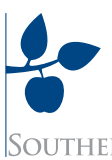
Modelling of Stereoisomerism for Generative Chemistries

Jakob Lykke Andersen, Christoph Flamm,
Daniel Merkle, Peter F. Stadler

Department of Mathematics and Computer Science
University of Southern Denmark

Bled, February 2015



University of Southern Denmark

A Note on Terminology

External Representation

The external storage format used for data exchange.

(E.g., a molecule is stored as a SMILES string, or InChI string)

Internal Representation (Implementation)

The data structures used to represent the the model, and the algorithms to manipulate the data.

(E.g., a molecule is an adjacency list with . . .)

Model

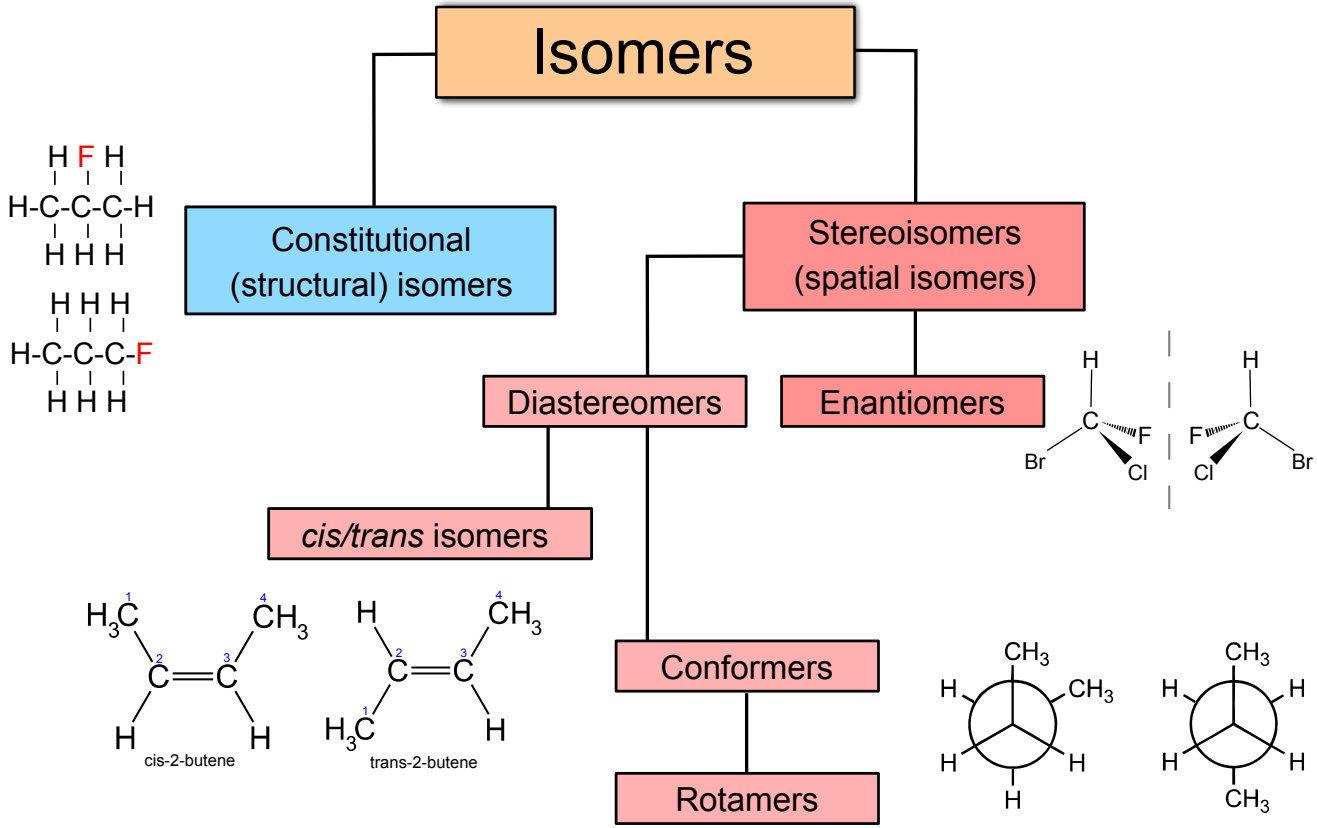
The abstract mathematical description of objects and their semantics.

(E.g., a molecule is an connected, undirected, simple, labelled graph.)

Reality

???

Isomers

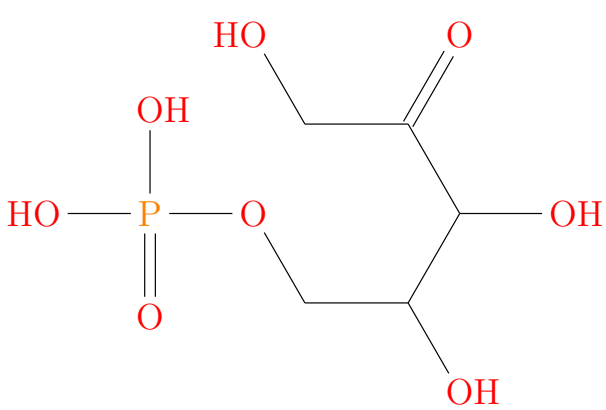


"Isomers are molecules with the same chemical formula but different chemical structures." [Wikipedia, Isomer]

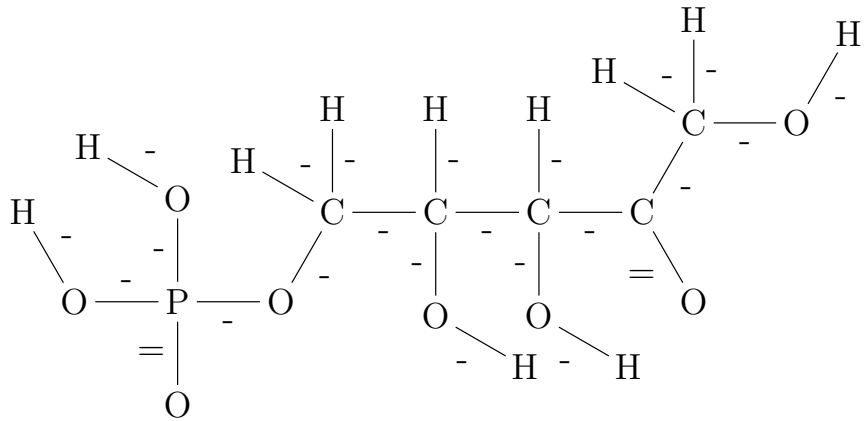
(not to be confused with the "structure" in "structural isomers")

Our Current Molecule Model

A molecule is a connected, undirected, simple, labelled graph.



(a) A molecule

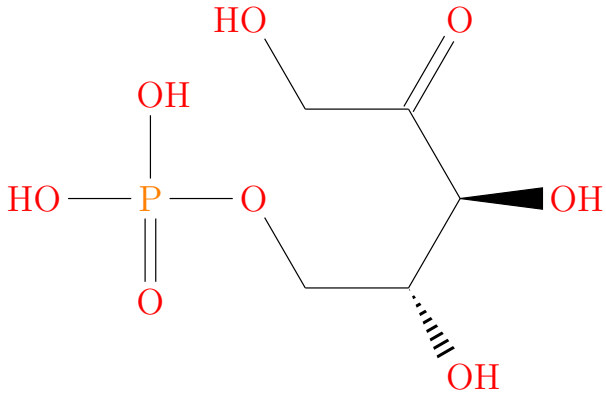


(b) A model of a molecule

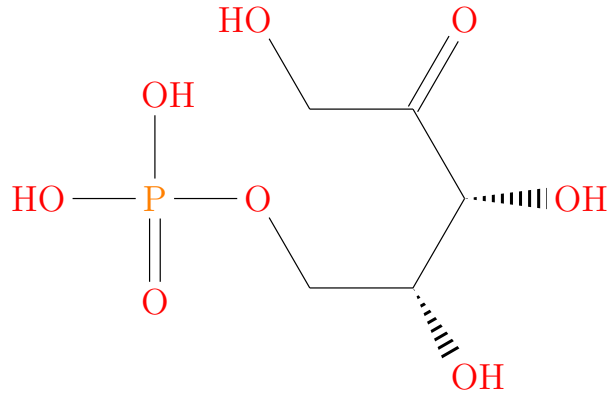
We can distinguish between constitutional (structural) isomers. . .

Our Current Molecule Model

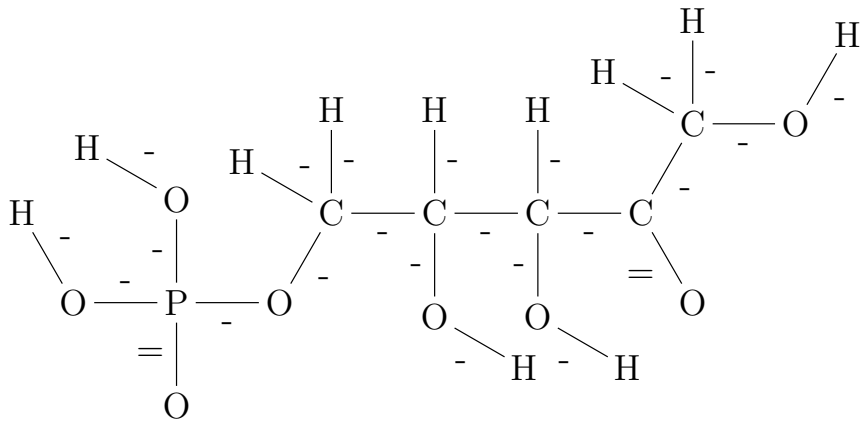
. . . but not stereoisomers.



(a) Ribulose 5-phosphate



(b) Xylulose 5-phosphate



(c) The model for both molecules.

Our Current Molecule Model

Data Structures

- Graphs with labels (adjacency lists with strings).

  Only local information about atoms and bonds.

Algorithms

- Graph isomorphism

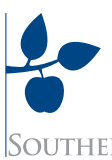
  Are two data structures representations of the same graph?
- Subgraph monomorphism

  Pattern matching for graphs. Substructure search.
- Composition of transformation rules

  Generalised graph transformation. Computing reactions.
- Graph canonicalisation

  Faster isomorphism check. Making "comfortable" storage formats.

We would like to model stereochemistry as well.



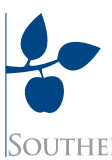
Extension for Modelling (Some) Stereochemistry

Goals

- Molecules are still graphs, but now with more information.
- Information is still localised on atoms and bonds.
- It should really be an extension: the current algorithms are simplifications of the new algorithms.

Limitation

- Only local geometry (or derived thereof) can be modelled.



Extension for Modelling (Some) Stereochemistry

Goals

- Molecules are still graphs, but now with more information.
- Information is still localised on atoms and bonds.
- It should really be an extension: the current algorithms are simplifications of the new algorithms.

Limitation

- Only local geometry (or derived thereof) can be modelled.

Data Structures

- Each edge (bond): A *behaviour* (usually the bond type)
- Each vertex (atom):
  - Number of incident lone pairs.
  - A geometry tag.
  - An ordered list of incident edges and lone pairs.

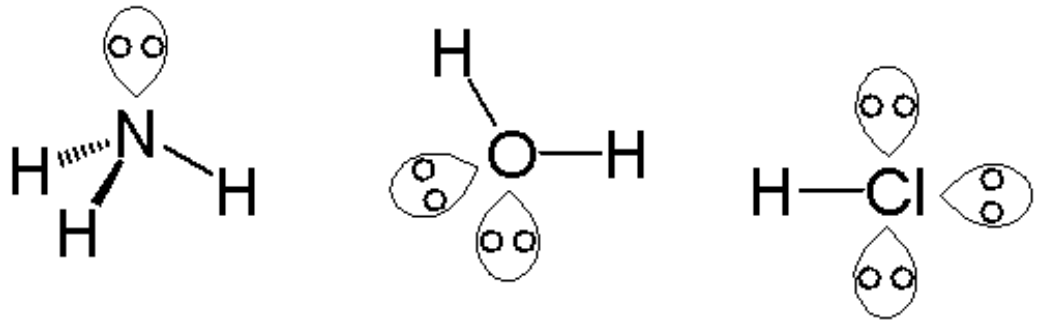
Based on "the ordered list method".

[Petrarca et al., J. Chem. Doc., 1967]

[Wipke and Dyott, J. Am. Chem. Soc., 1974]

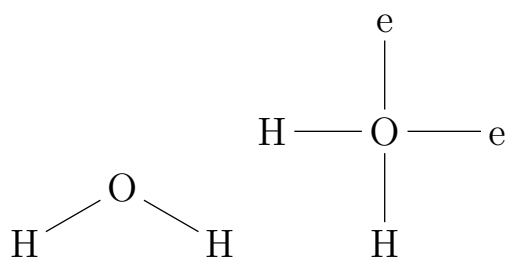
Lone Pair-Augmentation

Lone pairs contribute to the geometry.



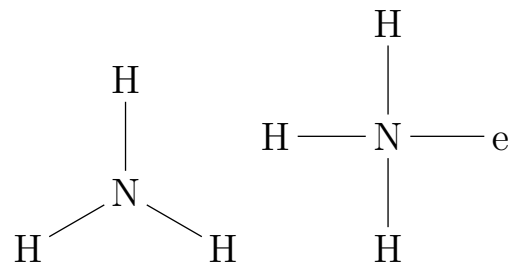
[Wikipedia, Lone Pair]

Add a virtual edge and vertex for each lone pair:



(a) Before (b) After

Oxygen in water.



(a) Before (b) After

Nitrogen in ammonia.

(A virtual edge has single bond behaviour as default.)

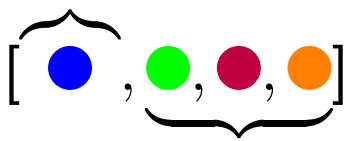
Example: TetrahedralFree

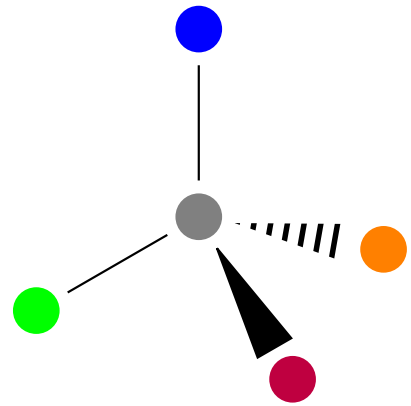
Tetrahedral $\equiv$ 4 neighbours, tetrahedron shape

Free $\equiv$ all have single bond behaviour

(E.g., a carbon with 4 bonds)

Ordering Semantics





Example: TetrahedralFree

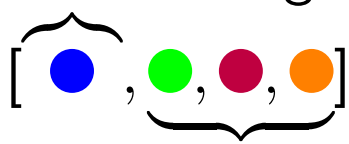
TETRAHEDRAL $\equiv$ 4 neighbours, tetrahedron shape

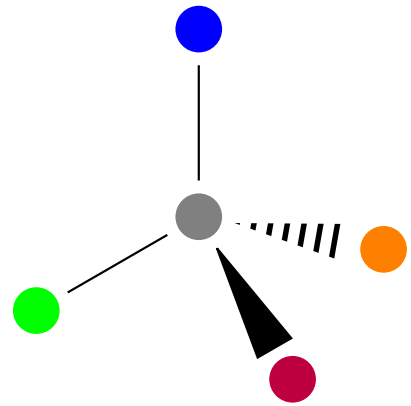
FREE $\equiv$ all have single bond behaviour

(E.g., a carbon with 4 bonds)

Ordering Semantics

"up" is where hte first neighbour points





Example: TetrahedralFree

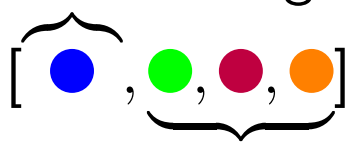
TETRAHEDRAL ≡ 4 neighbours, tetrahedron shape

FREE ≡ all have single bond behaviour

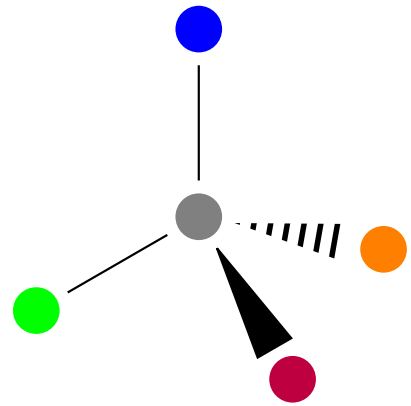
(E.g., a carbon with 4 bonds)

Ordering Semantics

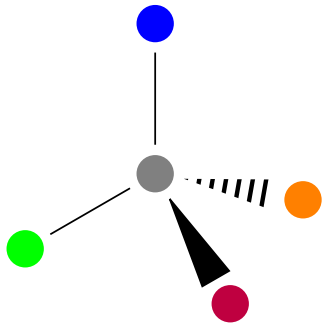
"up" is where hte first neighbour points

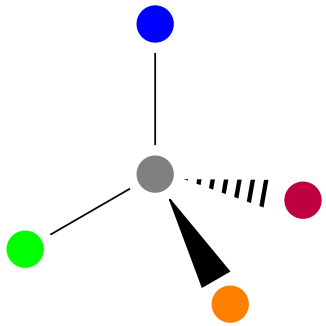


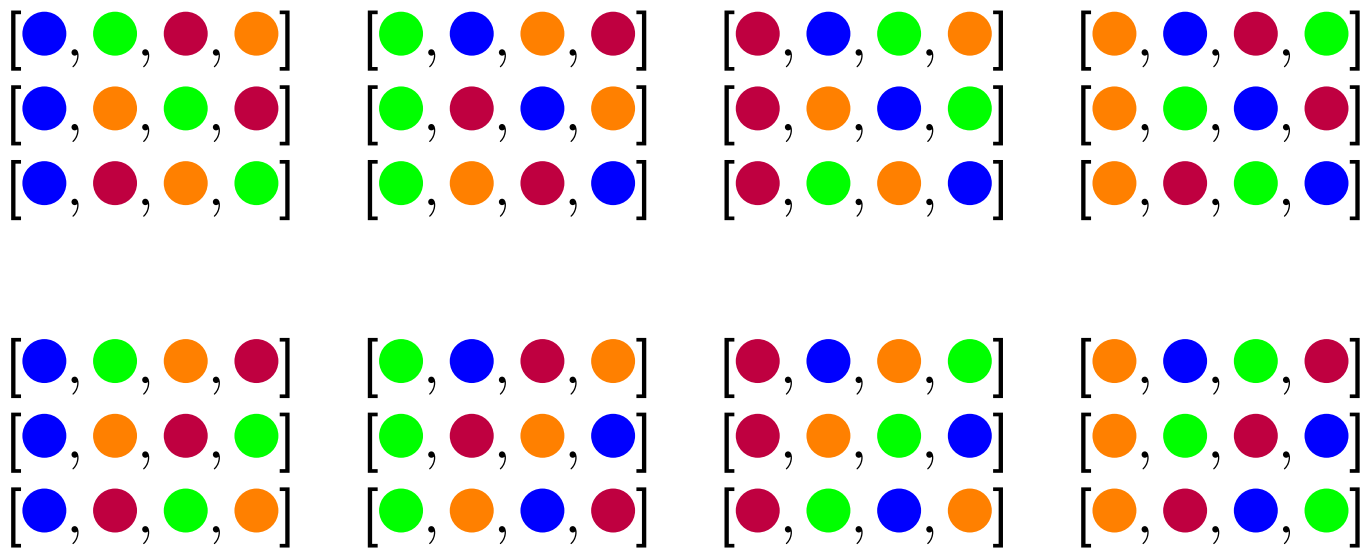
in positive order from above



Example: TetrahedralFree







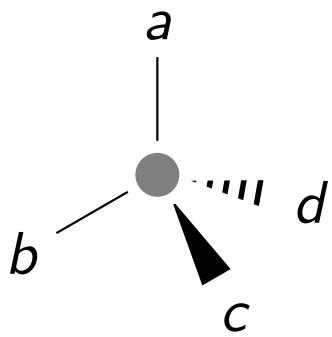
Equivalence permutation group: $G_{\equiv} = \langle (1)(2\ 3\ 4), (1\ 2)(3\ 4) \rangle$

Non-equivalence permutations: $G_{\not\equiv} = G_{\equiv} \circ (1)(2)(3\ 4)$

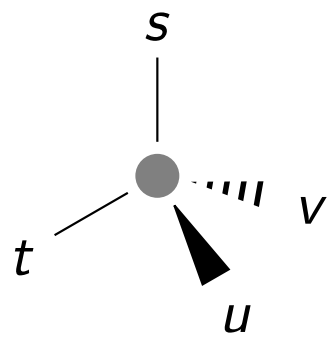
Example: TetrahedralFree, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



$[a, b, c, d]$

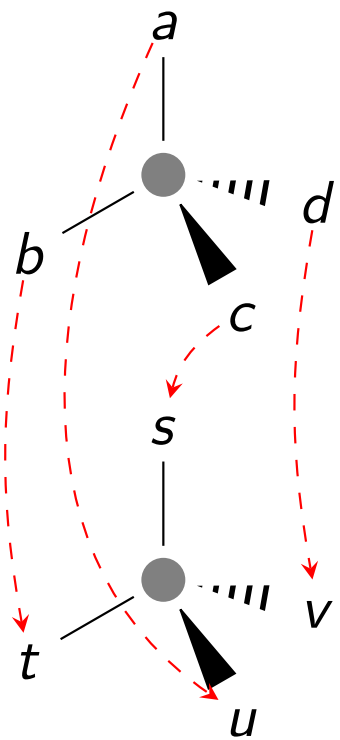


$[s, t, u, v]$

Example: TetrahedralFree, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



$[a, b, c, d]$

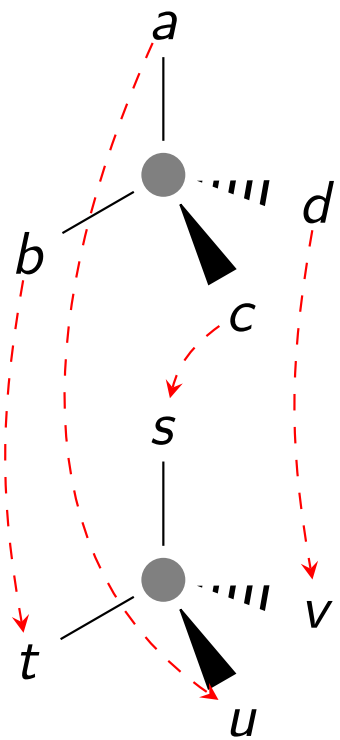
Induced permutation: (1 3)(2)(4)

$[s, t, u, v]$

Example: TetrahedralFree, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



$[a, b, c, d]$

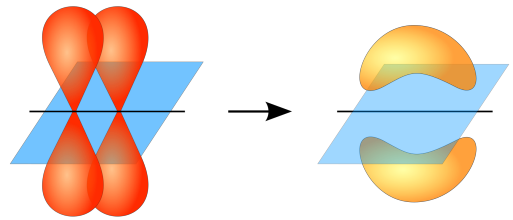
Induced permutation: $(1\ 3)(2)(4) \in G_{\not\equiv}$

$[s, t, u, v]$

Edge Behaviours (Bond Types)

- SINGLE: no rotational constraints.
- DOUBLE: inhibits rotation, 1 reference half-plane.

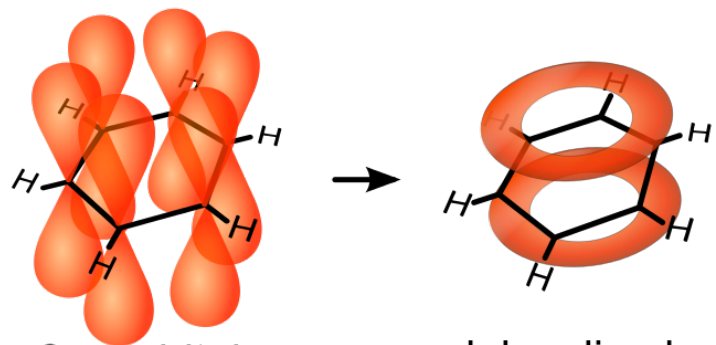
Formation of π-bond:



[Wikipedia, Pi bond]

- TRIPLE: inhibits rotation, 2 reference half-planes.
- CONJUGATED: inhibits rotation, 1 reference half-plane.

Formation of conjugated bonds:



[Wikipedia, Conjugated system]

Example: TrigonalD

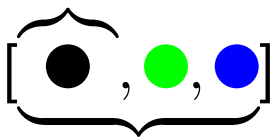
TRIGONAL $\equiv$ 3 neighbours, planar shape

D $\equiv$ 1 double bond, 2 single bonds

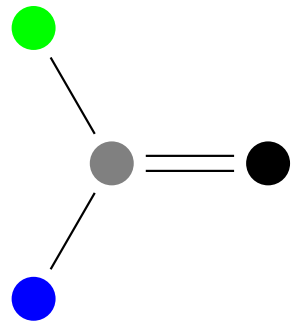
Ordering Semantics

The ordering defines a reference half-plane.

always the double bond



in positive order from above

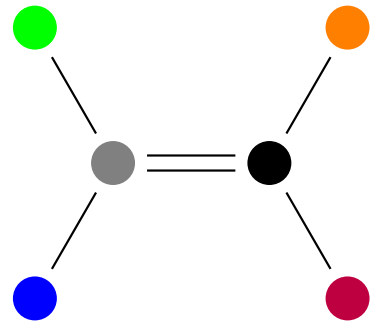


- Incident reference half-planes are equal.
- [●, ●, ●] and [●, ●, ●] have opposite half-planes.
  Half-plane-swapping permutation(s): $G_{\circlearrowleft} = \{(1)(2\ 3)\}$.
- $G_{\equiv} = \{(1)(2)(3)(4)\}$, $G_{\not\equiv} = \emptyset$

Example: TrigonalD

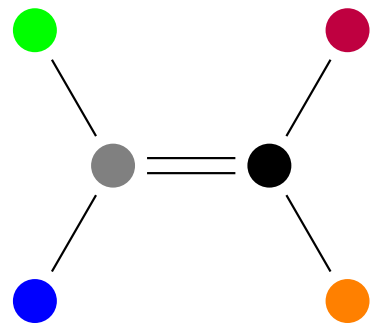
Either ●: [●, ●, ●] ●: [●, ●, ●]

or ●: [●, ●, ●] ●: [●, ●, ●]



Either ●: [●, ●, ●] ●: [●, ●, ●]

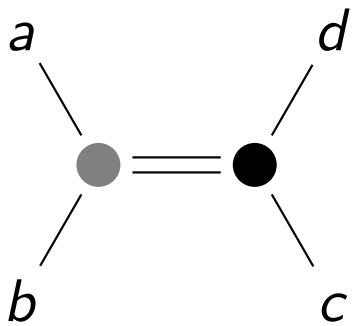
or ●: [●, ●, ●] ●: [●, ●, ●]



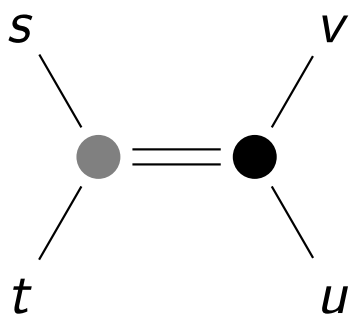
Example: TrigonalD, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



●: [●, a, b] ●: [●, c, d]

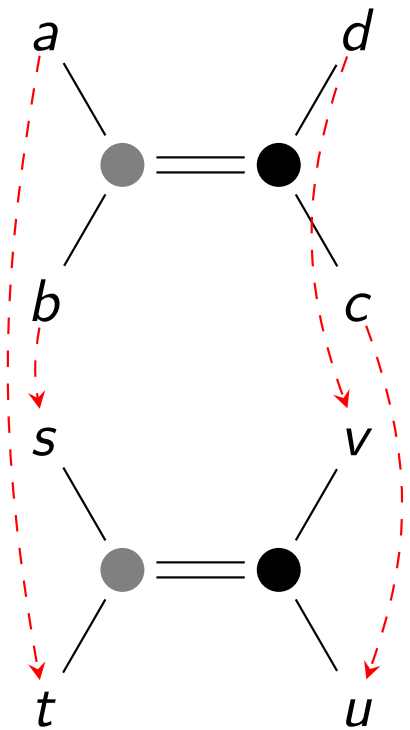


●: [●, s, t] ●: [●, u, v]

Example: TrigonalD, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



●: [●, a, b] ●: [●, c, d]

Induced permutations:

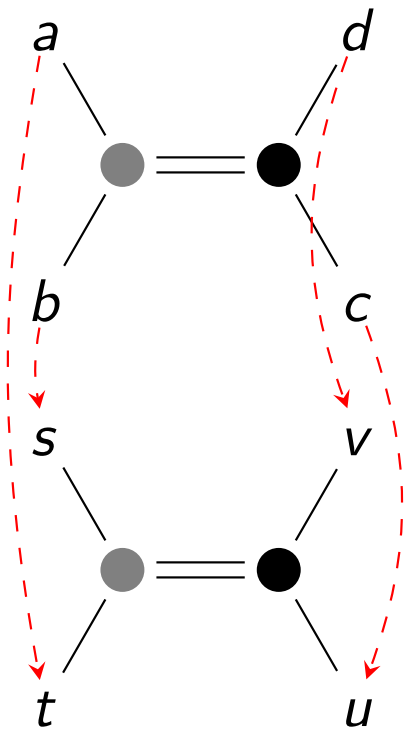
●: (1)(2 3) ●: (1)(2)(3)

●: [●, s, t] ●: [●, u, v]

Example: TrigonalD, Isomorphism

Given a graph isomorphism.

Decide if it is still valid with stereo information.



●: $[●, a, b]$ ●: $[●, c, d]$

Induced permutations:

●: (1)(2 3) ●: (1)(2)(3)

●: $[●, s, t]$ ●: $[●, u, v]$

● swaps half-plane, ● does not ⇒ invalid isomorphism.

Example: LINEARDD

LINEAR $\equiv$ 2 neighbours, linear shape

DD $\equiv$ 2 double bonds

Ordering Semantics

The ordering does not matter:

[●, ●] and [●, ●] mean the same.

I.e., $G_{\equiv} = \langle (1\ 2) \rangle$

Half-plane Propagation

The other half-plane is at 90°, seen from either end.

Example: LINEARDD

LINEAR $\equiv$ 2 neighbours, linear shape

DD $\equiv$ 2 double bonds

Ordering Semantics

The ordering does not matter:

[●, ●] and [●, ●] mean the same.

I.e., $G_{\equiv} = \langle (1\ 2) \rangle$

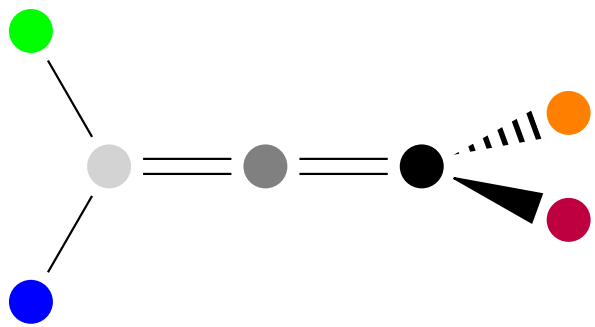
Half-plane Propagation

The other half-plane is at 90°, seen from either end.

Example

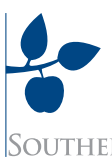
●: [●, ●, ●]

●: [●, ●, ●]



(Partially) Unspecified Stereo Information

- Our current molecules have no information.
  (absence of information $\equiv$ completely unspecified information)
- Parts of a molecule may have unspecified information.
- Part of a local configuration may be unspecified
  (e.g., in trigonal bipyramidal geometry).



(Partially) Unspecified Stereo Information

- Our current molecules have no information.
  (absence of information $\equiv$ completely unspecified information)
- Parts of a molecule may have unspecified information.
- Part of a local configuration may be unspecified
  (e.g., in trigonal bipyramidal geometry).

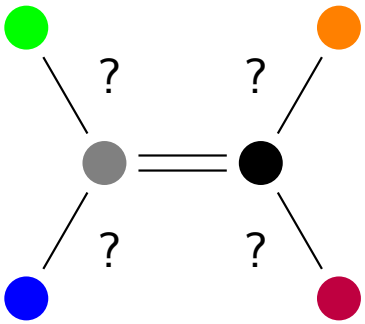
Data Structure

- Attach a "fixed"-flag to each ordering element.
  (they may not be mutually independent)

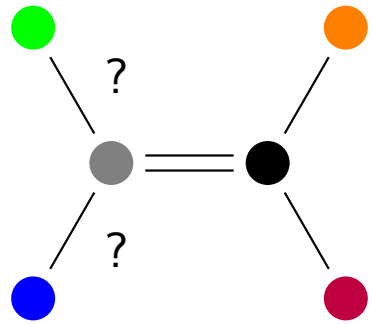
Semantics

- Fixed elements may still be moved by $G_{\equiv}$.
- Moves permutations with the non-stabilised elements all being non-fixed into $G_{\equiv}$.
- Variations of the same molecule are now partially ordered by generality/specificity.

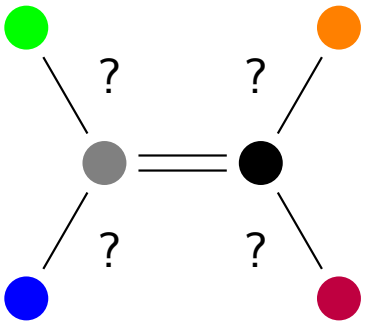
(Partially) Unspecified Stereo Information



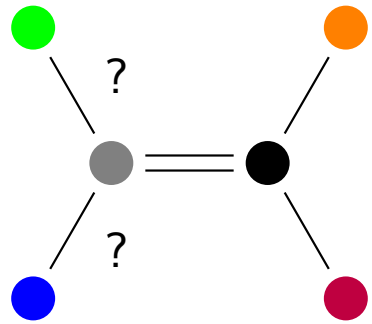
is less (specific) than
is not isomorphic to
can unify with

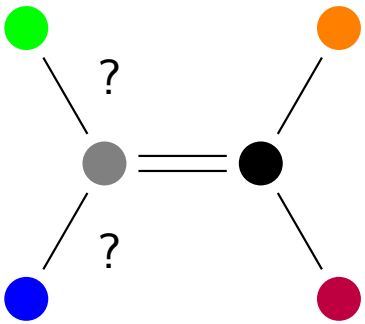


(Partially) Unspecified Stereo Information

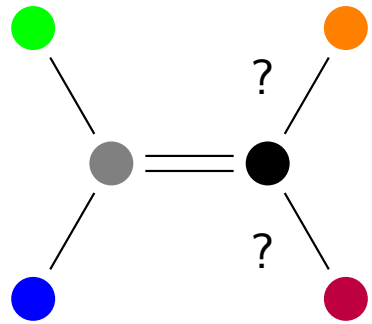


is less (specific) than
is not isomorphic to
can unify with

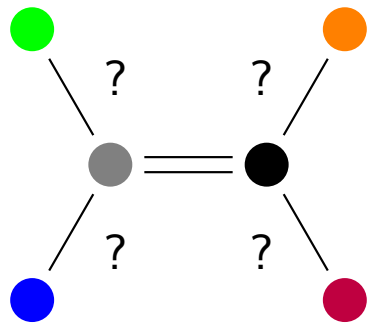




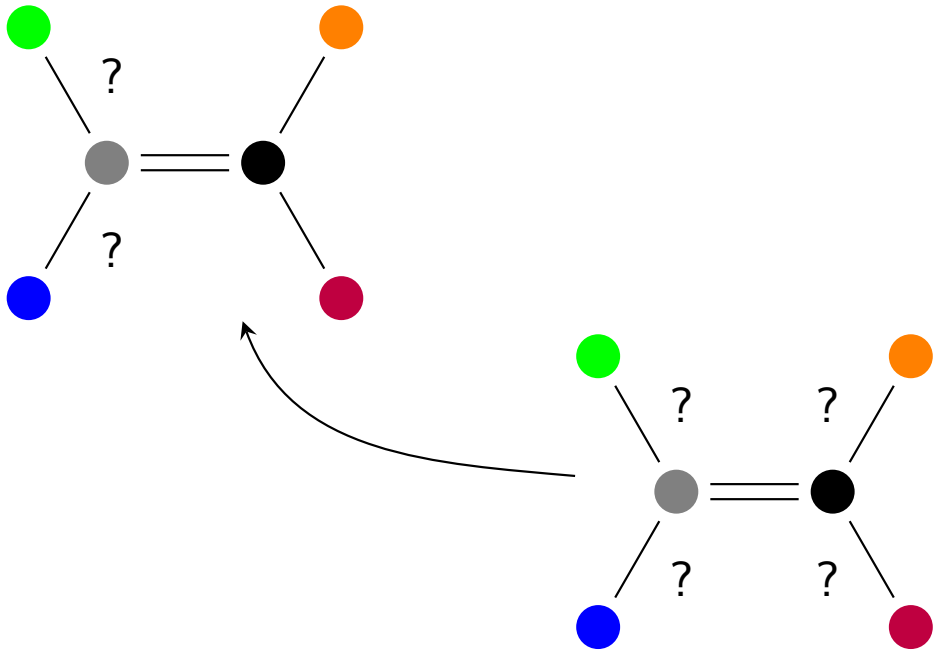
is incomparable (by specificity) to
is not isomorphic to
can unify with



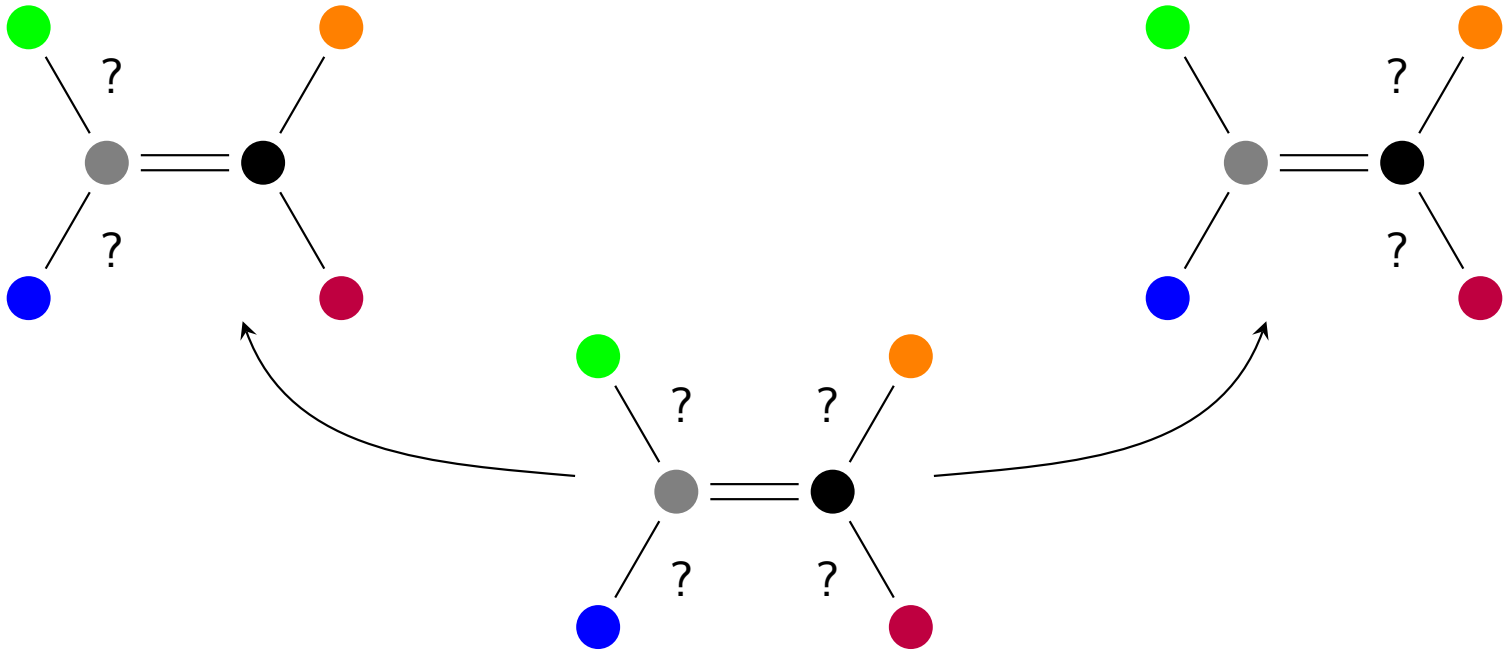
Partial Order of Graphs (Specificity)



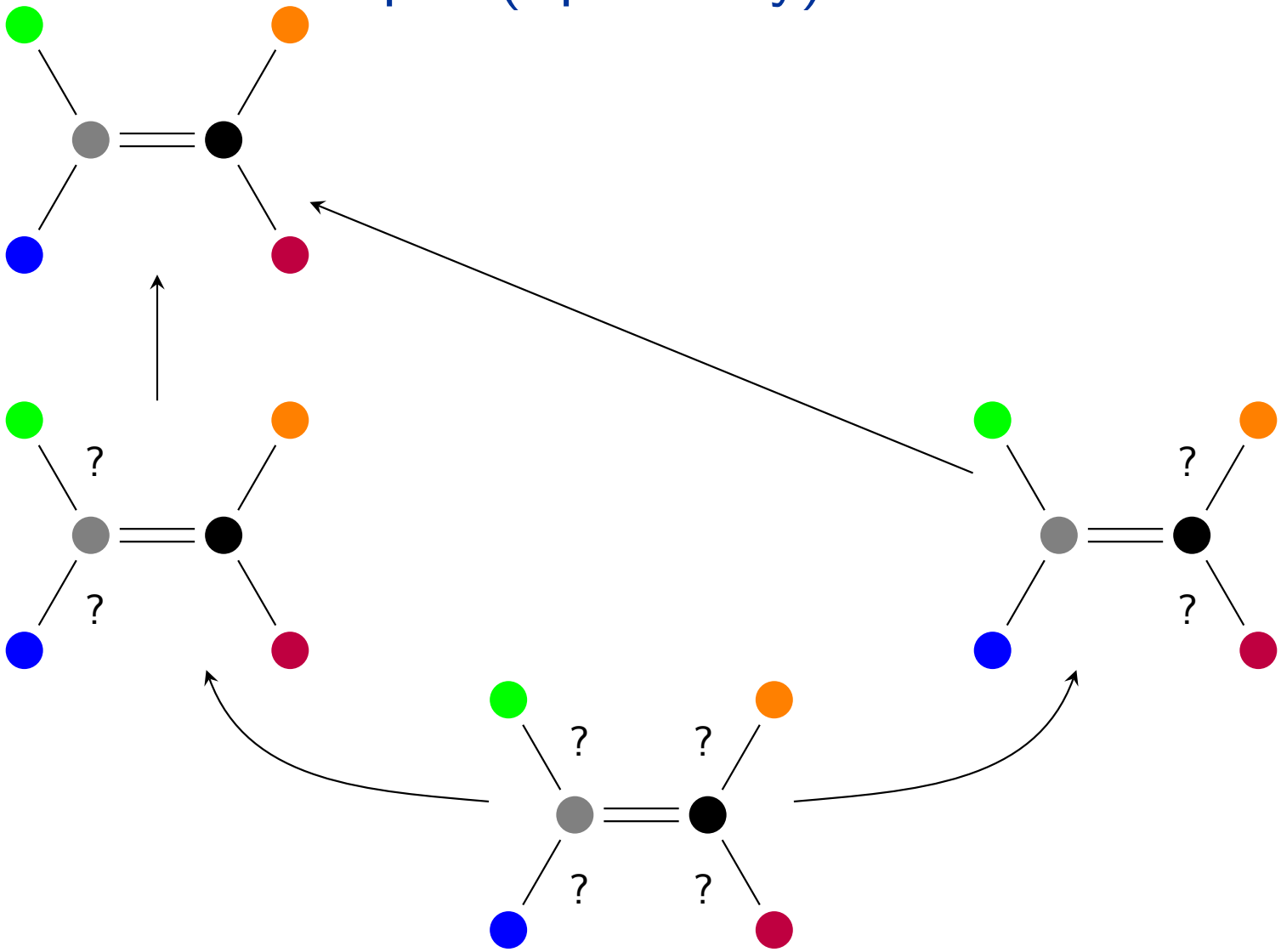
Partial Order of Graphs (Specificity)



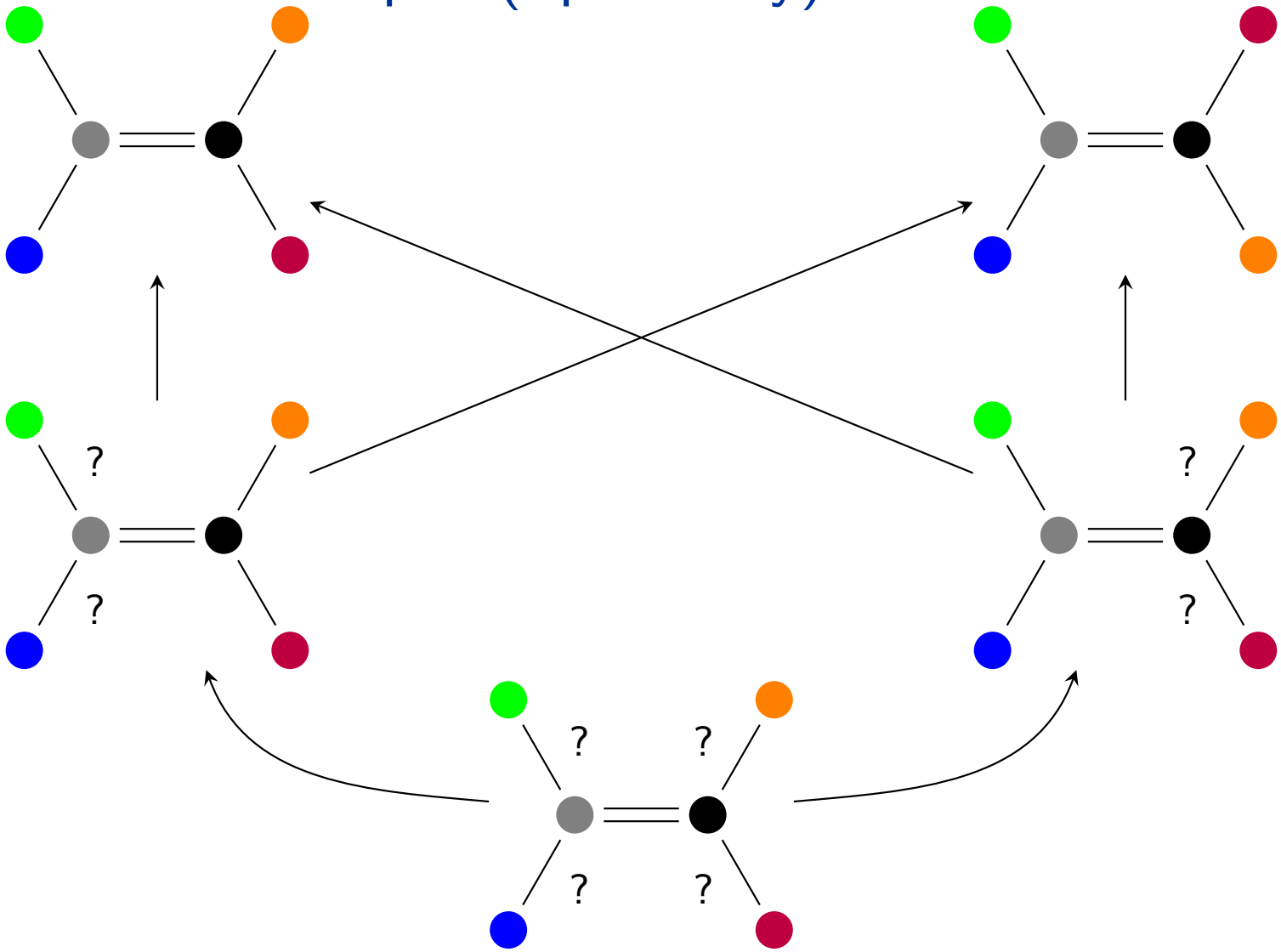
Partial Order of Graphs (Specificity)



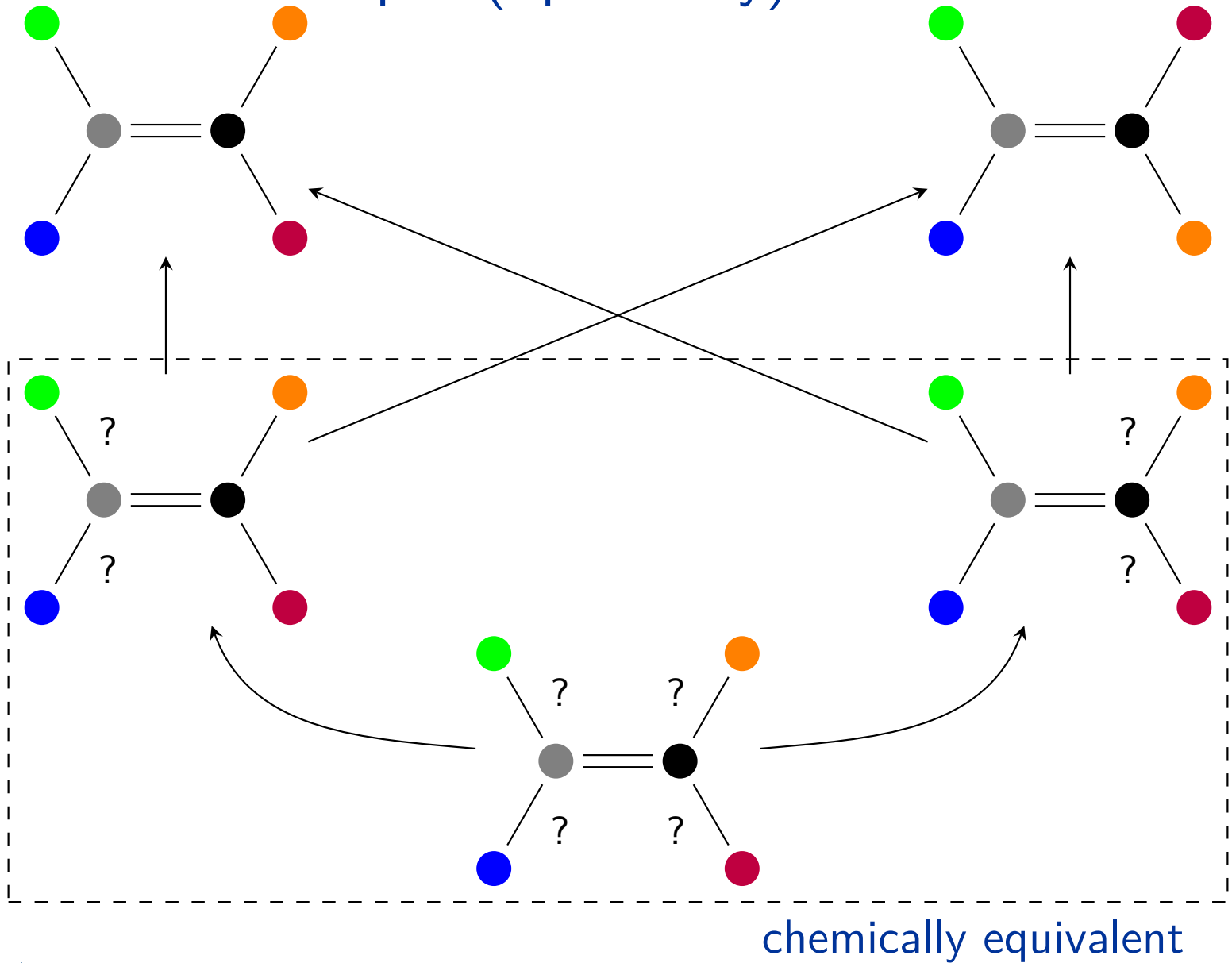
Partial Order of Graphs (Specificity)



Partial Order of Graphs (Specificity)



Partial Order of Graphs (Specificity)



Substructures with Stereo Information

Modelling status

- In progress.

Some thoughts

- Partially specified geometries (encoding of "I don't care").
- Partially specified orderings ("virtual neighbours").
- Transformation rules that change stereo information.
- Transformation rules that fix/relax stereo orderings.
- Heavy use of variations of techniques from term rewriting.

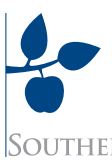
Some time in the future...

```
rule [  ruleID "Stereospecific Diels-Alder"
left [
    node [ id 1 label "C" stereo [ order "2, -_b, -_a" ] ]
    node [ id 4 label "C" stereo [ order "3, -_d, -_c" ] ]
    edge [ source 1 target 2 label "=" ]
    edge [ source 2 target 3 label "-" ]
    edge [ source 3 target 4 label "=" ]

    node [ id 5 label "C" stereo [ order "6, -_f, -_e" ] ]
    node [ id 6 label "C" stereo [ order "5, -_g, -_h" ] ]
    edge [ source 5 target 6 label "=" ]
]
context [
    ndoe [ id 2 label "C" stereo [ order "1, -, 3" ]  ]
    ndoe [ id 3 label "C" stereo [ order "4, 2, -" ]  ]
]
right [
    node [ id 1 label "C" stereo [ order "2, 5, -_a, -_b" ] ]
    node [ id 4 label "C" stereo [ order "3, 6, -_c, -_d" ] ]
    node [ id 5 label "C" stereo [ order "6, 1, -_e, -_f" ] ]
    node [ id 6 label "C" stereo [ order "5, 4, -_h, -_g" ] ]
    edge [ source 1 target 2 label "-" ]
    edge [ source 2 target 3 label "=" ]
    edge [ source 3 target 4 label "-" ]
    edge [ source 4 target 5 label "-" ]
    edge [ source 5 target 6 label "-" ]
    edge [ source 6 target 1 label "-" ]
]   ]
```

Summary and Current Status

- Modelling of stereochemistry is non-trivial.
- A lot of fun algorithmics.
- It seems is doable, using only local properties.

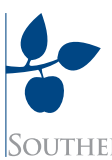


Summary and Current Status

- Modelling of stereochemistry is non-trivial.
- A lot of fun algorithmics.
- It seems is doable, using only local properties.

Implementation status

- Basic data structures for stereo information.
- System for first-order term unification.
- Basic information inference from GML specification.



Summary and Current Status

- Modelling of stereochemistry is non-trivial.
- A lot of fun algorithmics.
- It seems is doable, using only local properties.

Implementation status

- Basic data structures for stereo information.
- System for first-order term unification.
- Basic information inference from GML specification.

Subset of related challenges

- Canonicalisation algorithm (in progress).
- Deciding if something is a valid molecule.
- Visualisation in 2D and 3D.
- Inference from input (R/S, E/Z, . . .).
- Interconversion with SMILES (whatever that is).
- Interconversion with the “open” “standard” InChI.

Thank You for Listening

"Fortunately, the charging one has been solved now that we've all standardized on mini-USB. Or is it micro-USB? Shit."

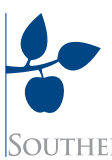
[XKCD: Standards]

Bonus Slide: SMILES

- The original version is proprietary, but OpenSMILES exist.
- Unclear/unfinished specification.
- Missing/unclear molecule model.
- No true support for conjugated bonds.
- Everyone seems to implement their own specification.
- Widespread belief that “the canonicalisation algorithm” works. (hint: it doesn’t)

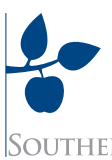
Bonus Slide: InChI

- "IUPAC International Chemical Identifier"
- Identity crisis: is it a standard, tool, or algorithm?
- Missing/unclear molecule model.
- No separation between standard/specification and implementation.



Bonus Slide: InChI

- "IUPAC International Chemical Identifier"
- Identity crisis: is it a standard, tool, or algorithm?
- Missing/unclear molecule model.
- No separation between standard/specification and implementation.
- "the InChI source code [...] acts as the final arbiter of the correctness" — [InChI Tech. FAQ]
- "Mathematical details of the algorithms used will not be presented. They have been derived from methods reported in the literature [...]. They will be made available in the form of tested and documented source code along with the final version of the InChI" — [InChI Tech. Manual]



Bonus Slide: InChI

- "IUPAC International Chemical Identifier"
- Identity crisis: is it a standard, tool, or algorithm?
- Missing/unclear molecule model.
- No separation between standard/specification and implementation.
- "the InChI source code [...] acts as the final arbiter of the correctness" — [InChI Tech. FAQ]
- "Mathematical details of the algorithms used will not be presented. They have been derived from methods reported in the literature [...]. They will be made available in the form of tested and documented source code along with the final version of the InChI" — [InChI Tech. Manual]
-
```
if ( k < r ) {
    goto L9; /* cannot understand it... */
}
```

[InChI source code, the canonicalisation code]