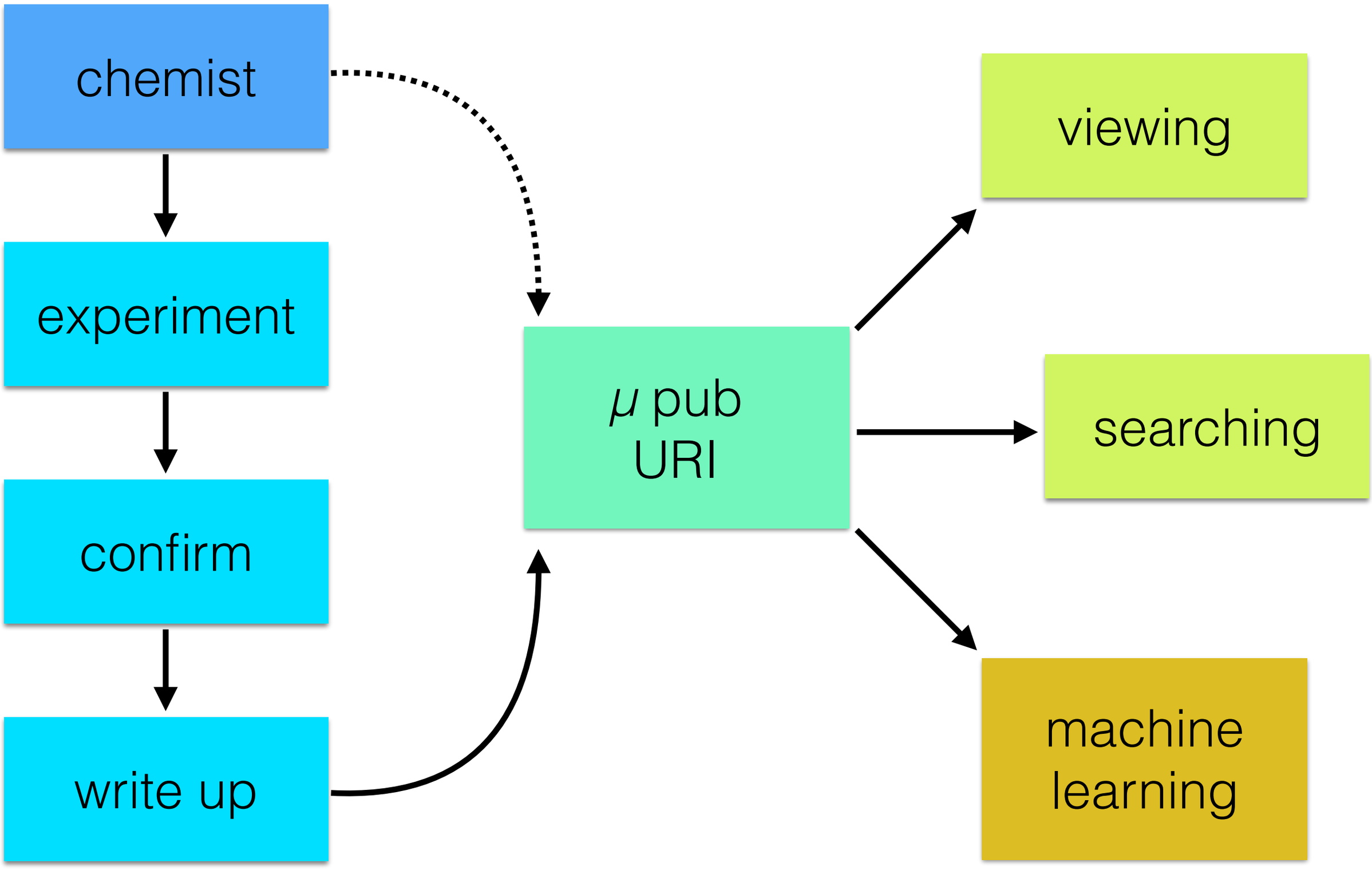
The anatomy of a chemical reaction: *Dissection by machine learning algorithms*

Alex M. Clark, Ph.D.

August 2014

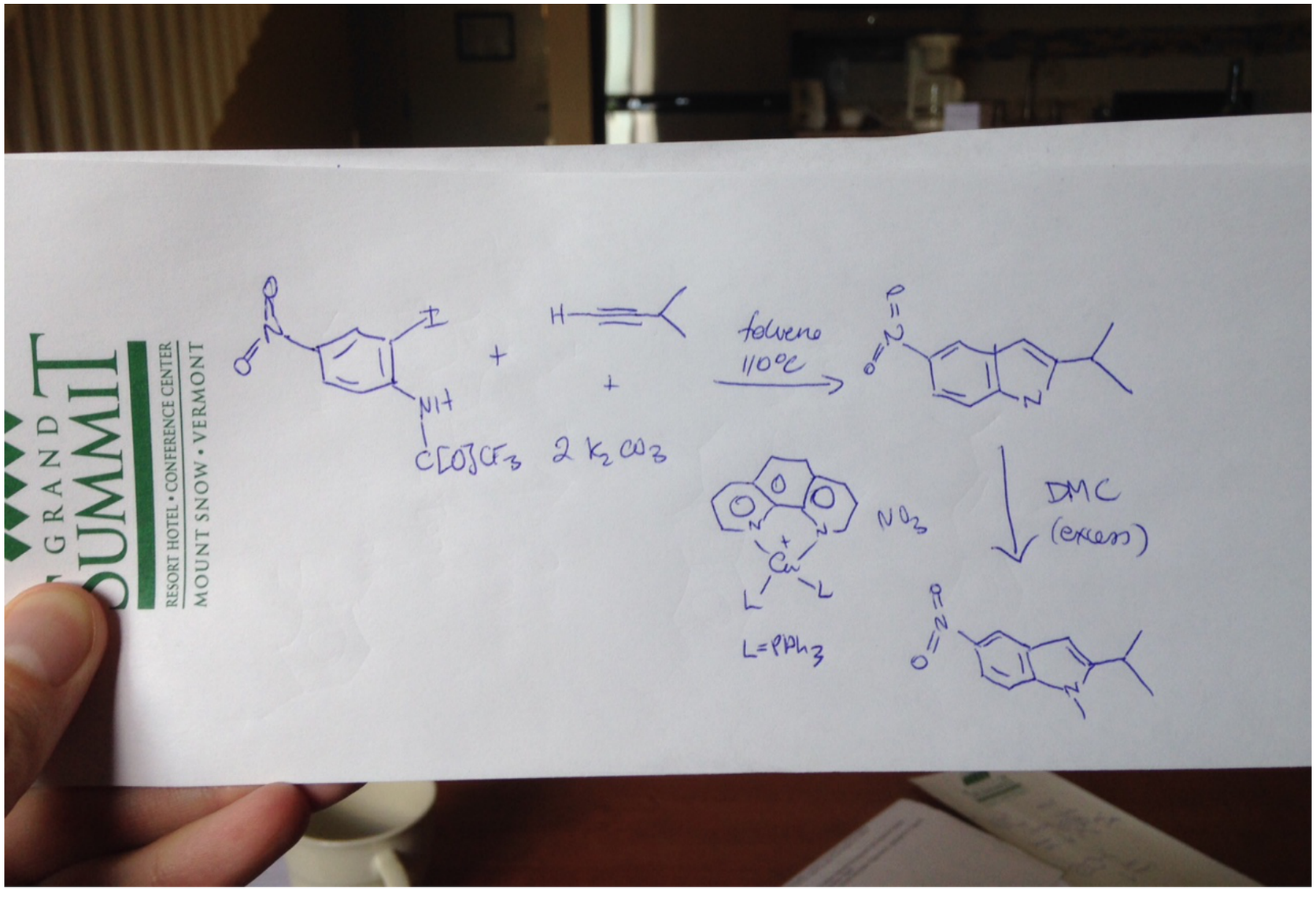
http://molmatinf.com

21st Century Publishing



All your byte are belong to us

- Just because a reaction scheme is digital…

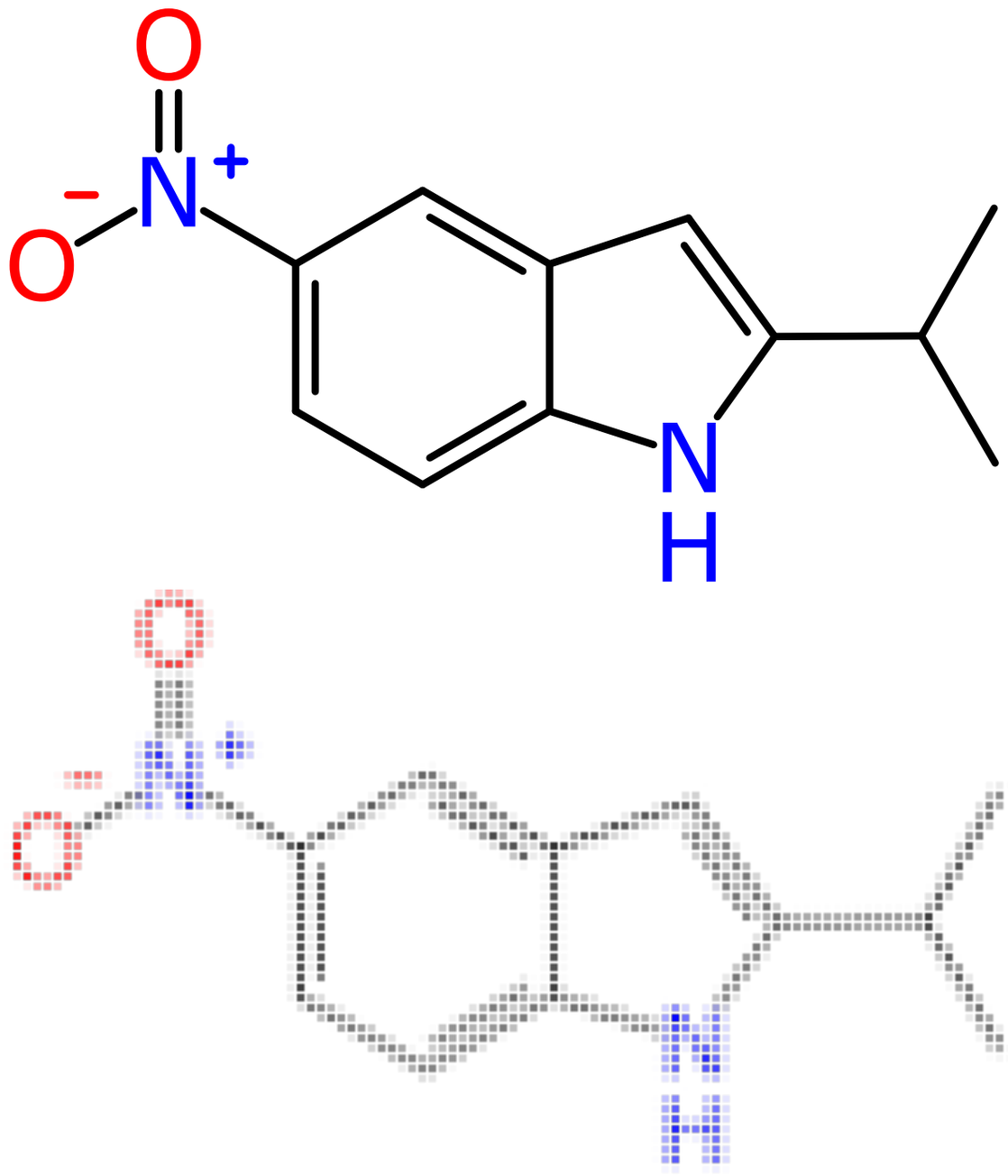


- … doesn't mean it's of any use to a computer.

Production Raster Graphics

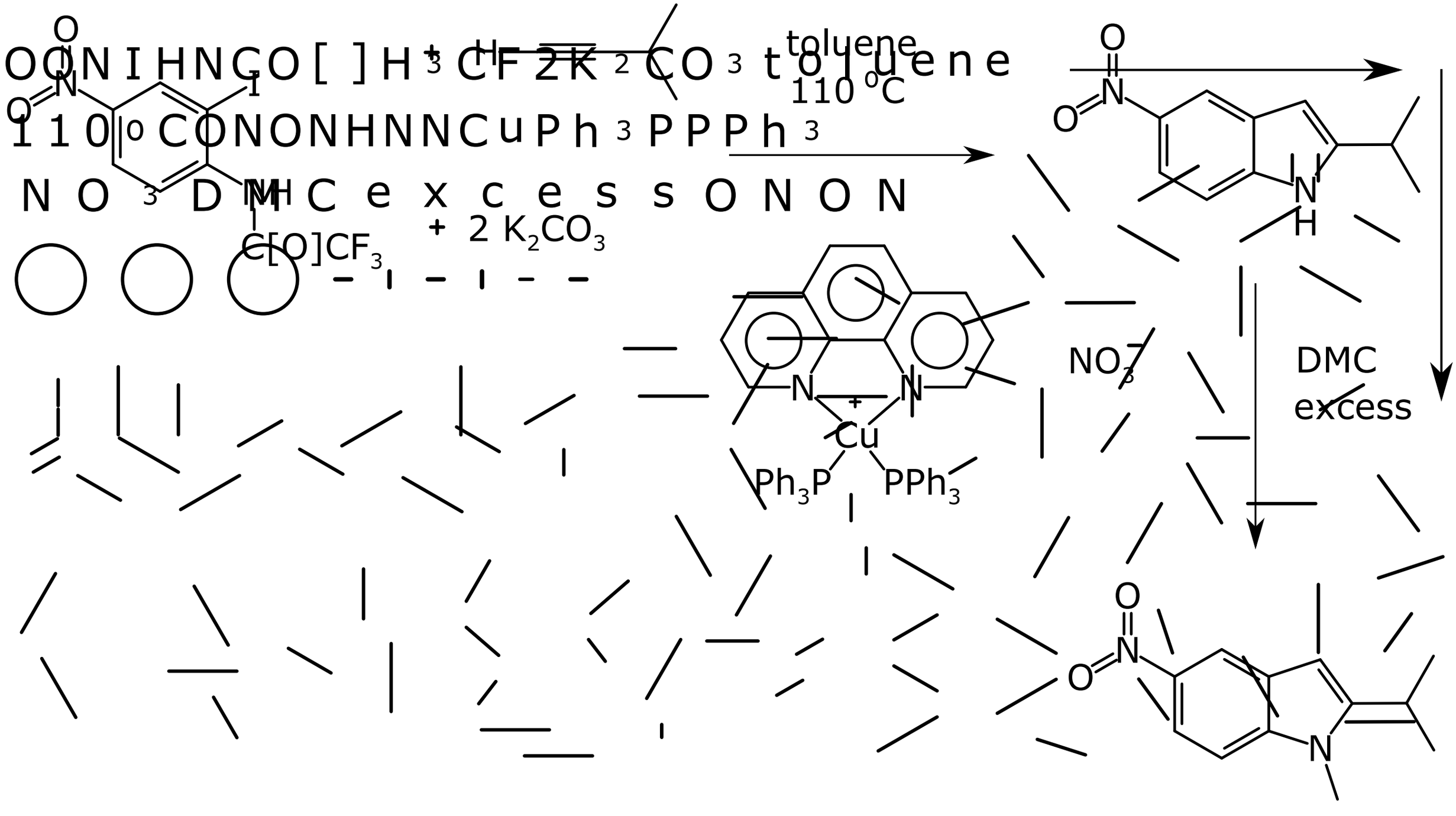
```
Generic molfile

 15 16  0  0  0  0  0  0  0  0999 V2000
   -3.9510    4.0500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -5.2500    3.3000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.6519    3.3000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -5.2500    1.8000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.9510    1.0500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.6519    1.8000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2306    3.7694    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3482    2.5611    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2240    1.3407    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -6.5490    4.0500    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
   -7.8481    3.3000    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
   -6.5490    5.5500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.1518    2.5673    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9072    1.2714    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.8964    3.8695    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  2  0  0  0  0
  2  4  2  0  0  0  0
  4  5  1  0  0  0  0
  5  6  2  0  0  0  0
  6  3  1  0  0  0  0
  3  7  1  0  0  0  0
  7  8  2  0  0  0  0
  8  9  1  0  0  0  0
  9  6  1  0  0  0  0
  2 10  1  0  0  0  0
 10 11  1  0  0  0  0
 10 12  2  0  0  0  0
  8 13  1  0  0  0  0
 13 14  1  0  0  0  0
 13 15  1  0  0  0  0
M  CHG  2  10   1  11  -1
M  END
```



Production Vector Graphics

- Manuscripts usually delivered as PDFs:



Spreadsheets

ssreaction.xlsx - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View

Paste Clipboard | Calibri 11 Font | Wrap Text Merge & Center Alignment | Custom Number | Conditional Formatting Format as Table Cell Styles Styles | Insert Delete Format Cells | AutoSum Fill Clear Sort & Filter Find & Select Editing

B2 10-08-2015

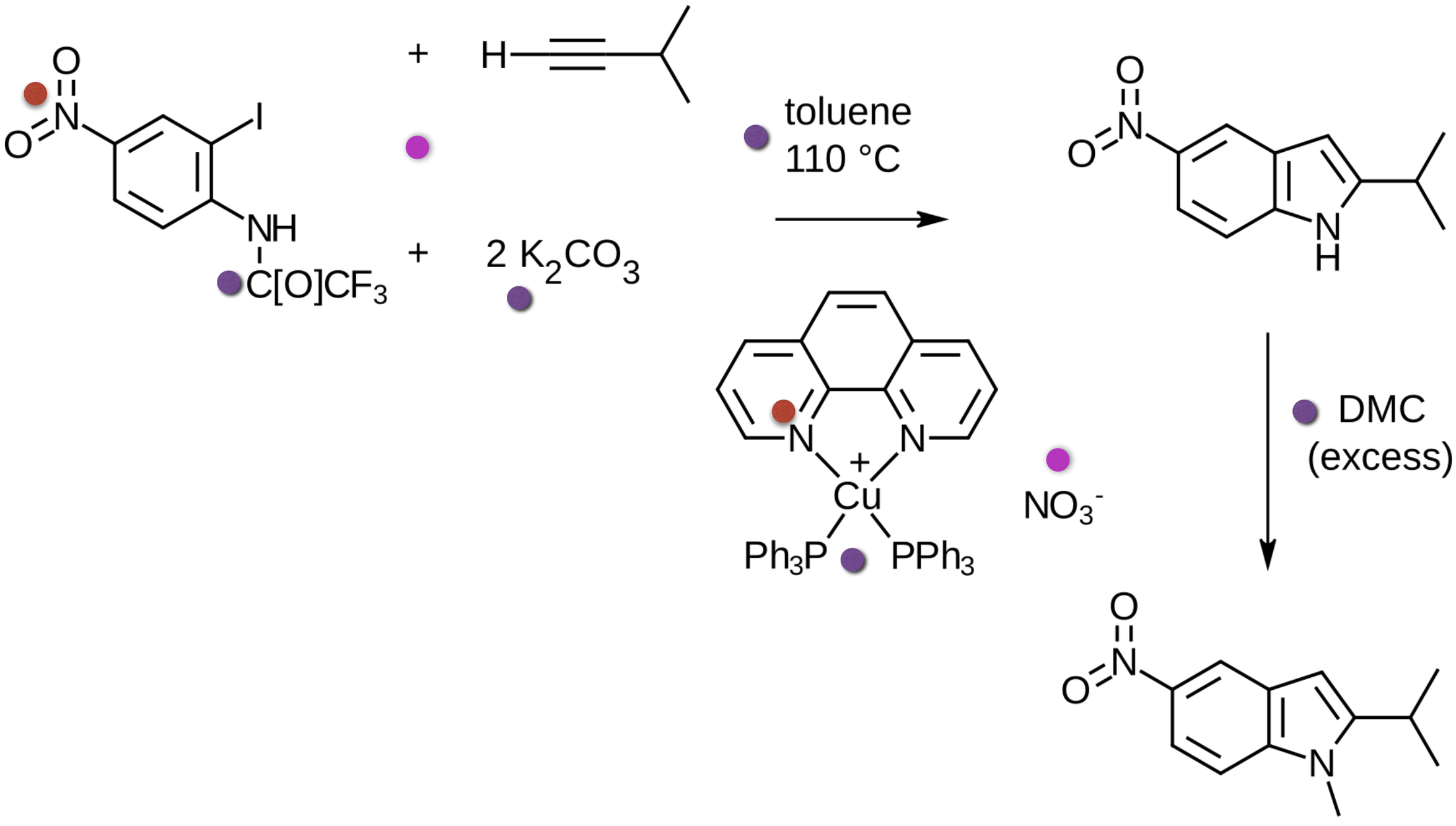
	A	B	C	D	E	F
1	**Cyclisation and Methylation**					
2	**Date:**	10-Aug-2015				
3						
4		**Reactants**	**Reagents**	**Solvents**	**Catalysts**	**Product**
5	**Step 1**	[O-][N+](=O)c1cc(I)c(cc1)NC(=O)C(F)(F)F	[K+].[K+].[O-]C([O-])=O	Cc1ccccc1	[O-][N+]([O-])=O.c1ccc2ccc3ccc[n+]4c3c2[n+]1[Cu-3]4([P+](c1ccccc1)(c1ccccc1)c1ccccc1)[P+](c1ccccc1)(c1ccccc1)c1ccccc1	CC(C)c1cc2cc(ccc2[nH]1)[N+]([O-])=O
6		CC(C)C#C				
7	**Step 2**		COC(=O)OC			CC(C)c1cc2cc(ccc2[n]1C)[N+]([O-])=O
8						
9						
10						

Sheet1 Sheet2 Sheet3

Ready 100%

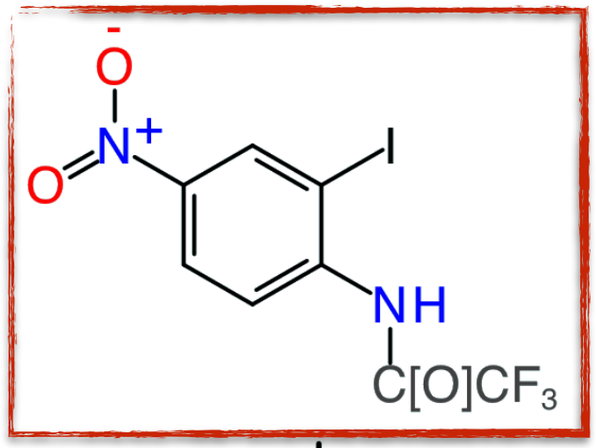
- Data gives the impression of organisation
- Very high degrees of freedom, nothing for structures

Common Scheme



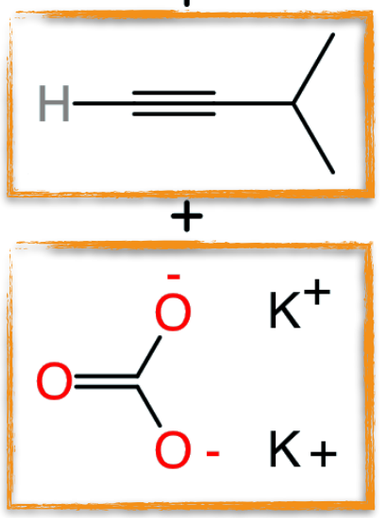
Digitally Friendly

primary reactant

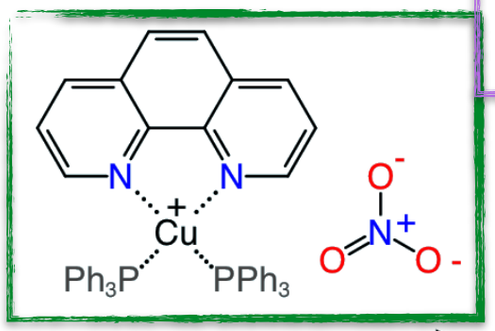


+

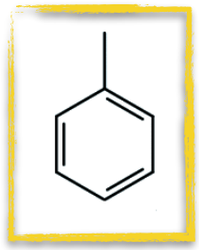
secondary reactants



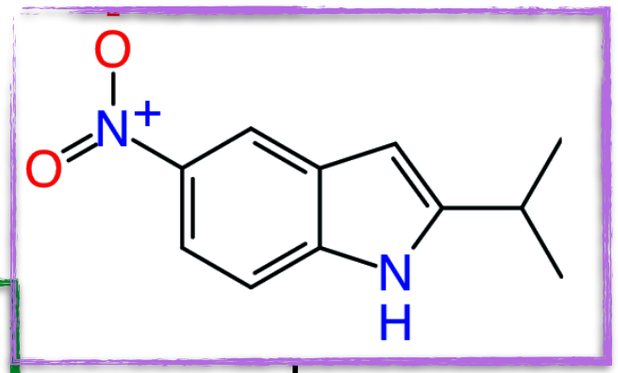
catalyst



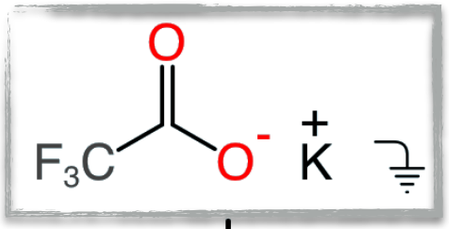
solvent



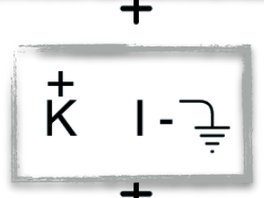
intermediate



+

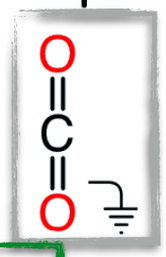


+

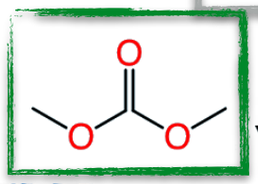


+

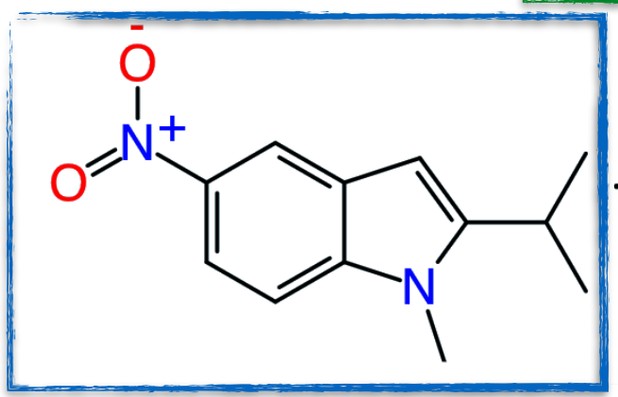
byproducts



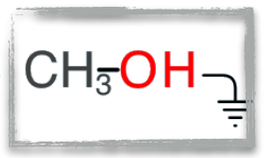
reagent



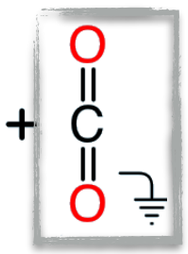
final product



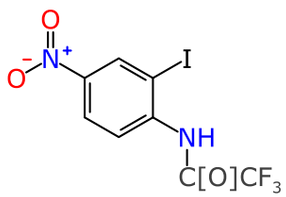
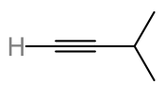
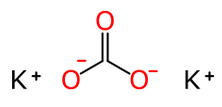
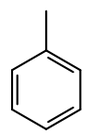
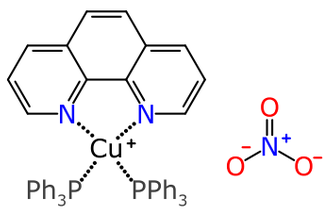
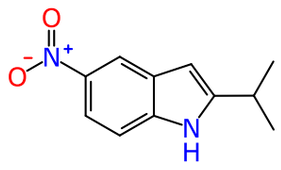
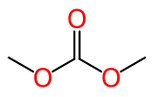
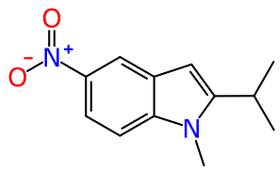
+



+



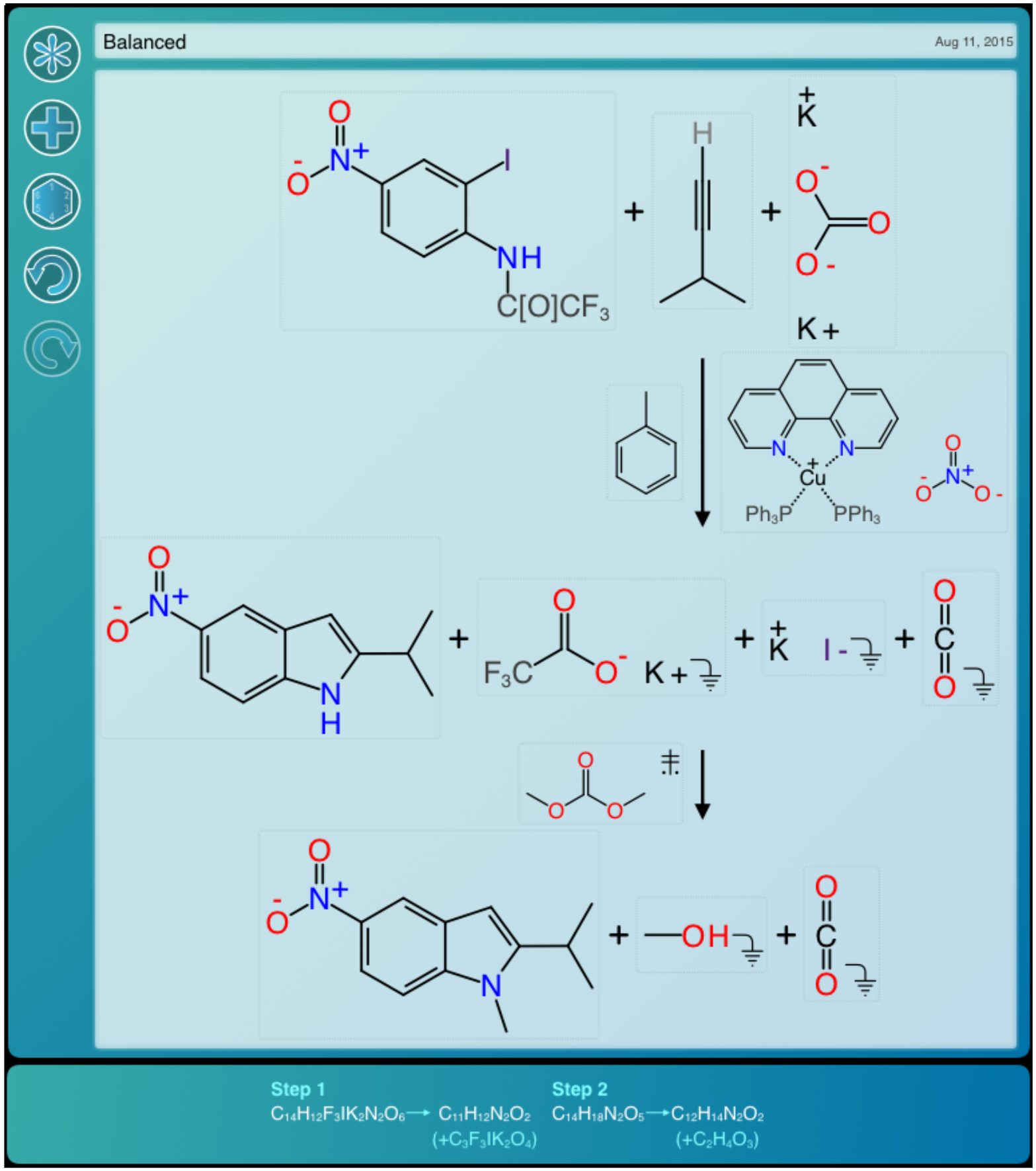
Representation

Step	Stoich.	Structure	Role
1	1		Reactant
1	1		Reactant
1	1		Reagent
1	1		Reagent
1	1		Reagent
1	1		Product
2	1		Reagent
2	1		Product

- For **machines**: representation must be very rigidly defined
- For **humans**: can generate diagram programmatically
- MDL RXN/RDfile ~50% there
- DataSheet XML with Experiment aspect

`http://molmatinf.com/fmtaspect.html`

Balancing


Balanced
Aug 11, 2015
Step 1
$C_{14}H_{12}F_3IK_2N_2O_6 \rightarrow C_{11}H_{12}N_2O_2$ $(+C_3F_3IK_2O_4)$
Step 2
$C_{14}H_{18}N_2O_5 \rightarrow C_{12}H_{14}N_2O_2$ $(+C_2H_4O_3)$

Quantities

Equiv:	1
MW:	360.029 g/mol
Mass:	1 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	

Equiv:	1
MW:	68.117 g/mol
Mass:	0.189199 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	

Equiv:	1
MW:	138.206 g/mol
Mass:	0.383874 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	

Equiv:	
MW:	830.327 g/mol
Mass:	
Volume:	
Moles:	
Density:	
Conc:	

Equiv:	
MW:	92.1384 g/mol
Mass:	
Volume:	
Moles:	
Density:	0.865 g/mL
Conc:	

Equiv:	1
MW:	204.225 g/mol
Mass:	0.567247 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	100 %

Equiv:	1
MW:	152.114 g/mol
Mass:	0.422504 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	

Equiv:	1
MW:	166.003 g/mol
Mass:	0.461082 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	

Equiv:	1
MW:	44.0095 g/mol
Mass:	0.122239 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	

Equiv:	
MW:	90.0779 g/mol
Mass:	
Volume:	
Moles:	
Density:	
Conc:	

Equiv:	1
MW:	218.252 g/mol
Mass:	0.606206 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	100 %

Equiv:	1
MW:	32.0419 g/mol
Mass:	0.0889981 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	

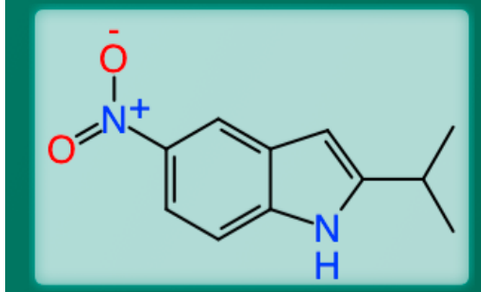
Equiv:	1
MW:	44.0095 g/mol
Mass:	0.122239 g
Volume:	
Moles:	0.00277756 mol
Density:	
Conc:	
Yield:	

Green Metrics

Σ reactants = 1 g + 0.189199 g + 0.829233 g + 0.115314 g + 17.34 g + 0.5 g = 19.9737 g

Σ products = 0.567247 g + 0.606206 g = 1.17345 g

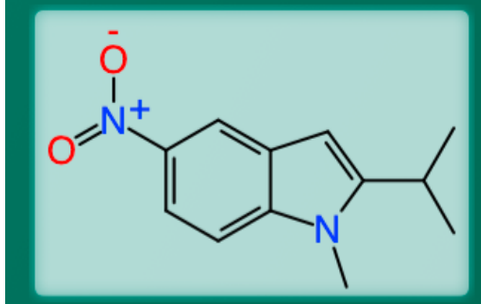
Σ waste = 0.422504 g + 0.461082 g + 0.122239 g + 0.0889981 g + 0.122239 g = 1.21706 g



$$\textbf{PMI} = \frac{1\text{ g} + 0.189199\text{ g} + 0.829233\text{ g} + 0.115314\text{ g} + 17.34\text{ g}}{0.567247\text{ g}} = 34.3303$$

$$\textbf{E-factor} = \frac{0.422504\text{ g} + 0.461082\text{ g} + 0.122239\text{ g}}{0.567247\text{ g}} = 1.77317$$

$$\textbf{Atom-E} = \frac{204.225}{360.029 + 68.117 + 138.206} = 36.0598\ \%$$



$$\textbf{PMI} = \frac{1\text{ g} + 0.189199\text{ g} + 0.829233\text{ g} + 0.115314\text{ g} + 17.34\text{ g} + 0.5\text{ g}}{0.606206\text{ g}} = 32.9487$$

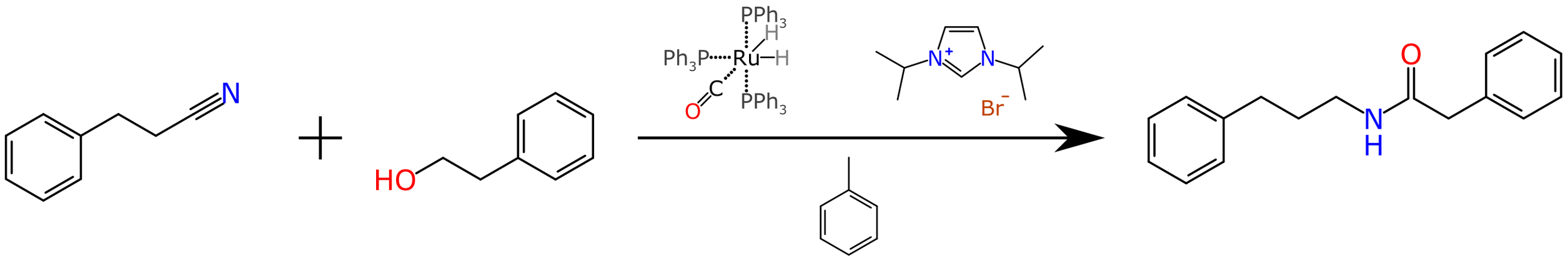
$$\textbf{E-factor} = \frac{0.422504\text{ g} + 0.461082\text{ g} + 0.122239\text{ g} + 0.0889981\text{ g} + 0.122239\text{ g}}{0.606206\text{ g}} = 2.00767$$

$$\textbf{Atom-E} = \frac{218.252}{204.225 + 90.0779} = 74.1588\ \%$$

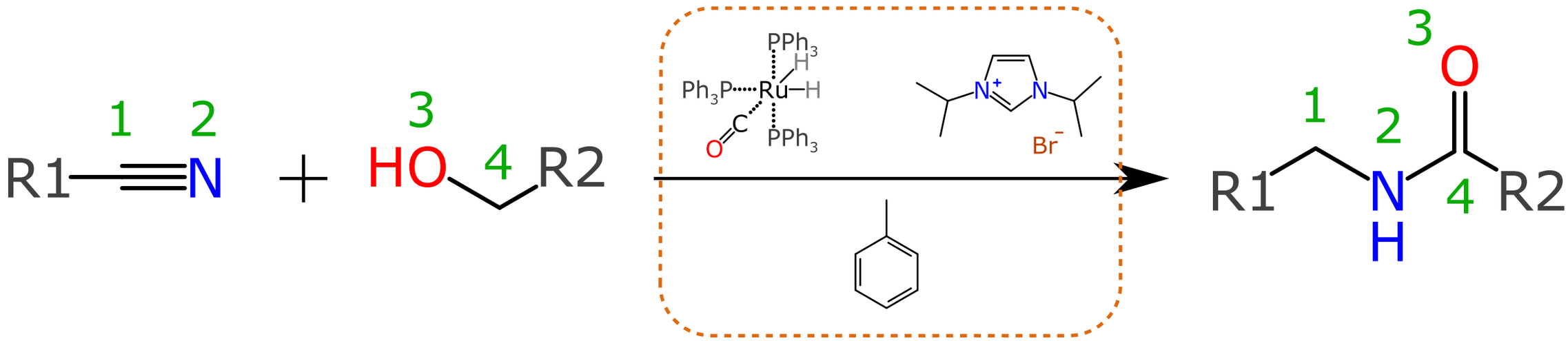
- Totals for reactants, products & waste
- For each non-waste product: **yield**, **PMI**, **E-factor**, **Atom-E**… always calculated, always recorded

Reaction Transforms

- **Reaction** = specific description of experiment

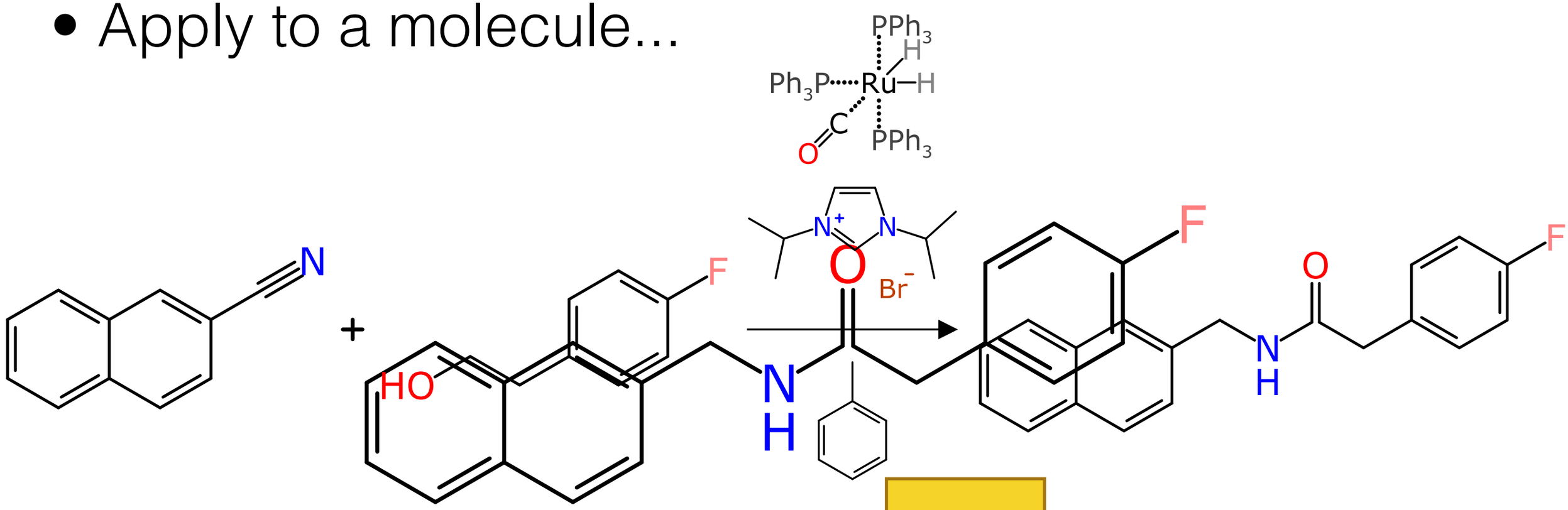


- **Transform** = the generic form of a reaction



Convenience

- Apply to a molecule...



Equiv:	1	Equiv:	1	Equiv:	0.1	Equiv:	0.1	Equiv:	137.692	Equiv:	1
MW:	153.18 g/mol	MW:	140.155 g/mol	MW:	917.952 g/mol	MW:	233.149 g/mol	MW:	92.1384 g/mol	MW:	293.335 g/mol
Mass:	5.22202 g	Mass:	4.77798 g	Mass:	3.12936 g	Mass:	0.794821 g	Mass:	432.5 g	Mass:	10 g
Volume:		Volume:		Volume:		Volume:		Volume:	500 mL	Volume:	
Moles:	0.0340907 mol	Moles:	0.0340907 mol	Moles:	0.00340907 mol	Moles:	0.00340907 mol	Moles:	4.69402 mol	Moles:	0.0340907 mol
Density:		Density:		Density:		Density:		Density:	0.865 g/mL	Density:	
Conc:		Conc:		Conc:		Conc:		Conc:		Conc:	
										Yield:	100 %

Σ reactants = 5.22202 g + 4.77798 g + 3.12936 g + 0.794821 g + 432.5 g = 446.424 g **Σ products** = 10 g = 10 g **Σ waste** = ? = 0

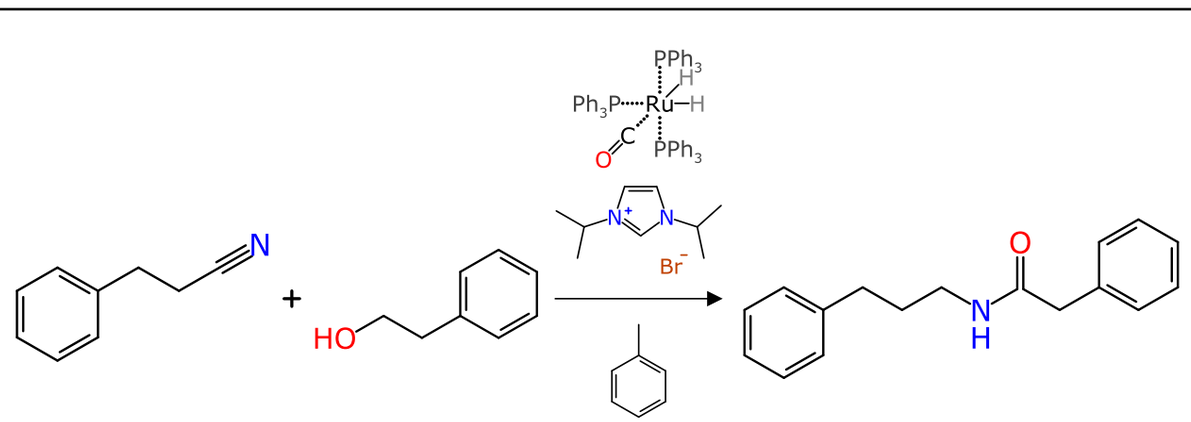
PMI = $\frac{5.22202\ g + 4.77798\ g + 3.12936\ g + 0.794821\ g + 432.5\ g}{10\ g}$ = 44.6424

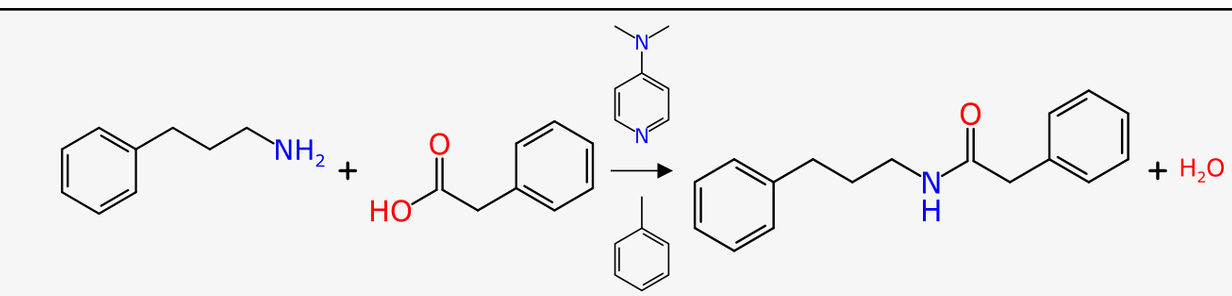
E-factor = $\frac{436.424\ g}{10\ g}$ = 43.6424

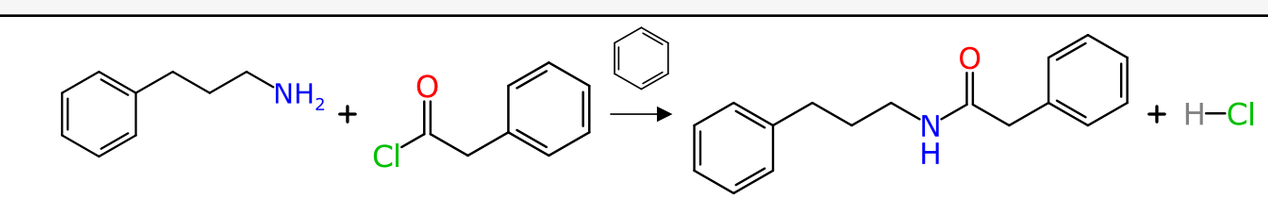
Atom-E = $\frac{293.335}{153.18 + 140.155}$ = 100 %

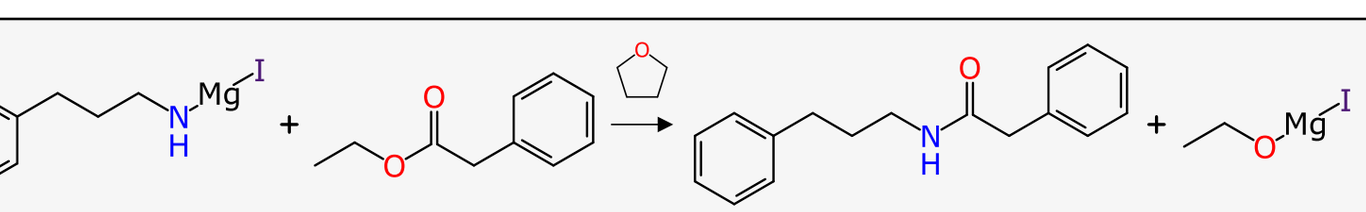
Decision Making

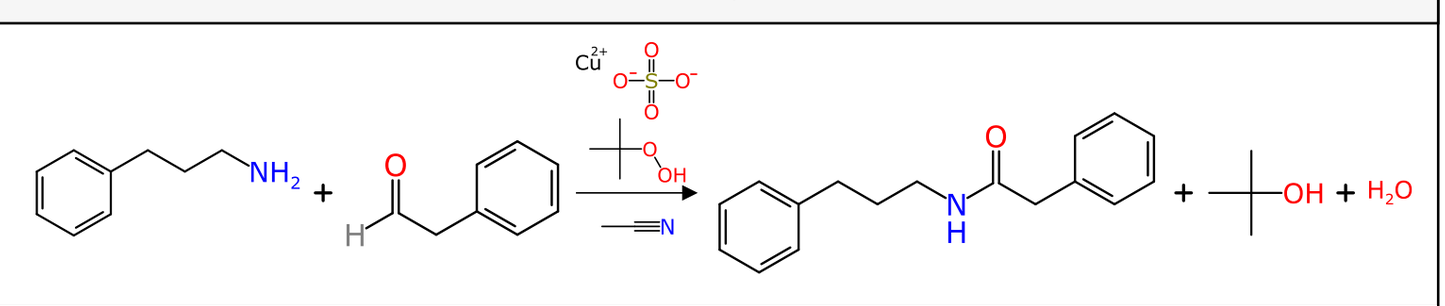
Product Search Results







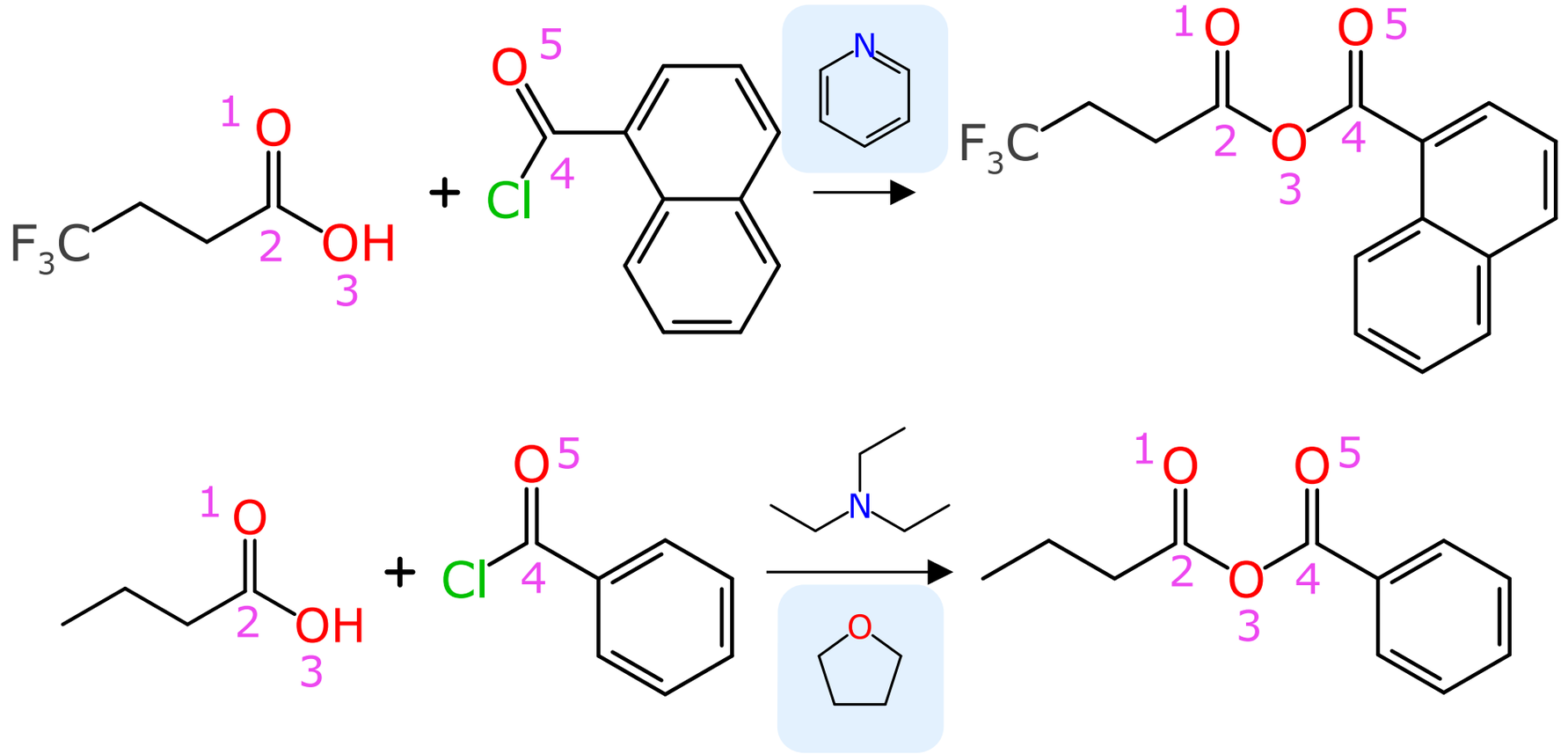




Yield	PMI	E-factor	Atom Economy
100%	2.18	1.18	100%
84%	12.49	11.49	93.3%
82%	19.17	18.17	87.4%
100%	8.26	7.26	56.3%
63%	8.93	7.93	73.3%

Model Building

- Most reaction data is noisy and incomplete
- Imagine opportunities with quantity & quality...



- For example: model solvent substitution

Conclusions & Future

- Most published reactions intractible to machines
- Most reaction informatics formats 50% complete
- Full description has immediate benefits...
- ... eventual large scale machine learning.
- µPublications with provenance: the path to open repositories - **but** requires attention to content

Acknowledgments

- Antony Williams
- Sean Ekins
- Leah McEwen
- Open data advocates
- Inquiries to info@molmatinf.com

MOLECULAR MATERIALS INFORMATICS

http://molmatinf.com
http://molsync.com
http://cheminf20.org

@aclarkxyz