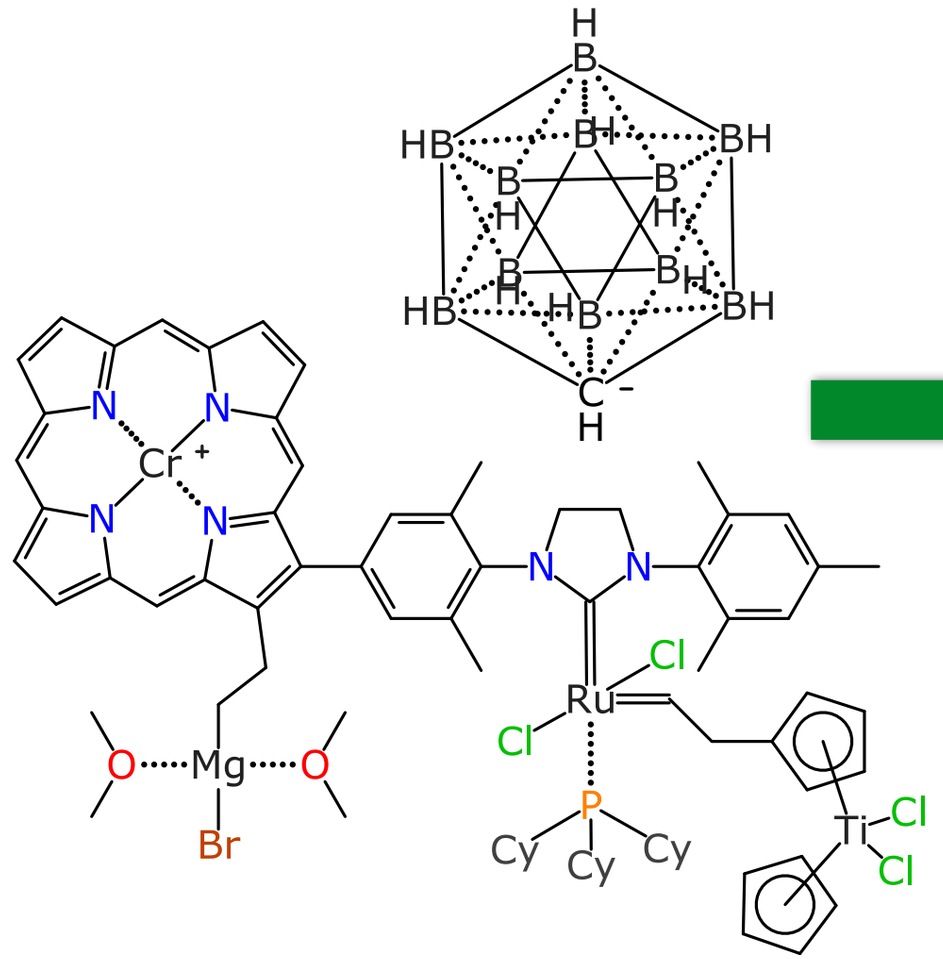
Coordination InChI
Preliminary survey of inorganic compounds

Alex M. Clark, Ph.D.

August 2019

Goal

- Ideally



{**coordination INCHI**}

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

Deliverable

- Training set for inorganic 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. intractible problems persist

- Use as a definitive pass/fail validation key

Source Data

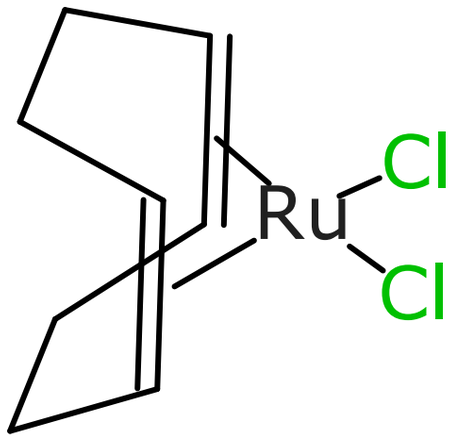
- **Cambridge Structural Database**:
  - ≤ 500K inorganics that aren't polymers
  - 2D coordinates, intelligent bonds, H-counts
  - selected ~500 by diverse clustering

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

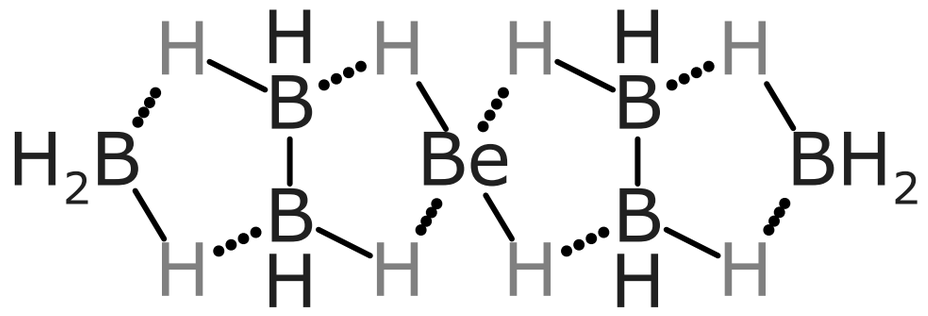
- **Miscellaneous**:
  - privately curated data ~500 compounds
  - carefully drawn inorganic valences

Exotic Bonding Types

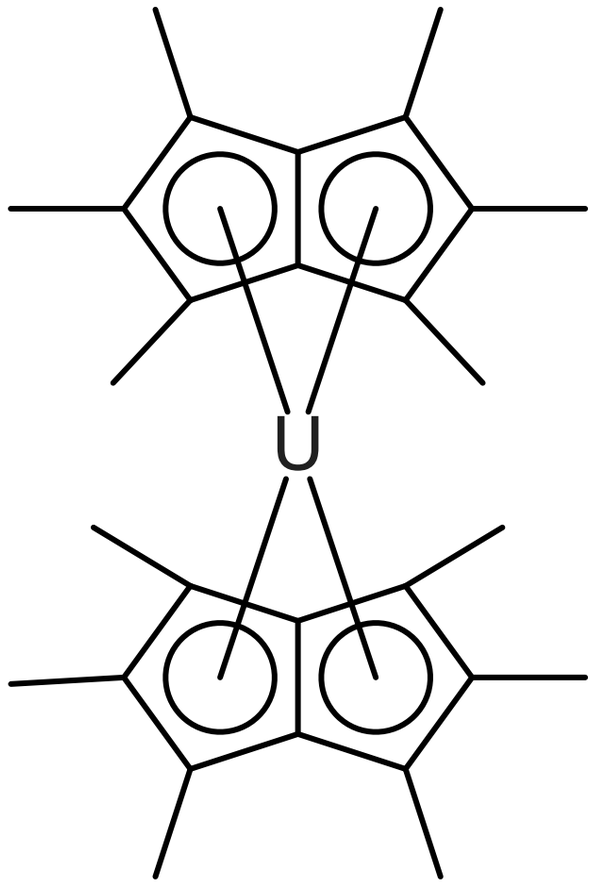
- Identified **17** types that need attention...



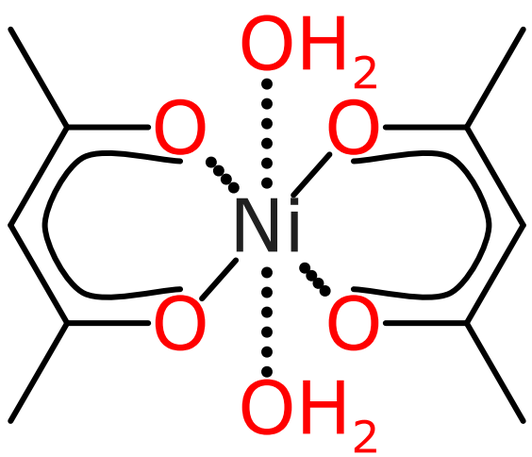
alkene



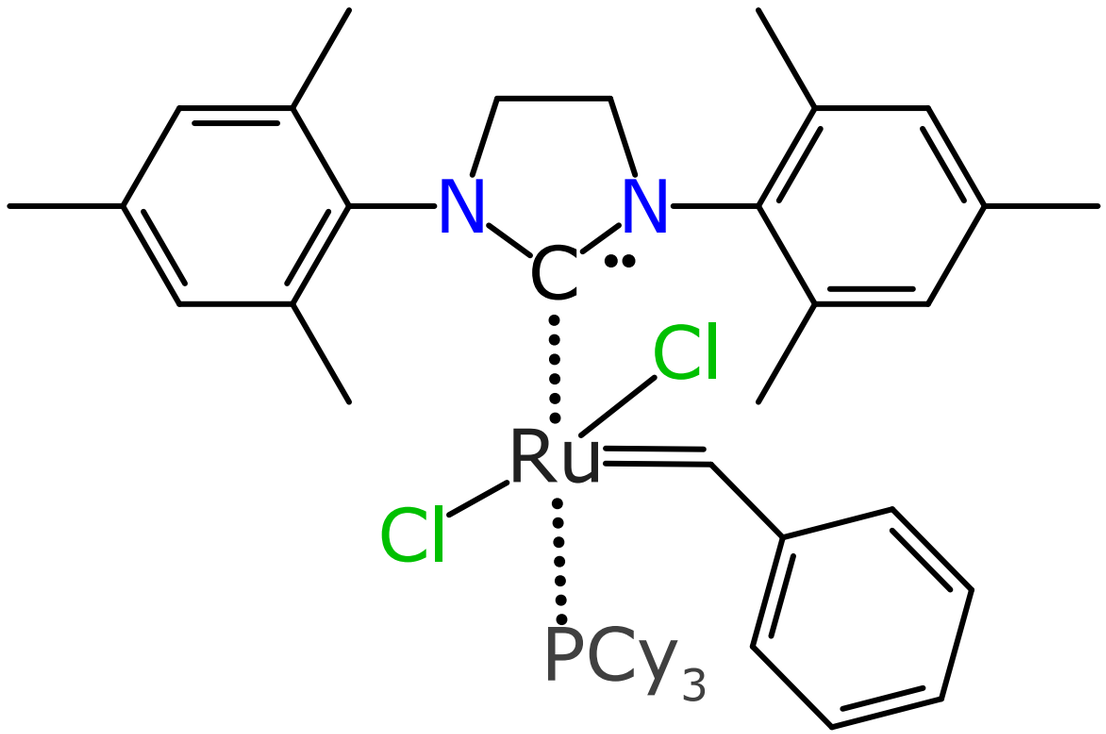
alternating



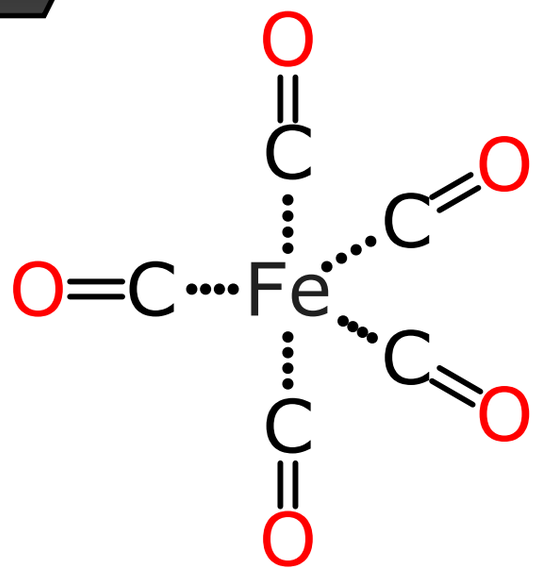
arene



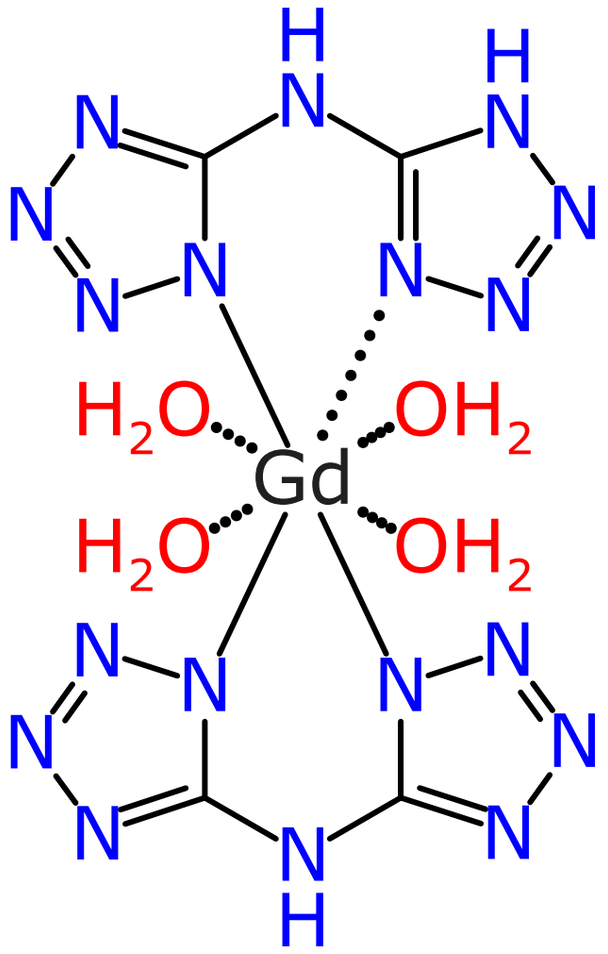
bidentate



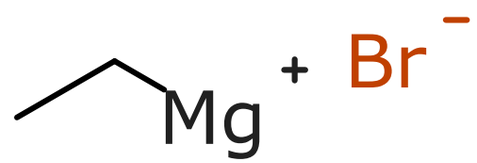
carbene



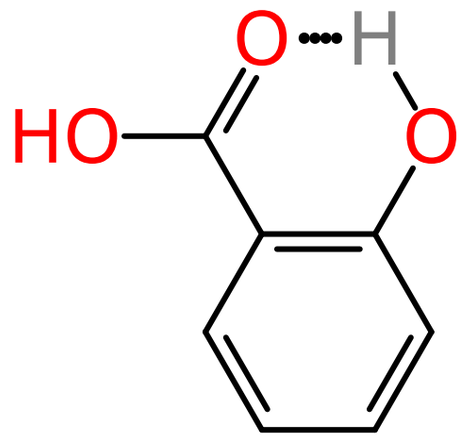
carbonyl



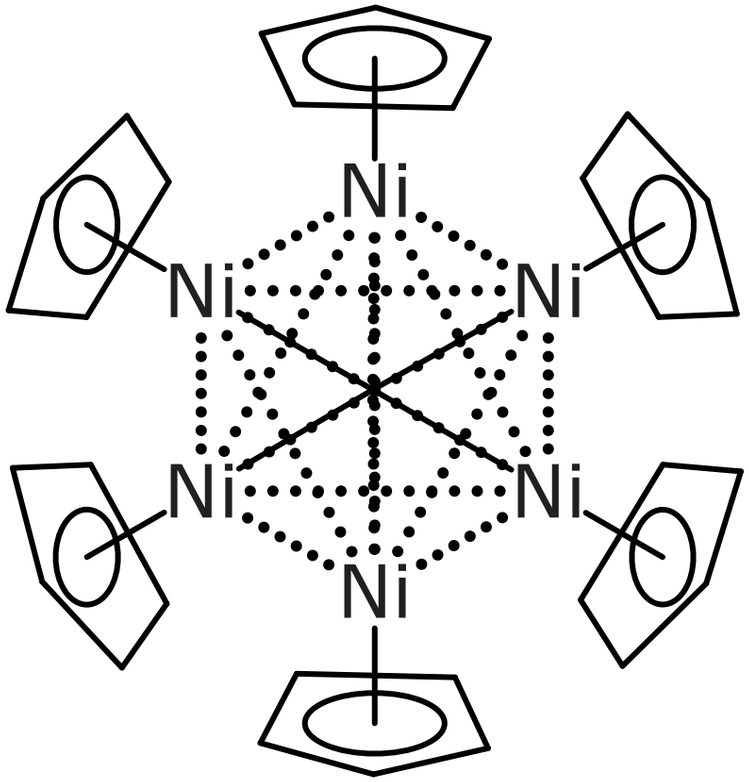
dative



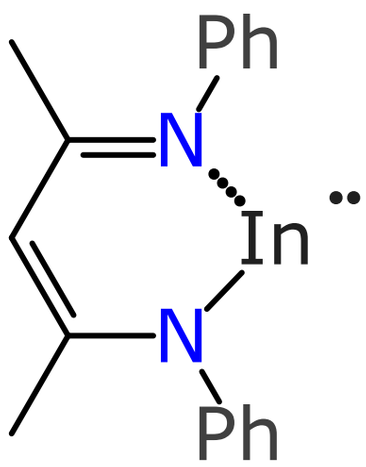
disconnected



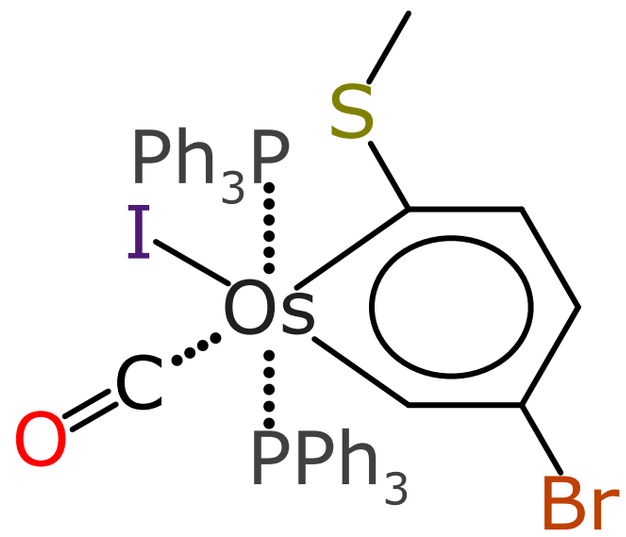
H-bond



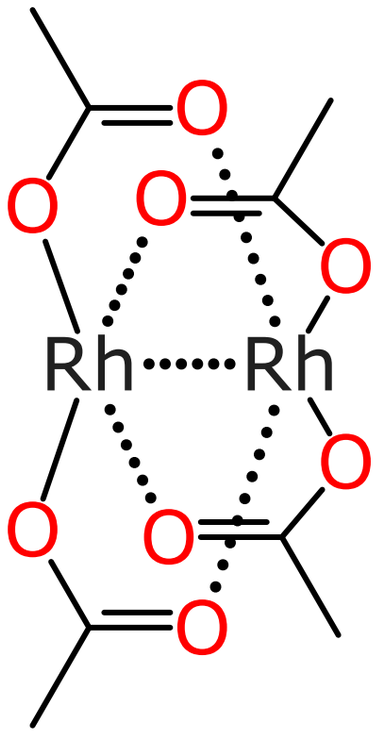
hypervalent



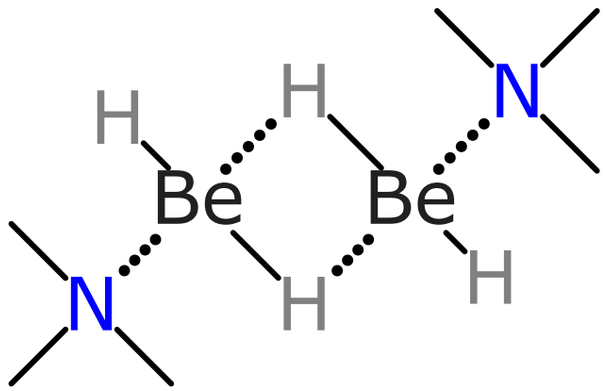
hypovalent



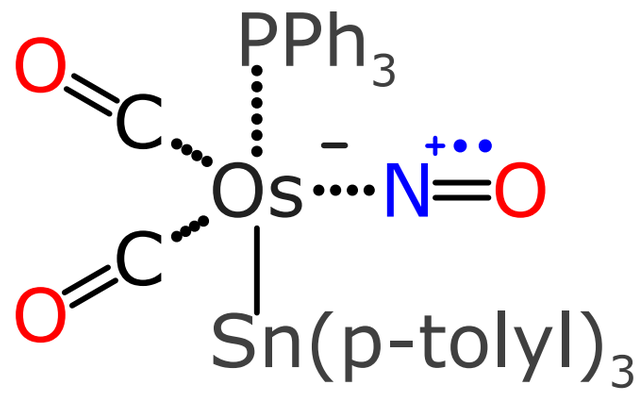
metallabenzene



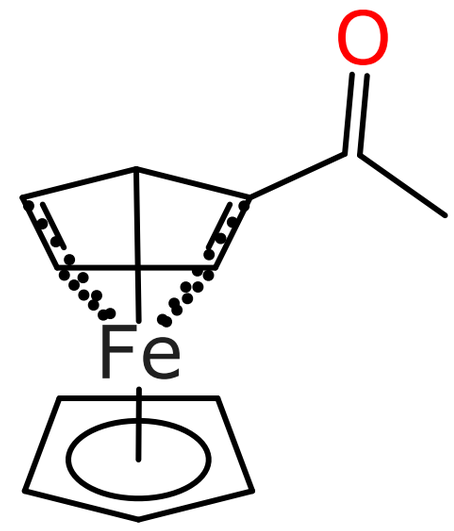
metal-metal



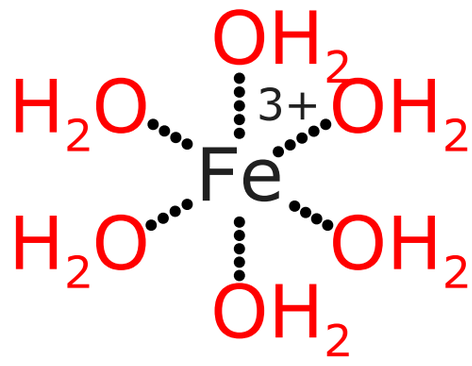
multicentre



nitrosyl



symmetry



terminal O

Core Datastructure

Rule 1

- If your representation does not imply the correct molecular formula

*then you are **wrong***

- Most cheminformatics formats/editors/use patterns fail this test for nontrivial inorganics

Rule 2

- In order of preference:

  (a) correct valence for early main groups
  (b) inferred electron delocalisation paths
  (c) realistic bond orders & formal charges
  (d) sensible oxidation states on metals
  (e) symmetry

- Usually possible to satisfy all conditions, with frequent exception of symmetry

Rule 3

- Non-trivial inorganics usually offer many correct ways to draw

- <u>Avoid overspecification</u>
  - more metadata can be added later

- Use only minimum information needed to:

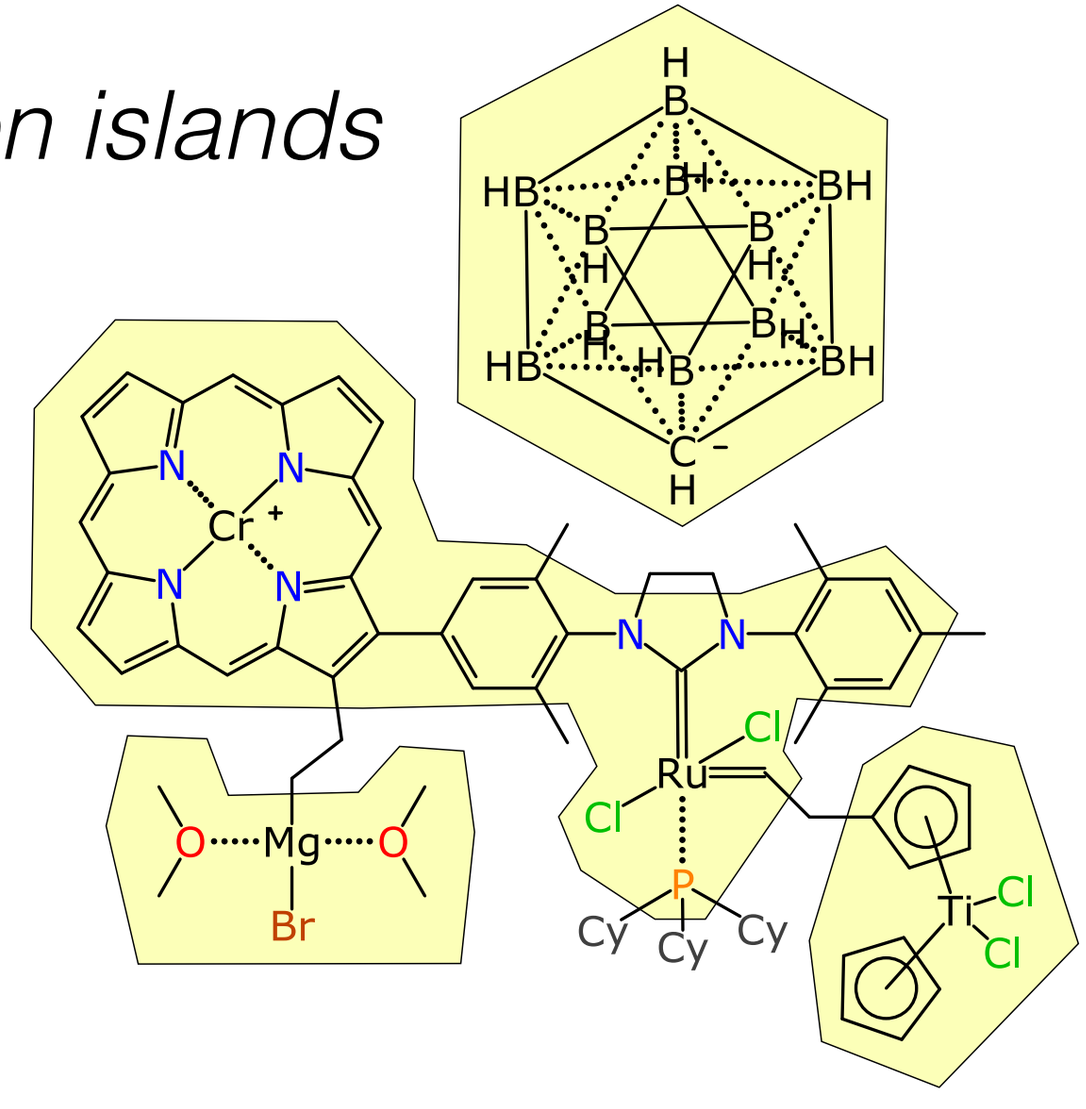
  - satisfy rule 1 (imply formula)
  - optimise for rule 2
  - resolve genuinely different molecules

Core Datastructure

Algorithm: Prerequisites

1. complete heavy atom graph
2. hydrogen counts
3. bond orders → *delocalisation islands*
4. net charges for each *island*

- **GIGO**



Algorithm: Implementation

- atom priority → [element, hcount, chg*]
- bond → <0, 0..1, 1, 1..2, 2, 2..3, 3+>
- iterate: atom priority → [a, ⇧{b_1, a_1}, {b_2, a_2}, …]
- if degenerate, *bump* lowest priority atom & repeat
- outcome: atom priority = walk order
- can now serialise in various different ways, e.g. SMILES-esque, InChI-esque

Algorithm: Outcome

- Algorithm weakest link is **detecting delocalisation islands**
- User weakest link is implying **correct hydrogen counts**
- Remarkably tolerant to multiple ways of drawing inorganic bonds
- Preliminary results are promising for disambiguating inorganics correctly