1; 2\*1-5H;/rC10H10Fe/c1-2-4-5-1)11(1,2,4,5)6-7(11)9(11)10(11)6)11/h1-10H

<https://github.com/IUPAC-InChI/InChI/issues>

ognition of InChI must be extended to the tips in inorganic compounds. While is may lead up to 2n stereoisomers (with n = s), just one octahedral centre leads up to 30



# Dynamic InChI: Canonical, unique and on-the-fly Symposium: Open Source Chemoinformatics Resources

Gerd Blanke,<sup>1</sup> Frank Lange,<sup>2</sup> Sonja H. Herres-Pawlis,<sup>2</sup> Jonathan M. Goodman<sup>3</sup>

1. StructurePendium Technologies GmbH, Essen, Germany;
2. Institute of Inorganic Chemistry, RWTH Aachen University, 52074 Aachen, Germany;
3. Yusuf Hamied Department of Chemistry, University of Cambridge, CB2 1EW, UK

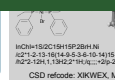
InChI TRUST

TEST THE INCHI DEVELOPMENTS



REFERENCES:

1. InChI Source Code: <https://github.com/IUPAC-InChI/InChI>
2. Making the InChI FAIR and sustainably by moving to open-source on GitHub: <https://chemrxiv.org/engage/chemrxiv/article-details/6a9584301103d79c547b086>



InChI=1S2C18H18N2O4  
InChI=1S2C18H18N2O4  
InChI=1S2C18H18N2O4  
CSD refcode: XRWEX, M.L. Capson, D.J. Nelson, M.W. Drover (2023) ACS Org. Inorg. Au.3.217

# WHERE IS THE INCHI NOW?

## ■ Developments

Organometallics

Inorganic molecules

Atropisomers

Extended stereochemistry

Isotopologues

Tautomers

Polymers (*? May need a registration authority ?*)

■ *[[ InChI and InChIeR? ]]*

*[[ (InChI everything Registration) ]]*

## ■ Applications

■ RInChI

■ MInChI

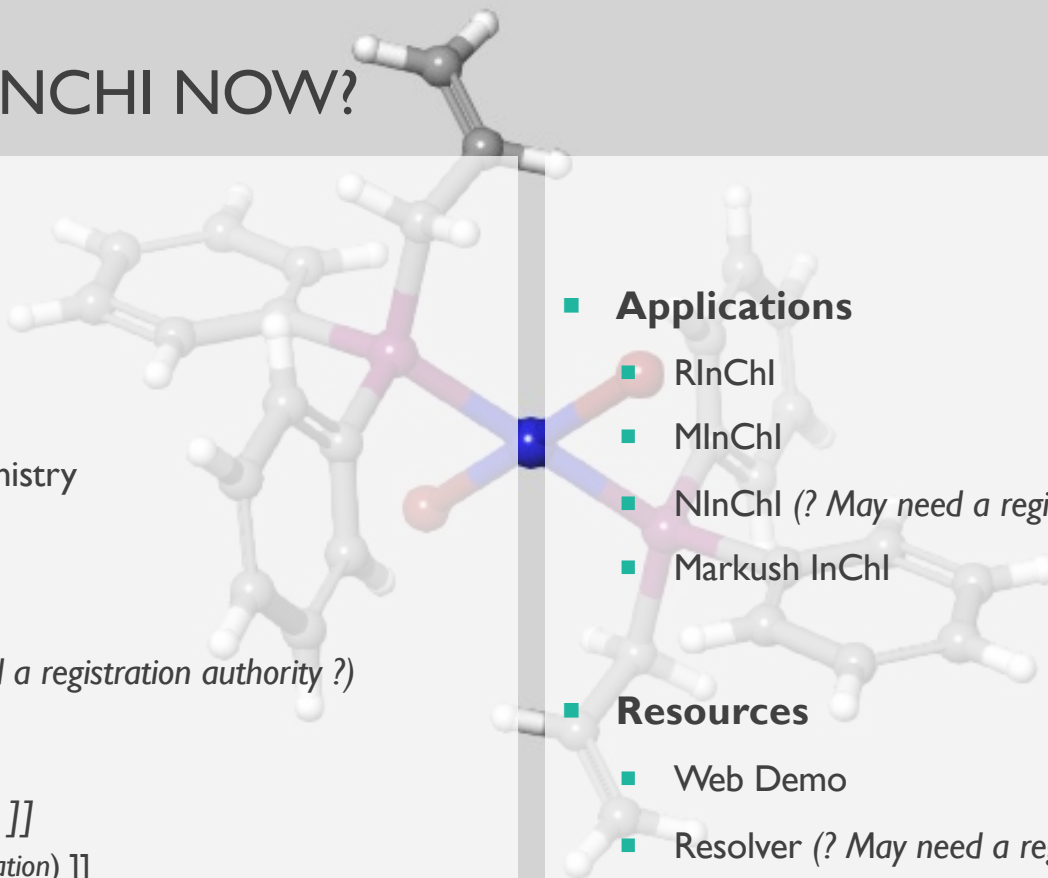
■ NInChI (*? May need a registration authority ?*)

■ Markush InChI

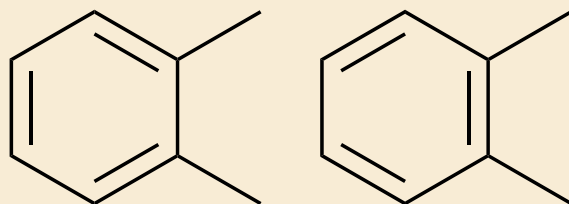
## ■ Resources

■ Web Demo

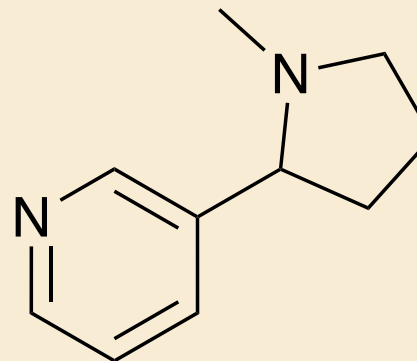
■ Resolver (*? May need a registration authority ?*)



# Drawing molecules is difficult and not perfect for organics



o-xylene



nicotine

*Much harder for  
organometallics  
and inorganics*

*Pure Appl. Chem.*, Vol. 78, No. 10, pp. 1897–1970, 2006.  
doi:10.1351/pac200678101897  
© 2006 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY  
CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION\*

## GRAPHICAL REPRESENTATION OF STEREOCHEMICAL CONFIGURATION

(IUPAC Recommendations 2006)

*Prepared for publication by*  
JONATHAN BRECHER

CambridgeSoft Corporation, 100 CambridgePark Drive, Cambridge, MA 02140, USA

*Pure Appl. Chem.*, Vol. 80, No. 2, pp. 277–410, 2008.  
doi:10.1351/pac200880020277  
© 2008 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY  
CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION\*

## GRAPHICAL REPRESENTATION STANDARDS FOR CHEMICAL STRUCTURE DIAGRAMS\*\*

(IUPAC Recommendations 2008)

*Prepared for publication by*  
JONATHAN BRECHER

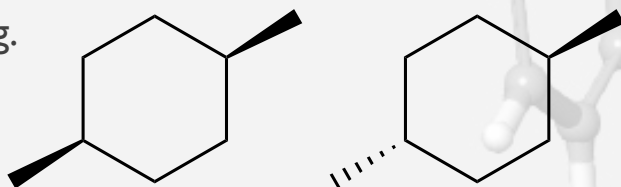
CambridgeSoft Corporation, 100 CambridgePark Drive, Cambridge, MA 02140, USA



## STEREOCHEMICAL LAYERS:

/T, /M, /S

The /t layer lists the tetrahedral stereogenic centres and labels them “+” or “-” depending on the configuration calculated using the canonical numbering.



Meso compounds and racemic mixtures are common, and do not have absolute stereochemistry.

How best to construct an InChI for these? If all of the “+” in the /t layer are switched for “-”, and all the “-” for “+”, the /t layer would look very different, but the substance would be the same.

Which should be the canonical InChI?

The /m and /s layers are a way to make the choice

## CAMPAIGN FOR REAL MOLECULES (CARM)

- We can imagine arbitrary geometrical constructs which are hard to distinguish

- *Keep CaRM*

- **The InChI need only distinguish real molecules**

- The InChI is a *representifier*<sup>TM</sup>  
(an incomplete representation and a good identifier)

- What are the example molecules?

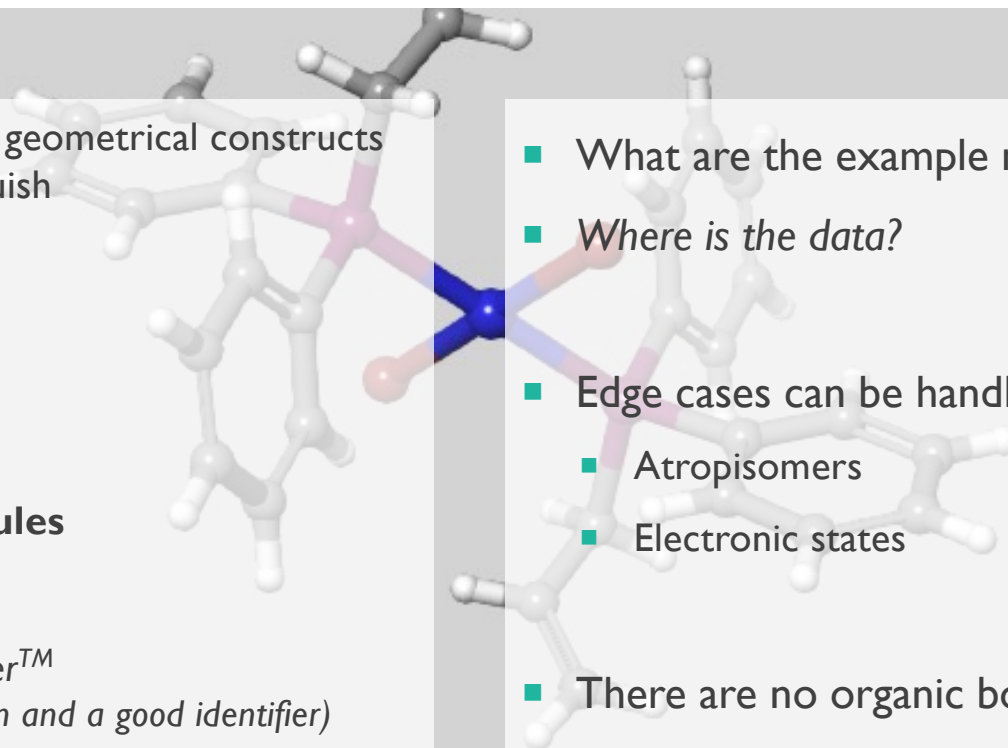
- *Where is the data?*

- Edge cases can be handled later

- Atropisomers

- Electronic states

- There are no organic bond-stretch isomers



## CAMPAIGN FOR REAL MOLECULES (CARM)



- **Always find examples**
- Molecules are small
- There are not many of them
- Not all geometric subtleties are important for identification
- What is the same and what is different?
- *Keep CaRM*
- Examples should be a pair of molecules that illustrate a distinction which needs to be made
- Not enough to have an example of a feature: also need an example of another molecule that would be identical but for this feature.

## CAMPAIGN FOR REAL MOLECULES (CARM)



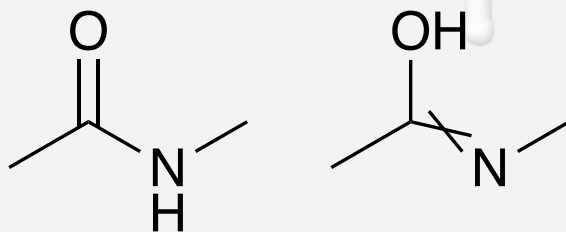
- **InChIKey**
- *Are they long enough?*
- **Absolute on difference:**
  - *A different InChIKey means certainty: a different molecule*
- **High probability on identity:**
  - *The same InChIKey means high probability: the same molecule*
- However long the InChIKey there will always be a finite possibility of a collision
- Must not assume there will be no collisions for practical purposes
- What is an acceptable collision rate?

## CAMPAIGN FOR REAL MOLECULES (CARM)

- **Tautomers**

- *The InChI is not designed to be human readable*

- *It is successful*



- Which tautomer do you prefer?

- Much agreement on secondary amides (and that ChemDraw gets this wrong)

- For many molecules, rather user dependent

- Lots of possible tautomers

- *Need a non-InChI program to learn each individual's preferences?*

# CAMPAIGN FOR REAL MOLECULES (CARM)



## ■ Developments

Organometallics

Inorganic molecules

Atropisomers

Extended stereochemistry

Isotopologues

Tautomers

Polymers (*? May need a registration authority ?*)

- *[[ InChI and InChIeR? ]]*
- *[[ (InChI everything Registration) ]]*

## ■ Applications

■ RInChI

■ MInChI

■ NInChI (*? May need a registration authority ?*)

■ MarkushInChI

## ■ Resources

■ Web Demo

■ Resolver (*? May need a registration authority ?*)

## NOTES:

### ■ InChI

The InChI is free and open source. Anyone with a molecular structure can generate a canonical InChI without charge.

### ■ It is might also be useful to have

- Identifiers for substances without known or defined molecular structures
- **InChIeR** – a property and partial-structure derived identifier

### ■ The InChIeR

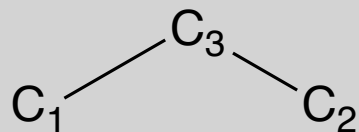
- Uniqueness can be guaranteed only with a registration step
- Registration identifier based on properties, so similar substances should have similar InChIeR
- Registration process will include check that something similar has not been registered before
- Initiate with well-known no-structure substances
- Registration could be charged

<https://iupac-inchi.github.io/InChI-Web-Demo/>

INCHI

Propane:

InChI=1S/C3H8/c1-3-2/h3H2,1-2H3





# IODIDE

*A few notes...*

- What do you do with incompletely characterized molecules?

- <https://doi.org/10.3390/metabo11070431>

- MInChI is also grappling with incompletely characterized components

- Identifier vs Representation

- The InChI is an identifier

- MInChI (and other things) are more descriptive

- Use cases are important!

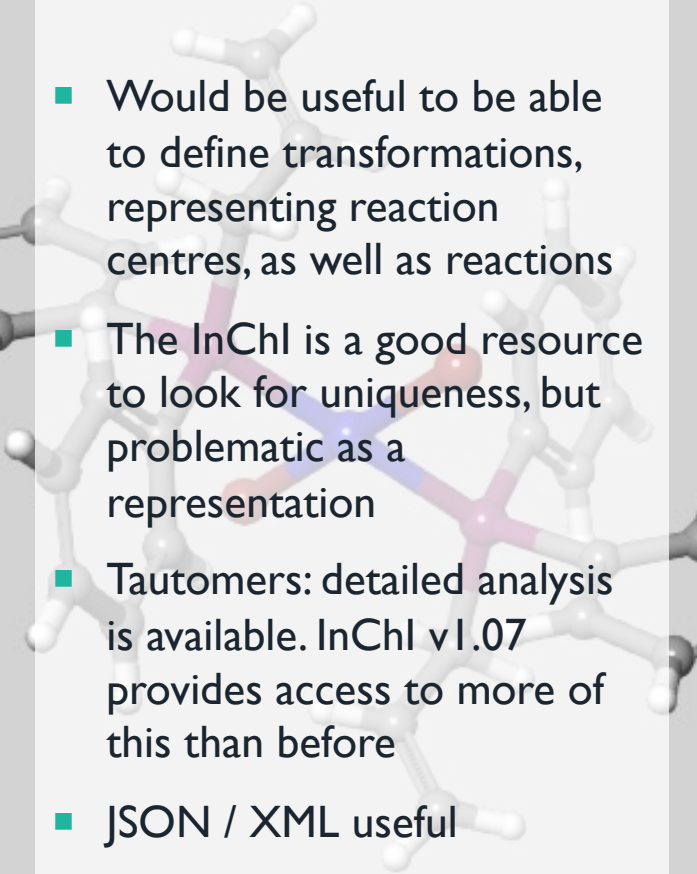
- It would be helpful to be able to describe families of reactions. RInChI, MInChI and MarkushInChI together might be able to do this.

- Integrating metabolic databases is a major challenge

- RInChI does not have atom mapping, but this is planned for a future release. Determining what the atom mapping is, particularly when there is uncertainty, is challenging.

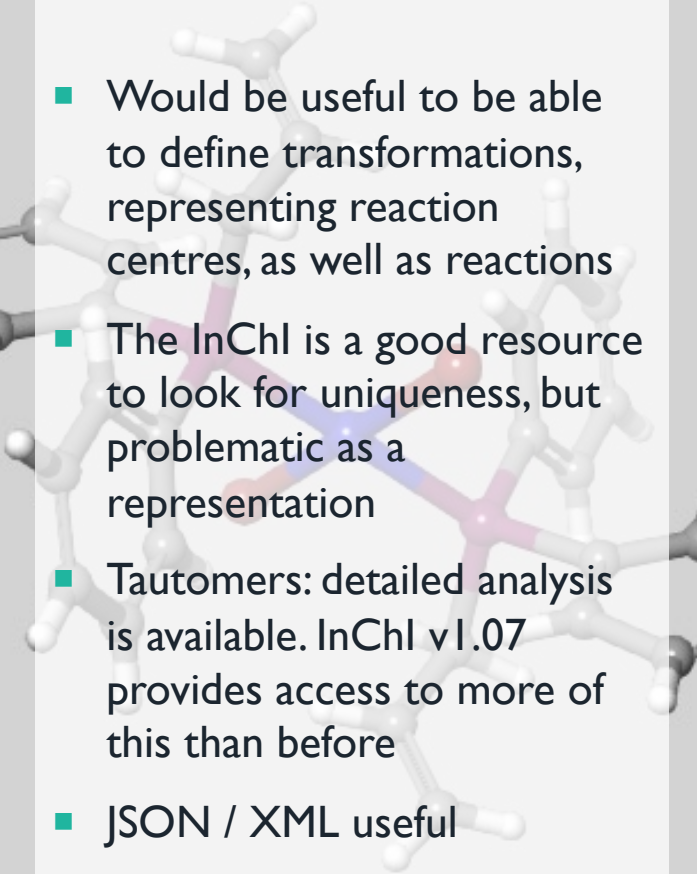
# IODIDE

*A few more notes...*

- 
- Would be useful to be able to define transformations, representing reaction centres, as well as reactions
  - The InChI is a good resource to look for uniqueness, but problematic as a representation
  - Tautomers: detailed analysis is available. InChI v1.07 provides access to more of this than before
  - JSON / XML useful
  - The Reaction InChI, Mixture InChI and Nano InChI are all rather different, but might benefit from a common code base: what do they have in common?
  - These need more information than just molecular structure, so data needed to supplement molfile input
  - An SD File can associate a molfile with other data
  - Different communities have different protocols: a modular approach may be helpful

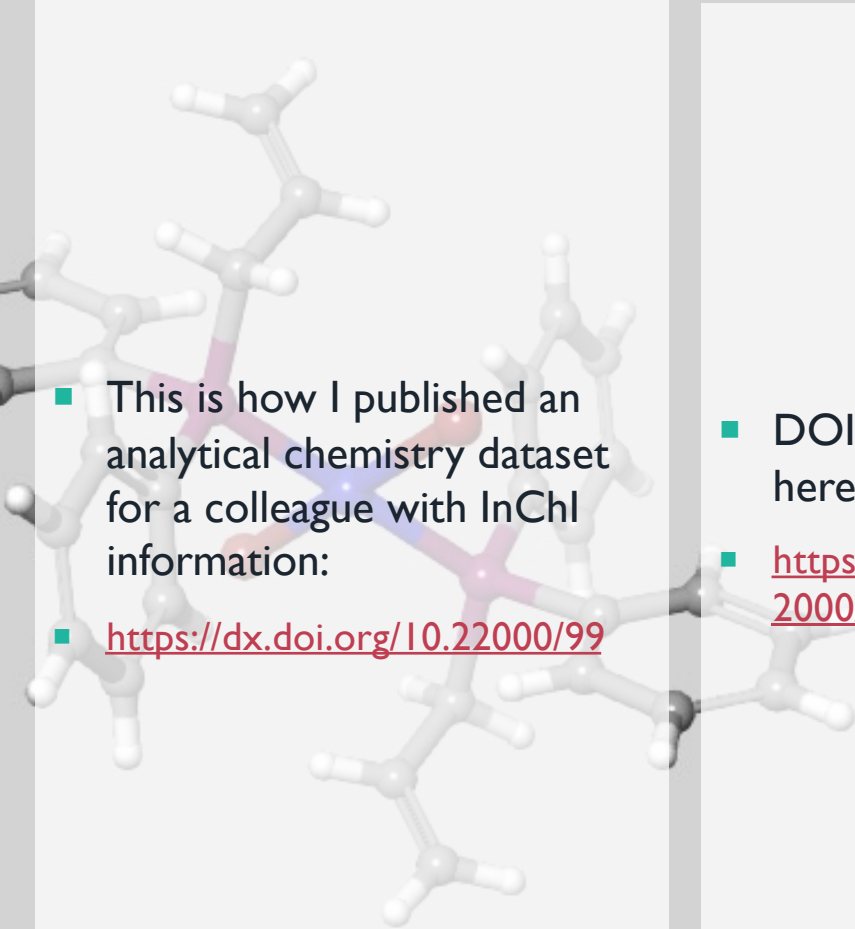
# IODIDE

*A few more notes...*

- 
- Would be useful to be able to define transformations, representing reaction centres, as well as reactions
  - The InChI is a good resource to look for uniqueness, but problematic as a representation
  - Tautomers: detailed analysis is available. InChI v1.07 provides access to more of this than before
  - JSON / XML useful
  - The Reaction InChI, Mixture InChI and Nano InChI are all rather different, but might benefit from a common code base: what do they have in common?
  - These need more information than just molecular structure, so data needed to supplement molfile input
  - An SD File can associate a molfile with other data
  - Different communities have different protocols: a modular approach may be helpful

# IODIDE

*Example dataset*



- This is how I published an analytical chemistry dataset for a colleague with InChI information:

- <https://dx.doi.org/10.22000/99>

- DOI metadata can be found here:

- <https://api.datacite.org/doi/10.22000/99>

# (RMN) InChI *very brief update*

August 17, 2024

Denver, CO USA

*Materials from Gerd Blanke, Thomas Exner, Leah McEwen, Evan Bolton*



# InChI Roadmap



All technical developments depend on further funding.

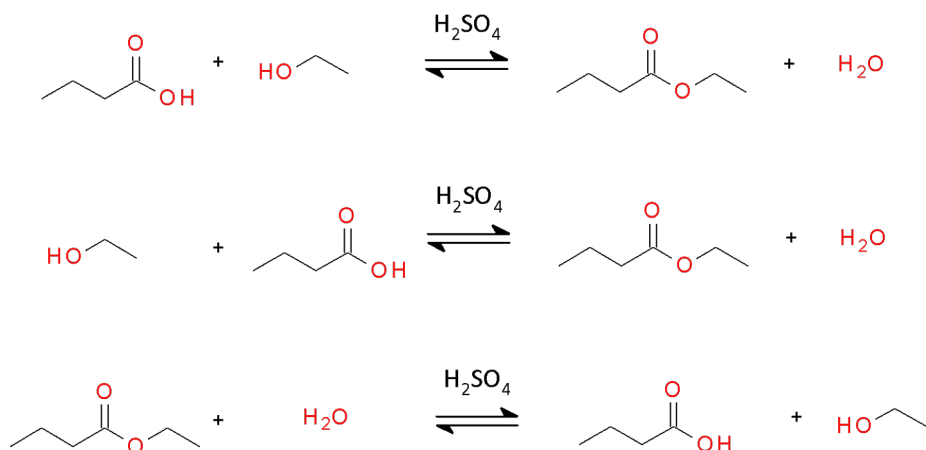
# The InChI "Application" projects

- RInChI, MInChI, NInChI
- All use (apply) InChI strings to identify chemical components in larger "real world" contexts
- All are handling multiple components (*most of the time*)
- Different use cases
  - Different component relationships and properties
  - Different goals for outputs beyond identifying components
  - Identify specific systems? Compare composition? Trace persistence?
- To formalize these nascent InChI related notations, can these projects share a common codebase?



# RInChI – What it is

- Example: Esterification, equilibrium reaction



- RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3!H2O/h1H2<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=
- RInChI is the unique identifier for chemical reactions

# MInChI: multi-component system layered notation

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

**MInChI=0.99.1S/  
CH2O/c1-2/h1H2&  
CH4O/c1-2/h2H,1H3&  
H2O/h1H2  
/n{{1&3}&2}  
/g{{37wf-2&}&{10:15pp0}}**

- "&" component separator
- "{}" mixture groups
- "/n" component index
- "/g" concentration

## 1. Component layer

- Components must be uniquely identifiable by their InChI(key)s or other standard open algorithms or referable standard descriptors
- Components are concatenated alphabetically per InChI convention, delimiter '&'

## 2. Index/hierarchy layer

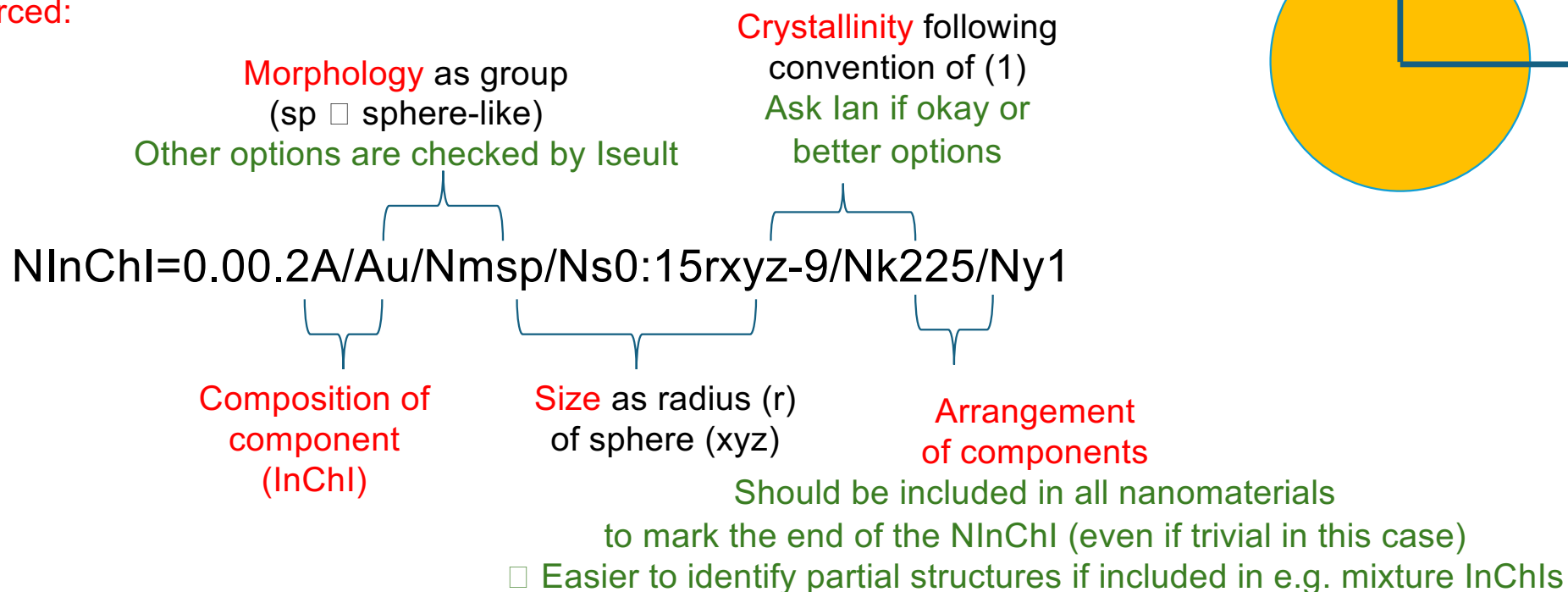
- Each component is indexed starting with 1 for the first component in the alphabetical order
- Sub-mixtures are identified with brackets around the related indices, e.g. {}

## 3. Property layer (e.g., concentration)

- List concentrations based on the order of components and groups
- Units as reported and noted with defined codes (next slide)

# NInChI example: 30nm gold nanoparticles (with citric acid stabilisation from synthesis)

If specifying stabilisation is not enforced:



(1) <https://pscf.readthedocs.io/en/latest/groups.html>

# NInChI example: 30nm gold nanoparticles (with citric acid stabilisation from synthesis)

If specifying stabilisation is enforced:

NInChI=0.00.2A/Au/Nmsp/Ns0:15rxyz-9/Nk225!

C6H8O7/c7-3(8)1-6(13,5(11)12)2-4(9)10/h13H,1-2H2,(H,7,8)(H,9,10)(H,11,12)

~~/Nmmol/No100~~

Individual  
molecules

Surface occupation  
= 100 %

~~/Ny1>2~~

Component 2 (citric acid)  
non-covalently bound to  
compound 1 (Au)

□ order is from inside to outside



# Architecture of InChI: RInChI, MInChI, NInChI

- InChI meeting in Aachen, July 2024:
  - Introduction to RInChI, MInChI and NInChI
    - One option: separate each notation in the strings and code
  - Is there a common data input format?
    - Relationships of components
    - Hierarchy, order
    - Properties, dependencies
  - Application InChI deliveries (technically seen: libraries)
  - Test environments and test data
- Need to articulate requirements for overall data model
  - Format agnostic
  - Consider enabling several different serializations (e.g., SDF, JSON)

# Handling multiple components

Project	Chemical components	Properties / Relationships (in string)
<b>RInChI</b> Reaction InChI	Clustered by role	Reaction role, reaction direction <i>others in AuxInfo</i>
<b>MInChI</b> Mixtures InChI	Compiled & indexed	Hierarchy, order, concentration (separate layer)
<b>NInChI</b> Nanoparticles InChI	Sequential with properties	Arrangement, association Numerous

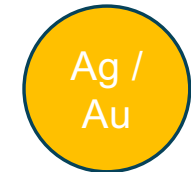
- ***What information do we need for each of these system notations?***
- ***Are there overlaps in information requirements?***
- ***Are there synergies in how each area handle components?***

# Gold / silver

Mixture InChI

$NInChI=0.00.2A/(Ag\&Au/n\{1\&2\}/g\{50wf-2\&\})/Nmsp/Ns0:20rxyz-9/Nk225/Ny1$

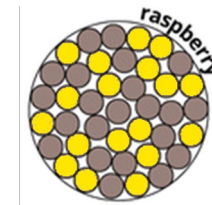
Alloy of 50% Au and 50 % Ag



**Multi-crystalline core of 50% Au and 50 % Ag crystals**

Mixture InChI of two NInChIs = morphology of multi-crystalline or raspberry

$NInChI=0.00.2A/(Ag/Nmsp/Ns0:1rxyz-9/Nk225/Ny1\&$   
 $Au/Nmsp/Ns0:1rxyz-9/Nk225/Ny1/n\{1\&2\}/g\{50wf-2\&\})$   
 $/Nmsp/Ns0:20rxyz-9/Ny1$



**Replaces:**  $NInChI=0.00.2A/Ag/Nmmc/Ns0:20rxyz-9/Nk(F\ m\ -3\ m)!$   
 $Au/Nmmc/Ns0:20rxyz-9/Nk(F\ m\ -3\ m)$   
 $/Ny\{1\&2\}/Ng\{50wf-2\&\}$

# PubChem use cases for [M,R] InChI

**PubChem** Tenamfetamine (Compound)

## 7.1 Drug Transformations

3,4-Methylenedioxyamphetamine is a known transformation product of 3,4-Methylenedioxy-N-methylamphetamine and 3,4-Methylenedioxy-N-ethylamphetamine.

S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag | DOI:10.5281/zenodo.3754448

► NORMAN Suspect List Exchange

**PubChem** Aspirin (Compound)

## 8.10 Biochemical Reactions

3 items [Download](#)

Reaction	PubChem Pathway	Source	Taxonomy
Acetylsalicylic acid + EDNRA → EDNRAcetComp	Nsp9 interactions (COVID-19 Disease Map)	COVID-19 Disease Map	Homo sapiens (human)
ASA + H2O → H+ + ASA- + H2O	Drug ADME	Reactome	Homo sapiens (human)
ASA + H2O → H+ + ASA- + H2O	Aspirin ADME	Reactome	Homo sapiens (human)

► PubChem

COMPOUND SUMMARY

## Benzene, toluene, ethylbenzene and xylene

PubChem Reference 482590225  
Collection SID

PubChem CID Not available because this is not a discrete structure.

See also: [Benzene](#) (has component); [Toluene](#) (has component); [Ethylbenzene](#) (has component) ... [View More](#) ...

## 1 Synonyms

Benzene, toluene, ethylbenzene and xylene  
BTEX  
BTEX compounds  
BTEX hydrocarbons  
Benzene, toluene, ethylbenzene and (p-, m- and o-) xylene  
Benzene, toluene, ethylbenzene and xylene (BTEX)  
Benzene, toluene, ethylbenzene, and xylene

► PubChem



# An InChI Application in the Publication Workflow of the Beilstein Journals:

## EXTRACTION AND DISSEMINATION OF FAIR CHEMICAL DATA WITH AN OPEN SOURCE WORKFLOW

Status August 2024



**BEILSTEIN** INSTITUT

Wendy Patterson

[wpatterson@beilstein-institut.de](mailto:wpatterson@beilstein-institut.de)

# State-of-the-Art in Publishing Machine-readable Chemistry

## Biphenylene-containing polycyclic conjugated compounds

Cagatay Dengiz

Figure 1: The correlation between stability and Clar's rule in acenes.

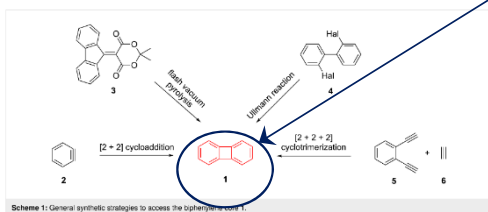
within the acene backbone [7,8], stabilization of the acene core structure through the integration of diverse units [9,10] and the introduction of bulky substituents [11]. These approaches aim to maintain the desirable electronic properties of acenes while mitigating the aforementioned challenges to the best possible extent. Our focus in this review is primarily on exploring the role of biphenylenes in stabilizing the core structures of acenes and other PAHs.

### Review

#### Biphenylenes and [N]phenylenes

Biphenylene (1), which consists of two aromatic benzene rings connected with a four-membered ring, is a highly intriguing compound in terms of its structure. It possesses a planar config-

uration, whereas approaches to the synthesis of acenes, such as vacuum pyrolysis [16–18], [2 + 2] cycloaddition [19,20], [2 + 2 + 2] cycloaddition [21], and the Ullmann reaction [15,22] (Scheme 1). Due to the observed low yields in flash vacuum pyrolysis, the difficulty in synthesizing starting materials, such as 3, and the impractical nature of scaling up the method for large quantities, the other three approaches have gained popularity for synthesizing biphenylene derivatives [23]. The utilization of in-situ aryne synthesis to generate biphenylene through the dimerization of arynes 2 from diverse substrates has gained popularity. However, this approach occasionally gives rise to the production of high-energy intermediates, such as benzene-diazonium-2-carboxylate, and yields that are comparatively low [20]. After the Ullmann reaction was successfully employed for the first reported synthesis of biphenylene [15], subsequent studies have explored various transition-metal-mediated coupling reactions using 2,2'-diiodinated biphenyls 4 as starting materials [24,25]. Although the cobalt-mediated alkyl-aryne trimerization route frequently used by Vollhardt and co-workers is not the first choice for the synthesis of the biphenylene itself, it has led to the synthesis of structurally demanding substituted biphenylenes and the emergence of a family of polycyclic hydrocarbons called [N]phenylenes.



Scheme 1: General synthetic strategies to access the biphenylene core 1.

Journal articles currently contain no digital or machine-readable chemical information!

SOTA:

- images embedded in PDFs and HTML
- Some examples of linking to repositories in DAS or ref list (both are typically improperly implemented)

### Confirmation of found DOIs

The DOIs below were taken from your data availability statement. Please look at each one carefully and confirm if the information is correct. If a DOI could not be found, please check for errors. If the DOI is correct but not registered yet, please also confirm this.

To make changes, please select "Edit" and repeat the process. After confirming all entries, please continue.

DOI	Title	Publisher	Info	Status	Actions
10.14272/collection/RAJ_2022-08-25	RAJ_2022-08-25	chemotion.net	DOI registered via DataCite	VALID	
10.5517/ccxk6bz	CCDC 850681: Experimental Crystal Structure Determination	Cambridge Crystallographic Data Centre	DOI registered via DataCite	VALID	
10.22000/986			No registered DOI found. Either there is a syntax error, or the DOI is not registered yet (i.e., the dataset is under review/embargo).	INVALID	

Edit

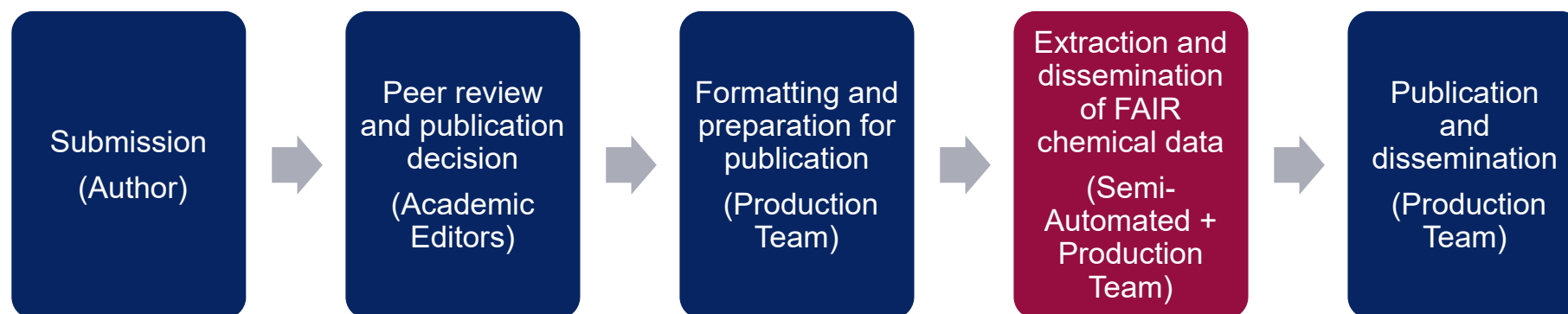
### Data Availability Statement

The data generated and analyzed during this study are openly available in the Chemotion repository at [https://doi.org/10.14272/collection/RAJ\\_2022-08-25](https://doi.org/10.14272/collection/RAJ_2022-08-25).

Crystallographic data has been deposited at CCDC under <https://doi.org/10.5517/ccxk6bz>.

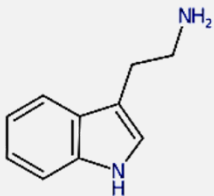
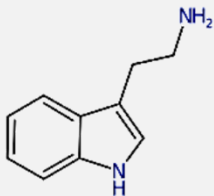
Further data will be openly available in RADAR at <https://doi.org/10.22000/986> following an embargo from the date of publication.

## Proposed Workflow to Enable FAIR Data in Publishing



1. **Extraction** of chemical structures
2. **Conversion** to and **validating** using [InChI](#)
3. **Embedding** of [InChIs](#) and further information back into the article → machine-readable chemical information in the article
4. **Dissemination** as FAIR Data

# Using InChI within the Beilstein Journals Publishing Workflow

No.	Found in	Structure	Validation Structure	Tani	Molecular Formula/IUPAC Name	InChIKey / Validation InchiKey	SMILES / Validation SMILES
174/203	<a href="#">Scheme 16</a>			1,0	<div style="display: flex; flex-direction: column; align-items: center;"> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-bottom: 5px;">3</div> <p><b>C<sub>10</sub>H<sub>12</sub>N<sub>2</sub></b> 2-(1H-indol-3-yl)ethanamine</p> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-top: 5px;">4</div> </div>	<div style="display: flex; flex-direction: column; align-items: center;"> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-bottom: 5px;">2</div> <p><b>APJYDQYYACXCRM-UHFFFAOYSA-N</b> APJYDQYYACXCRM-UHFFFAOYSA-N</p> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-top: 5px;">5</div> </div>	<div style="display: flex; flex-direction: column; align-items: center;"> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-bottom: 5px;">2</div> <p><b>C1=CC2=C(C=C1)NC=C2CCN</b> N1C=C(C2=CC=CC=C12)CCN</p> <div style="border: 1px solid black; border-radius: 50%; width: 30px; height: 30px; display: flex; align-items: center; justify-content: center; margin-top: 5px;">5</div> </div>

1. Beilstein ChemXtract (F. Bänisch): Extract structures from article
2. CDK: Convert structure to InChI + SMILES
3. CDK: Convert SMILES to structure + molecular formula
4. STOUT: AI tool (Steinbeck group) that converts SMILES to IUPAC name
5. OPSIN: Validation check for IUPAC name – IUPAC name to validation InChI + SMILES
6. CDK: Convert Validation SMILES to Validation Structure
7. RDKit: Calculates Tanimoto coefficient to test chemical similarity between SMILES and Validation SMILES

# Extraction of Chemical Structures during Article Publication Workflow

## 1. Beilstein ChemXtract (F. Bänisch): Extracts structures from article

Tool is developed in Java and will be published and maintained open source

Current limitations of structure extraction tool:

- Abbreviations (work in progress) – will be in MVP and an abbreviations library will be continuously updated, released on GitHub with invitation to community to contribute
- Markush/variable structures (especially difficult when provided as tables) – V2 will tackle some of the easier cases
- Reactions
- Author drawing errors not caught by our Production Team cannot be corrected
- Sgroup data (polymers etc.)

# Dissemination of Chemical Information (InChIs!) Collected during the Publishing Workflow – Overview

## Possibilities for Dissemination of InChIs in our MVP Release (Dec. 2024)

1. **JSON-LD** – embedded in HTML metadata using schema.org “Chemical Substance”. Google highly prefers this format and this will increase our chances for more visibility (of InChIs and the article).
2. **XML/JATS XML** – Will embed in our XML but likely not the JATS XML that we send to PMC, etc.
3. **PubChem** – In work, looking for a way to automatically generate IUPAC name or some other human identifiable name; looking into bulk delivery options, etc.; may not make it into MVP.
4. Crossref Metadata – Under evaluation but unlikely to implement as this would require the substance to also have a DOI/URL; no chemistry-specific metadata possible at this time.

# Dissemination of Chemical Information within the Publishing Workflow – JSON-LD

In order to be indexed and findable by Google, the substance data must go into the metadata of the full text HTML. For this, JSON-LD can be used, either using the schema type ChemicalSubstance or MolecularEntity or both.

## Exemplary JSON-LD

```
{
  "@context": "https://schema.org",
  "@type": "ChemicalSubstance",
  "@id": "??? What goes in here, maybe DOI of the substance ???",
  "identifier": "BQJCRHHNABKAKU-KBQPJGBKSA-N-1860-5390-20-2",
  "url": "https://www.beilstein-journals.org/bjoc/articles/20/2",
  "name": "Morphine",
  "alternateName": "InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1",
  "image": "https://www.beilstein-journals.org/bjoc/content/smiles/CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O",
  "description": "",
  "hasBioChemEntityPart": {
    "@type": "MolecularEntity",
    "smiles": "CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O",
    "inChIKey": "BQJCRHHNABKAKU-KBQPJGBKSA-N",
    "inChI": "InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1",
    "name": "Morphine",
    "molecularFormula": "C17H19NO3",
    "iupacName": "(4R,4aR,7S,7aR,12bS)-3-methyl-2,4,4a,7,7a,13-hexahydro-1H-4,12-methanobenzofuro[3,2-e]isoquinoline-7,9-diol"
  }
}
```

# Dissemination of Chemical Information within the Publishing Workflow – XML

Will be stored in our XML. PMC does allow for use of the JATS XML tags <chem-struct-wrap> and <chem-struct> but they only use this for display and the tags are not ideal (more info available if anyone is interested). **Who is on the NISO-chemistry committee that we could work with to improve the JATS XML tags for chemical information?**

## Exemplary substance entry for article 1860-5397-20-2

```
...
<substance id="BQJCRHHNABKAKU-KBQPJGBKSA-N-1860-5390-20-2">
  <inchi>InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-
11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1</inchi>
  <inchi-key>BQJCRHHNABKAKU-KBQPJGBKSA-N</inchi-key>
  <smiles>CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O</smiles>
  <iupac-name>(4R,4aR,7S,7aR,12bS)-3-methyl-2,4,4a,7,7a,13-hexahydro-1H-4,12-methanobenzofuro[3,2-e]
isoquinoline-7,9-diol</iupac-name>
  <trivial-name>Morphine</trivial-name>
  <molecular-formula>C17H19NO3</molecular-formula>
  <backref>1860-5390-20-2-1</backref>
  <backref>1860-5390-20-2-i2</backref>
</substance>
<substance>
...
</substance>
...
```



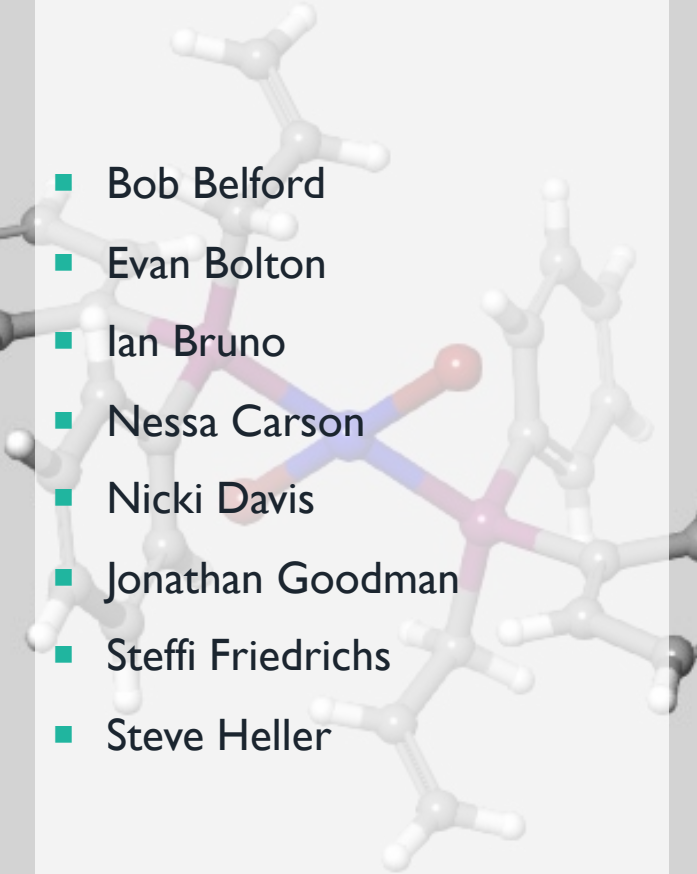
## Next Steps for Future Releases under Consideration

1. Working on the limitations of the Beilstein ChemXtract tool
2. Make human-readable/Visible for readers/authors – HTML/webpage with additional chemical info, search function on webpage, possibly involve author in the validations steps, JMol/JSMol for an interactive 3D-model of the substance, links to PubChem
3. Collaborations with the Lens, EuropePMC (EMBL-EBI) and others interested in the dissemination of chemical information
4. Encouraging publishers to incorporate this open source workflow into their production work and to join us in future development for a truly FAIR chemical data ecosystem!
5. Using the workflow to extract structures for previously published articles

## INCHI MEETING

Thank you to everyone who attended, in-person and on-line



- 
- Bob Belford
  - Evan Bolton
  - Ian Bruno
  - Nessa Carson
  - Nicki Davis
  - Jonathan Goodman
  - Steffi Friedrichs
  - Steve Heller
  - Frank Lange
  - Leah McEwen
  - Timur Madzhidov
  - Hunter Moseley
  - Marc Nicklaus
  - Wendy Patterson
  - Vin Scalfini
  - Wendy Warr