Reaction Informatics with RDKit
An explosive combination

Roger Sayle

NextMove Software, Cambridge, UK

3rd RDKit User Group Meeting, Merck KGaA, Darmstadt, Germany, Thursday 23rd October 2014

RDKIT TALKTORIAL: WE'RE WATCHING

- Programmers who write:

```
double avg-sim = 0.0;
for (i=0; i<len(ref_id_list); i++) {
  sim = …
  avg_sim = ((avg_sim*i) + sim) / (i+1);
}
printf(“avg_sim: %g\n”,avg_sim);
```

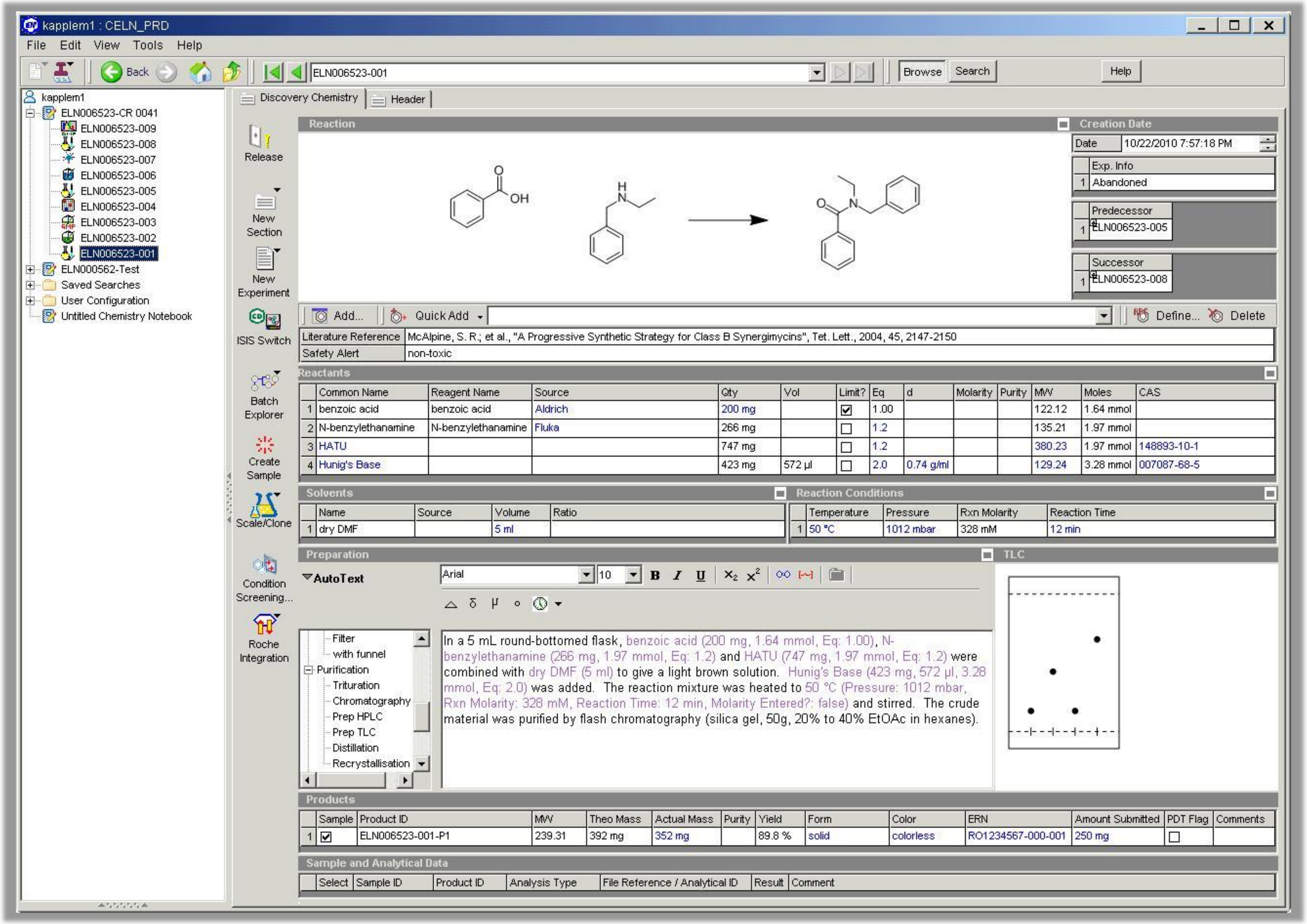
- Should prefer to write:

```
double total = 0.0;
int count = len(ref_id_list);
for (i=0; i<count; i++) {
  sim = …
  total += sim;
}
printf(“avg_sim”: %g\n”,double/count);
```

INTRODUCTION

- Reaction Informatics is the discipline of representing and analyzing chemical reactions in a computer.
- Reactions in this sense are instances of physical experiments, often with quantities and conditions, not virtual transformations.
- Elsevier’s Reaxys, Infochem’s SPRESI and in-house electronic lab notebooks (ELNs) are examples for reaction databases.

EXAMPLE REACTION IN AN ELN



ELN006523-001

Creation Date

Date	10/22/2010 7:57:18 PM

	Exp. Info
1	Abandoned

	Predecessor
1	ELN006523-005

	Successor
1	ELN006523-008

Literature Reference	McAlpine, S. R.; et al., "A Progressive Synthetic Strategy for Class B Synergimycins", Tet. Lett., 2004, 45, 2147-2150
Safety Alert	non-toxic

Reactants

	Common Name	Reagent Name	Source	Qty	Vol	Limit?	Eq	d	Molarity	Purity	MW	Moles	CAS
1	benzoic acid	benzoic acid	Aldrich	200 mg		☑	1.00				122.12	1.64 mmol	
2	N-benzylethanamine	N-benzylethanamine	Fluka	266 mg		☐	1.2				135.21	1.97 mmol	
3	HATU			747 mg		☐	1.2				380.23	1.97 mmol	148893-10-1
4	Hunig's Base			423 mg	572 µl	☐	2.0	0.74 g/ml			129.24	3.28 mmol	007087-68-5

Solvents

	Name	Source	Volume	Ratio
1	dry DMF		5 ml	

Reaction Conditions

	Temperature	Pressure	Rxn Molarity	Reaction Time
1	50 °C	1012 mbar	328 mM	12 min

In a 5 mL round-bottomed flask, benzoic acid (200 mg, 1.64 mmol, Eq: 1.00), N-benzylethanamine (266 mg, 1.97 mmol, Eq: 1.2) and HATU (747 mg, 1.97 mmol, Eq: 1.2) were combined with dry DMF (5 ml) to give a light brown solution. Hunig's Base (423 mg, 572 µl, 3.28 mmol, Eq: 2.0) was added. The reaction mixture was heated to 50 °C (Pressure: 1012 mbar, Rxn Molarity: 328 mM, Reaction Time: 12 min, Molarity Entered?: false) and stirred. The crude material was purified by flash chromatography (silica gel, 50g, 20% to 40% EtOAc in hexanes).

Products

	Sample	Product ID	MW	Theo Mass	Actual Mass	Purity	Yield	Form	Color	ERN	Amount Submitted	PDT Flag	Comments
1	☑	ELN006523-001-P1	239.31	392 mg	352 mg		89.8 %	solid	colorless	RO1234567-000-001	250 mg	☐	

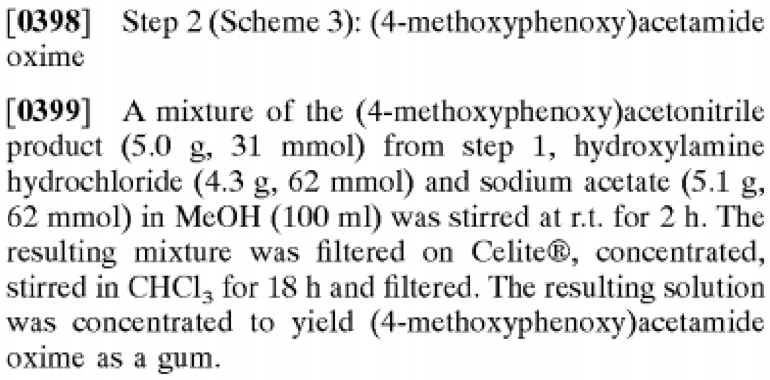
Sample and Analytical Data

	Select	Sample ID	Product ID	Analysis Type	File Reference / Analytical ID	Result	Comment

PHARMACEUTICAL PATENT EXAMPLE

US2002/0103226 [Merck]

Phosphodiesterase-4 inhibitors

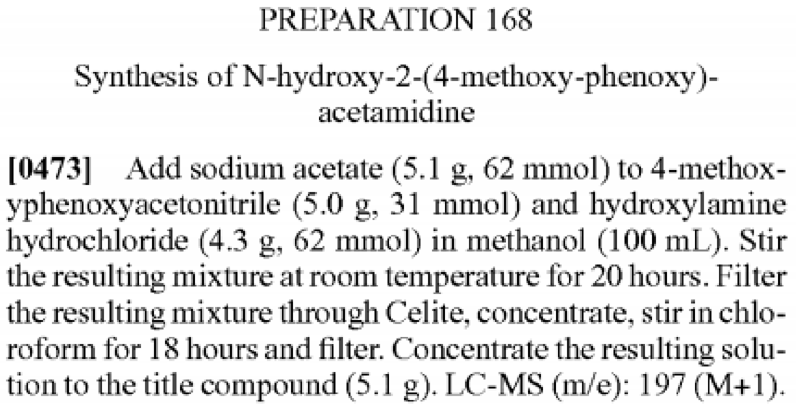


[0398] Step 2 (Scheme 3): (4-methoxyphenoxy)acetamide oxime

[0399] A mixture of the (4-methoxyphenoxy)acetonitrile product (5.0 g, 31 mmol) from step 1, hydroxylamine hydrochloride (4.3 g, 62 mmol) and sodium acetate (5.1 g, 62 mmol) in MeOH (100 ml) was stirred at r.t. for 2 h. The resulting mixture was filtered on Celite®, concentrated, stirred in $CHCl_3$ for 18 h and filtered. The resulting solution was concentrated to yield (4-methoxyphenoxy)acetamide oxime as a gum.

US2008/0139505 [Lilly]

Glutamate receptor potentiators



PREPARATION 168

Synthesis of N-hydroxy-2-(4-methoxy-phenoxy)-acetamidine

[0473] Add sodium acetate (5.1 g, 62 mmol) to 4-methoxyphenoxyacetonitrile (5.0 g, 31 mmol) and hydroxylamine hydrochloride (4.3 g, 62 mmol) in methanol (100 mL). Stir the resulting mixture at room temperature for 20 hours. Filter the resulting mixture through Celite, concentrate, stir in chloroform for 18 hours and filter. Concentrate the resulting solution to the title compound (5.1 g). LC-MS (m/e): 197 (M+1).

Both InChI=1S/C9H12N2O3/c1-13-7-2-4-8(5-3-7)14-6-9(10)11-12/h2-5,12H,6H2,1H3,(H2,10,11)

I SAY TOMAYTO, YOU SAY TOMAHTO

- **SMILES**: A line notation for molecules
- **SMARTS**: A pattern notation for molecules.
- **Reaction SMILES**: A line notation for reactions
  - Components annotated as reactants, agents or products.
- **Reaction SMARTS**: SMARTS for reactions.
  - “>[Pd]>” Find palladium catalyzed reactions
- **SMIRKS**: A molecular transformation notation.
  - “[Pb:1]>>[Au:1]” Transform lead into gold.

RDKit has both kinds of Reaction

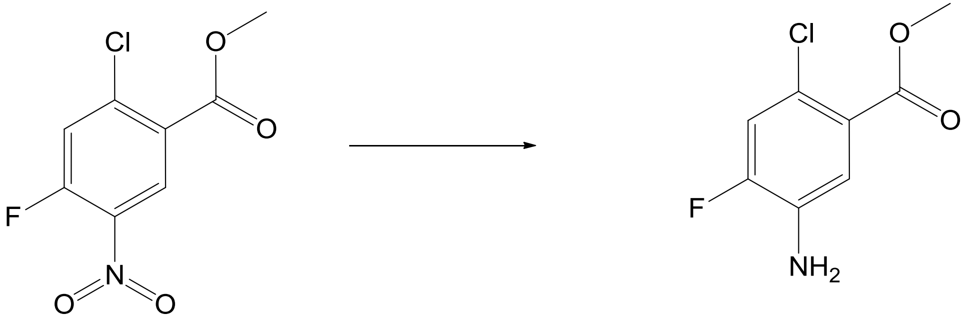
- RDKit::ChemicalReaction
  - ChemicalReaction::addReactantTemplate
  - ChemicalRecation::addProductTemplate
  - ChemicalReaction::addAgentTemplate
- RDKit::RWMol/RDKit::ROMol
  - RDKit::Atom::getProp<int>(“molRxnRole”)
- Representation Interconversion
  - RxnMolToChemicalReaction [ROMol→ChemicalReaction]
  - ChemicalReactionToRxnMol [ChemicalReaction→ROMol]

FILE FORMAT CLEVERNESS

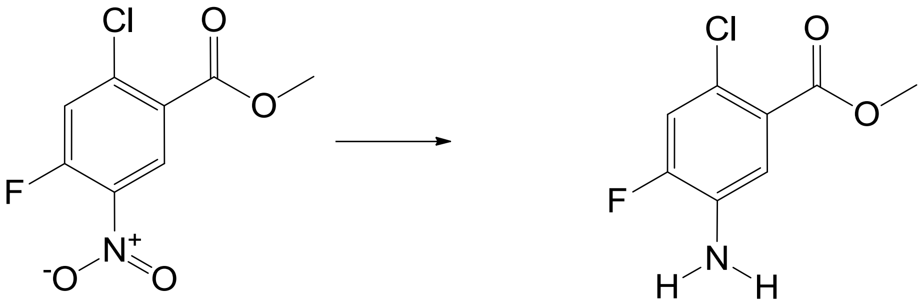
- RDKit uses ChemAxon extensions to MDL RXN files.
  - ChemicalReactionToRxnBlock(ChemicalReaction &rxn,
    bool separateAgents)
  - RxnBlockToChemicalReaction [and friends]
- SMILES
  - RxSmartsToChemicalReaction(…, bool useSmiles)
  - ChemicalReactionToRxnSmiles
  - ChemicalReactionToRxnSmarts
- SD and Mol files
  - MolToMolBlock
  - MolBlockToMol

MOLECULE NORMALIZATION

- Duplicate reactions can be caused by alternate chemistry representations requiring normalization.
- This problems can be solved by Reaction InChIs.
- EN01585-15



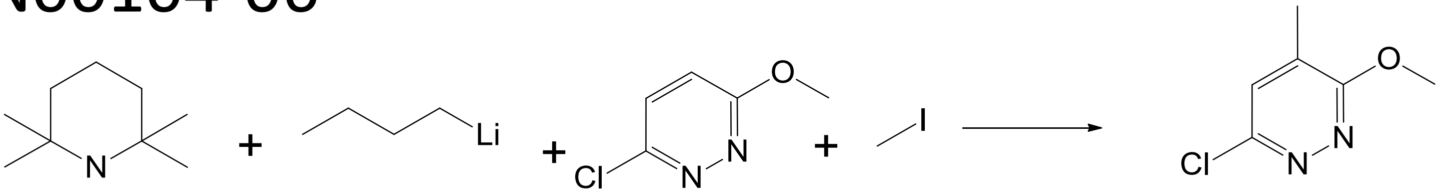
- EN01995-47



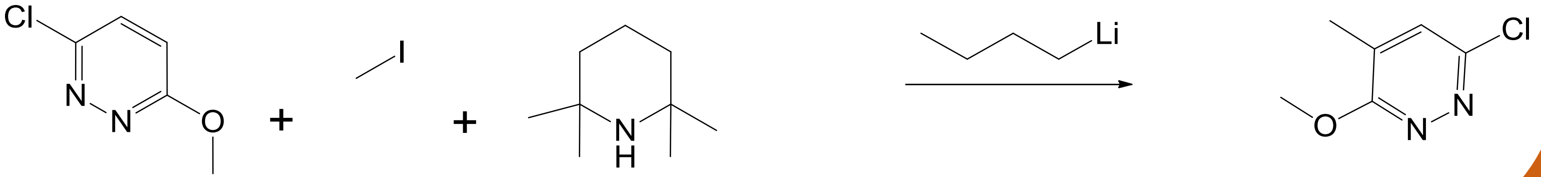
REACTION ROLE NORMALIZATION

- Some duplicates result from inconsistent reaction roles (reactants vs. agents) in the chemist's sketch.

- EN00104-06

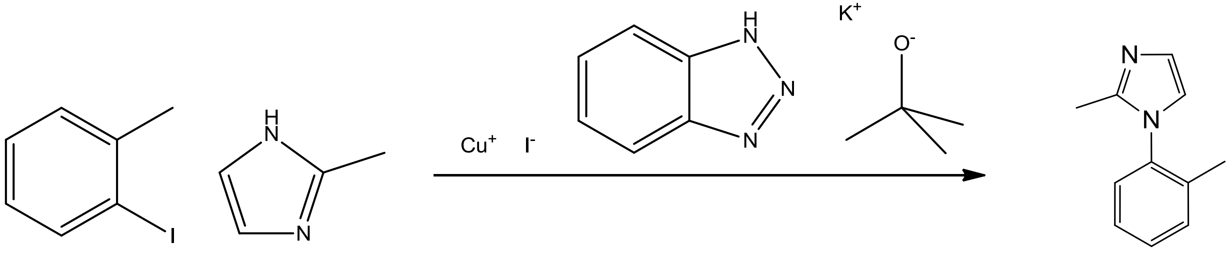


- EN00104-47

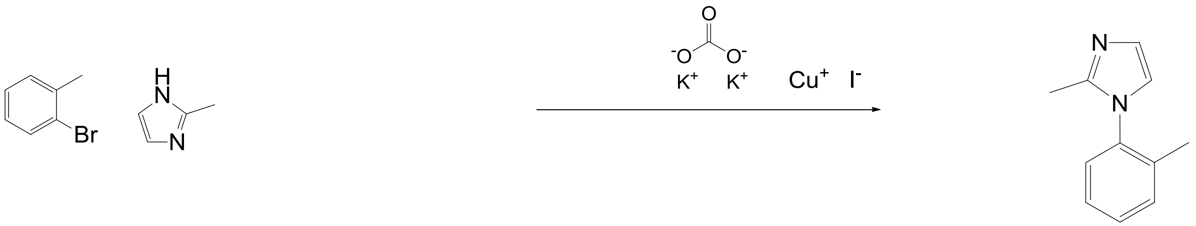


REACTION ROLE NORMALIZATION

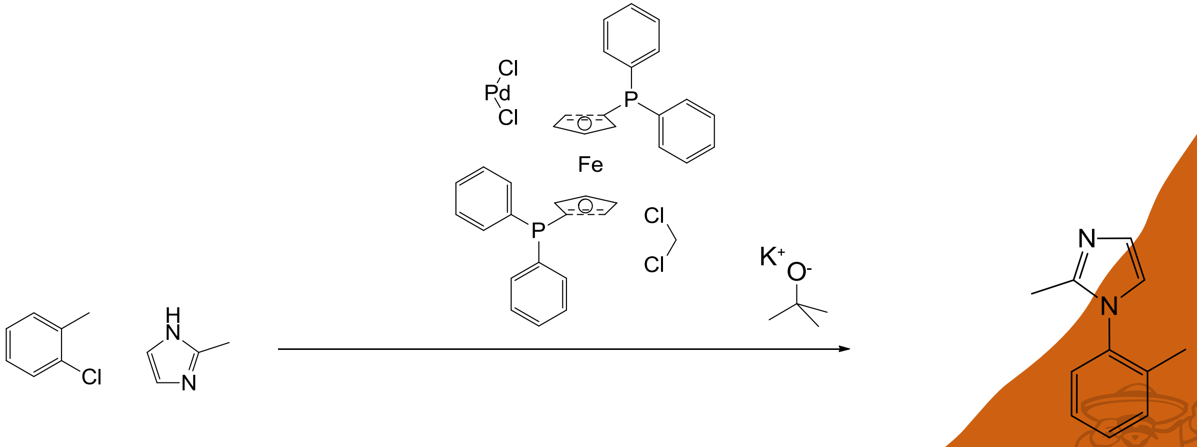
- EN00930-16



- EN00930-25



- EN00930-60



CHEMICAL HAZARD ANALYSIS

- Elemental Composition Analysis
- Oxygen Balance
- Heat of Formation Prediction
- Maximum Heat of Decomposition
- Maximum Heat of Combustion

ADDING GHS DIAMONDS

LEONARD'S RULE

- "In general, compounds with structures that contain a high proportion of nitrogen and/or oxygen atoms, relative to carbon atoms ([Number of C+N+O atoms] / [Number of N+O] < 3), tend to be unstable."

- John Leonard, Barry Lygo and Garry Procter, "**Advanced Practical Organic Chemistry**", Third Edition, CRC Press, 2013.

Energetic Substructures

- UN recommendations on “Transport of Dangerous Goods” lists the following function groups as being associated with explosive properties:
  - **C-C unsaturation**: Acetylenes, acetylides, 1,2-dienes.
  - **C-Metal, N-Metal**: Grignard reagents, organo-lithium compounds.
  - **N-N**: Azides, aliphatic azo compounds, diazonium salts, hydrazines, sulfonylhydrazides.
  - **O-O**: Peroxides, ozonides.
  - **N-O**: Hydroxylamines, nitrates, nitro compounds, nitroso compounds, N-oxides, 1,2-oxazoles.
  - **N-Halogen**: Chloramines, fluoramines.
  - **O-Halogen**: Chlorates, perchlorates, iodosyl compounds.

OXYGEN BALANCE

- Oxygen Balance (OB%) is used to indicate the degree to which an explosive can be oxidized.
- $OB\% = \frac{-1600}{Mol.wt.of\ compound} \times (2X + \left(\frac{Y}{2}\right) + M\ - Z)$
  - X = number of carbon atoms
  - Y = number of hydrogen atoms
  - Z = number of oxygen atoms
  - M = number of metallic atoms
- TNT ($C_7H_5N_3O_6$) has MW of 227.1 and OB of -74%.

HEAT OF FORMATION APPROXIMATION

- Vatani *et al.* propose a simple estimate of ΔH_f°

$$\Delta H_f^\circ \sim= 50 - 80\mathit{nSK} + 53\mathit{SCBO} - 169\mathit{nO} - 175\mathit{nF} - 267\mathit{nHM}$$

nSK	Number of non-H atoms
SCBO	Sum of conventional bond orders
nO	Number of Oxygen atoms
nF	Number of Fluorine atoms
nHM	Number of Inorganic atoms

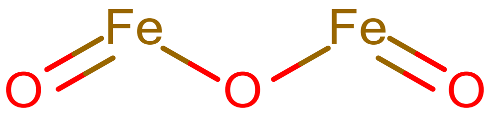
Training $R^2 = 0.983$ (n=892), Test $Q^2_{ext} = 0.9894$ (n=223)

Ali Vatani, Mehdi Mehrpooya and Farhad Gharagheizi, "**Prediction of Standard Enthalpy of Formation by a QSPR Model**", *International Journal of Molecular Sciences*, Vol. 8, pp. 407-432, 2007.

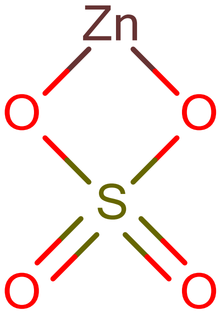
SOME MODELS CAN BE TOO SIMPLE

- Interestingly, some approximations can be over simplified, and useless for some applications.
- Alas, the model of Vatani *et al.* 2007 can’t be used to categorize a reaction as endothermic vs. exothermic.
- Descriptors based upon atomic composition, which are conserved during a reaction, result in identical heat of formations of both reactants and products.

A SIDE NOTE ON REPRESENTATION



Ferric oxide
Eisen(III)-oxid



Zinc sulfate
Zinksulfat

Predicting Calorimetry

- Legally, explosive hazard is quantified by experimental calorimetry to determine the maximum heat of decomposition/combustion of a compound.
- Heats of combustion are determined from the pyrolysis of a compound, into combustion products, in the presence of oxygen.
- Decomposition is similar, but with no additional oxidation, reflecting deflagration during detonation.
- These value determine whether a container must display an explosive hazard diamond during shipping.

Combustion/Deflagration Products

ΔHf(CO2) = -393.5 kJ/mol

ΔHf(H2SO4) = -735.13 kJ/mol

ΔHf(H2O) = -240.6 kJ/mol

ΔHf(SO3) = -395.77 kJ/mol

ΔHf(CO) = -111.8 kJ/mol → CO2 on combustion

ΔHf(SO2) = -296.81 kJ/mol → SO3 on combustion

ΔHf(H2S) = -20.6 kJ/mol → H2O + SO3 on combustion

ΔHf(CH4) = -74.87 kJ/mol → CO2 + H2O on combustion

ΔHf(H3N) = -45.94 kJ/mol → H2O + N2 on combustion

ΔHf(SCl2) = -17.57 kJ/mol → SO3 + Cl2 on combustion

ΔHf(SO2Cl2) = -354.80 kJ/mol → SO3 + Cl2 on combustion

ΔHf(HF) = -273.30 kJ/mol

ΔHf(HCl) = -92.31 kJ/mol → H2O + Cl2 on combustion

ΔHf(HBr) = -36.29 kJ/mol → H2O + Br2 on combustion

Example: Trinitrotoluene (TNT)

- Decomposition
  - $C_7H_5N_3O_6 \rightarrow 2.5H_2O + 1.75CO_2 + 5.25C + 1.5N_2$

    H_f=-63.2 →H_f=2.5*-240.6 + 1.75*-393.5 = -1290.12kJ/mol

    ΔH_d = -1226.92 kJ/mol
- Combustion
  - $C_7H_5N_3O_6 + 5.25O_2 \rightarrow 2.5H_2O + 7CO_2 + 1.5N_2$

    H_f=-63.2 →H_f=2.5*-240.6 + 7*-393.5 = -3356kJ/mol

    ΔH_c = -3292.8 kJ/mol

Setting Thresholds

- The “Guidelines for Chemical Reactivity Evaluation and Application to Process Design” by the Center for Chemical Process Safety (CCPS) advises
  - above 2.93 kJ/g heat of decomposition → High hazard
  - 1.26 to 2.93 kJ/g heat of decomposition → Medium hazard
  - 0.42 to 1.26 kJ/g heat of decomposition → Low hazard
  - Below 0.42 kJ/g heat of decomposition → Very low hazard

- Original limits in Kcal/g [1Kcal/mol = 4.184kJ/mol]
- Use molecular weight (g/mol) to covert from kJ/mol.

Thresholds in Practice

- In practice, the American Society for Testing Materials (ASTM) recommend a more complex categorization based both on maximal heat of decomposition, and the difference between the maximal heats of combustion and decomposition.
- This captures the degree to which a compounds is oxidized (detonation vs. deflagration).

ASTM CHETAH CLASSIFICATIONS

Name	MW	Hf	Hd	Hc	Hazard
Hexane	86.18	-199.4	2.31	-44.63	Very low
Acetone	58.08	-250.0	0.16	-28.45	Very low
Acetic acid	60.05	-483.9	0.04	-13.06	Very low
p-Nitroaniline	138.12	-5.6	-3.44	22.28	Medium*
2,4-Dinitrophenol	184.11	-235.5	-4.54	-14.16	High
2,4,6-Trinitrotoluene	227.13	-63.2	-5.40	-14.50	High
Picric acid	229.10	-217.9	-5.35	-10.93	High
Nitroglycerin	227.09	-370.0	-6.22	-6.22	High

H_f Heat of Formation (kJ/mol)
H_d Heat of Decomposition (kJ/g)
H_c Heat of Combustion (kJ/g)

ACCOUNTING FOR SCALE

- Clearly the scale of a reaction influences the chemical hazard associated with a reaction.
- This also allows us to unify safety thresholds between process development and MedChem R&D.
- Fortunately, having thresholds specified in kJ/mol or kJ/g allows us to account for quantities.
- Hence, safety thresholds are in energy (Joules).

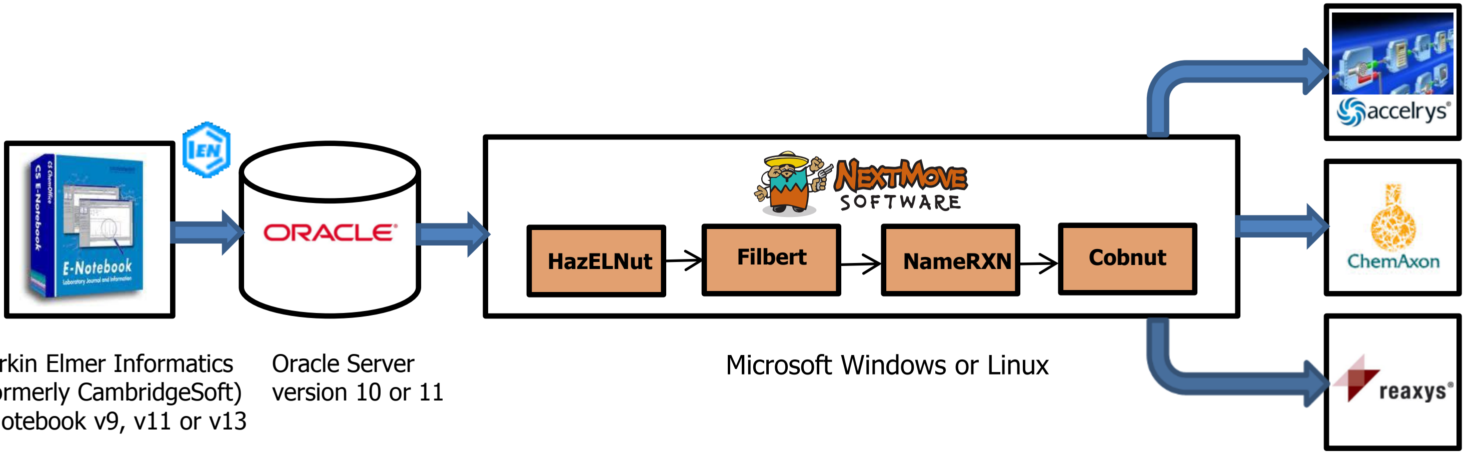
CONCLUSIONS/INSIGHTS

- Currently reviewing algorithms for estimating predicting compound heats of formation.
- However, early results show that the classification of hazardous materials is relatively insensitive to heat of formation of the starting material.
- The chemical hazard reflects the elemental composition rather than it’s functional groups.

ACKNOWLEDGEMENTS

- The team at NextMove Software
  - Daniel Lowe
  - Noel O’Boyle
  - John May
- The generosity of the big pharma
- Many thanks for your time.

CONTEXT: THE HAZELNUT PIPELINE



Perkin Elmer Informatics (formerly CambridgeSoft) eNotebook v9, v11 or v13

Oracle Server version 10 or 11

Microsoft Windows or Linux

Accelrys Pipeline Pilot (AstraZeneca, AbbVie & Hoffmann-La Roche)

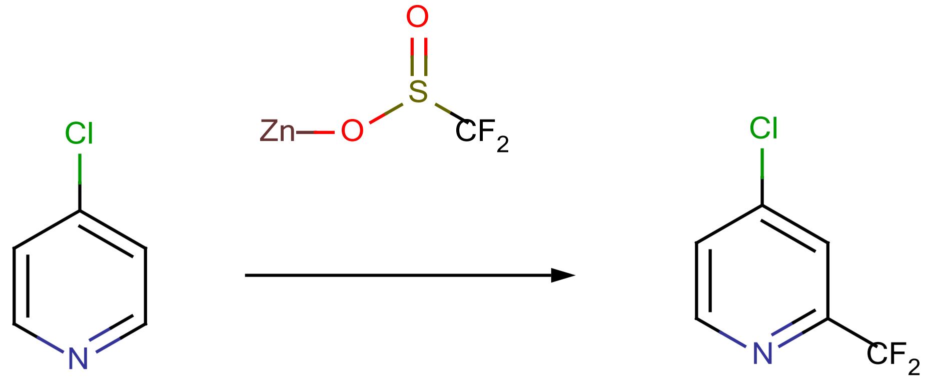
ChemAxon JChem Cartridge (GlaxoSmithKline & Novartis)

Elsevier Reaxys (Hoffmann-La Roche, AstraZeneca)

The pharmaceutical industry is increasingly making use of reaction data warehouses from ELN data to better share and learn from the experience of in-house and CRO chemists.

HIGH IMPACT REACTIONS

- In the introduction to Dial-a-Molecule yesterday, Richard Whitby introduced the concept of a high impact reactions. Here's a financial example.



$1.10 USD per g

$3,500 USD per g

Yuta Fujiwara et al., "Practical and innate carbon-hydrogen functionalization of heterocycles", *Nature* Vol. **492**, pp. 95-99, 6th December 2012.

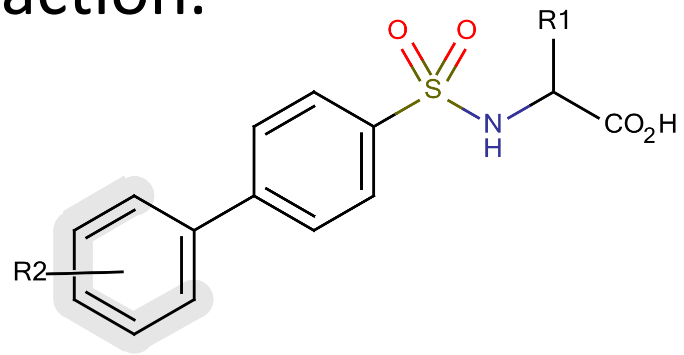
SYNTHESIS FAILURES AT LILLY

- At the 2013 Sheffield Cheminformatics conference, Christos Nicolaou highlighted the technical challenge with predicting compounds potentially accessible by the Lilly’s Advanced Synthesis Lab (ASL).
- In a proof-of-concept pilot project, only 25 of 90 compounds suggested by Lilly’s Annotated Reaction Repository (LARR) rule-set could be successfully synthesized in practice.

- `http://cisrg.shef.ac.uk/shef2013/talks/14.pdf`

SYNTHESIS FAILURES AT GSK

- Fortunately, poor success rates are not unique to Lilly or flow-chemistry. Is any reaction reliable?
- For example, Picket et al. 2011 describe the parallel synthesis of a 50x50 library of MMP-12 inhibitors by an iodo-Suzuki coupling reaction.



- Only 1704 of 2500 could be assayed [566 not made]

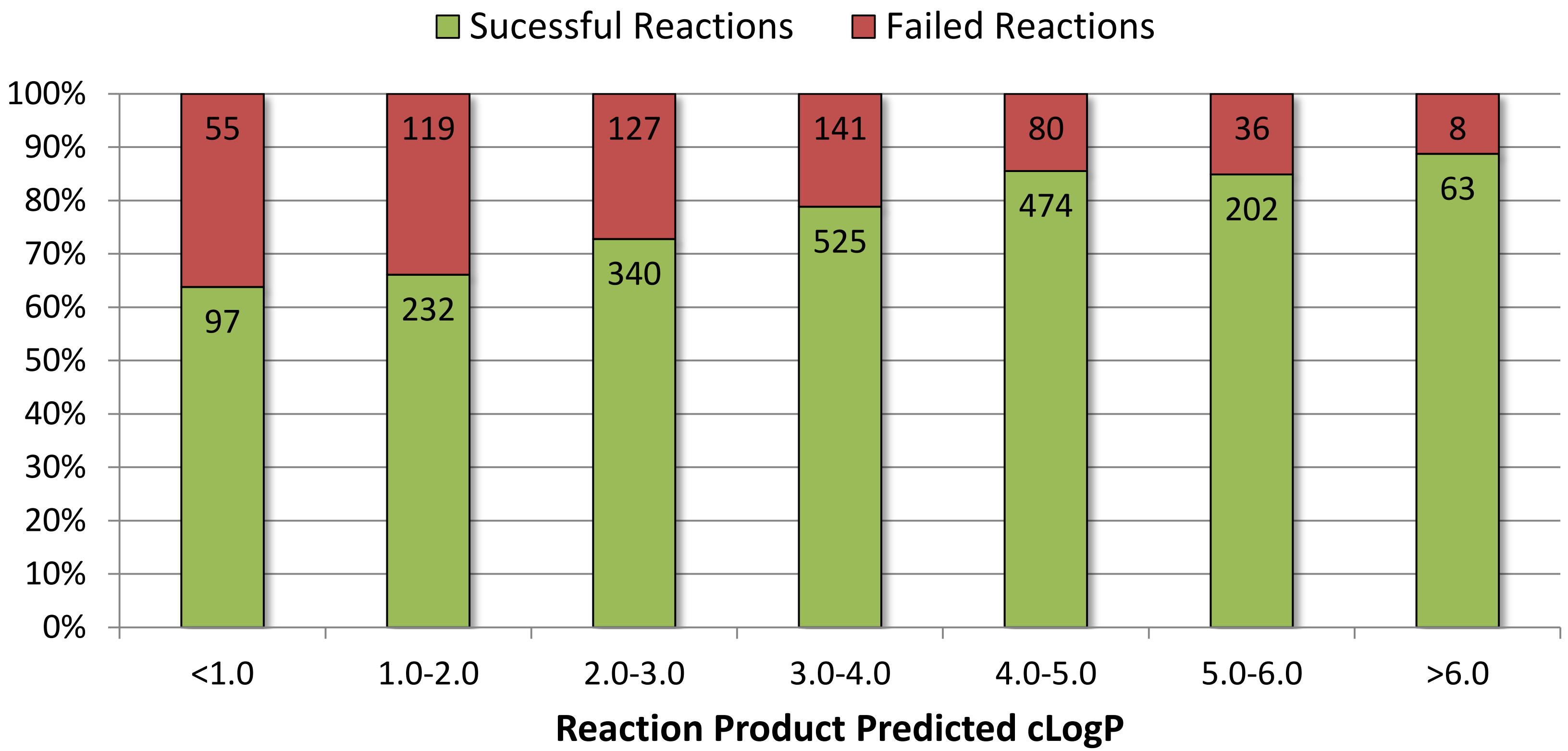
Pickett et al., *ACS Med. Chem. Lett.* 2(1):28, 2011

LEARNING FROM FAILURE

- Nadine *et al.* 2012 [1] hypothesize that low LogP is a major cause of synthesis failure in parallel synthesis of combinatorial libraries.
- Analysis confirms that this is indeed a significant factor for the GSK MMP-12 library.
  - 1704 compounds measured, mean logP = 3.56 (1.44)
  - 566 compounds not made, mean logP = 2.83 (1.52)
  - Student’s t-test for different distributions, $p<2x10^{-22}$.

1. Nadine, Hattotuwagama and Churcher ,“Lead-Oriented Synthesis: A New Opportunity for Synthetic Chemistry”, *Angew. Chem. Int. Ed*, 51:1114 2012.

NADINE-CHURCHER HYPOTHESIS



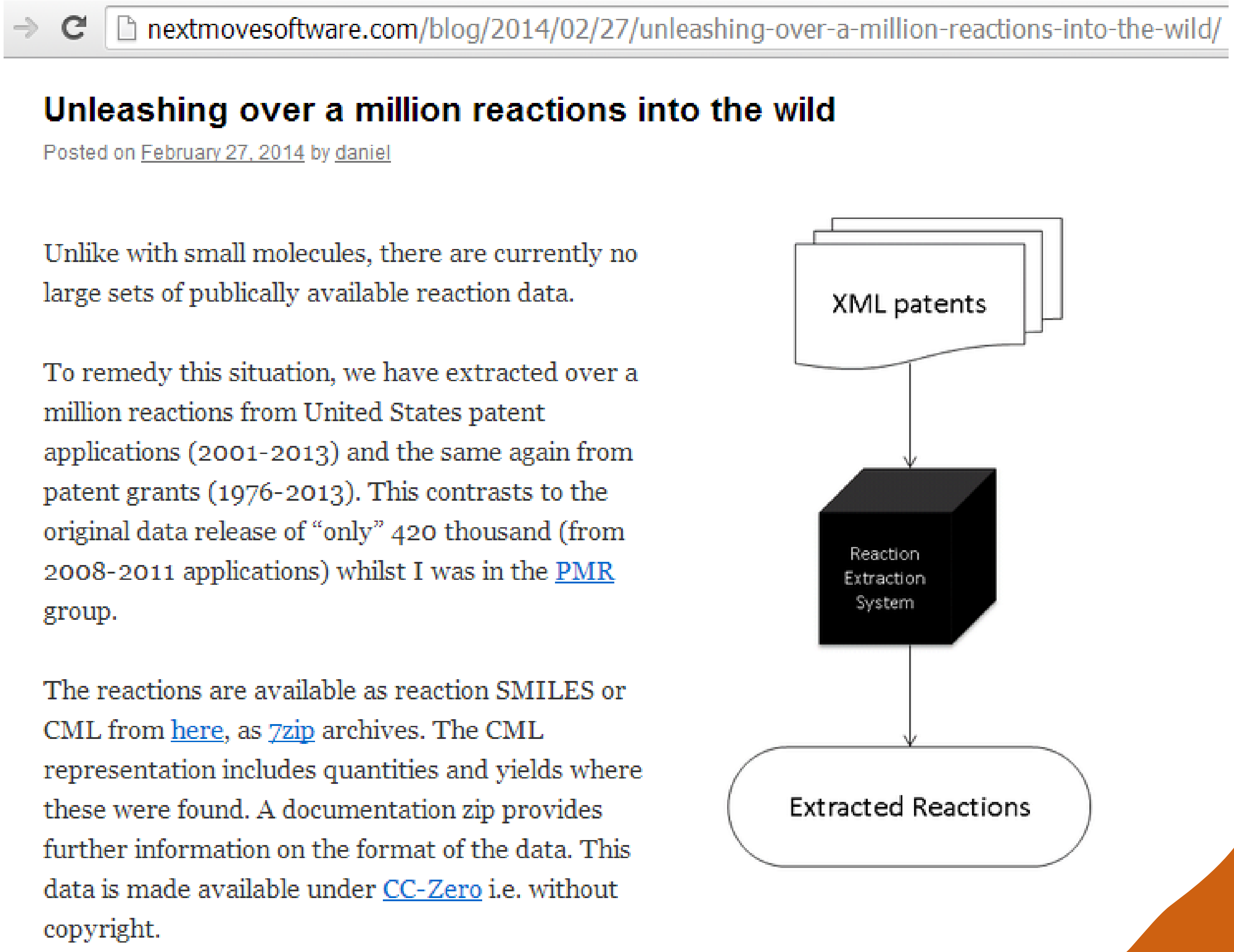
The clear trend between Suzuki coupling success rate and predicted octanol-water partition co-efficient.

NADINE-CHURCHER HYPOTHESIS

On 16,335 Suzuki coupling reactions extracted from US patent applications between 2001 and 2012.

LogP	Mean Yield	N Obs
< 1.0	52.89%	196
1.0 – 2.0	56.02%	1155
2.0 – 3.0	56.72%	2881
3.0 – 4.0	58.14%	4071
4.0 – 5.0	57.26%	3186
5.0 – 6.0	59.25%	2126
> 6.0	63.83%	2720

CHEMICAL REACTIONS FOR FREE



nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/

Unleashing over a million reactions into the wild

Posted on February 27, 2014 by daniel

Unlike with small molecules, there are currently no large sets of publically available reaction data.

To remedy this situation, we have extracted over a million reactions from United States patent applications (2001-2013) and the same again from patent grants (1976-2013). This contrasts to the original data release of "only" 420 thousand (from 2008-2011 applications) whilst I was in the PMR group.

The reactions are available as reaction SMILES or CML from here, as 7zip archives. The CML representation includes quantities and yields where these were found. A documentation zip provides further information on the format of the data. This data is made available under CC-Zero i.e. without copyright.

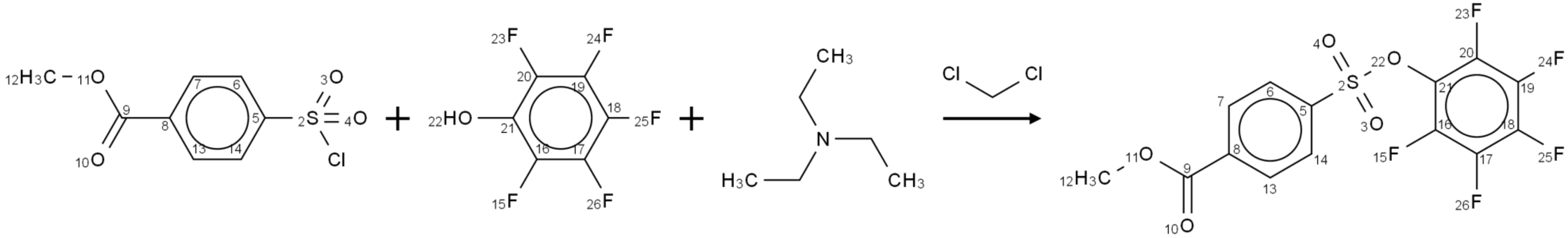
EXAMPLE REACTION MINING INPUT

Methyl 4-[(pentafluorophenoxy)sulfonyl]benzoate

To a solution of methyl 4-(chlorosulfonyl)benzoate (606 mg, 2.1 mmol, 1 eq) in DCM (35 ml) was added pentafluorophenol (412 mg, 2.2 mmol, 1.1 eq) and Et_3N (540 mg, 5.4 mmol, 2.5 eq) and the reaction mixture stirred at room temperature until all of the starting material was consumed. The solvent was evaporated in vacuo and the residue redissolved in ethyl acetate (10 ml), washed with water (10 ml), saturated sodium hydrogen carbonate (10 ml), dried over sodium sulphate, filtered and evaporated to yield the title compound as a white solid (690 mg, 1.8 mmol, 85%).

Unilever Cambridge
Centre For Molecular Science Informatics

EXAMPLE REACTION MINING OUTPUT



Unilever Cambridge

Centre For Molecular Science Informatics

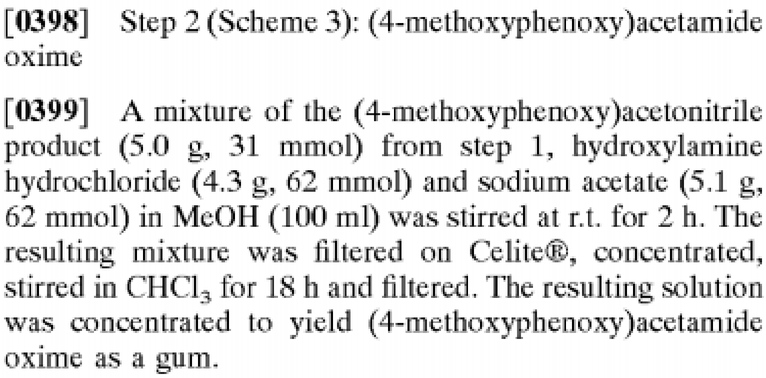
PHARMACEUTICAL PATENT EXAMPLE

US2002/0103226 [Merck]

Phosphodiesterase-4 inhibitors

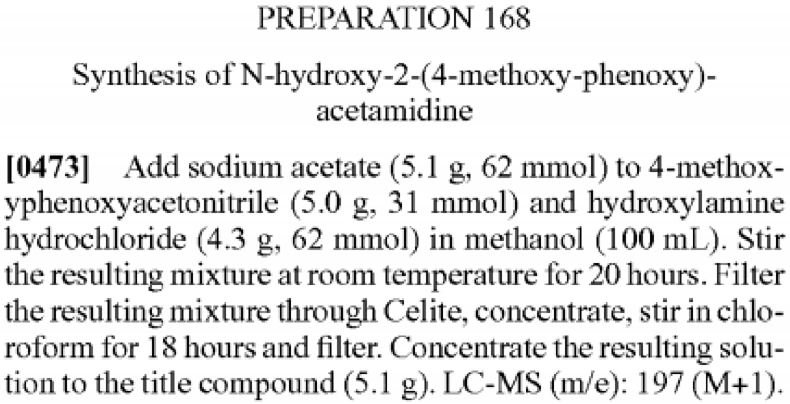
US2008/0139505 [Lilly]

Glutamate receptor potentiators



[0398] Step 2 (Scheme 3): (4-methoxyphenoxy)acetamide oxime

[0399] A mixture of the (4-methoxyphenoxy)acetonitrile product (5.0 g, 31 mmol) from step 1, hydroxylamine hydrochloride (4.3 g, 62 mmol) and sodium acetate (5.1 g, 62 mmol) in MeOH (100 ml) was stirred at r.t. for 2 h. The resulting mixture was filtered on Celite®, concentrated, stirred in $CHCl_3$ for 18 h and filtered. The resulting solution was concentrated to yield (4-methoxyphenoxy)acetamide oxime as a gum.



PREPARATION 168

Synthesis of N-hydroxy-2-(4-methoxy-phenoxy)-acetamidine

[0473] Add sodium acetate (5.1 g, 62 mmol) to 4-methoxyphenoxyacetonitrile (5.0 g, 31 mmol) and hydroxylamine hydrochloride (4.3 g, 62 mmol) in methanol (100 mL). Stir the resulting mixture at room temperature for 20 hours. Filter the resulting mixture through Celite, concentrate, stir in chloroform for 18 hours and filter. Concentrate the resulting solution to the title compound (5.1 g). LC-MS (m/e): 197 (M+1).

Both InChI=1S/C9H12N2O3/c1-13-7-2-4-8(5-3-7)14-6-9(10)11-12/h2-5,12H,6H2,1H3,(H2,10,11)

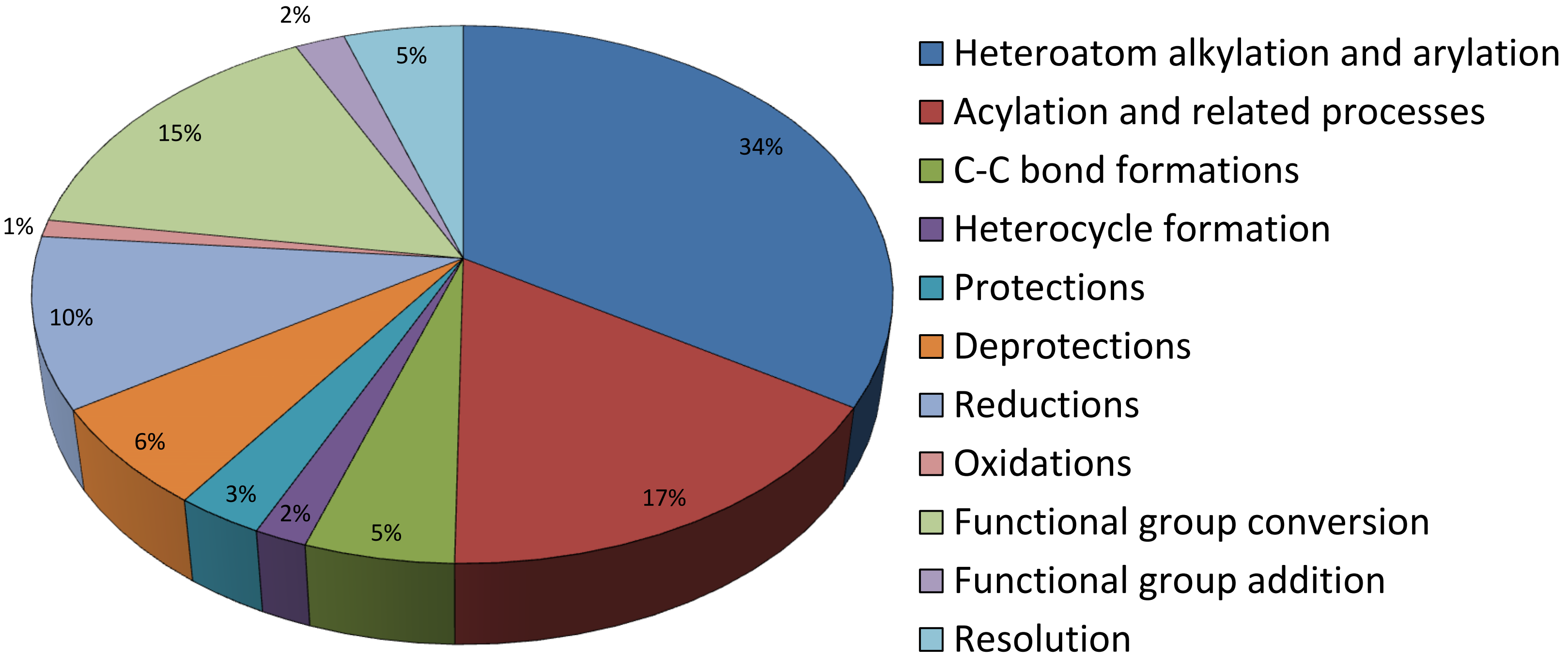
10 MOST POPULAR REACTIONS

ID	Name	Count
2.1.2	Carboxylic acid + amine	26,040
1.3.1	Buchwald-Hartwig amination	22,048
3.1	Suzuki coupling	16,508
1.7.6	Williamson ether synthesis	15,665
2.1.1	Amide Schotten-Baumann	11,016
7.1	Nitro to amino	10,234
6.1.1	N-Boc deprotection	9,821
6.2.2	CO2H-Me deprotection	9,487
6.2.1	CO2H-Et deprotection	6,749
2.2.3	Sulfonamide Schotten-Baumann	6,223

REACTION ONTOLOGY

- Reactions are classified into a common subset of the Carey et al. classes and the RSC's RXNO ontology.
- There are 12 super-classes
  - e.g. 3 C-C bond formation (RXNO:0000002).
- These contain 84 class/categories.
  - e.g. 3.5 Pd-catalyzed C-C bond formation (RXNO:0000316)
- These contain ~300 named reactions/types.
  - e.g. 3.5.3 Negishi coupling (RXNO:0000088)
- These require >675 SMIRKS-like transformations.

CATEGORIZATION OF ELN REACTIONS

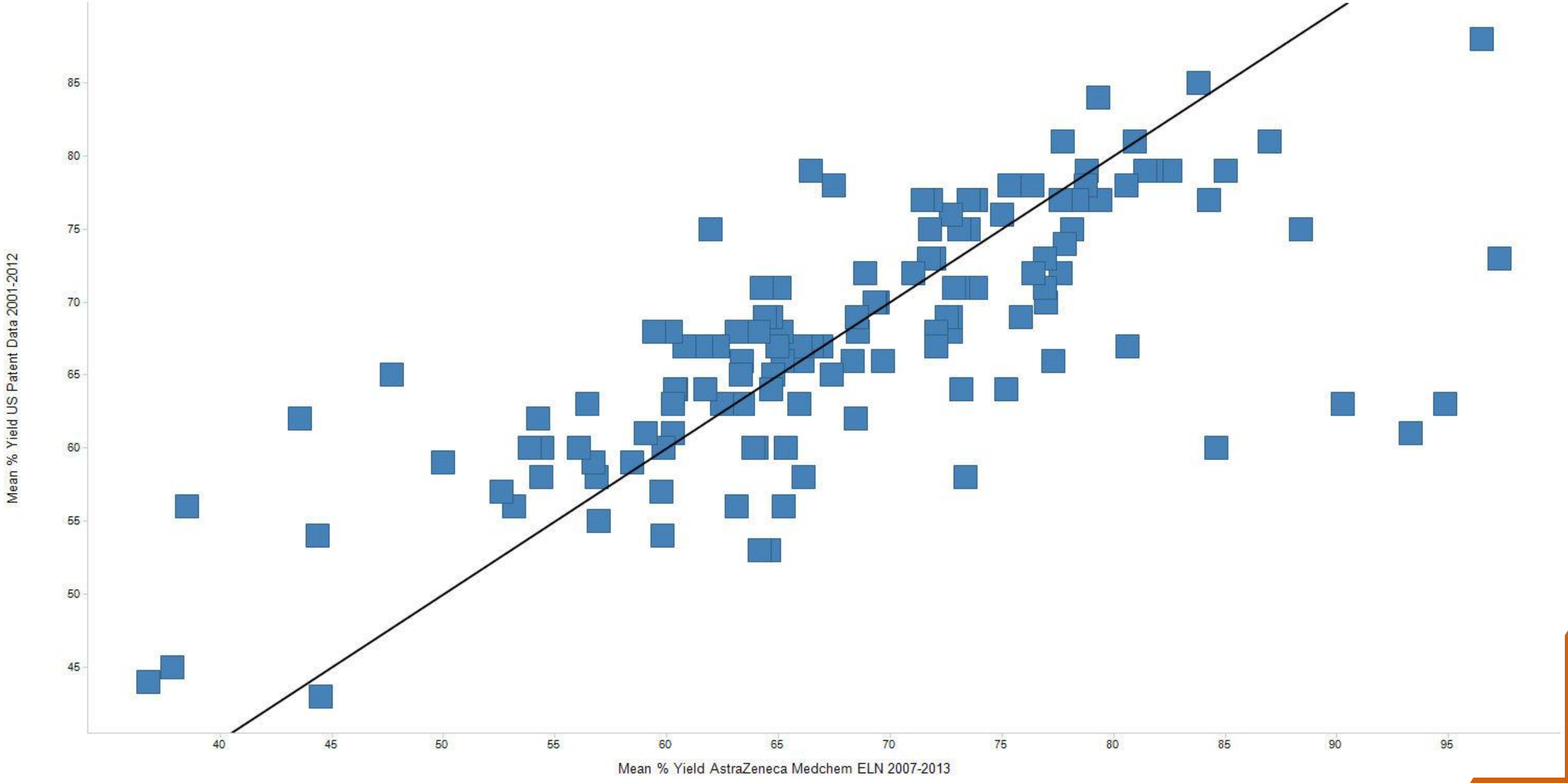


1. J. Carey, D. Laffan, C. Thomson, M. Williams, *Org. Biomol. Chem.* 2337, 2006.
2. S. Roughley and A. Jordan, *J. Med. Chem.* 54:3451-3479, 2011.

MOST/LEAST SUCCESSFUL REACTIONS

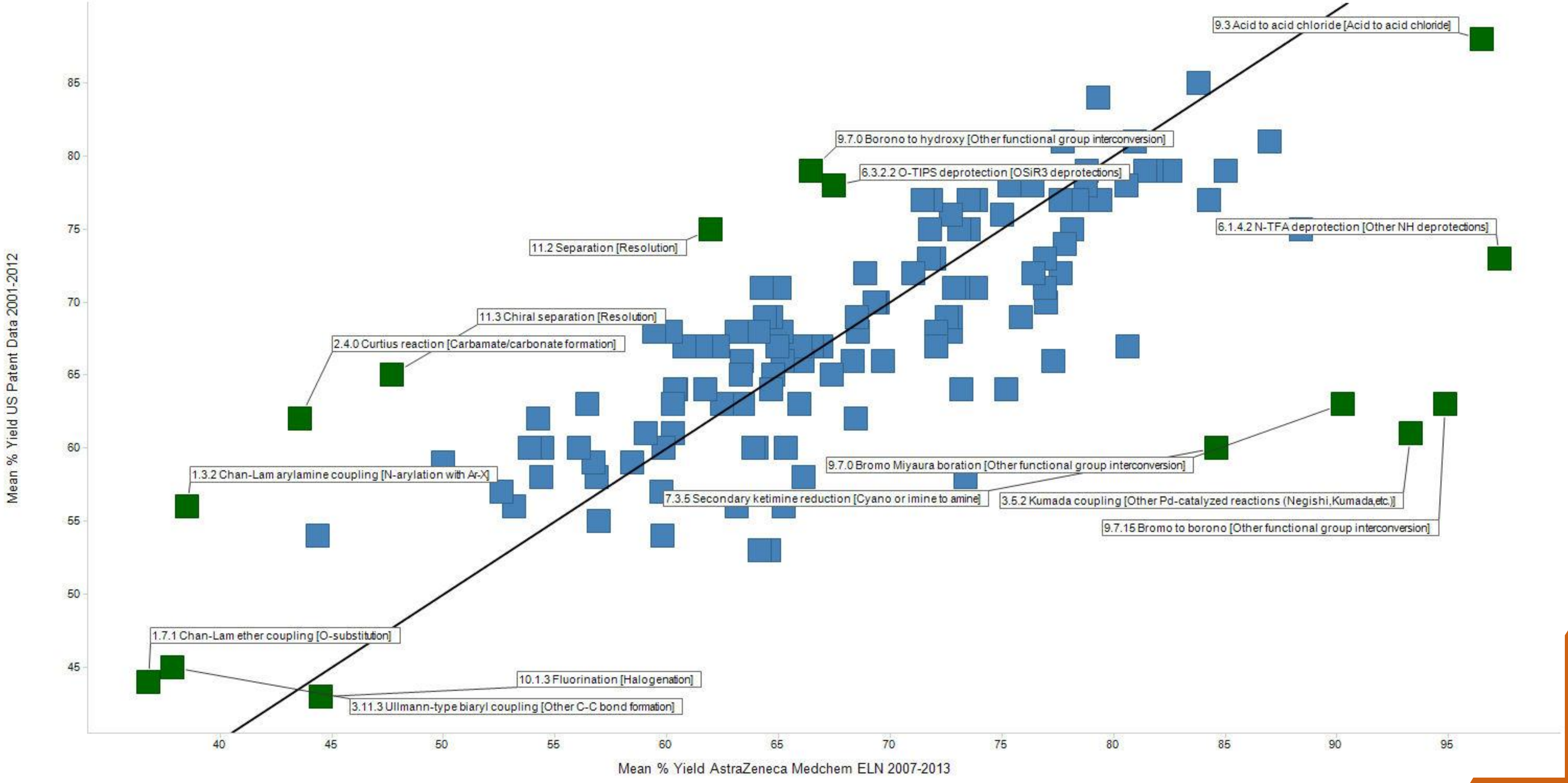
ID	Name	Mean Yield	Count
1.7.2	Diazomethane esterification	91%	41
9.3.1	Carboxylic acid to acid chloride	88%	704
9.7.14	Bromo to azido	85%	235
1.7.5	Methyl esterification	84%	2918
9.7.19	Bromo to iodo Finkelstein reaction	82%	116
6.1.3	N-Cbz deprotection	81%	1359
	…		
4.1.11	Larock indole synthesis	47%	55
3.11.3	Ullmann-type biaryl coupling	44%	407
1.7.1	Chan-Lam ether coupling	44%	154
4.1.4	Pinner pyrimidine synthesis	39%	47

"BIG DATA" REACTION YIELD ANALYSIS



AZ Data courtesy of Nick Tomkinson, AstraZeneca RDI, Alderley Park, UK.

"BIG DATA" REACTION YIELD ANALYSIS

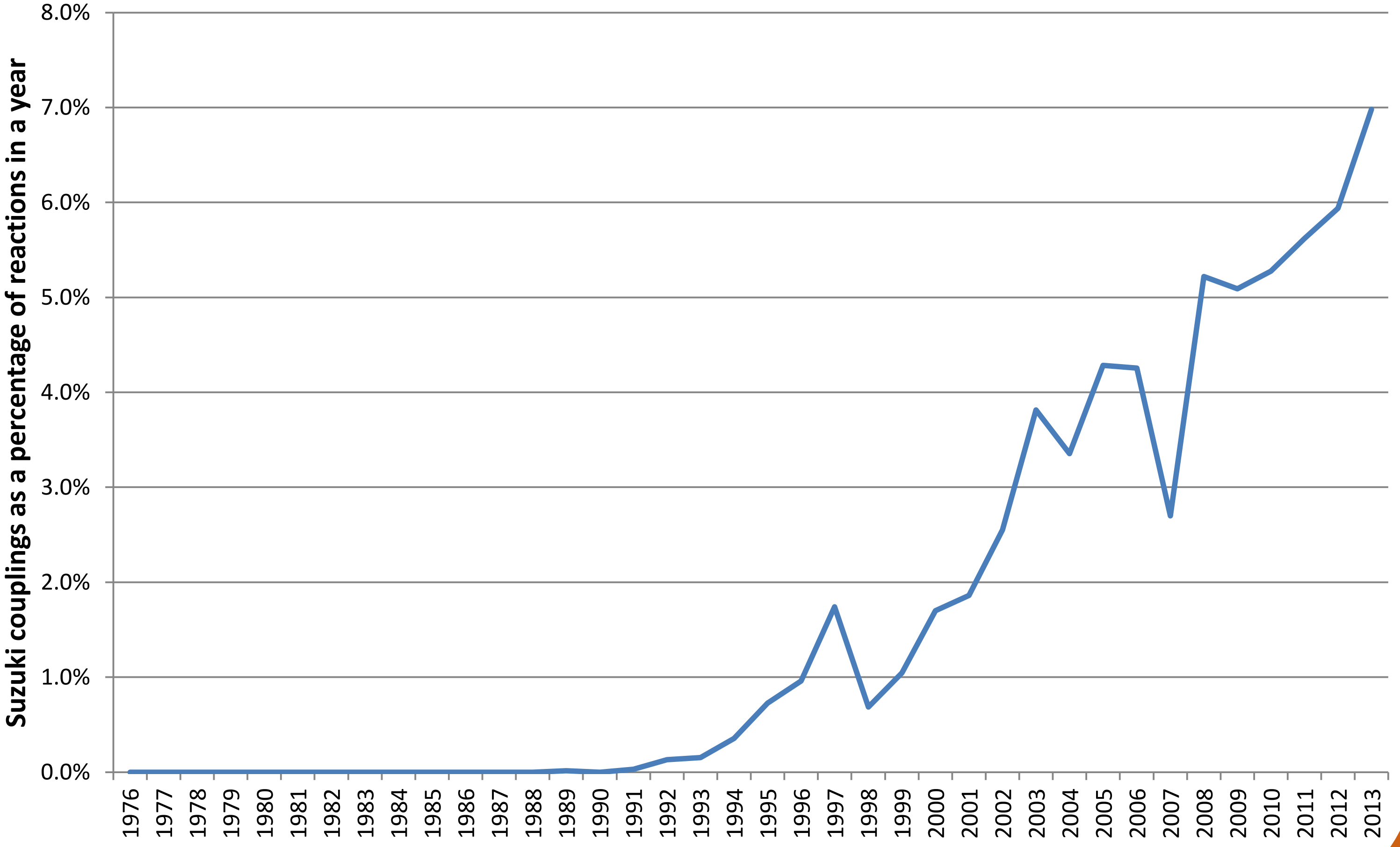


AZ Data courtesy of Nick Tomkinson, AstraZeneca RDI, Alderley Park, UK.

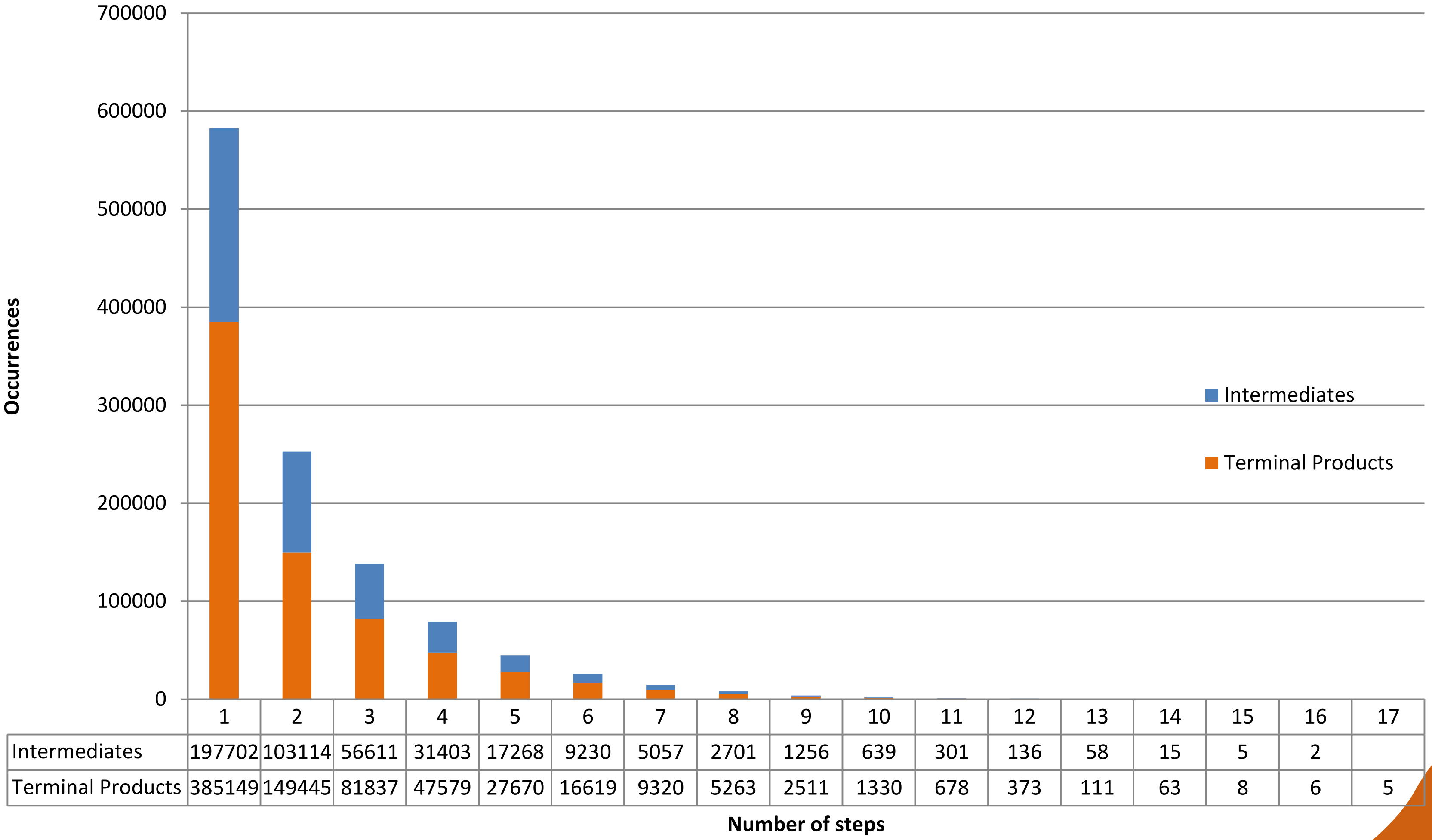
SUZUKI COUPLING LEAVING GROUPS

Leaving Group	Mean Yield	N Observations
Bromo	58.80%	10817
Chloro	57.96%	2752
Iodo	57.21%	2049
Triflyloxy	65.48%	717

Trends in Reaction Types



IDENTIFY SYNTHETIC ROUTES

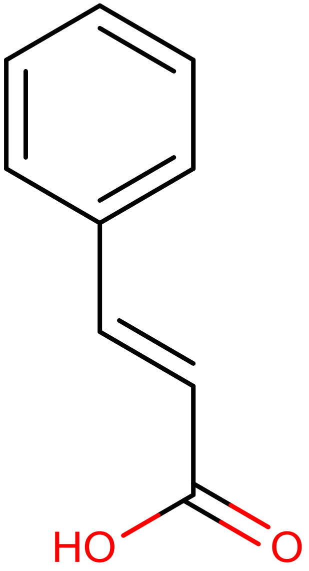


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Intermediates	197702	103114	56611	31403	17268	9230	5057	2701	1256	639	301	136	58	15	5	2	
Terminal Products	385149	149445	81837	47579	27670	16619	9320	5263	2511	1330	678	373	111	63	8	6	5

APPLICATION TO PLANNING 1

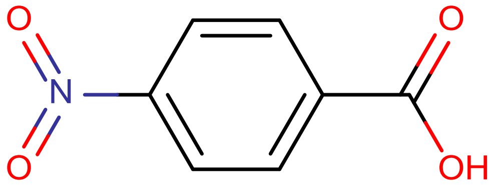
- Cinnamic Acid (PhCHCHCO2)

1. Bromo Heck reaction (272)
2. Horner-Wadsworth-Emmons reaction (268)
3. Wittig olefination (129)
4. Bromo Heck-type reaction (62)
5. Iodo Heck reaction (49)
6. Triflyloxy Heck[-type] reaction (43)
7. Ester Schotten-Baumann (10)
8. Bromo Suzuki coupling (5)
9. Stille reaction (2)
10. Olefin metathesis (1)



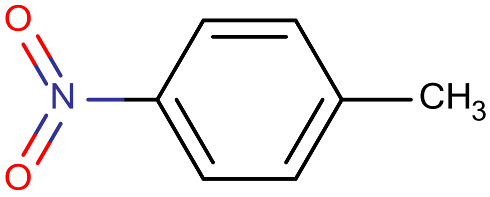
APPLICATION TO PLANNING 2

- p-Nitrobenzoic acid



1. Nitrile to carboxy (12)
2. CO2H-Me deprot (8)
3. CO2H-Et deprot (5)
4. Ester hydrolysis (1)
5. Nitration (1)

- p-Nitrotoluene



1. Nitration (96)
2. Bromo Suzuki-type (1)
3. Chloro Suzuki (1)

EXPERIMENTAL VALIDATION

- Synthesis of a novel aromatic heterocycle previously unreported in the literature.
- William Pitt et al., “Heteroaromatic Rings of the Future”, Journal of Medicinal Chemistry, 52(9):2952-2963, 2009.

Cautionary tale: Hidden effects

- In her 2009 PhD thesis, Hina Patel at the university of Sheffield, under the supervision of Val Gillet and Michael Bodkin of Eli Lilly UK, failed to observe any relationship between yield and product similarity.
- These conclusions were derived from the study of the 96 yields reported in just two papers published by the Chemistry department at Sheffield in 2006.

1. H. Cope et al. “Synthesis and SAR study of acridine, 2-methylquinoline and 2-phenylquinazoline analogues as anti-prion agents”, European Journal of Medicinal Chemsitry, Vol. 41, pp. 1124-1143, 2006.
2. T.R.K Reddy et al. “Library Design, synthesis and Screening: Pyridine dicarbonitriles as potential Prion Disease Therapeutics”, Journal of Medicinal Chemistry, Vol. 49, pp. 607-615, 2006.

COPE ET AL. 2006 YIELD DATA (1)

Reactant	Series 1 yield	Series 3 yield
3-Et	59%	92%
3-OMe	31%	99%
4-OMe	37%	86%
3-MeOH	42%	96%
3-F	79%	89%
3-OPh	31%	70%
3-CN	83%	80%
4-CO2H	18%	???

A reasonable initial hypothesis is that a carboxy group interferes with the acid catalyzed SNAr N-arylation.

COPE ET AL. 2006 YIELD DATA (2)

Reactant	Series 1 yield	Series 3 yield
3-Et	59%	92%
3-OMe	31%	99%
4-OMe	37%	86%
3-MeOH	42%	96%
3-F	79%	89%
3-OPh	31%	70%
3-CN	83%	80%
4-CO2H	18%	98%!

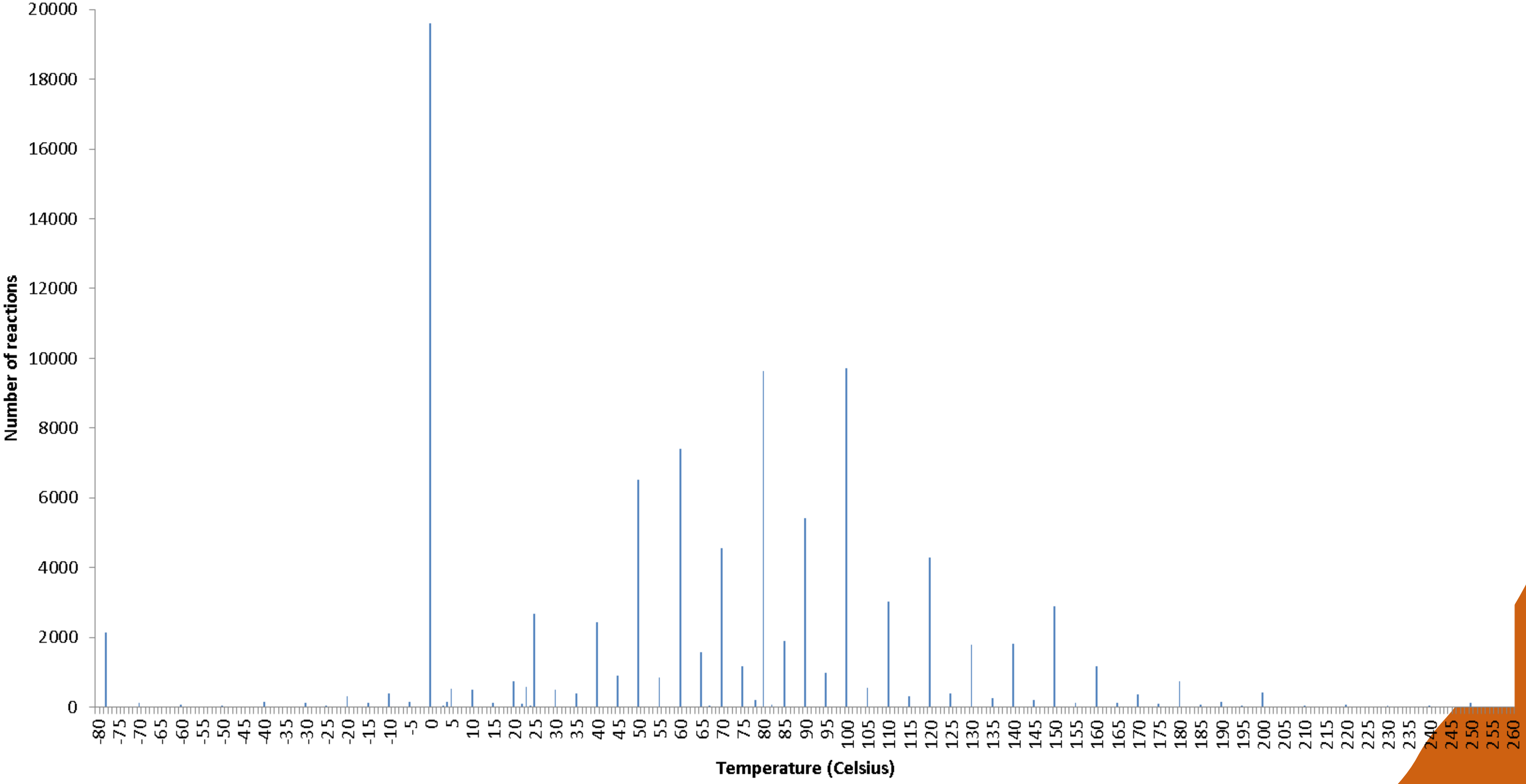
“Big data” reaction yield analysis reveals that the carboxylic acid is well tolerated in this reaction class.

COPE ET AL. 2006 YIELD DATA (3)

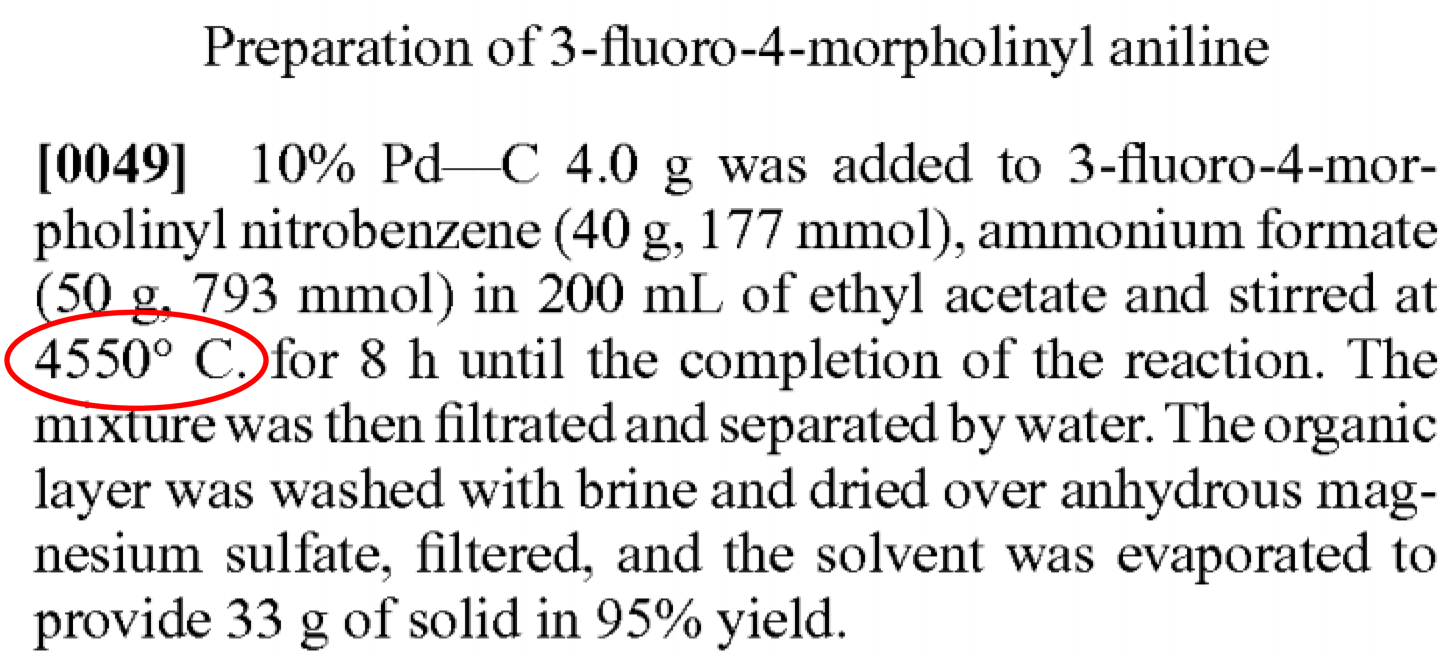
Reactant	Series 1 yield	Series 3 yield
3-Et	59% (24hrs)	92% (5hrs)
3-OMe	31% (24hrs)	99% (18hrs)
4-OMe	37% (24hrs)	86% (18hrs)
3-MeOH	42% (24hrs)	96% (18hrs)
3-F	79% (24hrs)	89% (18hrs)
3-OPh	31% (24hrs)	70% (18hrs)
3-CN	83% (24hrs)	80% (18hrs)
4-CO2H	18% (90 mins)	98% (26hrs)

The bigger picture is that Patel didn’t attempt to account for other significant factors in the Cope data, including reaction duration and temperature.

REACTION TEMPERATURES



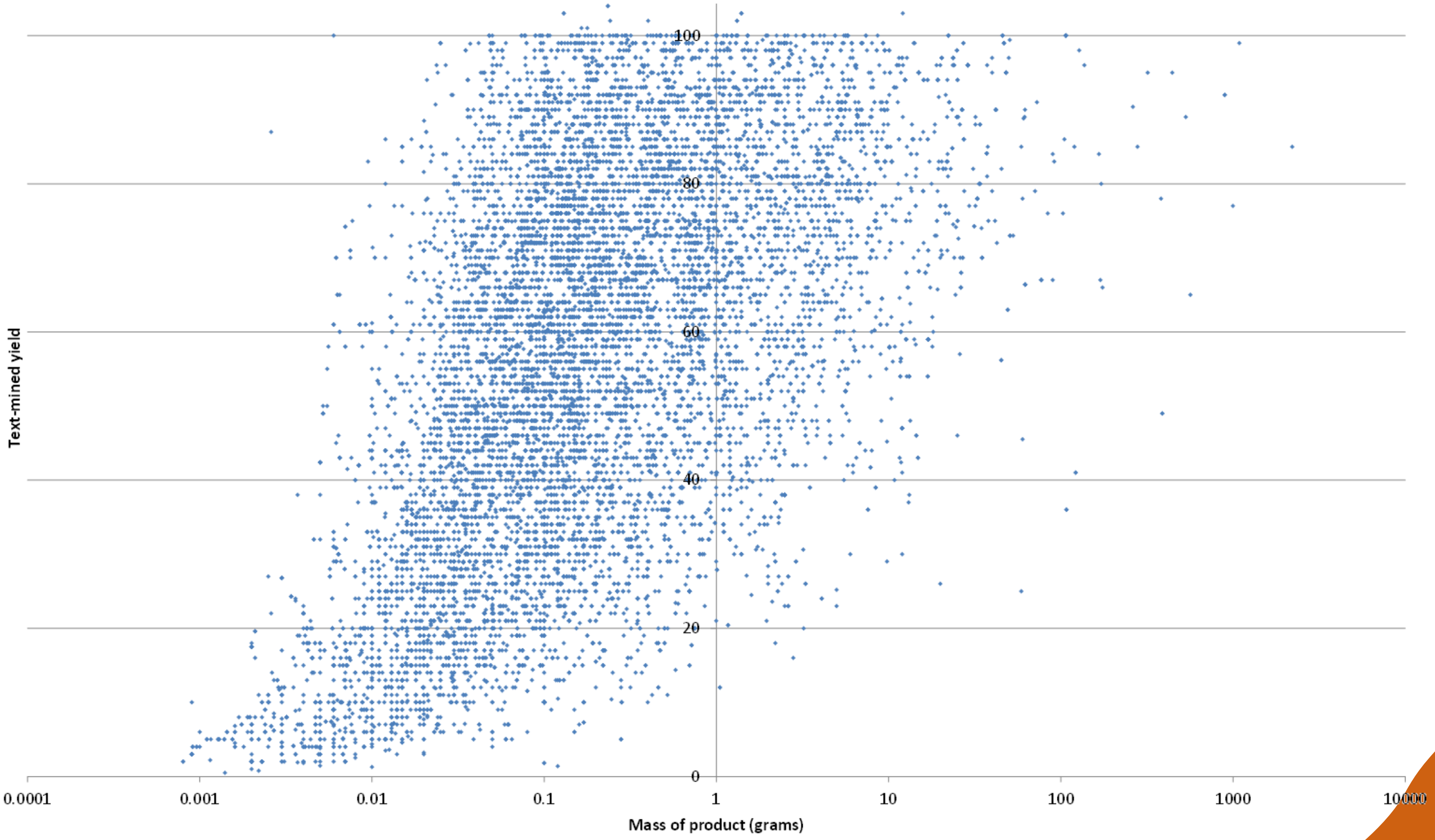
OUTLIERS INEVITABLE



Preparation of 3-fluoro-4-morpholinyl aniline

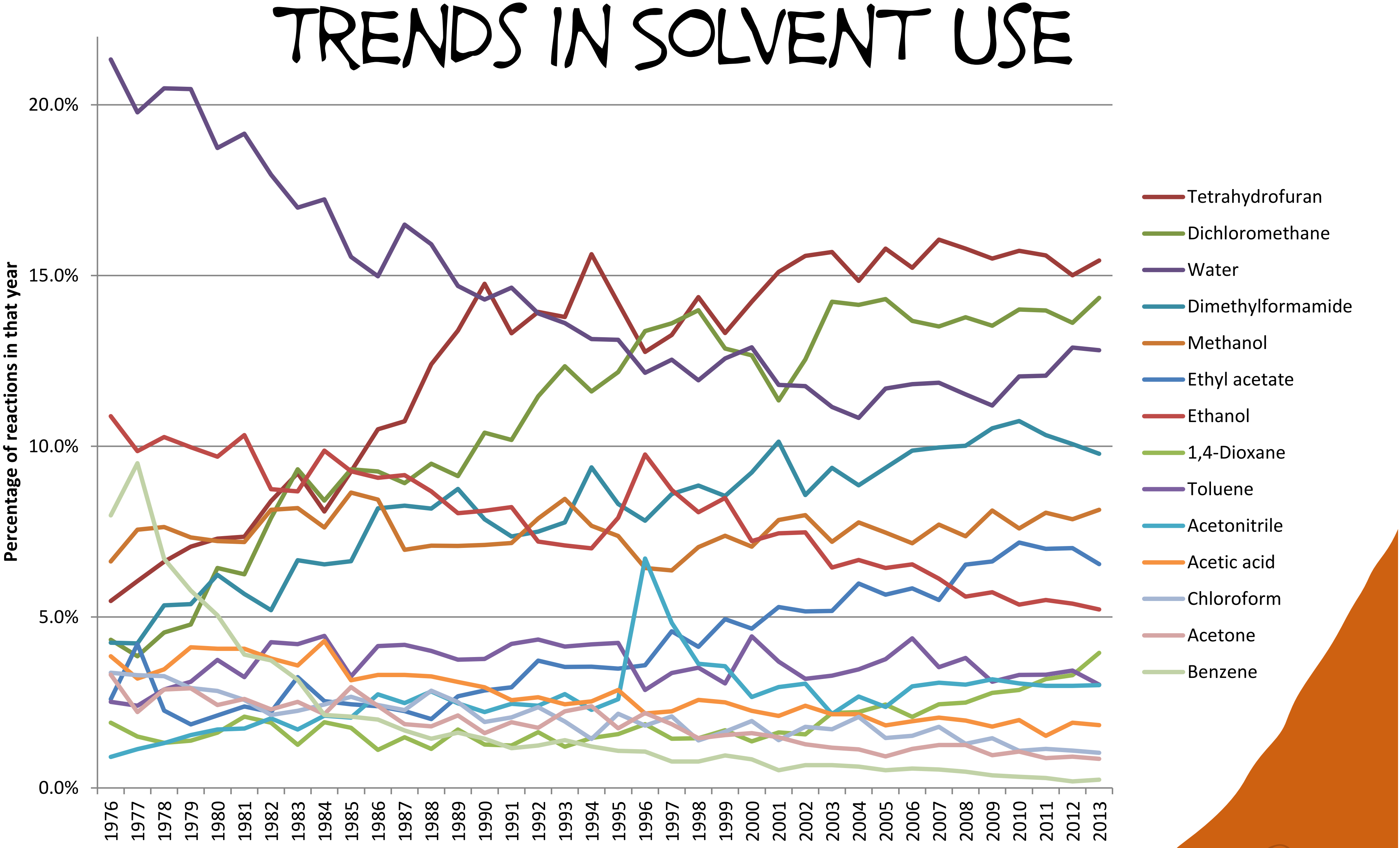
[0049] 10% Pd—C 4.0 g was added to 3-fluoro-4-morpholinyl nitrobenzene (40 g, 177 mmol), ammonium formate (50 g, 793 mmol) in 200 mL of ethyl acetate and stirred at 4550° C. for 8 h until the completion of the reaction. The mixture was then filtrated and separated by water. The organic layer was washed with brine and dried over anhydrous magnesium sulfate, filtered, and the solvent was evaporated to provide 33 g of solid in 95% yield.

Scale vs Yield



2001-2013 US applications, Suzuki couplings

TRENDS IN SOLVENT USE



ARE SOLVENTS GETTING GREENER?

1976	2013
Water (21%)	Tetrahydrofuran (15%)
Ethanol (11%)	Dichloromethane (14%)
Benzene (8%)	Water (13%)
Methanol (7%)	Dimethylformamide (10%)
Tetrahydrofuran (5%)	Methanol (8%)
Dichloromethane (4%)	Ethyl acetate (7%)
Dimethylformamide (4%)	Ethanol (5%)
Acetic acid (4%)	1,4-Dioxane (4%)
Chloroform (3%)	Toluene (3%)
Acetone (3%)	Acetonitrile (3%)

Total for top 10: 71% 82%

CONCLUSIONS

- Aggregating large experimental data sets is more than an dream, it can provide valuable scientific insights not otherwise visible.
- A major challenge is dealing with different representations and data models.
- Understanding these challenges and how this data can ultimately be used, can help guide the methods used to collect and store it.

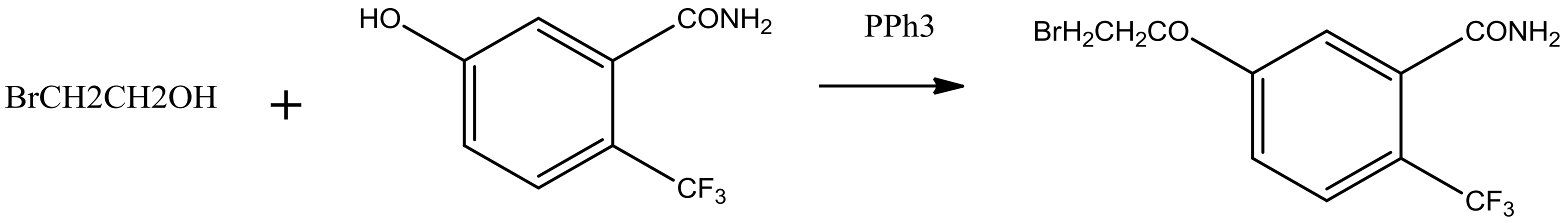
ACKNOWLEDGEMENTS

- NextMove Software
  - Daniel Lowe
  - Noel O’Boyle

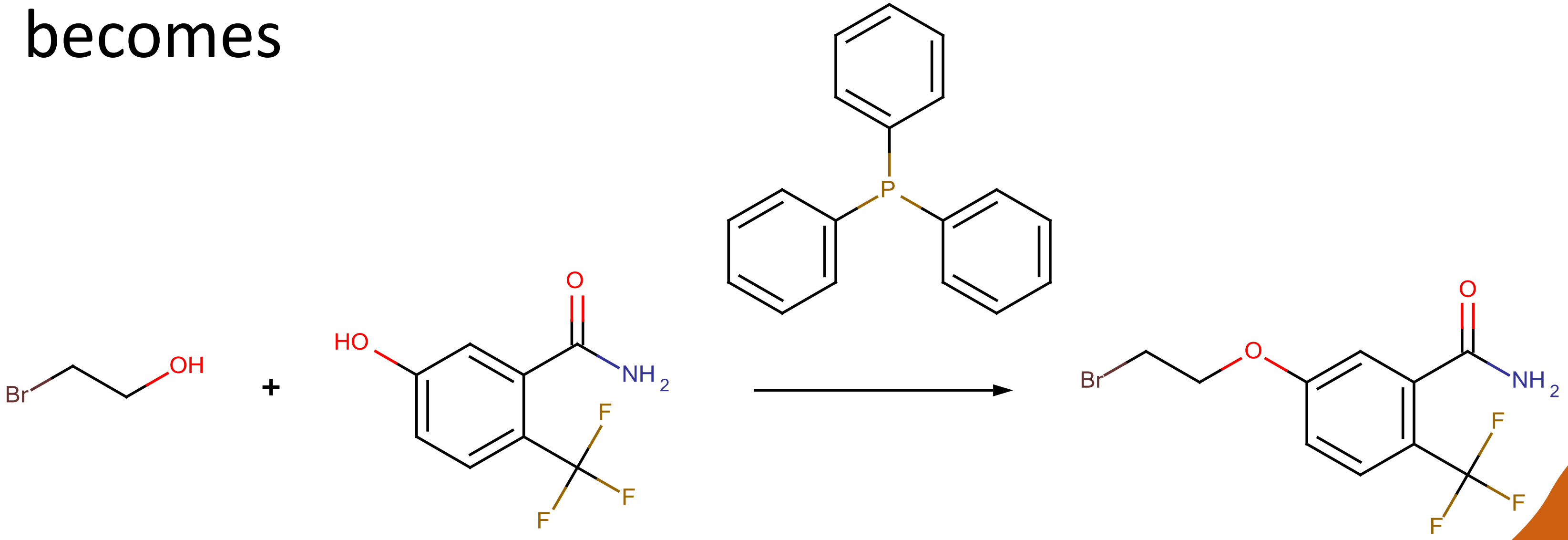
- Thank you for you time.
- Questions?

- AbbVie
- AstraZeneca
- Bristol-Myers Squibb
- GlaxoSmithKline
- Hoffmann-La Roche
- Novartis
- Royal Society of Chemistry
- Vernalis
- Vertex Pharmaceuticals

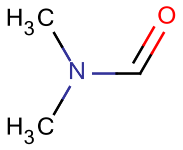
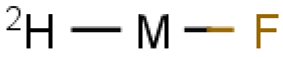
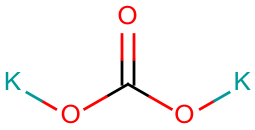
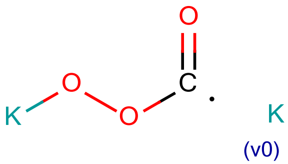
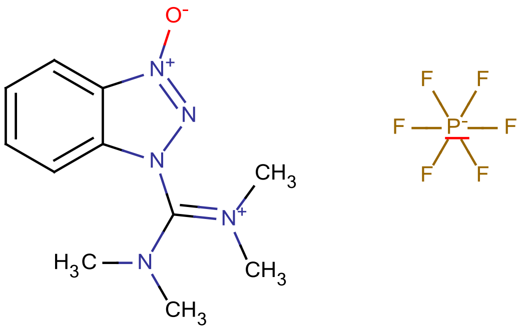
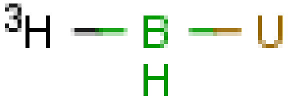
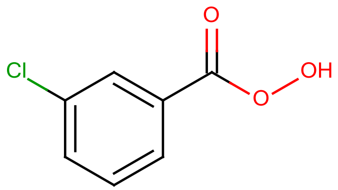
SKETCH SUPERATOM EXPANSION



becomes



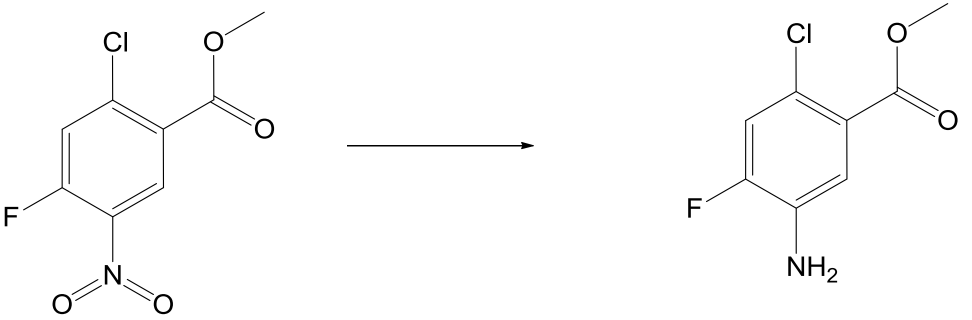
CHEMDRAW SUPERATOM CORRECTION

Chemist drew	Chemist Intended	Chemist got
DMF		
K2CO3		
HBTU		
mCPBA		

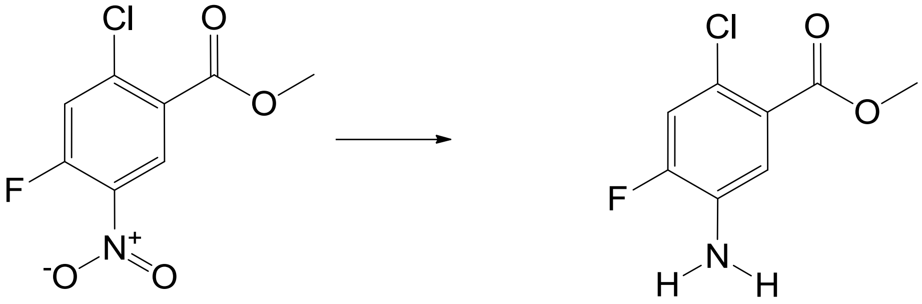
ROLE OF MOLECULE NORMALIZATION

- Some duplicates are caused by alternate chemistry representations requiring normalization of SMILES.

- EN01585-15

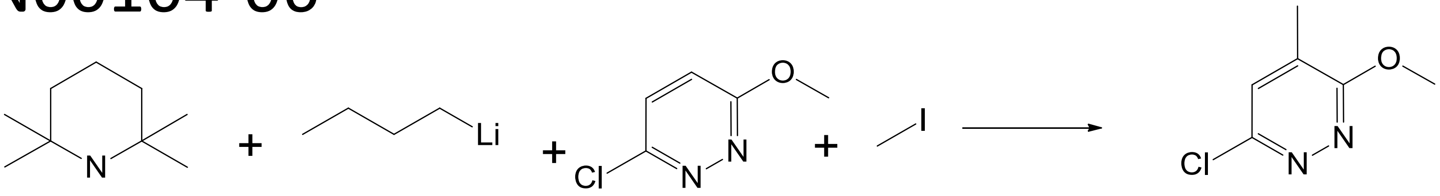


- EN01995-47

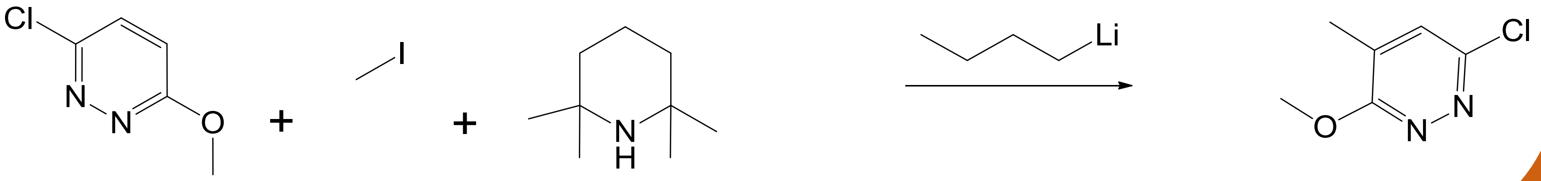


ROLE OF REACTION NORMALIZATION

- Some duplicates are caused by inconsistent reaction roles in the chemist's sketch.
- EN00104-06

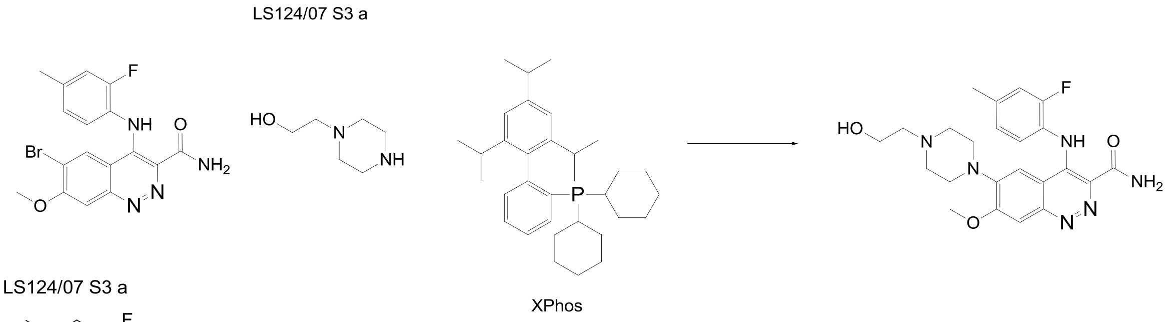


- EN00104-47

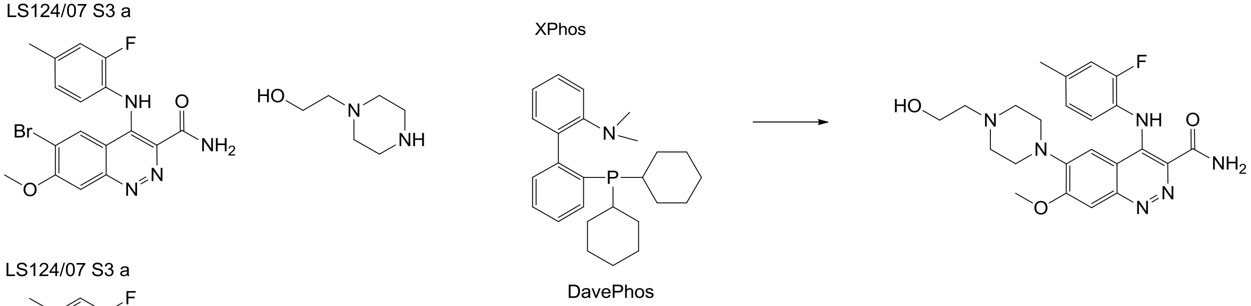


AZ EXAMPLES WITH SAME PARENT

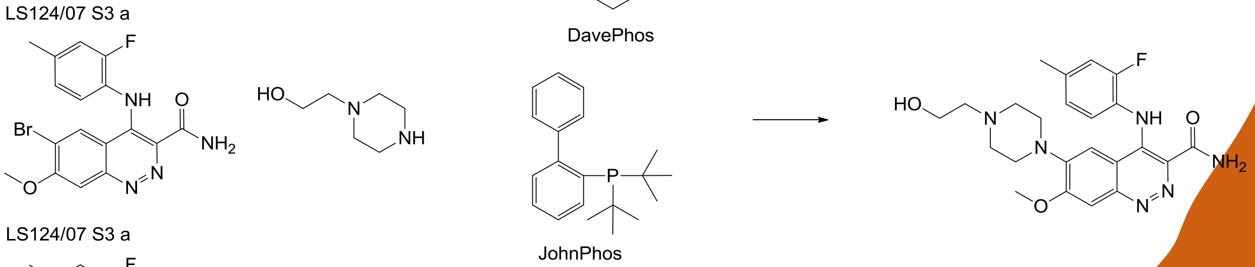
- EN01325-20



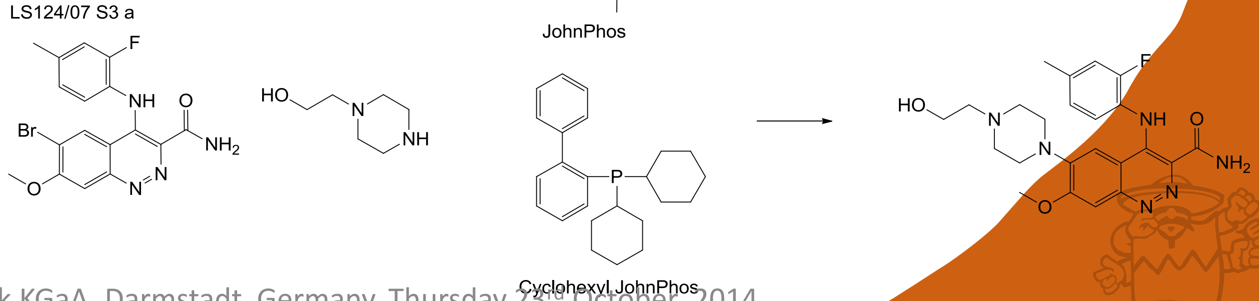
- EN01325-22



- EN01325-25

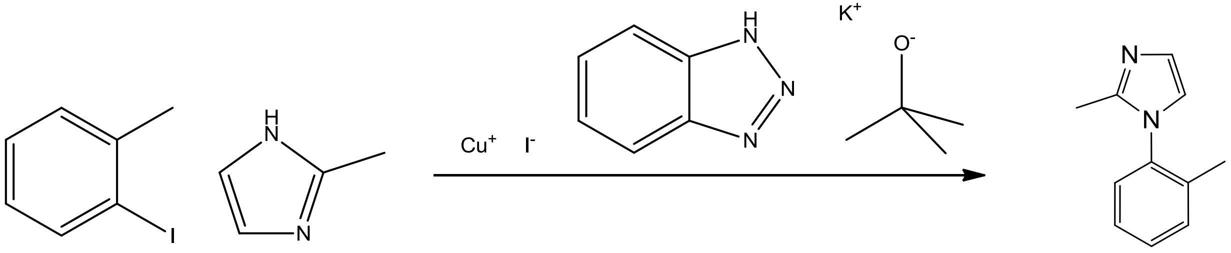


- EN01325-27

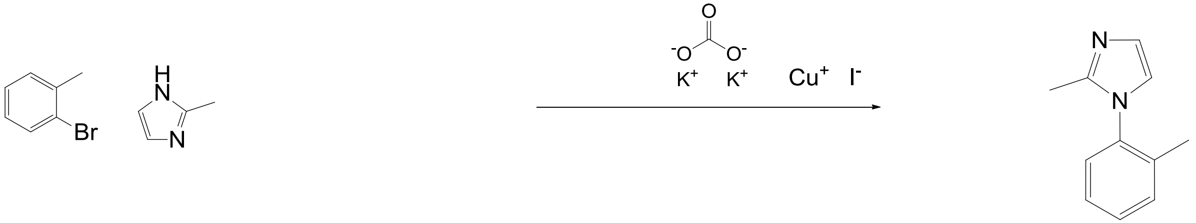


EXAMPLES WITH SAME GRANDPARENT

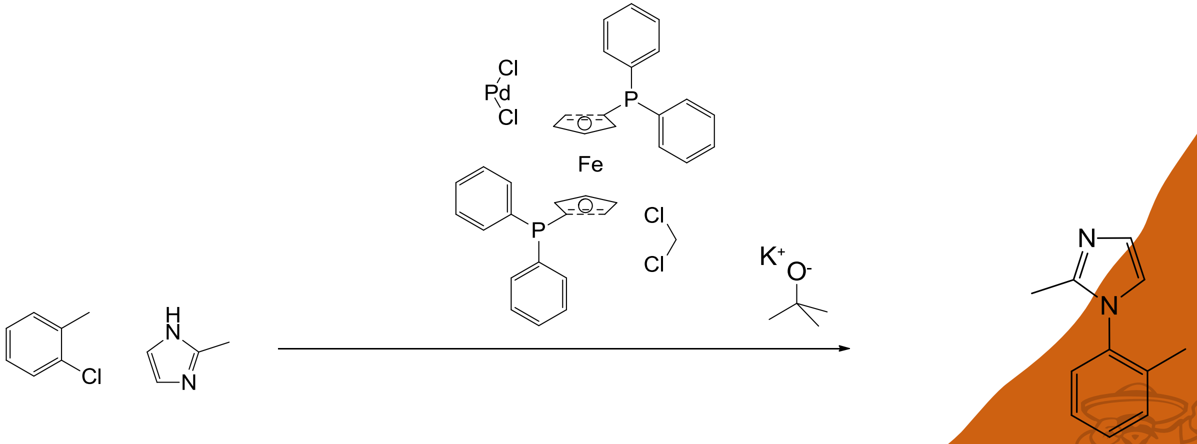
- EN00930-16



- EN00930-25



- EN00930-60



Latest Developments

- The very latest developments have been on alerting potential health & safety issues in ELNs.
  - Detecting potentially explosive and flammable energetic materials, and the scale on which they are prepared.
  - Identifying incompatible reagent combinations, such metal hydrides and acids, leading to hydrogen exotherm issues.
  - Other hazards, such as microwave-compatible solvents.
  - Heats of formation determine classify reactions as exo- or endo-thermic.
  - Heats of decomposition/combustion for hazards.
  - Stability prediction in DMSO at room temperature.