Representing molecules with minimalism: A solution to the entropy of informatics

Alex M. Clark

Collaborative Drug Discovery

Introduction

- InChI is great if...
  - applied within its domain
  - used for the problems it is designed to solve
  - supplied with appropriate data

J. Cheminf. ***7***:9 (2015)

Encoding Types

- 1D → ***identifier*** (InChI, SMILES)
- 2D
  - → ***diagram*** (ChemDraw etc.) optimised for <u>presentation</u>
  - → ***representation*** (Molfile) optimised for <u>information</u>
- 3D → ***conformation*** (QC, MM, crystals)

Sketch vs. Cheminformatics

- Most cheminformatics data comes in through *sketches*:
  - chemists use software designed to match conventions dating back to 19th century
- Software is given a choice:
  1. capture the stylistic choices (**diagram** primacy)
  2. capture the fundamental meaning (**data** purity)
- We need to do both...

Minimalism

- Want to describe the *intention* of the scientist who has made a discovery
  - capture the fact without destroying information
  - display it faithfully and recognisably
  - compare to other molecules
  - manipulate easily with many software tools
- Start with extremely minimalistic **primitives**...
- ... extend that with semi-optional **metadata**

Primitives

- Connection table style closely aligned with editor tools

- Atoms
  - **Symbol**
  - **Isotope**
  - **Charge** (integral)
  - Virtual **hydrogens** (explicit vs. computed)
  - **X**,**Y** coordinates

- Bonds
  - **From**, **To**
  - **Order** (0, 1, 2, 3, 4, ...)
  - **Wedge**: quasi-3D (stereochemistry in situ)

- Anything else is a nice-to-have...

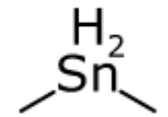
Practical Solution

- An actual format is necessary...
- ... MDL Molfile format is the *ad hoc* standard
  - originally poorly specified, though now much better
  - but too late: literally *thousands* of scripts/products
  - doesn't really matter what the spec says (cf. V3000)
- Effectively the lowest common denominator feature set of software in current use...
- ... need to expand this subset, while deprecating.

OpenMolecule Idea

- Determine a *level* of nonstandard features necessary
  - level **1.0**: features that every MDL Molfile reader & writer can handle with no information loss
  - level **1.x**: increasing likelihood of confusion
- Lowest level possible
  - mark internally
  - *or* quarantine

```
 1:
 2: OpenMolecule:1.3
 3:
 4:   3  2  0  0  0  0  0  0  0  0999 V2000
 5:    -0.6000    3.8500    0.0000 Sn  0  0  0  0  0  0  0  0  0  0  0  0
 6:    -1.8990    3.1000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 7:     0.6990    3.1000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 8:   1  2  1  0  0  0  0
 9:   1  3  1  0  0  0  0
10: M  HYD  1   1   2
11: M  END
12:
```



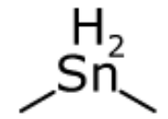
OpenMolecule Level 1.3
HydogenCounting Atoms:1 Level:1.3

- Keep using Molfiles *compatibly* until replacement...

OpenMolecule Idea

- Determine a *level* of nonstandard features necessary
  - level **1.0**: features that every MDL Molfile reader & writer can handle with no information loss
  - level **1.x**: increasing likelihood of confusion
- Lowest level possible
  - mark internally
  - *or* quarantine

```
1:
2: OpenMolecule:1.3
3:
4:   3  2  0  0  0  0  0  0  0  0999 V2000
5:    -0.6000    3.8500    0.0000 Sn  0  0  0  0  0  0  0  0  0  0  0  0
6:    -1.8990    3.1000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7:     0.6990    3.1000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8:   1  2  1  0  0  0  0
9:   1  3  1  0  0  0  0
10:M  HYD  1   1   2
11:M  END
12:
```



OpenMolecule Level 1.3
HydogenCounting Atoms:1 Level:1.3

- Keep using Molfiles *compatibly* until replacement...

Level 1.0: Minimal

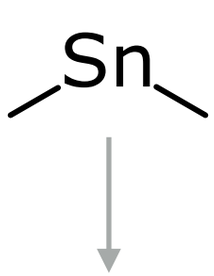
- Lowest common denominator of **all readers/writers**
  - the drug-like subset of organic chemistry
- To guarantee round-trip survival:
  - Atoms are atomic symbols
  - Bonds are 1, 2, 3
  - Chirality by wedges (not parity)
  - E/Z always specified by coordinates
  - Implied hydrogens never in doubt

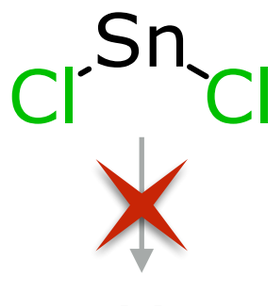
Level 1.1: Bigger

- 1000 atoms / bonds is a Molfile V2000 limit
- V3000 fixes this
- Limit rarely exceeded for most use cases
- V3000 is ugly, adoption slow
- Compliance means reading/writing V3000

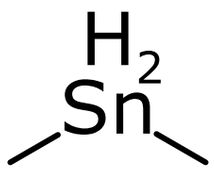
Level 1.2: H-control, zero bonds

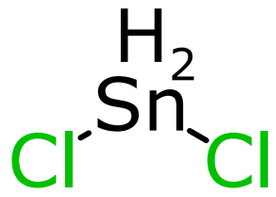
- Need to override virtual hydrogens *and* define a strict rule for implicit hydrogen counting





(II)





(IV)

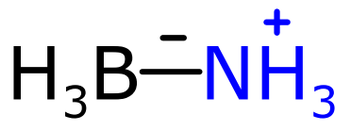
J. Chem. Inf. Model. **52**, 3149-3157 (2011)

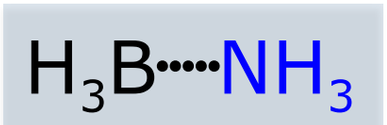
- Need a bond type that doesn't affect valence counting: allow 0-order

- Minimal changes - can describe a *lot* of inorganic chemistry - see later for beautification

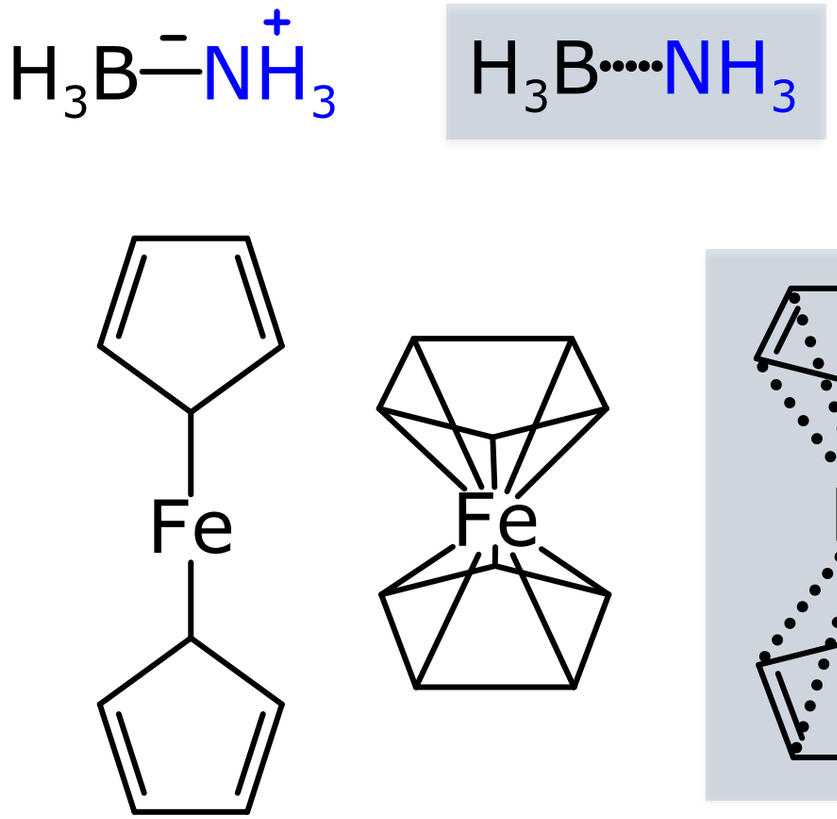
Inorganics

Inorganics

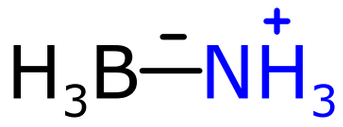


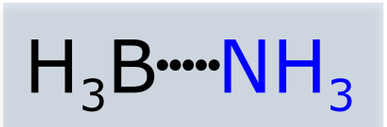


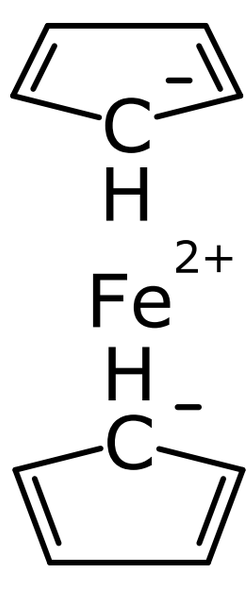
Inorganics

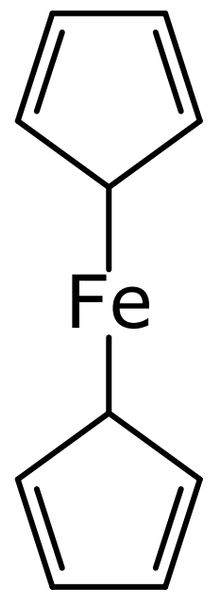


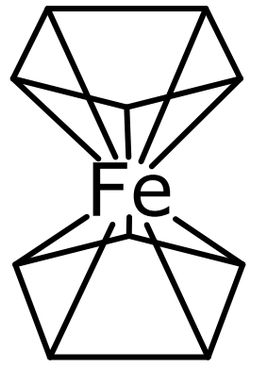
Inorganics

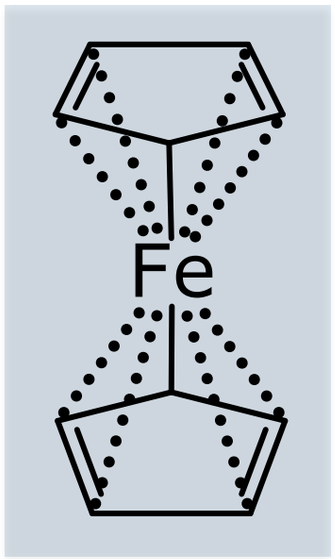

H3B—NH3

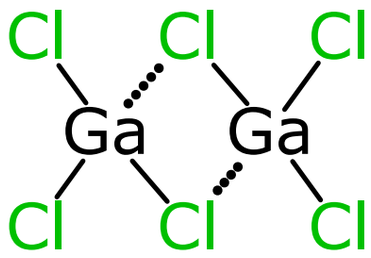

H3B····NH3

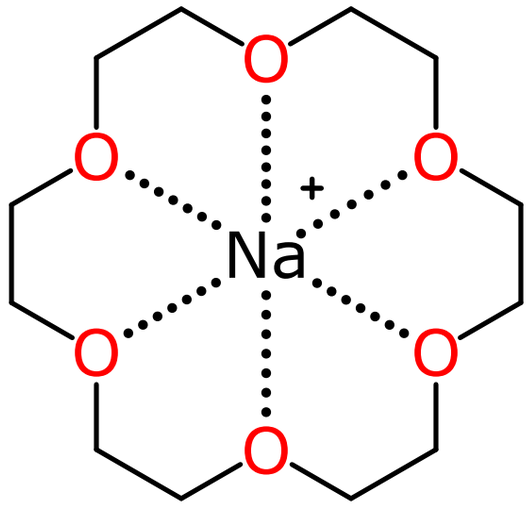


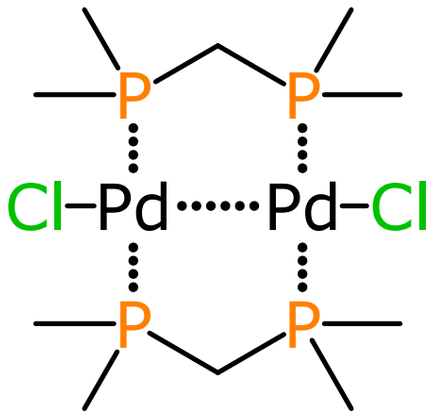


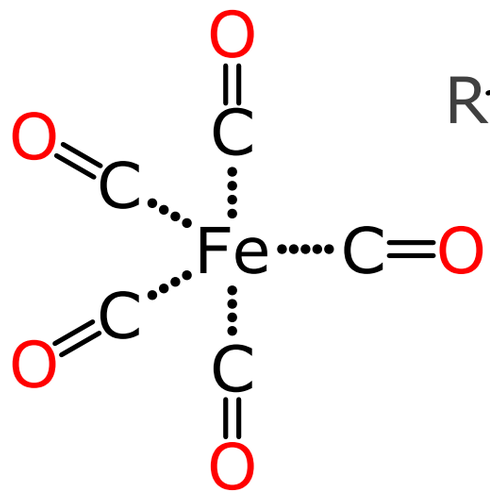

Fe

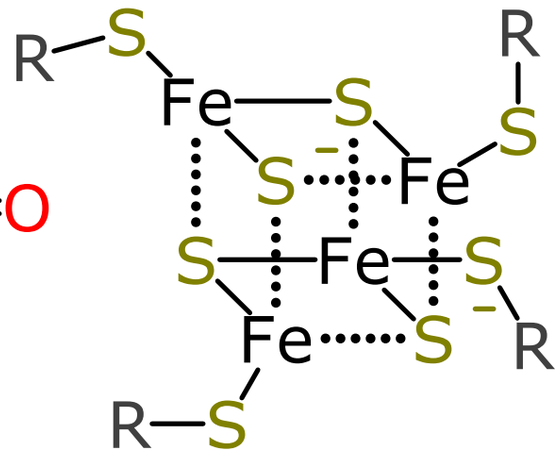

Fe


Cl
Cl
Cl
Ga
Ga
Cl
Cl
Cl


O
O
O
Na
+
O
O
O


P
P
Cl–Pd····Pd–Cl
P
P

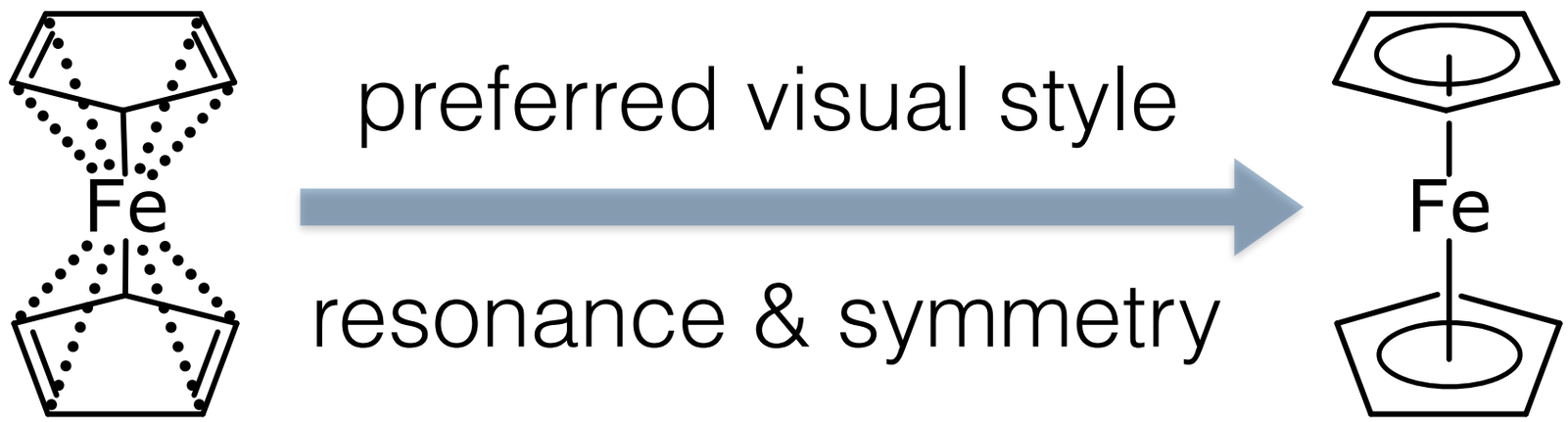

O
C
O=C
Fe····C=O
C
O
C
O


R
S
Fe
S
R
S
S
Fe
S
Fe
S
Fe
S
R
R—S

Molecule Metadata

- Metadata = optional information
  - can beautify the structure, or assist interpretation
  - **but** core representation is still meaningful & renderable

- Can help interpretation and/or aesthetics, e.g.

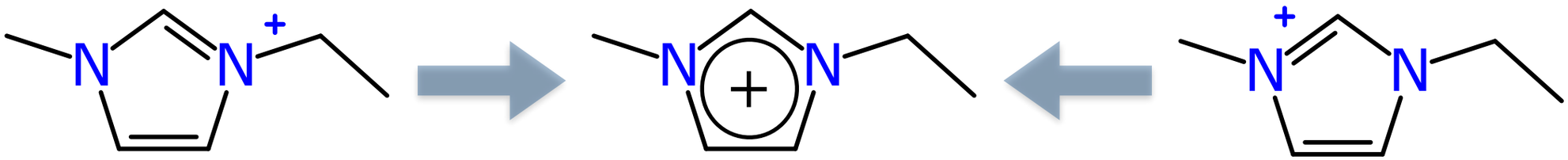


Informatics Metadata

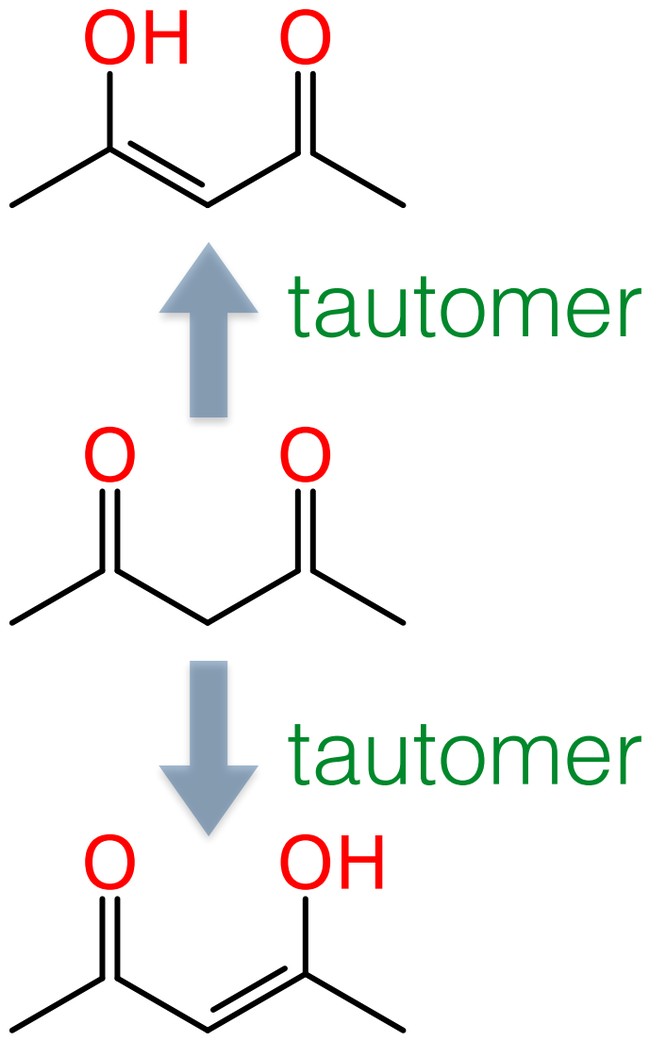
Informatics Metadata

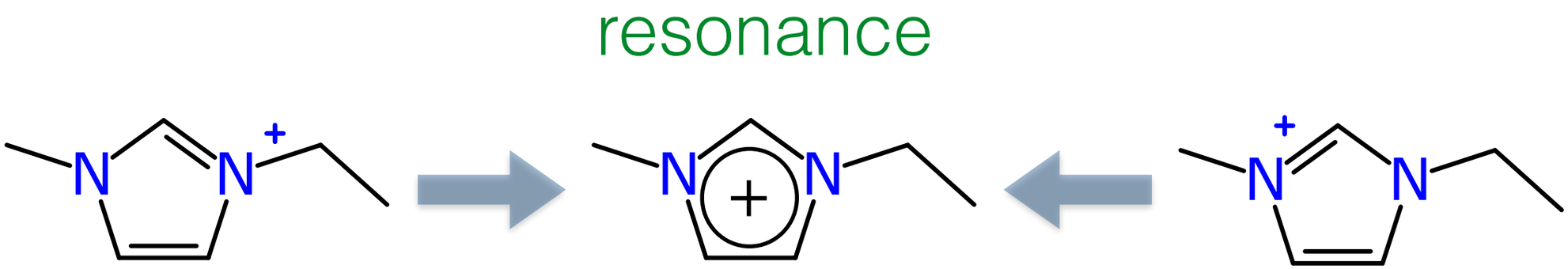
Informatics Metadata

resonance



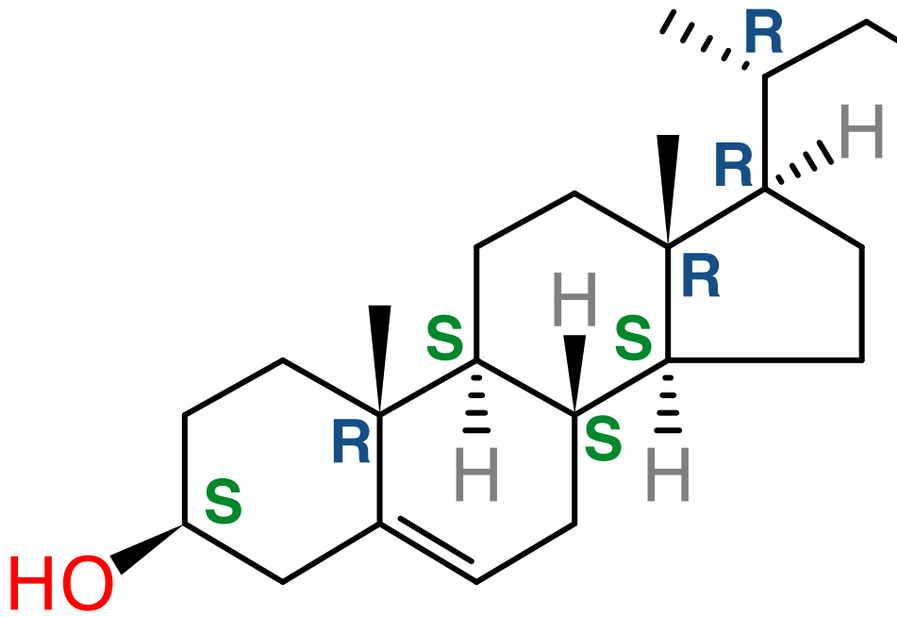
Informatics Metadata

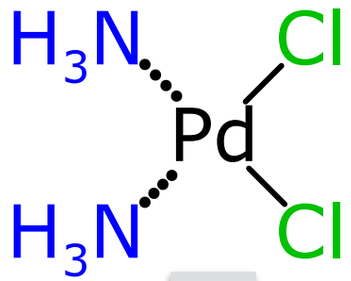




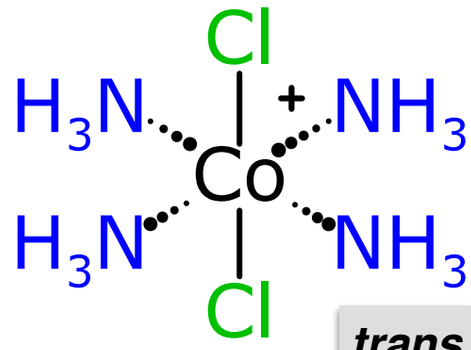
Stereochemistry

- Best to use 2D coordinates + wedges to derive...
- ... label/parity is metadata

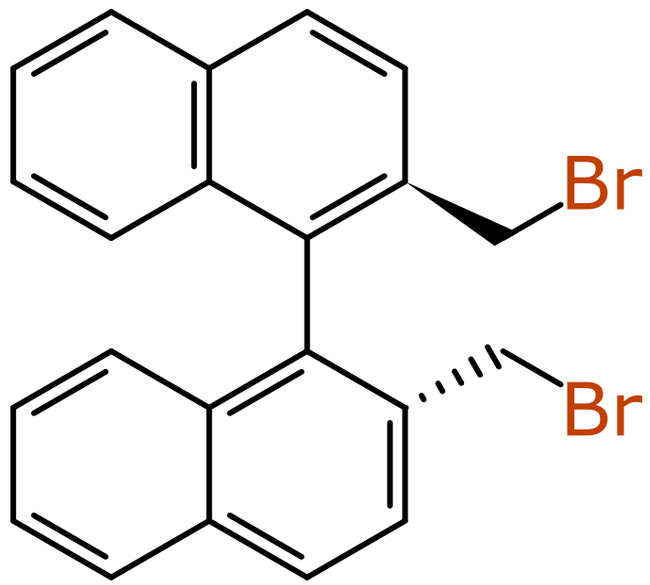




cis



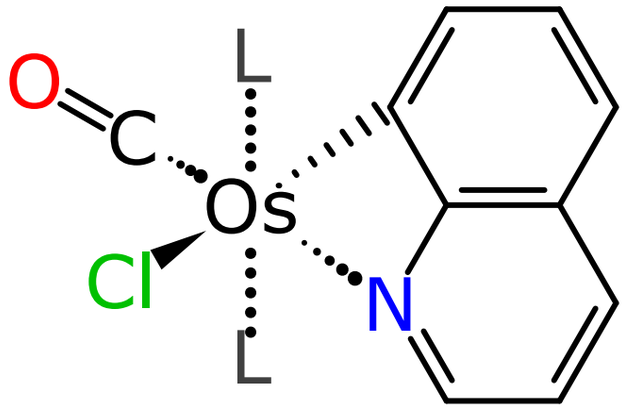
trans

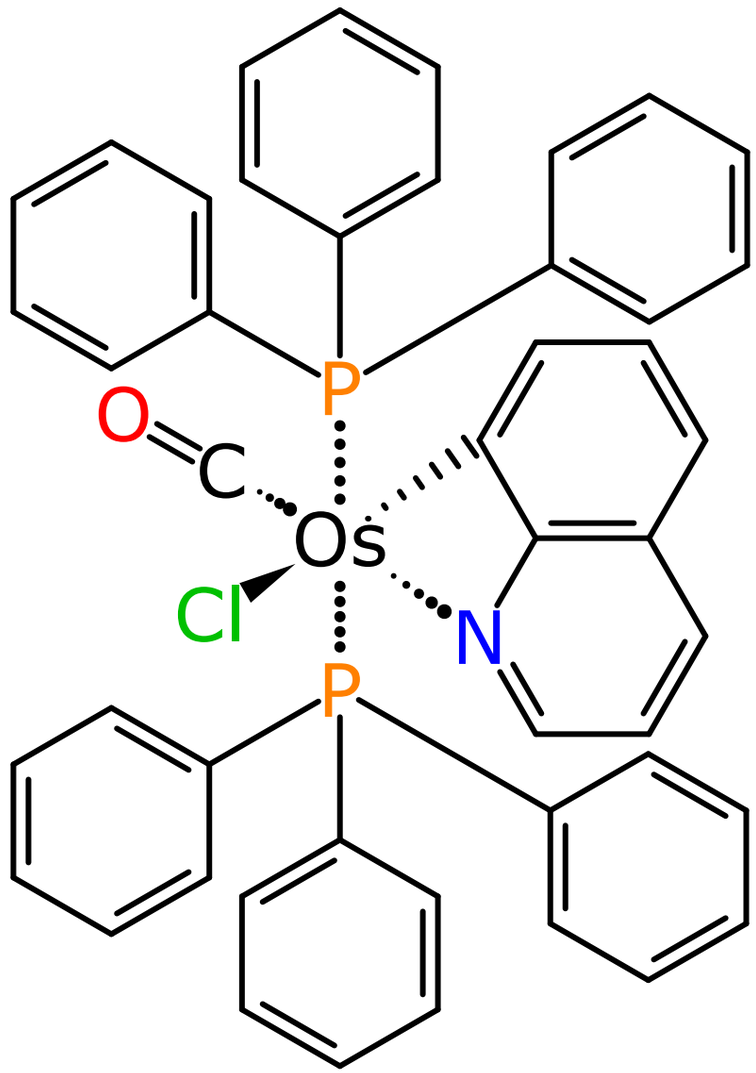


(R)-2,2′-Bis(bromomethyl)-1,1′-binaphthyl

Abbreviations

- Using abbreviations instead of all heavy atoms must be accompanied by a definition...
- ... everyone has their own
- Popular but nice-to-have in drug discovery; essential elsewhere
- Support it via S-group in *Molfile*, inline abbreviations in *SketchEl*
- *Must* be able to expand to get all atom

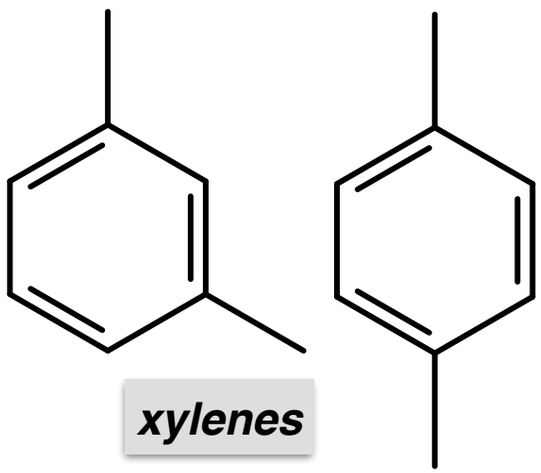




Enumeration Rules

- Formulae for multiple structures are *not* safe for naïve interpretation (exception to minimalist approach)

Mixtures

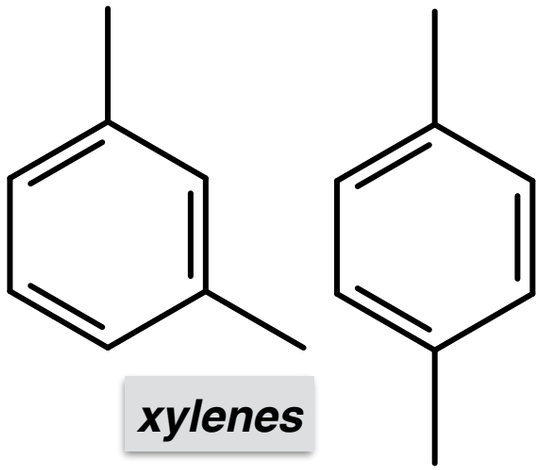


xylenes

Enumeration Rules

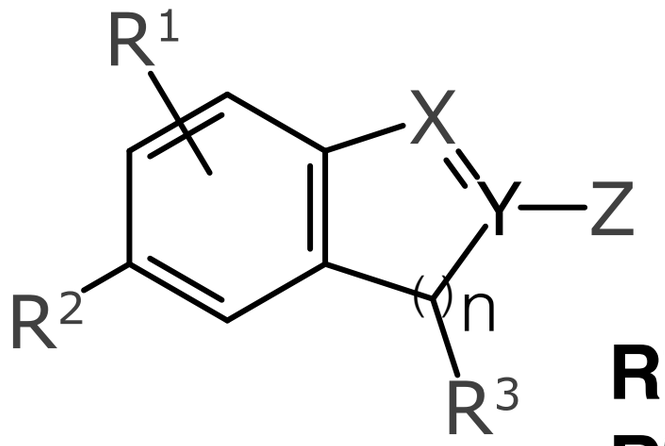
- Formulae for multiple structures are *not* safe for naïve interpretation (exception to minimalist approach)

Mixtures



xylenes

Markush

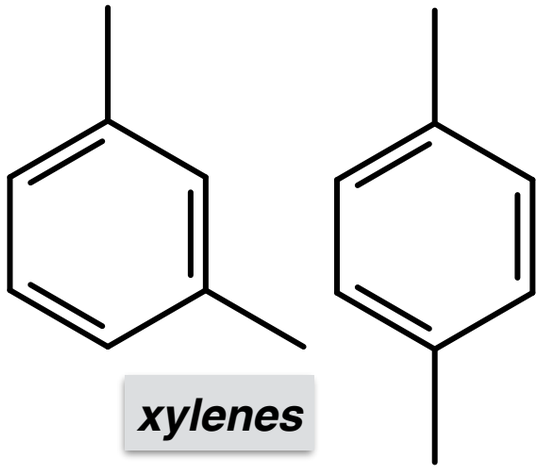


R¹ = H, CH_3
R² = Me, Et, Pr, Bu
R³ = Ph, tolyl
X = N, CH
Y = C, N^+
Z = Cl, CH_2Cl

Enumeration Rules

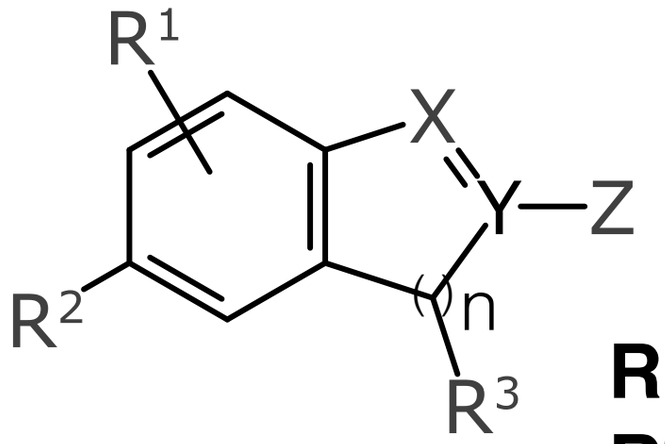
- Formulae for multiple structures are *not* safe for naïve interpretation (exception to minimalist approach)

Mixtures



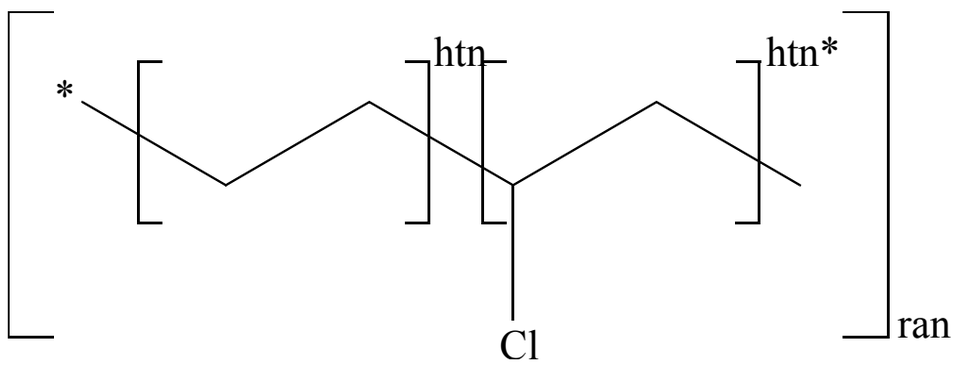
xylenes

Markush



R¹ = H, CH_3
R² = Me, Et, Pr, Bu
R³ = Ph, tolyl
X = N, CH
Y = C, N^+
Z = Cl, CH_2Cl

Polymers



Future Work

- Slumming it with molfiles:
  - agree on valid extensions
  - self-awareness of limitations
- Lossless conversions between different formats
- Chart a course away from Molfiles...