RMG-Py: Transitioning to the New Style Adjacency List

Connie Gao

10/22/2014

Green Group Meeting

Outline

- Motivation for the new adjacency list
- The new style adjacency list
  - Molecular adjlists
  - Group adjlists
  - Isomorphism checks
- Restricting cross spin reactions in reaction families
- Special treatment: CO
- Treatment of ions
- Unit tests for the database
- User-friendly alterations the database

The role of adjlists in RMG

Initial species and conditions

Thermodynamic and kinetic Libraries

Reaction Mechanism Generator

Detailed Reaction Mechanism

Molecules

```
CH2(S)
multiplicity 1
1  C u0 p1 c0 {2,S} {3,S}
2  H u0 p0 c0 {1,S}
3  H u0 p0 c0 {1,S}
```

Defines a single isomer of a species

Groups

```
R!Hx3
1  R!H u1 p1
```

Defines a functional group that describes part of any number of molecules

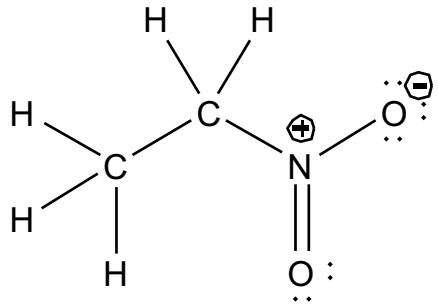
Used to estimate thermo via group additivity and reaction rates between sites

Motivation for the new adjacency list

- Makes tracking electrons simpler
- Adjusts for elements with variable valency
  - Motivated by addition of Nitrogen chemistry, but makes RMG more extensible towards other chemistries in the future
- Accounting for spin multiplicity as a species property
- Labeling of partial charges can lead to ion chemistry in the future

The new standard Molecule adjlist

```
nitroethane
multiplicity 1
1  C u0 p0 c0   {2,S} {4,S} {5,S} {6,S}
2  C u0 p0 c0   {1,S} {3,S} {7,S} {8,S}
3  N u0 p0 c+1  {2,S} {9,D} {10,S}
4  H u0 p0 c0   {1,S}
5  H u0 p0 c0   {1,S}
6  H u0 p0 c0   {1,S}
7  H u0 p0 c0   {2,S}
8  H u0 p0 c0   {2,S}
9  O u0 p2 c0   {3,D}
10 O u0 p3 c-1  {3,S}
```



Requires explicit hydrogens

multiplicity: *optional*, is computed on molecule creation based on the formula 2s+1, otherwise checked for consistency

u – unpaired electrons: always required

p – lone electron pairs: *optional*, if not defined is assumed to be zero

c – formal charges: *optional*, if not defined is assumed to be zero

All attributes printed

Consistency checks in place for:

- user-defined multiplicities based on conservation of angular momentum and hund's rule
- user-defined lone pairs/charges based on conservation of valence electrons

Molecules: conversion from old adjlists

Old Style

```
CH2(S)
1   C 2S
```

Intermediate Style

```
CH2(S)
1   C 2S 0 {2,S} {3,S}
2   H 0  0 {1,S}
3   H 0  0 {1,S}
```

Step 1 Identify number of lone pairs

Use a default value:
Carbon = 0, Oxygen = 2, Sulfur = 2, etc.

Use given value

Step 2 Identify number of unpaired electrons and additional lone pairs based on electronic state

2S converts to 0 unpaired electrons and 1 additional lone pair

Step 2b (old style only) Saturate H

Step 3 Determine the partial charge on each atom based on the element's valency

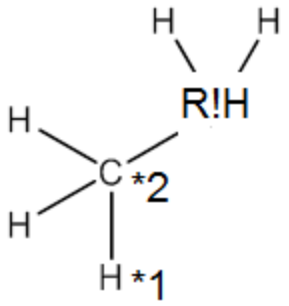
Step 4 Determine a default maximum multiplicity based on mult = 2s + 1, where s = number of unpaired electrons

Output

```
CH2(S)
multiplicity 1
1   C u0 p1 c0 {2,S} {3,S}
2   H u0 p0 c0 {1,S}
3   H u0 p0 c0 {1,S}
```

The new standard Group adjlist

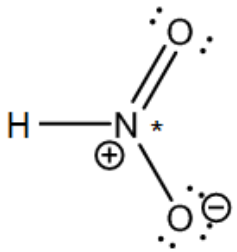
```
RH2CH3
1     R!H u0    {2,S} {3,S} {4,S}
2     H   u0    {1,S}
3     H   u0    {1,S}
4  *2 C   u0    {1,S} {5,S} {6,S} {7,S}
5  *1 H   u0    {1,S}
6     H   u0    {1,S}
7     H   u0    {1,S}
```



multiplicity: *optional*, is assumed to be a wildcard unless specified by user
u – unpaired electrons: always required
p – lone electron pairs: *optional*, if not defined is assumed to be a wildcard
c – formal charges: *optional*, if not defined is assumed to be a wildcard

Not all attributes printed – multiplicity, lone pairs, or charges are only printed when they are not wildcards.

```
nitro
multiplicity 1
1  * N   u0 p0 c+1 {2,D} {3,S} {4,S}
2    O   u0 p2 c0  {1,D}
3    O   u0 p3 c-1 {1,S}
4    H   u0 p0 c0  {1,S}
```



More notes on Group adjlist formatting

New Style (implied)

```
R4
multiplicity [1,2,3,4,5]
1 *1 R!H              u1 px cx {2,[S,D,T,B]}
2 *4 R!H              u0 px cx {1,[S,D,T,B]} {3,S}
3 *2 [Cd,Ct,CO,N]     u0 px cx {2,S} {4,[D,T]}
4 *3 [Cd,Ct,Od,Cdd,N] u0 px cx {3,[D,T]}
```

New Style (printed)

```
R4
1 *1 R!H              u1 {2,[S,D,T,B]}
2 *4 R!H              u0 {1,[S,D,T,B]} {3,S}
3 *2 [Cd,Ct,CO,N]     u0 {2,S} {4,[D,T]}
4 *3 [Cd,Ct,Od,Cdd,N] u0 {3,[D,T]}
```

Not all attributes printed – multiplicity, lone pairs, or charges are only printed when they are not wildcards.

A union of Atomtypes or Bondtypes is now written with square brackets [ ] rather than braces { }

Groups: conversion from old adjlists

Old Style

```
R!Hx3
1   R!H 3D
```

Intermediate Style

```
R!Hx3
1   R!H 3D 0
```

Step 1 Identify number of lone pairs

Set as wildcard

Use given value

Step 2 Identify number of unpaired electrons and additional lone pairs based on electronic state

3D converts to 1 unpaired electrons and 1 additional lone pair

Step 3 Set the formal charges and overall multiplicity to wildcards

Output

```
R!Hx3
1   R!H u1
```

Output

```
R!Hx3
1   R!H u1 p1
```

Conversion currently leads to broadly defined groups, which can be **problematic**

Isomorphism checks extended

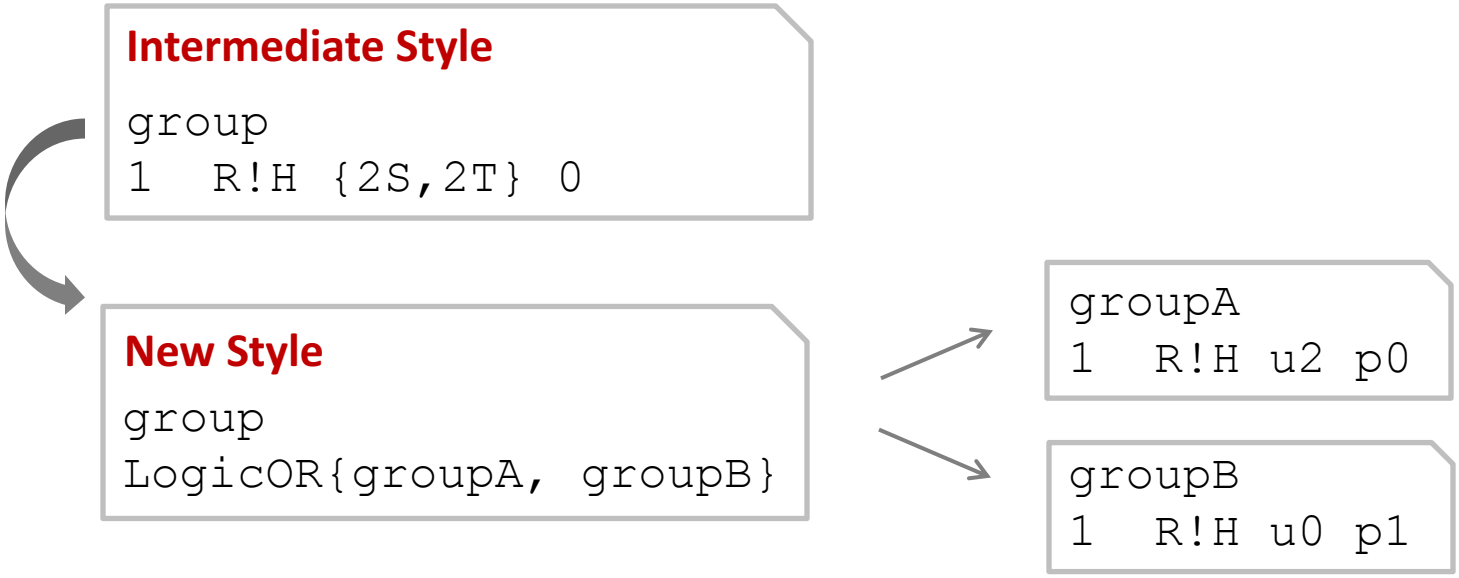
Now checks for:

- Connectivity and atoms
- Same number on each atom:
  - Unpaired electrons
  - Lone pairs
  - Formal charges
- Overall multiplicity

Suite of unit tests have been developed by Nick

Obstacles encountered: 2S vs. 2T

Many groups defined in a way that would require new LogicOR groups to be defined

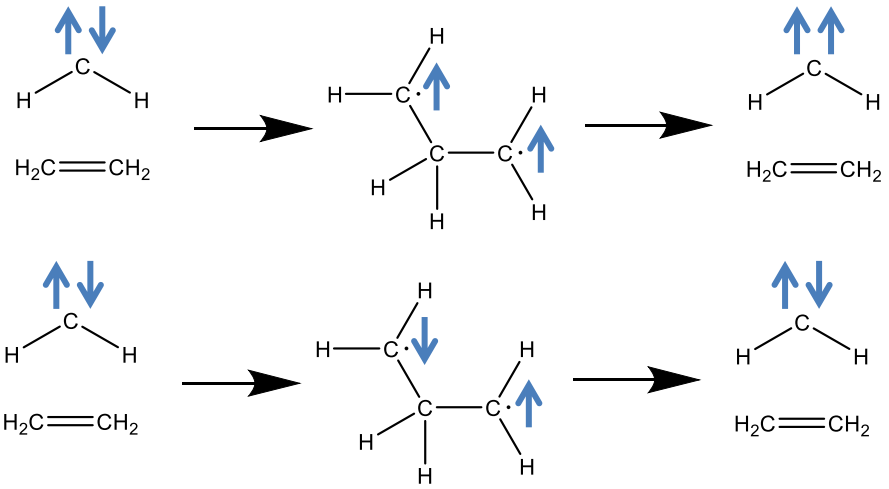


However…. Should so many mixed electronic state groups exist? Do we expect similar thermo and kinetics for such groups?

The answer is NO. Many such mixed electronic state groups can be eliminated by restricting them in families….

Restricting cross spin reactions in Reaction Families

RMG previously allowed cross spin reactions



Pros

- Finds all possible reactions

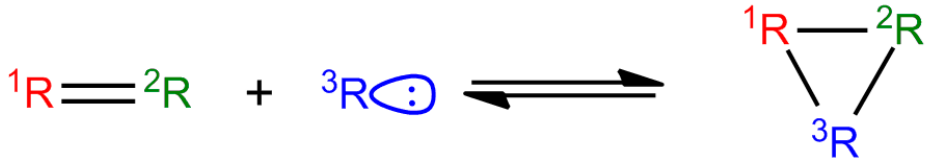
Cons

- We do not know the thermo of different spin state species accurately nor the distribution of the products
- Increased computation time

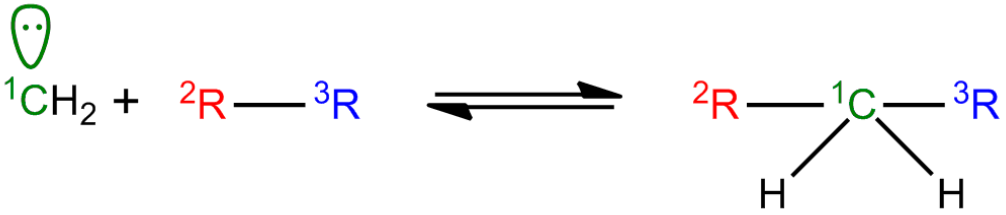
→ **We have reverted Beat's changes by forcing multiplicities of products to be determined by the Reaction's Family's template**

Families with singlet reactants

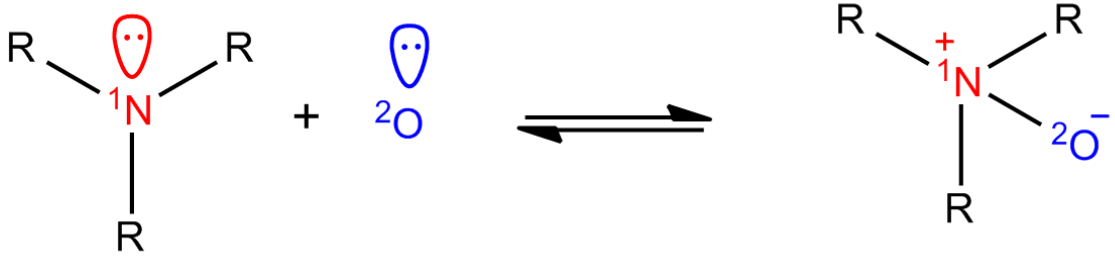
1+2_Cycloaddition



1,2_Insertion_carbene



lone_electron_pair_bond



Other families now restrict reactant sites to non-singlets

* More drawing (Credit: Nathan) can be found in `RMG-database/families`

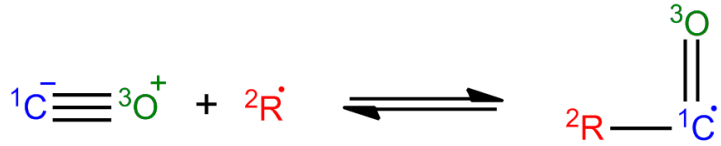
Special treatment: CO

```
1 C u0 p1 c-1 {2,T}
2 O u0 p1 c+1 {1,T}
```

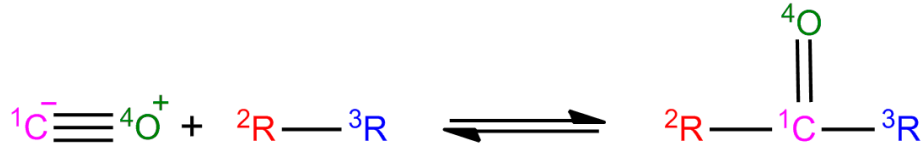
Ground state of CO is triple bonded singlet.

Previously, double bonded and triple bonded versions were both found in RMG. Now, all reaction libraries and families use the **triple bonded form** for consistency. No more finding 2 versions of CO in your mechanism!

R_Addition_COm

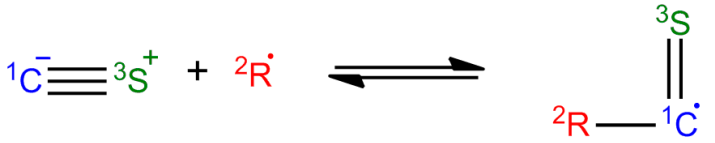


1,2_Insertion_CO



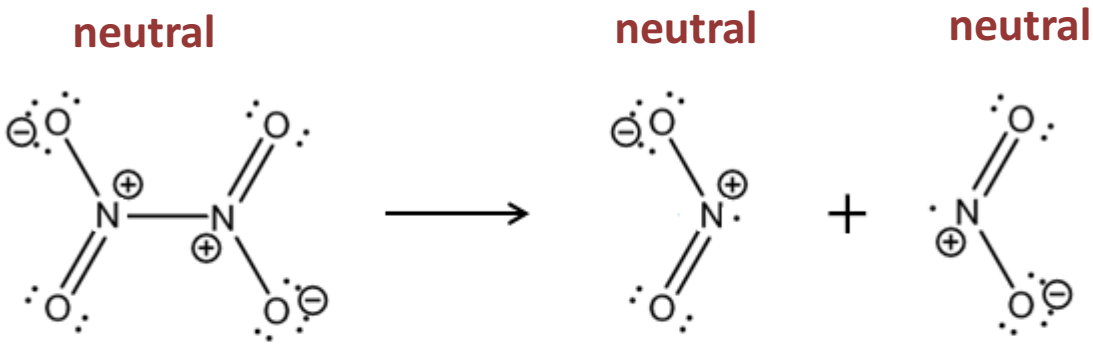
CS also converted to solely triple bonded singlet form

R_Addition_CSm



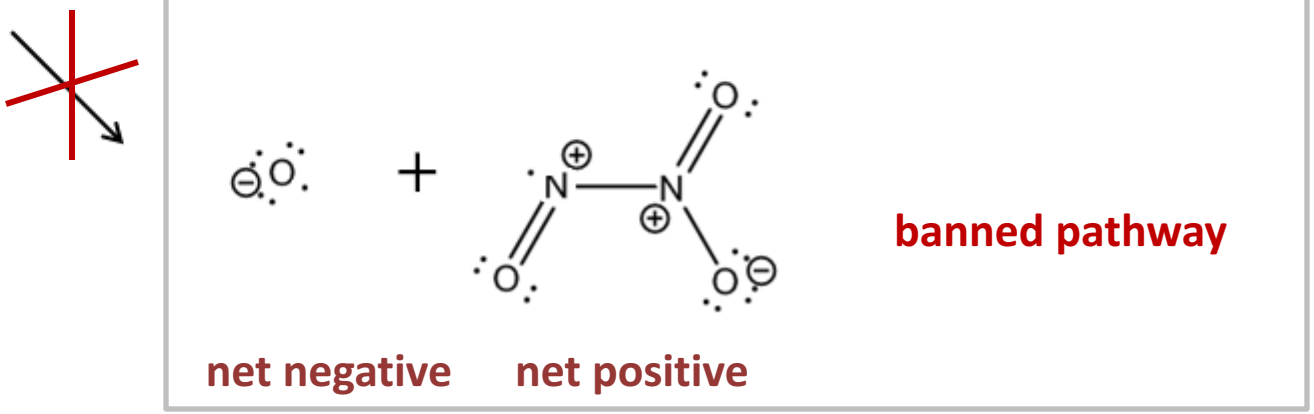
Treatment of ions

RMG currently bans all non-neutral molecules although most groups treat charges as wildcards



Why?

No kinetic or thermo data available for ions



`isMoleculeForbidden(self, molecule)`
`bans non-neutral species`

Unit tests for the database

Helps detect human error!

- ❑ Nodes are in the tree with proper parents?
- ❑ Nodes are nonidentical?
- ❑ Parent-child relationships are correct? Parent nodes must be more general than their children. (currently kinetics only)

With conversion to new adjlist, many groups may be too broadly specified. **Need to check if groups overlap.**

User-friendly alterations to the database

- Separation of kinetics from dictionaries for reaction libraries, as well as training and NIST depositories
  - Saves from rewriting dictionaries again and again
  - More scroll searchable

Future tasks

- Export database to new format
- Unit tests for detecting overlapping children in trees
- Create read-only output files for comparing kinetics and thermo values for rate rules and libraries more easily

Acknowledgments

- Many thanks to:
  - Nathan Yee
  - Nick Vandewiele
  - Enoch Dames
  - Shamel Merchant
- Professor Bill Green
- Professor Richard West
- Green Group & ComoCheng Group