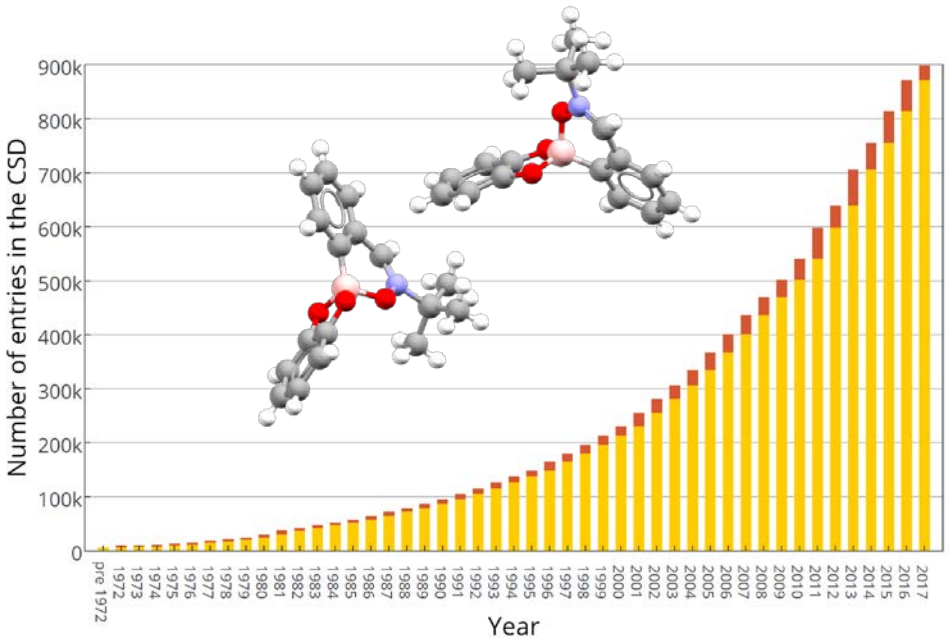
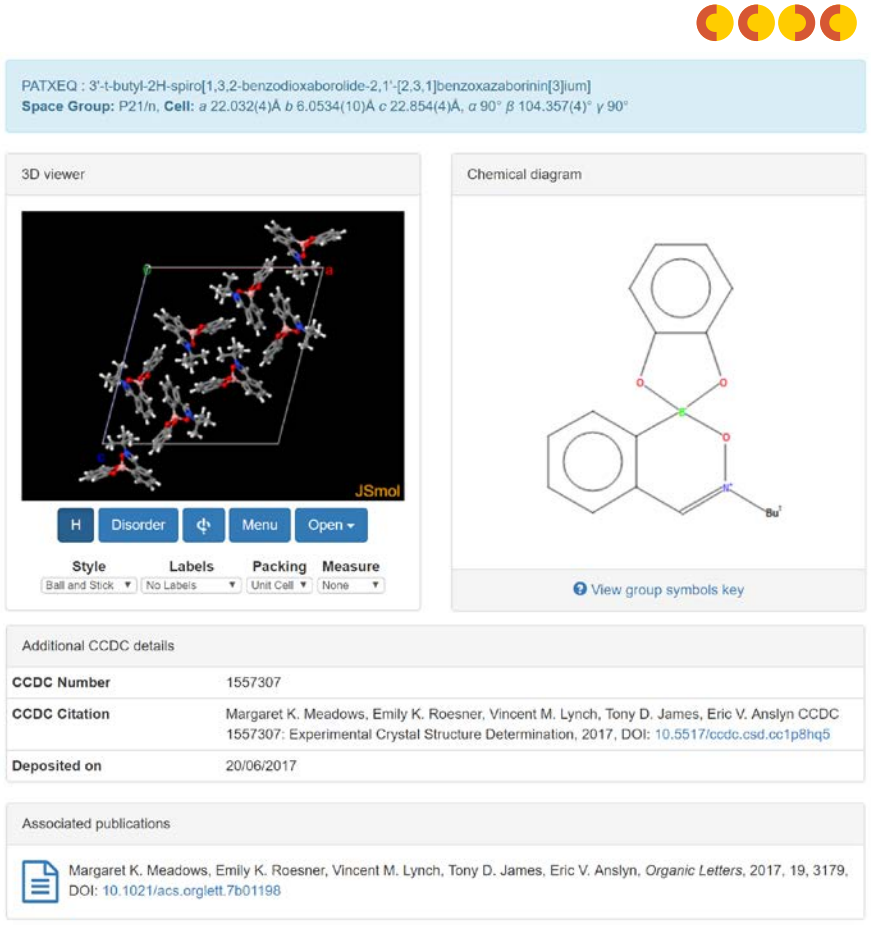
Organometallics & InChI

August 2017

The Cambridge Structural Database



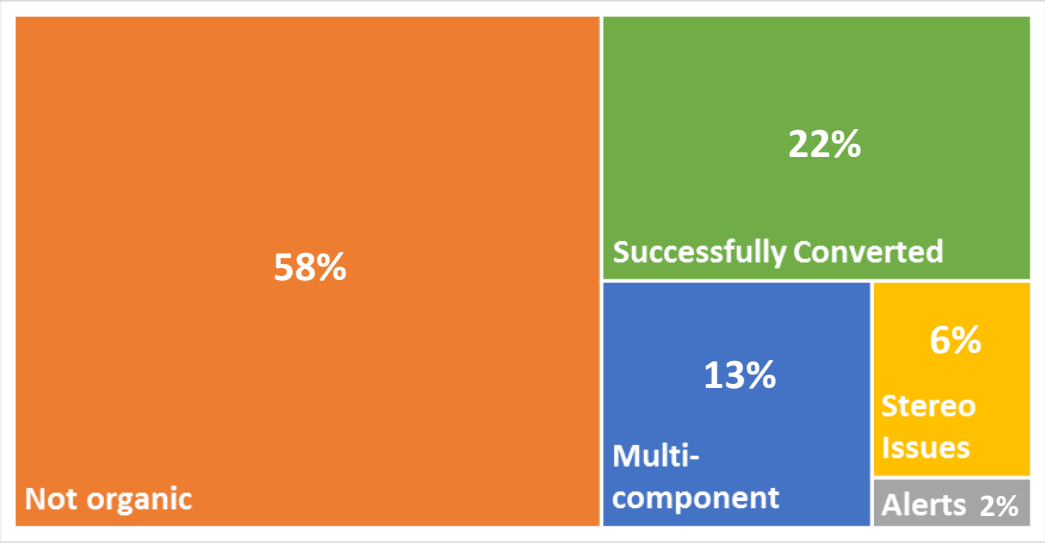


- **900,000+ small-molecule crystal structures**
- Over 60,000 datasets deposited annually
- Enriched and annotated by experts
- Structures available for anyone to download
- Links to over 1,000 journals

https://dx.doi.org/10.5517/ccdc.csd.cc1p8hq5

The CSD & InChIs

- **GOAL:** Reliable standard InChI representations to enable intersection of the CSD with e.g. ChemSpider and PubChem
- Based on a subset of 495,751 entries from CSD V5.36, InChIs successfully generated for 108,570 entries - **that's just 22%**

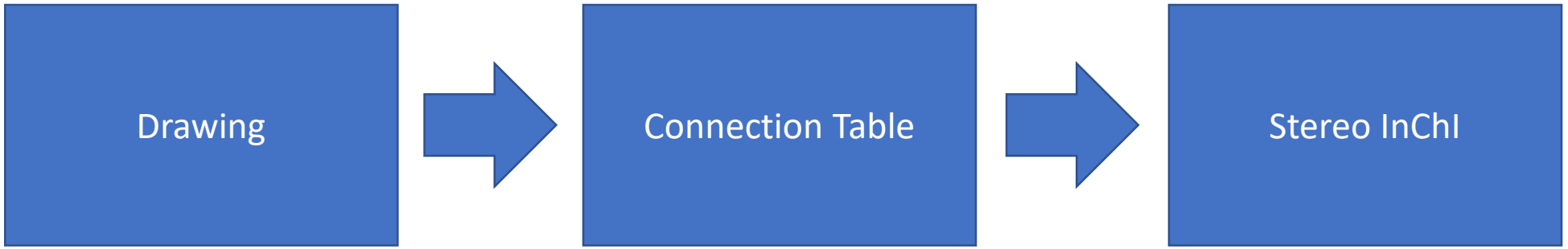


Order of filtering:

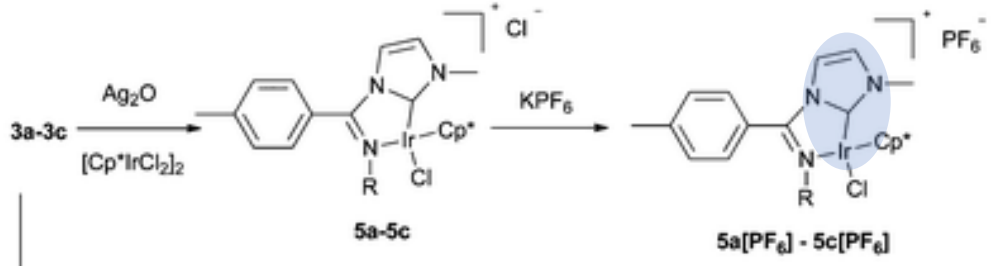
- *Not organic*
- *Multi-component*
- *InChi alerts*
- *Stereochemistry Issues*

Previously, at the EBI

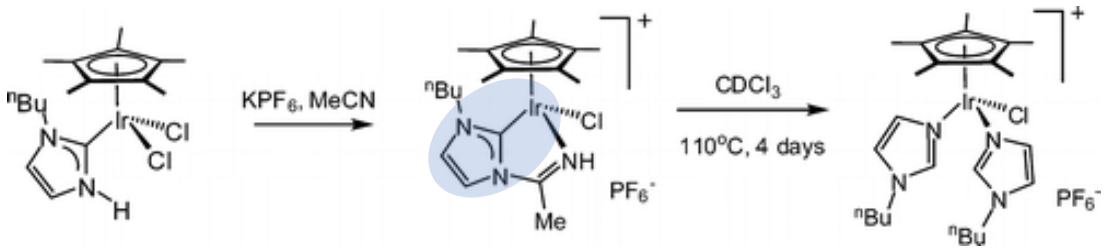
- Organometallic and coordination compounds: What can we achieve by August?
  - **Colin:** Determine tasks for stereo package for developer
  - **John:** Identify supported V2000 extensions, idioms and bugs
  - **Andrey:** Review examples of advanced normalization and suggest drawing rules



Different Expressions

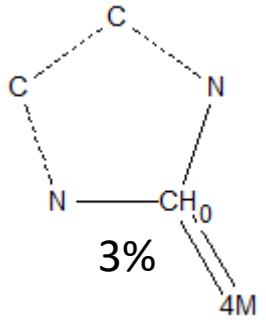


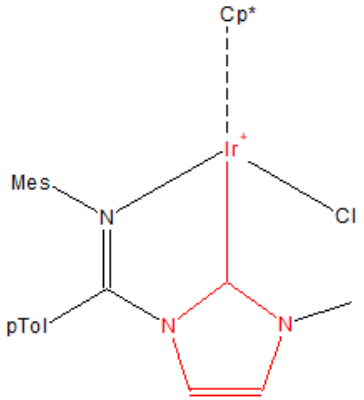
http://pubs.rsc.org/en/content/articlehtml/2012/dt/c2dt31989f



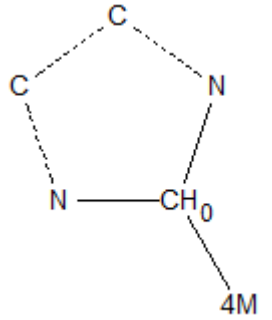
http://pubs.acs.org/doi/abs/10.1021/om700498w

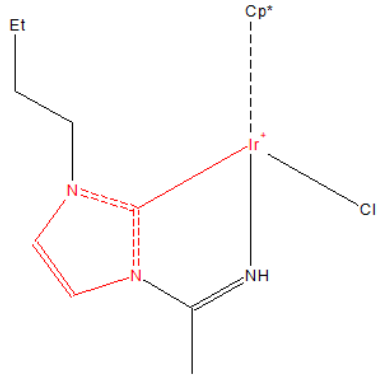
N = 9468 (CSD 5.38)



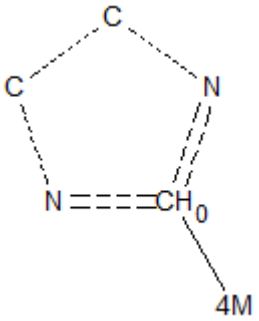


ECIWUK





LIMXAH



Reliable Representation

- How best to reliably represent organometallics?
  - dative vs covalent bonds?
  - explicit hydrogens/valencies?
  - dummy atoms?
  - zero-order bonds?

MOL V3000

```
1 = single
2 = double
3 = triple
9 = coordination
10 = hydrogen
```

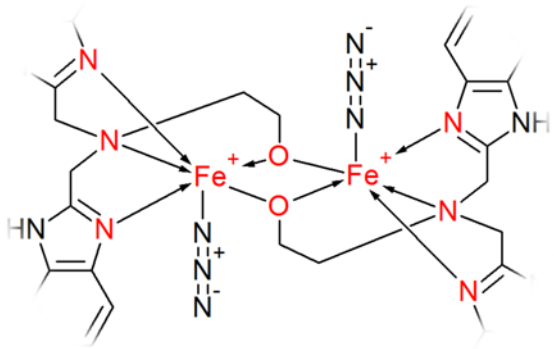
* excluding query bond types

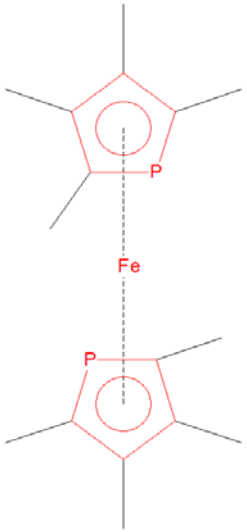
ACD/Labs MOL V2000 Extensions

```
M  ZZF  3   1  41   2  42   3  43
M  ZZH   1  5   2   3   4   5   6
M  ZZH   2  5   7   8   9  10  11
M  ZZH   3  5  12  14  15  16  17
M  ZZE  2  42  18  43  18
```

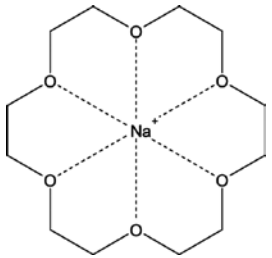
PubChem SDF
PUBCHEM_NONSTANDARDBOND

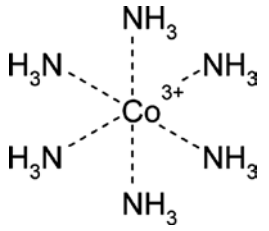
```
1 Single Bond
2 Double Bond
3 Triple Bond
4 Quadruple Bond
5 Dative Bond
6 Complex Bond
7 Ionic Bond
```

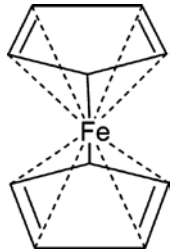




Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting. **Alex M. Clark.**
J. Chem. Inf. Model., 2011, 51 (12), 3149. doi:10.1021/ci200488k





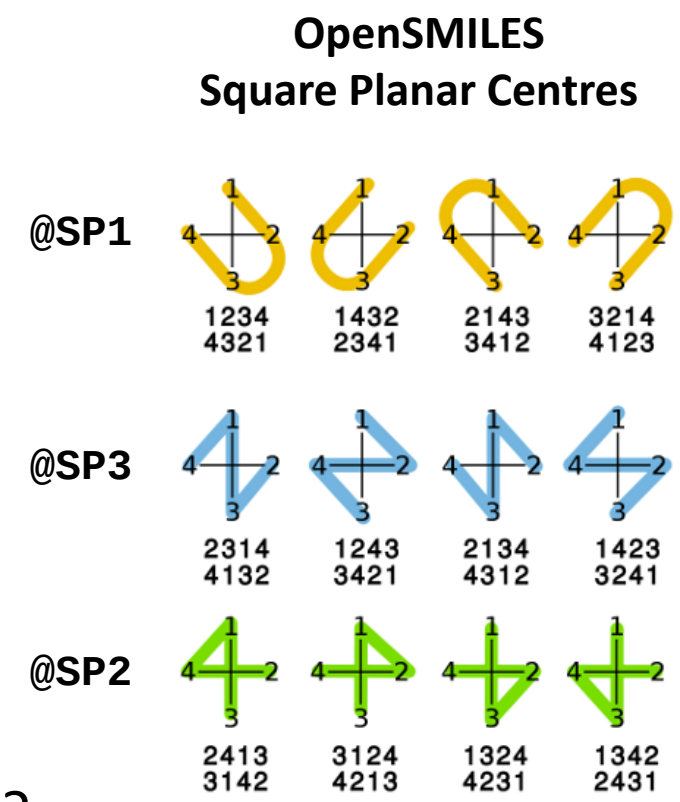


CSD Bond Types

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

Stereochemistry

- Task 1: Identify geometries from 3D and 2.5D
  - octahedral
  - square planar
  - trigonal bipyramidal
  - square pyramidal
- Task 2: Generate priority ordering
  - using CIP?
  - or any stable ordering?
  - relevance of IR-9.3.5.3 Red Book Priming Convention?



Putative Principles

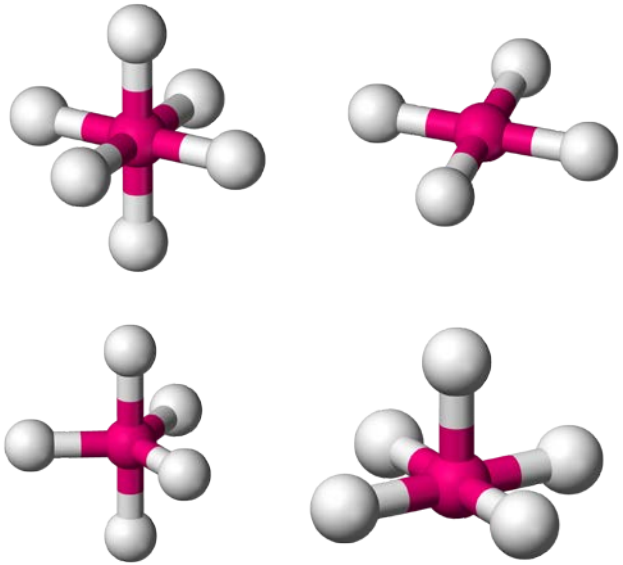
- Chemically-distinct systems should produce distinct InChIs
- A reliable representation for input into the InChI generator is possible without the need for a “delocalised” bond type
- Assume we don’t want to reconstruct original structure from the InChI (except *via* AuxInfo layer)
- Can we retain disconnection and avoid ambiguity?

Session Goals

- Update on progress since EBI
  - normalization/disconnection (Andrey)
  - representation (John)
  - stereochemistry (Ian channelling Colin)
- Identification of next steps
  - do we have enough information to start coding?
  - do we have sufficient test cases?
- Motivational use-cases
  - what positive transformation could organometallic InChI enable?

Stereochemistry

TASK 1: Identify Geometries



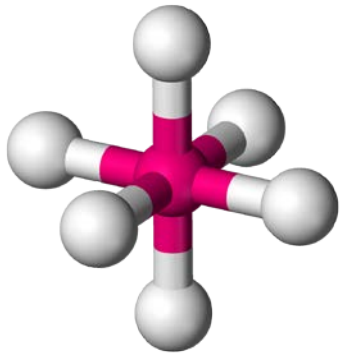
- Identify the following geometries around atom centres:
  - octahedral
  - square planar
  - trigonal bipyramidal
  - square pyramidal
- Cover three-dimensional structures first
  - there are test structures provided by CCDC and Elsevier
  - we should be careful about what we distribute with any code developed
- When we cover two-dimensional structures, only cover those drawn with wedged and dashed bonds
  - do not try to infer stereochemistry from flat bonds.

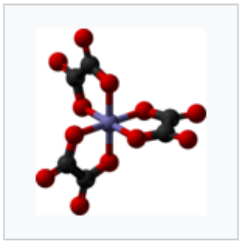
cis-$[CoCl_2(NH_3)_4]^+$

trans-$[CoCl_2(NH_3)_4]^+$

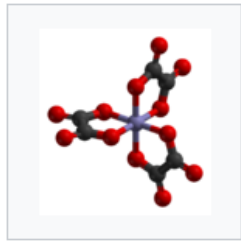
fac-$[CoCl_3(NH_3)_3]$

mer-$[CoCl_3(NH_3)_3]$

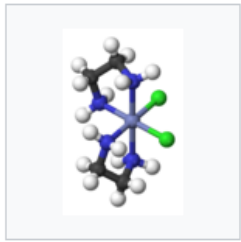




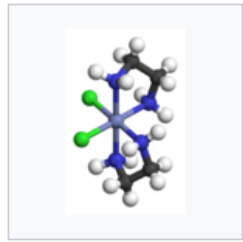
Λ-$[Fe(ox)_3]^{3-}$



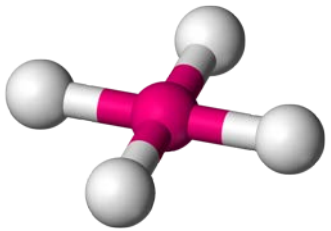
Δ-$[Fe(ox)_3]^{3-}$

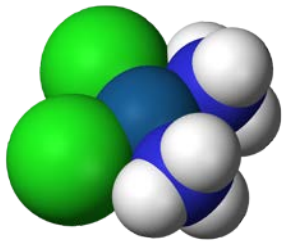


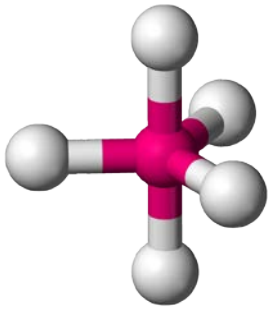
Λ-*cis*-$[CoCl_2(en)_2]^+$

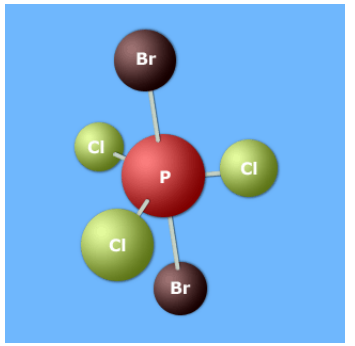


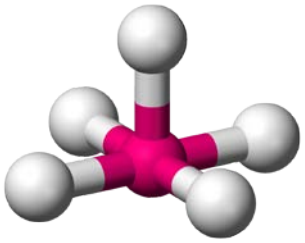
Δ-*cis*-$[CoCl_2(en)_2]^+$

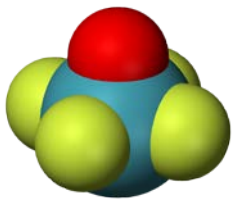










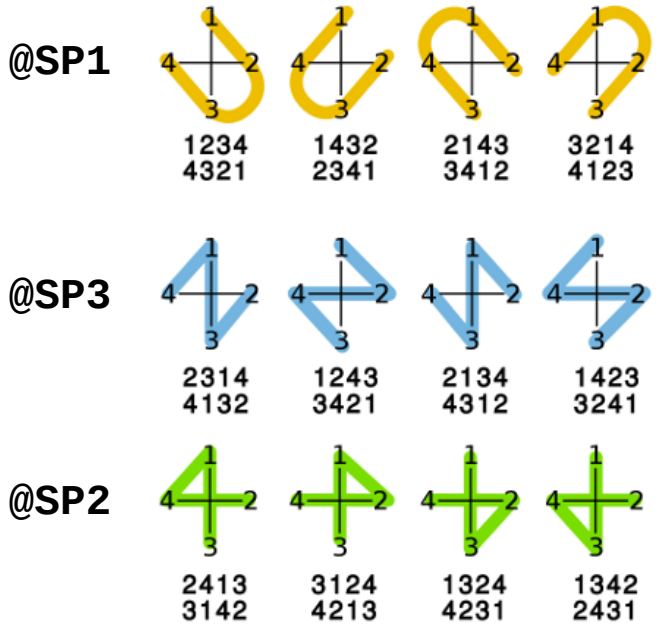


TASK 2: Identify configurations

- There is a labelling scheme that has been documented by OpenSMILES
  - This involves assigning the label for the configuration around a centre based on the order in which the atoms appear in the SMILES string.
- A single square-planar complex can therefore have three different labels according to the order in which the atoms are specified
  - Can this can be extended to the InChI by using the canonical numbering of each atom as its order?

http://opensmiles.org/opensmiles.html

OpenSMILES
Square Planar Centres



Trigonal Bipyramidal Centers

@TB1 - @TB18

Octahedral Centers

@OH1 - @OH28

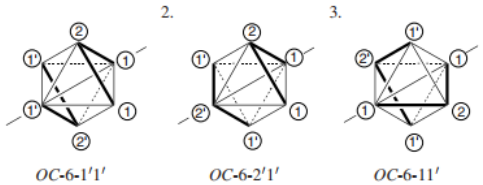
TASK 2: Identify configurations

- OpenSMILES approach not the same as implementing IUPAC nomenclature
  - potentially a harder task that involves assigning Cahn-Ingold-Prelog (CIP) priorities
  - there is code in RDKit to yield the CIP priorities but it isn't accessible via the Python wrapper.
- What effect does priming have if we take the SMILES approach?

IR-9.3.5.3 *Priming convention*

The priming convention is required in order to avoid ambiguity when using the configuration index to describe the stereochemistry of systems that contain either more than one polydentate ligand of a particular kind, or a polydentate ligand that contains more than one coordinating fragment of a particular kind. This situation is found commonly with bis(tridentate) complexes, but also arises in more complicated cases. The need for this convention is best illustrated by example.

Examples:



https://www.iupac.org/fileadmin/user_upload/databases/Red_Book_2005.pdf