Organometallics

InChIng forwards to better representations and happier chemists

Ian Bruno (CCDC)       Jonathan Goodman (University of Cambridge)
Colin Batchelor (RSC)   Gerd Blanke (StructurePendium Technologies GmbH)
Alex Clark (Molecular Materials Informatics)
Catalysing Chemistry

Unusual Highly Regioselective Direct Aldol Additions with a Moisture-Resistant and Highly Efficient Titanium Catalyst
Mahrwald and Schetter (2005), Org. Lett., 8(2), 281
https://doi.org/10.1021/ol052637z

[Rh₂(COD)₂(Dppm)(μ₂-Cl)]BF₄⁻: Precursor for a Selective Hydrogenation Catalyst and Its Recycling by Silica Entrapment
Lorenzini et al. (2006), Organometallics, 25(16), 3912
https://doi.org/10.1021/om060172q

• Nearly half of the reactions in the NextMove patent reaction database include a metal
• There are probably a lot more people working in pharmaceuticals and fine chemicals who regularly use organometallics than there are specialist organometallic chemists

Source: Jonathan Goodman
Bioorganometallic Chemistry

PDB Chemical Component HEM bound to PDB 19HC

PDB Chemical Component RU7 bound to PDB 1T3P

Structure of the FeMo cofactor of nitrogenase

By Smokefoot - Own work, CC BY-SA 4.0
https://commons.wikimedia.org/w/index.php?curid=37239342

PDB Chemical Component CFN bound to PDB 3K1A
Functional Materials

Semi-conductors

Electrical Conductivity in Platinum-Dimer Columns
https://doi.org/10.1021/ic801211m

Photoswitches

Understanding solid-state photoswitching in [Re(OMe2-bpy)(CO)3(η1-NO2)] crystals via in situ photocystallography
https://doi.org/10.1039/C8CE00774H
The Cambridge Structural Database

- Over 1 Million small-molecule crystal structures
- Over 80,000 datasets deposited annually
- Enriched and annotated by experts
- Each CSD entry includes a 2D chemical representation of the substance studied

Organic: 43%
Not Organic: 57%

At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Ti, Ge, Sn, Pb, Sb, Bi, Po

CSD V5.38  Organic: *BTEST 57

250,000th – IBEZUK
Conducting metal-dinitrogen polymer
Challenges Representing Organometallics

- Bond Perception
- Author Expression
- Stereochemistry
- Digital Representations
When is a Bond not a Bond?

doi:10.1107/S1600536807037804
a 7.3754(18) b 8.925(2) c 13.812(3);
α 92.347(2) β 97.596(2) γ 96.311(3)

doi:10.1002/zaac.201100181
a 7.3025(4) b 8.8332(5) c 13.7017(8)
α 92.919(1) β 97.319(1) γ 96.456(1)

AFGEY (orange) over-layed with AFGEY01 (cyan)
Different Expressions

http://pubs.rsc.org/en/content/articlehtml/2012/dt/c2dt31989f

http://pubs.acs.org/doi/abs/10.1021/om700498w

ECIWUK  85% of CSD Entries

LIMXAH  12% of CSD Entries

3% of CSD Entries (N = 9468)
Organometallics & Stereochemistry

Octahedral

Square Planar

Trigonal Bipyramidal

Square Pyramidal
Digital Representation

- How best to reliably represent organometallics?
  - dative vs covalent bonds?
  - delocalised bonds?
  - dummy atoms?
  - quadruple bonds?
  - zero-order bonds?

<table>
<thead>
<tr>
<th>MOL V3000</th>
<th>PubChem SDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Bond</td>
<td>1 Single Bond</td>
</tr>
<tr>
<td>Double Bond</td>
<td>2 Double Bond</td>
</tr>
<tr>
<td>Triple Bond</td>
<td>3 Triple Bond</td>
</tr>
<tr>
<td>Aromatic Bond</td>
<td>4 Quadruple Bond</td>
</tr>
<tr>
<td>Coordination Bond</td>
<td>5 Dative Bond</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>6 Complex Bond</td>
</tr>
<tr>
<td>7 Ionic Bond</td>
<td></td>
</tr>
</tbody>
</table>

* Strictly a query type

### CSD Bond Types

- Single
- Aromatic
- Double
- Pi
- Triple
- Delocalised
- Quadruple
- Polymeric

### ACD/Labs MOL V2000 Extensions

<table>
<thead>
<tr>
<th>Type</th>
<th>M</th>
<th>ZZ</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>M</td>
<td>ZZF</td>
<td>3</td>
<td>1</td>
<td>41</td>
<td>2</td>
<td>42</td>
<td>3</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>M</td>
<td>ZZH</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple</td>
<td>M</td>
<td>ZZE</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadruple</td>
<td>M</td>
<td>ZZF</td>
<td>3</td>
<td>5</td>
<td>12</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polymeric</td>
<td>M</td>
<td>ZZE</td>
<td>2</td>
<td>42</td>
<td>18</td>
<td>43</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### RDKit SMILES Extension

- and -> create a dative bond between the atoms
- c1cccn->2c1-c1n->3cccc1.[Pt]23(C1)C1

### OpenSMILES Extension

- quadruple bonds are represented by '$'
- [Rh–] (C1) (C1) (C1) (C1) $[Rh–] (C1) (C1) (C1) (C1) C1

10.1002/chem.201805800
10.5517/ccdc.csd.cc2035gs
InChI: The International Chemical Identifier

Standard InChI:
InChI=1S/C23H24N4O9/c24-16(22(31)32)9-10-36-15-7-3-12(4-8-15)18(26-35)20(29)25-17-11-27(21(17)30)19(23(33)34)13-1-5-14(28)6-2-13/h1-8,16-17,19,28,35H,9-11,24H2,(H,25,29) (H,31,32) (H,33,34)/b26-18+/t16-,17+,19-/m1/s1

InChI Layers

- Formula
- Connectivity
- Hydrogen positions

- Stereochemistry – atom and bond
- Isotopic and Fixed Hydrogens (tautomerism)
- Charge

Standard InChIKey:
CTNZOGJNVIFEBA-TWTPMLPMSA-N

A one-way hash of the Standard InChI
InChI & Organometallics

**GOAL:** Reliable standard InChI representations of compounds in the Cambridge Structural Database (CSD) to enable intersection of this with other data resources

**OUTCOME:** Could confidently generate reliable InChIs for **22%** of a subset of 495,751 entries from CSD 5.36

*Order of filtering:*
- Not organic
- Multi-component
- Inchi alerts
- Stereochemistry Issues

Success rate for organic portion was **51%**.

Multi-component entries were excluded by choice – if all these converted successfully then success rate would be ~80% of organics.

 Biggest Limitation: inability to reliably generate an InChI for metal-containing compounds
14

**Objective**

To establish requirements for extending the applicability of the IUPAC International Chemical Identifier to organometallic and coordination compound structures, so that developers contracted to the InChI Trust can enhance the InChI software accordingly.

> Link to InChI subcommittee

**July 2015 update:** The requirements statement is almost complete and includes over a hundred structures that will be more than detailed enough for the developer to work on. Over the last month, the chair has held a series of meetings with Ian Bruno (Cambridge Crystallographic Data Centre, CCDC) in order to discuss in more detail CCDC's requirements. In addition to the V2000 and V2000 files, CCDC's preparation for the IUPAC's InChI, a dozen or so structures that were found in the code, and that code will correspond to the InChI of these structures, has been discussed.

**Dec 2017 update:** We have held two meetings in 2017 as part of the two broader InChI workshops in Hinxton and Washington. This has uncovered important disagreements between people who are familiar with the code as to the best way to proceed. We are looking into a mechanism for finding people to make a proper investigation into disconnection and normalisation of coordination and organometallic structures, potentially using the CSD APIs as a source of structures.
Frequent Discussion Points

• Disconnection
  • Currently, standard InChI disconnects bonds to metal then normalises
  • It is possible to generate a non-standard InChI with a reconnection layer
  • Do we need to disconnect? Can we exploit existing reconnection layer?

• Stereochemistry
  • Do we need to be able to work out CIP priorities?
  • Do we need to be able to distinguish all stereochemistries?

• External challenges
  • Chemists are not consistent in how they draw metal-organic structures
  • Tools may not internally represent metal-organic structures the same way
  • Identical structures may end up represented different ways in MOL files
InChI Workshop, Boston, August 2018

• Questions considered:
  • What positive transformation could organometallic InChI enable?
  • What are the motivational use cases?

• Organometallic InChI drivers:
  • Organometallic InChI fundamental to integrity and perception of InChI: reputational risk if left unaddressed – possible barrier to adoption
  • Community don’t appear to be beating down the door: but perhaps patiently waiting for the working group to deliver
  • Potentially significant opportunities for application: but need something to be built to explore if this is the case
• Discussion of a Proposal from Jonathan Goodman for an Organometallic InChI Layer

Retains a degree of backward compatibility – current InChI layers are preserved

As a general rule, the organometallic layer should include as many inorganic bonds as reasonably possible
Stereochemistry Proposal

• Proposal from Jonathan Goodman:

Phase 1
In phase one, we assume that all the information we need about a molecule is available. We concentrate on constructing the InChI. The connectivity has already been defined by the /om layer. The /ma layer (metal architecture) now defines the stereochemistry.

Simple Stereochemistry
Assume that the metal-centred stereocentre is a clearly defined shape and it is straightforward to assign priorities to the substituents. The stereochemistry can be defined by a symbol for the shape, here ‘o’ for octahedral. This situation has been studied before. There are thirty different arrangements of the six substituents, and the desired one can be given a number from one to thirty. If, in this example, the central metal atom has canonical number 99

/ma99o1 or /ma99o2 or ... /ma99o30
Tetrahedral, trigonal bipyramidal, cubic, etc., structures can be defined in a similar way. Square planar compounds can either be treated as a separate category, or as a special case of...
Stereochemistry Proof of Concept

- Proof of Concept from John Mayfield: [https://github.com/johnmay/inchi-ma](https://github.com/johnmay/inchi-ma)

**InChI Metal Architecture Proof of Concept**

**InChI Metal Architecture**

Proof of concept to generate InChI's that distinguish structures with different Square Planar, Trigonal Bipyramidal, and Octahedral configurations.

It adds a '/ma' (metal architecture) layer as proposed by Jonathan Goodman to the end of an InChI. Currently the following geometries are supported:

- Square Planar: `/ma<atom><order>` where order is 1,2,3
- Trigonal Bipyramidal: `/ma<atom>tb<order>` where order is 1-20
- Octahedral: `/ma<atom>o<order>` where order is 1-30
Stereochemistry Proof of Concept

- **Proof of Concept from John Mayfield:** [https://github.com/johnmay/inchi-ма](https://github.com/johnmay/inchi-ма)

Note the numbers here refer to the input atom order in the SMILES, not the canonical InChI numbers.
Taking Stock…

- Community discussion across a number of workshops has established:
  - Reasons why it was still worth pursuing organometallic InChI
  - Confidence that stereochemistry could be handled
  - Opinions on what is necessary and where there could be compromise
  - Ideas about how to handle connectivity but no clear conclusion

- It was time to seek an expert willing to do some prototyping
The InChI Organometallic RFP*

*Request for Proposal – issued by the InChI Trust – refined following community feedback

• Be able to compare three prototype InChIs:
  • An organometallic InChI with a Goodman-esque coordination layer
  • As above but without initially disconnecting bonds to metal atoms
  • The InChI generated with the InChI metal reconnection layer

• To aid with this:
  • The InChI Trust has supplied a small set of test structures compiled earlier
  • The CCDC is providing access to chemical structures in the CSD through their Python API
Enter Alex Clark...

Scientist at Collaborative Drug Discovery
Founder of Molecular Materials Informatics


Initial Deliverables

- Training set of compounds:
  - real-world compounds (CSD, PubChem, misc)
  - some drawn well, others drawn badly

- Prognosis for issues to expect:
  a. current InChI works fine, or
  b. new layer is required, or
  c. intractable problems persist

- Use as a definitive pass/fail validation key
Goal

- Ideally

All drawings of a chemical entity produce the same InChI
- One InChI can never match two drawings of different molecules
- Probably impossible, but can we get close enough to be useful?

(coordination InChI)
Source Data

- **Cambridge Structural Database:**
  - $\leq 500K$ metal-organics that aren't polymers
  - 2D coordinates, intelligent bonds, H-counts
  - selected ~500 by diverse clustering

- **PubChem:**
  - picked ~200 from large subset of metal-containing substances
  - most had to be redrawn

- **Miscellaneous:**
  - privately curated data ~500 compounds
  - carefully drawn inorganic valences
Categorisation of bonding types that need attention [1/2]

- **alkene**
- **disconnected**
- **carbonyl**
- **arene**
- **metal-metal**
- **carbene**
- **nitrosyl**
- **bidentate**
Categorisation of bonding types that need attention [2/2]

- Dative
- Hypervalent
- Metallabenzenes
- Symmetry
- Multicentre
- Hypovalent
- Terminal O
- Alternating
Comparing InChIs for Different Representations of the same Species

InChI codes differ

\[
\text{InChI=1/C2H4Cl2Pt.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
3/C2H4.H,Ct.h3'1H/qm3'1;1/c2H4\ \\
\text{ClPt}3(6,4,5)2-6H1-2H2;2'1\ \\
\text{[C.2.-1/6][C.2.-1/6][Cl0.-1/6][Cl.0.-1/6]Cl.0.-1/6}\ \\
\text{[Cl.0.-1/6]/Pt.0.-1/6;[1:2=3][1:6=3]}\ \\
\text{[2:6=3][3:6=3][4:6=3][5:6=3]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H4.Cl2Pt.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
3/C2H4.H,Ct.h3'1H/qm3'1;1/c2H4\ \\
\text{ClPt}3(6,4,5)2-6H1-2H2;2'1\ \\
\text{[C.2.-1/6][C.2.-1/6][Cl0.-1/6][Cl.0.-1/6]Cl.0.-1/6}\ \\
\text{[Cl.0.-1/6]/Pt.0.-1/6;[1:2=3][1:6=3]}\ \\
\text{[2:6=3][3:6=3][4:6=3][5:6=3]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]

InChI codes differ

\[
\text{InChI=1/C2H5Be2.c1-2;::;/h1-2H2:3'1H;O=2;::;/}\ \\
2/C2H5Be.c1-2;r1=1/r0=1;\text{Be}2\ \\
\text{Cl1(6,7,8)1[1:2=1][1:3=1]}\ \\
\text{[1:4=1][1:5=1][1:6=1][1:7=1][1:8=1]}\ \\
\text{[2:3=1][2:4=1][3:4=1][4:5=1][5:6=1]}\ \\
\text{[6:7=1][7:8=1][8:10=1][9:11=1][10:11=1]}\ \\
\]
Comparing InChIs for Different Representations of the same Species

Prototype algorithm for canonical representation of metal-organic connectivity being explored by Alex
Prototype Connectivity Algorithm

Algorithm Prerequisites

1. Complete heavy atom graph
2. Hydrogen counts
3. Bond orders $\rightarrow$ *delocalisation islands*
4. Net charges for each *island*

Algorithm Outcomes

- Appears to be remarkably tolerant to multiple ways of drawing inorganic bonds
- Preliminary results are promising for disambiguating metal-organics correctly

Alex Clark, Work in Progress
Organometallic InChI: Current Status

• About two-thirds of the way through Phase 1 of RFP
  • Focus will now be on consolidating findings and creating a report

• There will then be various questions we may need to consider
  • Balance between getting the structure “right” and a tractable outcome
  • How faithfully we need to encode what is drawn
  • How rigorous canonicalization needs to be
  • How to implement – modifying the core code vs externally adding a layer

• Better representations and happier chemists hopefully not far off
  • Organic chemists using organometallics
  • Expert organometallic chemists
  • Cheminformaticians wrangling chemistry data
Acknowledgements

• Colin Batchelor – IUPAC Project lead
• Jonathan Goodman – Proposer-in-chief
• John Mayfield – Overnight Prototyper
• Alex Clark – Enterprising Explorer
• Andrey Yerin – RFP Critique and Challenge
• Gerd Blanke – Presentation Feedback
• Clare Tovee – Curation Insights and Examples
• IUPAC and InChI Trust – Project Funding and Support