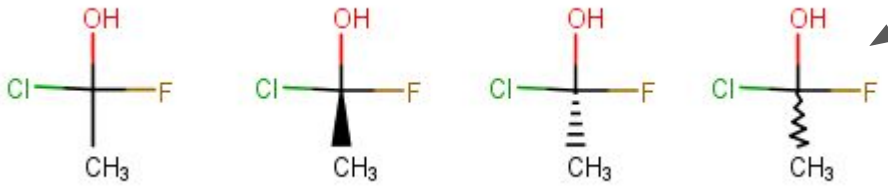
Thinking about the RDKit representation of stereochemistry

Greg Landrum

RDKit UGM 2019

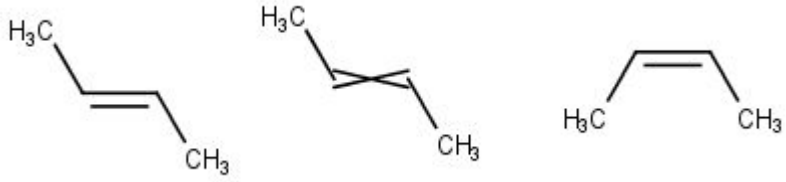
Status quo

Tetrahedral stereo



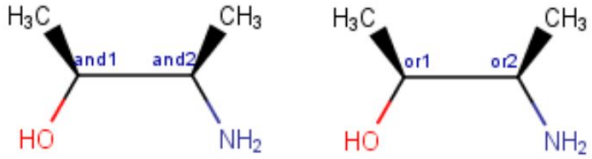
We don't really have any way to represent this

Double bond stereo



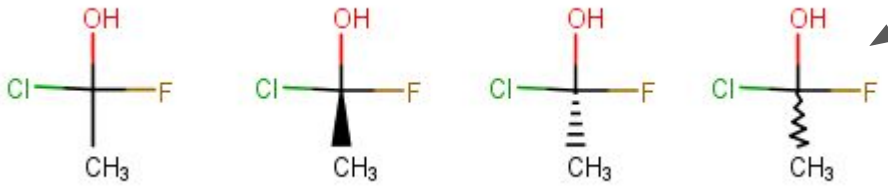
One more form you can't draw:
no information present

Stereo groups



Status quo

Tetrahedral stereo



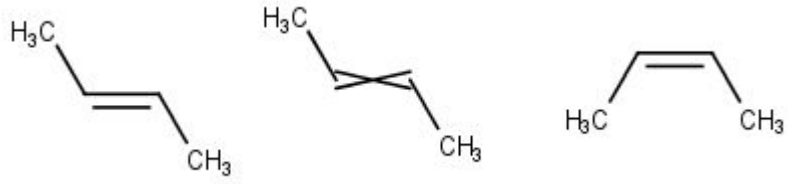
We don't currently have any way to represent this

Representation: atom attribute implicitly using the order of bonds around the atom

```
CHI_TETRAHEDRAL_CW, CHI_TETRAHEDRAL_CCW
```

Status quo

Double bond stereo



One more form you can't draw:
no information present

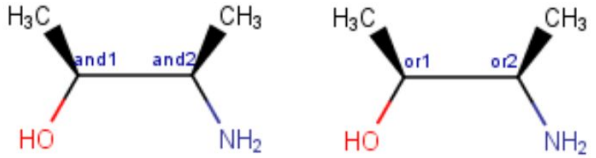
Representation: Bond attribute(s)

`STEREO_E, STEREO_Z`

`STEREO_CIS, STEREO_TRANS` (use stereoAtoms attribute)

Status quo

Stereo groups

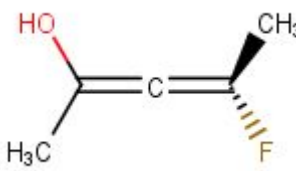


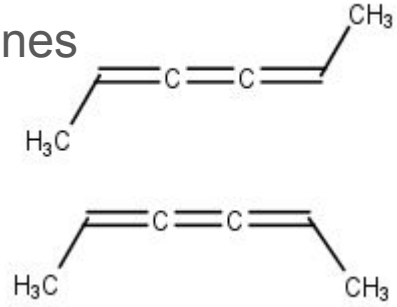
Representation: vector of `StereoGroup` objects on `ROMol`

More information in github:
https://github.com/rdkit/rdkit/blob/master/Docs/Code/EnhancedStereo.md

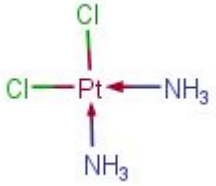
What's not handled?

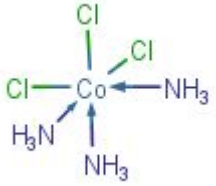
Allenes/cumulenes

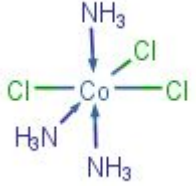




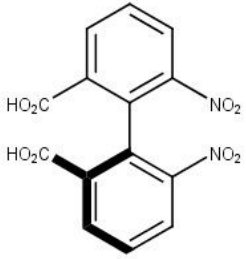
Non-tetrahedral atomic stereochemistry

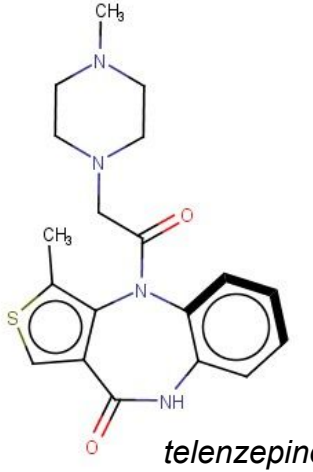






Atropisomers





telenzepine

Helicenes, etc.

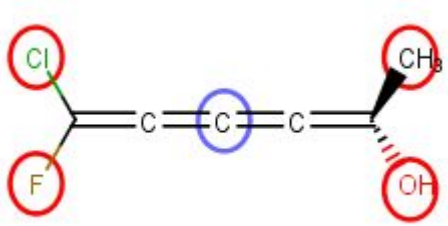
Things to think about

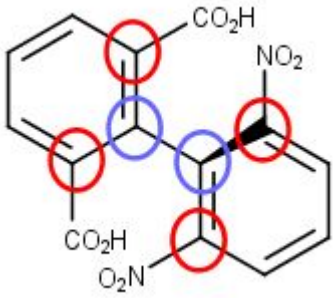
- What the input looks like: how do we know how to interpret the input correctly?
- What the internal representation is: how do we store it? At what level of granularity? Atom/bond property? Stored at the molecule level?
- Can we canonicalize it? How?
- What the public API is: how can client code interrogate/use stereochemistry?
- When it matters: when do we use the stereochemistry info? Just I/O? Drawing? Conformer generation? Substructure search? Fingerprints/descriptors? Elsewhere?
- How do we make clear the difference between when information has not been provided and when it is explicitly marked as unknown?

Simplifying the implementation

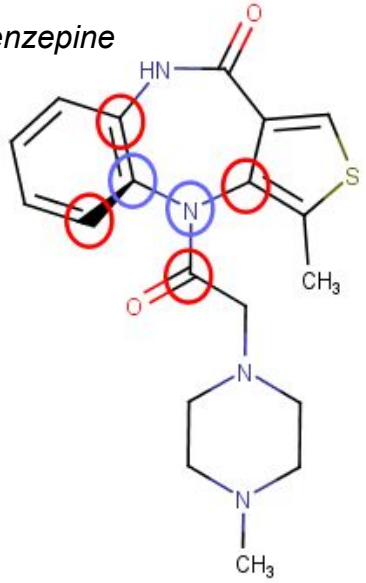
- Switch to using an explicit internal representation:
  - Directly store the atom order around a generalized stereocenter
  - Always store the atoms determining the stereochemistry of a double bond (currently done for cis/trans only)
  - ”generalized stereocenter”: an atom, set of atoms (position is defined by the mean if it’s a set of atoms), or bond
- Store stereo information on the molecule instead of on atoms/bonds
  - Still provide API functionality to retrieve whatever stereo info is present about atoms/bonds
  -

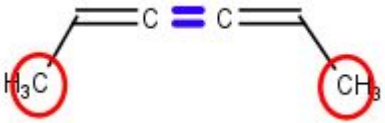
“Generalized stereocenters”





telenzepine





This isn’t really a generalized stereocenter. It’s a bond whose stereochemistry is determined by non-adjacent atoms