Automated mining a database of 9.4M reactions from the patent literature, and its application to synthesis planning

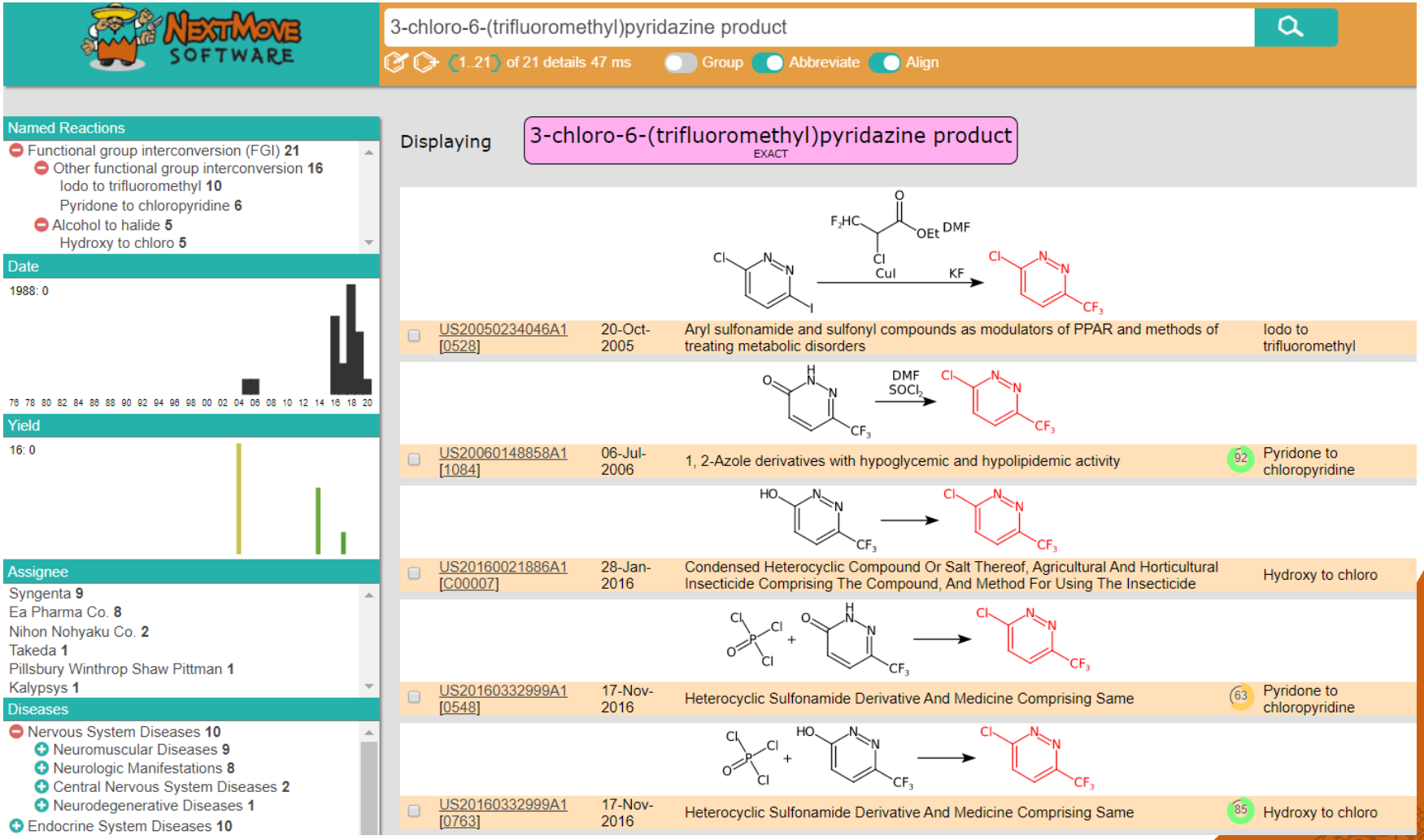
Roger Sayle, John Mayfield and Ingvar Lagerstedt

NextMove Software, Cambridge, UK

Daniel Lowe, *Minesoft, Cambridge, UK*

Strasbourg Summer School in Chemoinformatics 2020, Strasbourg, 28th June-2nd July 2020

PISTACHIO: A DATABASE OF 9.3M RXNS



TEXT MINING: US20160332999A1

US 20160332999A1

(19) **United States**

(12) **Patent Application Publication**

KOBAYASHI et al.

(10) **Pub. No.: US 2016/0332999 A1**

(43) **Pub. Date: Nov. 17, 2016**

(54) **HETEROCYCLIC SULFONAMIDE DERIVATIVE AND MEDICINE COMPRISING SAME**

(71) Applicant: **EA PHARMA CO., LTD.**, Tokyo (JP)

(72) Inventors: **Kaori KOBAYASHI**, Kawasaki-shi (JP); **Tamotsu SUZUKI**, Kawasaki-shi (JP); **Mizuki KAWAHIRA**, Kawasaki-shi (JP); **Tomohiro FUJII**, Kawasaki-shi (JP); **Masayuki SUGIKI**, Kawasaki-shi (JP); **Koji OHSUMI**, Kawasaki-shi (JP); **Tatsuya OKUZUMI**, Tokyo (JP)

(73) Assignee: **EA PHARMA CO., LTD.**, Tokyo (JP)

(21) Appl. No.: **15/222,178**

(22) Filed: **Jul. 28, 2016**

Related U.S. Application Data

(63) Continuation of application No. PCT/JP2015/052415, filed on Jan. 28, 2015.

(30) **Foreign Application Priority Data**

Jan. 28, 2014 (JP) 2014-013729
Aug. 6, 2014 (JP) 2014-160250

Publication Classification

(51) **Int. Cl.**

C07D 405/14	(2006.01)
C07D 405/12	(2006.01)
C07D 413/14	(2006.01)
C07D 407/12	(2006.01)
C07D 491/048	(2006.01)
C07D 498/04	(2006.01)
C07D 307/82	(2006.01)
C07D 409/14	(2006.01)

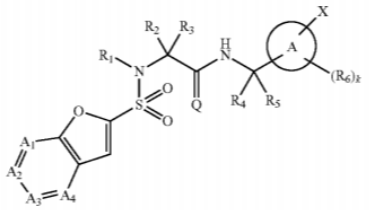
(52) **U.S. Cl.**

CPC ***C07D 405/14*** (2013.01); ***C07D 307/82*** (2013.01); ***C07D 405/12*** (2013.01); ***C07D 409/14*** (2013.01); ***C07D 407/12*** (2013.01); ***C07D 491/048*** (2013.01); ***C07D 498/04*** (2013.01); ***C07D 413/14*** (2013.01)

(57) **ABSTRACT**

The present invention provides a compound represented by the formula (I):

(I)



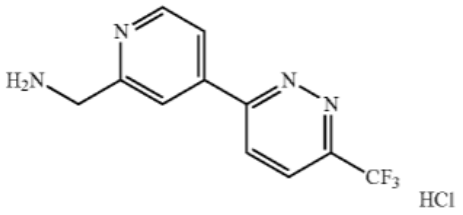
wherein each symbol is as defined in the DESCRIPTION, or a pharmaceutically acceptable salt thereof. The compound has a superior TRPA1 antagonist activity, and can provide a medicament useful for the prophylaxis or treatment of diseases involving TRPA1 antagonist and TRPA1.

TEXT MINING: US20160332999A1

Reference Example D-35

Synthesis of [4-[6-(trifluoromethyl)pyridazin-3-yl]-2-pyridyl]methylamine hydrochloride (D-35)

[0547]



(step 1) Synthesis of 3-chloro-6-(trifluoromethyl)pyridazine

[0548] To 3-(trifluoromethyl)-1H-pyridazin-6-one (1.1 g, 6.7 mmol) was added phosphorus oxychloride (10 mL) and the mixture was stirred at 100° C. for 2.5 hr, and concentrated under reduced pressure. To the obtained residue were added dichloromethane and water, and the mixture was stirred at room temperature for 5 min. The mixture was alkalified by adding potassium carbonate to partition the mixture. The organic layer was washed with saturated brine, dried over sodium sulfate, and the desiccant was filtered off, and the solvent was evaporated and the obtained residue was purified by silica gel column chromatography (petroleum ether/ethyl acetate) to give the title compound (0.77 g, 4.2 mmol, 63%).

[0549] MS (ESI) m/z 182 $(M+H)^+$

[0550] ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.82 (d, J=8.8 Hz, 1H), 7.74 (d, J=8.8 Hz, 1H).

(step 2) Synthesis of [4-[6-(trifluoromethyl) pyridazin-3-yl]-2-pyridyl]methylamine hydrochloride (D-35)

[0551] Using the compound obtained in step 1 instead of 5-bromo-2-(trifluoromethyl)pyrimidine, and by an operation similar to that in Reference Example D-32, step 2, the title compound was obtained (yield 54%).

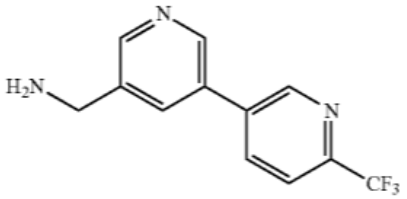
[0552] MS (ESI) m/z 255 $(M+H)^+$

[0553] ^{1}H NMR (400 MHz, CD_3OD) δ 8.92 (d, J=5.3 Hz, 1H), 8.59 (d, J=9.0 Hz, 1H), 8.37 (br s, 1H), 8.33 (d, J=9.0 Hz, 1H), 8.24 (dd, J=5.3, 1.4 Hz, 1H), 4.49 (s, 2H).

Reference Example D-36

Synthesis of [5-[6-(trifluoromethyl)-3-pyridyl]-3-pyridyl]methylamine (D-36)

[0554]



[0555] Using 5-bromo-2-trifluoromethylpyridine and (5-cyano-3-pyridyl)boronic acid instead of 5-bromo-2-trif-

TEXT MINING: US20160332999A1

US20160332999A1 Go

e.g. US6356863 or US2007129372

[+] Show advanced options

Navigation | Chemical Data

Abstract | Description | Claims

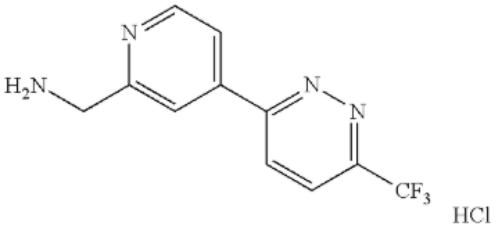
Table navigation: ← →

Entity type	Count (unique)	
Physquant	3919 (1206)	← →
Mol	3210 (1276)	← →
Reference	946 (354)	← →
Generic	645 (105)	← →
Prefix	546 (237)	← →
NMR	542 (542)	← →
MechanismOfAction	522 (34)	← →
Formula	449 (28)	← →
DictMol	369 (27)	← →
SolventMixture	222 (28)	← →
MassSpec	221 (141)	← →
DiseaseGrammar	195 (55)	← →
Exp. Procedure	195 (7)	← →
GeneOrProtein	132 (15)	← →
SolventMixtureNew	127 (8)	← →
Markush	122 (33)	← →

Reference Example D-35

Synthesis of [4-[6-(trifluoromethyl)pyridazin-3-yl]-2-pyridyl]methylamine hydrochloride (D-35)

[0547]



(step 1) Synthesis of 3-chloro-6-(trifluoromethyl)pyridazine

[0548] To 3-(trifluoromethyl)-1H-pyridazin-6-one (1.1 g, 6.7 mmol) was added phosphorus oxychloride (10 mL) and the mixture was stirred at 100° C. for 2.5 hr, and concentrated under reduced pressure. To the obtained residue were added dichloromethane and water, and the mixture was stirred at room temperature for 5 min. The mixture was alkalified by adding potassium carbonate to partition the mixture. The organic layer was washed with saturated brine, dried over sodium sulfate, and the desiccant was filtered off, and the solvent was evaporated and the obtained residue was purified by silica gel column chromatography (petroleum ether/ethyl acetate) to give the title compound (0.77 g, 4.2 mmol, 63%).

[0549] MS (ESI) m/z 182 $(M+H)^+$

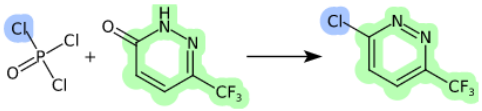
[0550] ^1H NMR (400 MHz, $CDCl_3$) δ 7.82 (d, J=8.8 Hz, 1H), 7.74 (d, J=8.8 Hz, 1H).

TEXT MINING: US20160332999A1

Pistachio Reaction Detail

Atom Mapping: Color None Number Abbreviate: Align:

NMSR:3216924



Pyridone to chloropyridine (9.7.112)

Name	Role	Formula	MW	Amount	Mass	Volume	Density	Yield
3-chloro-6-(trifluoromethyl)pyridazine	Product	$C_5H_2ClF_3N_2$	182.531 g/mol	4.2 mmol	770 mg			63 %
3-(trifluoromethyl)-1H-pyridazin-6-one	Reactant	$C_5H_3F_3N_2O$	164.086 g/mol	6.7 mmol	1.1 g			
phosphorus oxychloride	Solvent	Cl_3OP	153.332 g/mol			10 mL		

Info

Source	US20160332999A1 [0548]
Document	Kaori Kobayashi, Tamotsu Suzuki, Mizuki Kawahira, Tomohiro Fujii, Masayuki Sugiki, Koji Ohsumi, Tatsuya Okuzumi **Heterocyclic Sulfonamide Derivative And Medicine Comprising Same** U.S. Application (17-Nov-2016)
Affiliation	Ea Pharma Co.
IPC Codes	C07D 307/82, C07D 405/12, C07D 405/14, C07D 407/12, C07D 409/14, C07D 413/14, C07D 491/048, C07D 498/04
Diseases	Acute Pain, Asthma, Atopic Dermatitis, Chronic Obstructive Pulmonary Disease, Chronic Pain, Cough, Diabetic Neuropathies, Inflammatory Bowel Diseases, Irritable Bowel Syndrome, Osteoarthritis, Pancreatitis, Peptic Esophagitis, Pruritus
Targets	Transient receptor potential cation channel subfamily A member 1

Procedure

(step 1) Synthesis of 3-chloro-6-(trifluoromethyl)pyridazine

To 3-(trifluoromethyl)-1H-pyridazin-6-one (1.1 g, 6.7 mmol) was added phosphorus oxychloride (10 mL) and the mixture was stirred at 100° C. for 2.5 hr, and concentrated under reduced pressure. To the obtained residue were added dichloromethane and water, and the mixture was stirred at room temperature for 5 min. The mixture was alkalified by adding potassium carbonate to partition the mixture. The organic layer was washed with saturated brine, dried over sodium sulfate, and the desiccant was filtered off, and the solvent was evaporated and the obtained residue was purified by silica gel column chromatography (petroleum ether/ethyl acetate) to give the title compound (0.77 g, 4.2 mmol, 63%).

Identifiers

SMILES	O=[c:1]1[cH:2][cH:3][c:4]([n:5][nH:6]1)[C:7]([F:8])([F:9])[F:10].O=P(Cl)(Cl)[Cl:11]>>[F:8][C:7]([F:9])([F:10])[c:4]1[cH:3][cH:2][c:1]([n:6][n:5]1)[Cl:11]
RInChI	RInChI=1.00.1S/C5H2ClF3N2/c6-4-2-1-3(10-11-4)5(7,8)9/h1-2H<>C5H3F3N2O/c6-5(7,8)3-1-2-4(11)10-9-3/h1-2H,(H,10,11)<>Cl3OP/c1-5(2,3)4/d-

TEXT MINING: US20160332999A1

{"data":{"paragraphText":"To 3-(trifluoromethyl)-1H-pyridazin-6-one (1.1 g, 6.7 mmol) was added phosphorus oxychloride (10 mL) and the mixture was stirred at 100° C. for 2.5 hr, and concentrated under reduced pressure. To the obtained residue were added dichloromethane and water, and the mixture was stirred at room temperature for 5 min. The mixture was alkalified by adding potassium carbonate to partition the mixture. The organic layer was washed with saturated brine, dried over sodium sulfate, and the desiccant was filtered off, and the solvent was evaporated and the obtained residue was purified by silica gel column chromatography (petroleum ether/ethyl acetate) to give the title compound (0.77 g, 4.2 mmol, 63%).","headingText":"(step 1) Synthesis of 3-chloro-6-(trifluoromethyl)pyridazine", "documentId":"US20160332999A1","title":"Heterocyclic Sulfonamide Derivative And Medicine Comprising Same", "assignees":["Ea Pharma Co."],"authors":["Kaori Kobayashi","Tamotsu Suzuki","Mizuki Kawahira","Tomohiro Fujii","Masayuki Sugiki","Koji Ohsumi","Tatsuya Okuzumi"], "source":"USPTO_TEXT_A","reactionSmiles":"O=[c:1]1[cH:2][cH:3][c:4]([n:5][nH:6]1)[C:7]([F:8])([F:9])[F:10].O=P(Cl)(Cl)[Cl:11]>>[F:8][C:7]([F:9])([F:10])[c:4]1[cH:3][cH:2][c:1]([n:6][n:5]1)[Cl:11]","smiles":"O=[c:1]1[cH:2][cH:3][c:4]([n:5][nH:6]1)[C:7]([F:8])([F:9])[F:10].O=P(Cl)(Cl)[Cl:11]>>[F:8][C:7]([F:9])([F:10])[c:4]1[cH:3][cH:2][c:1]([n:6][n:5]1)[Cl:11]","ipcCodes":["C07D 307/82","C07D 405/12","C07D 405/14","C07D 407/12","C07D 409/14","C07D 413/14","C07D 491/048","C07D 498/04"],"date":"17-Nov-2016", "namerxndef":"Pyridone to chloropyridine (9.7.112)","namerxn":"9.7.112", "diseases":["Acute Pain","Asthma","Atopic Dermatitis","Chronic Obstructive Pulmonary Disease","Chronic Pain","Cough","Diabetic Neuropathies","Inflammatory Bowel Diseases","Irritable Bowel Syndrome","Osteoarthritis","Pancreatitis","Peptic Esophagitis","Pruritus"], "targets":["Transient receptor potential cation channel subfamily A member 1"], "location":"548", "rinchi":"RInChI=1.00.1S/C5H2ClF3N2/c6-4-2-1-3(10-11-4)5(7,8)9/h1-2H<>C5H3F3N2O/c6-5(7,8)3-1-2-4(11)10-9-3/h1-2H,(H,10,11)<>Cl3OP/c1-5(2,3)4/d-"}, "components":[{"role":"Product","name":"3-chloro-6-(trifluoromethyl)pyridazine", "smiles":"ClC=1N=NC(=CC1)C(F)(F)F", "quantities":[{"type":"Amount","value":0.0042,"text":"4.2 mmol"}, {"type":"Mass", "value":0.77,"text":"0.77 g"},{"type":"Yield","value":63.0,"text":"63%"}], "molWeight":182.5311146438476, "molFormula":"C₅H₂ClF₃N₂"},{"role":"Reactant","name":"3-(trifluoromethyl)-1H-pyridazin-6-one","smiles":"FC(C1=NNC(C=C1)=O)(F)F", "quantities": [{"type":"Amount", "value":0.0067, "text":"6.7 mmol"},{"type":"Mass","value":1.1,"text":"1.1 g"}], "molWeight":164.08552274705877,"molFormula":"C₅H₃F₃N₂O"}, {"role":"Solvent","name":"phosphorus oxychloride", "smiles":"O=P(Cl)(Cl)Cl", "quantities": [{"type":"Volume","value":0.01,"text":"10 mL"}], "molWeight":153.3319792921214, "molFormula":"Cl₃OP"}]}

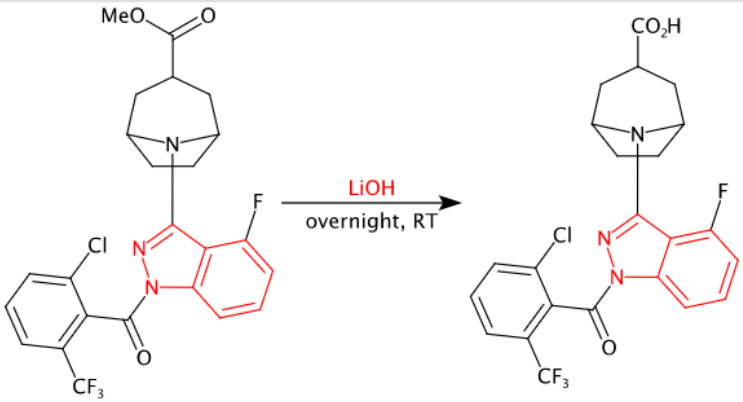
INTRODUCTION SUMMARY

- NextMove Software use text mining to automatically extract a database of chemical reactions from the patent literature.

PISTACHIO: ALEXA/SIRI FOR CHEMISTS

30-40% yield Merck LiOH deprotections in the last six months indazole substructure

Displaying 30-40% yield (YIELD) Merck (ASSIGNEE) LiOH (EXACT) deprotections (REACTION_TYPE) in the last six months (BEG_DATE) indazole substructure (SUBSTRUCTURE)



US09663522B2 [0316] 30-May-2017 3-Aminocycloalkyl Compounds As Rorgammat Inhibitors And Uses Thereof 35 CO2H-Me deprotection

Automated Extraction of Reactions from the Patent Literature

Daniel Lowe
Unilever Centre for Molecular Science Informatics
University of Cambridge

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

NEXTMOVE'S FREE "USPTO" DATA SET

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

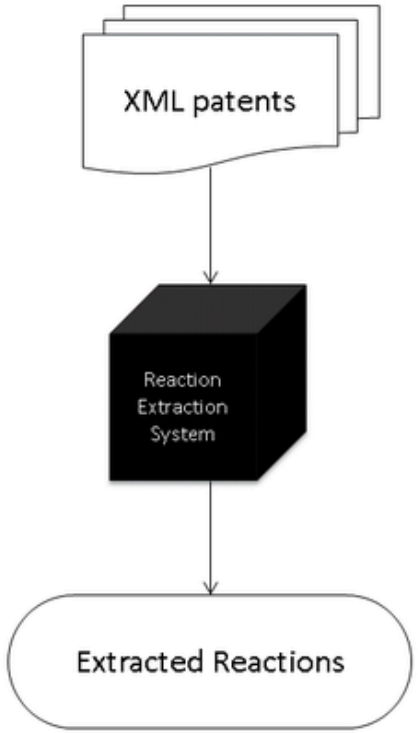
Unleashing over a million reactions into the wild

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.

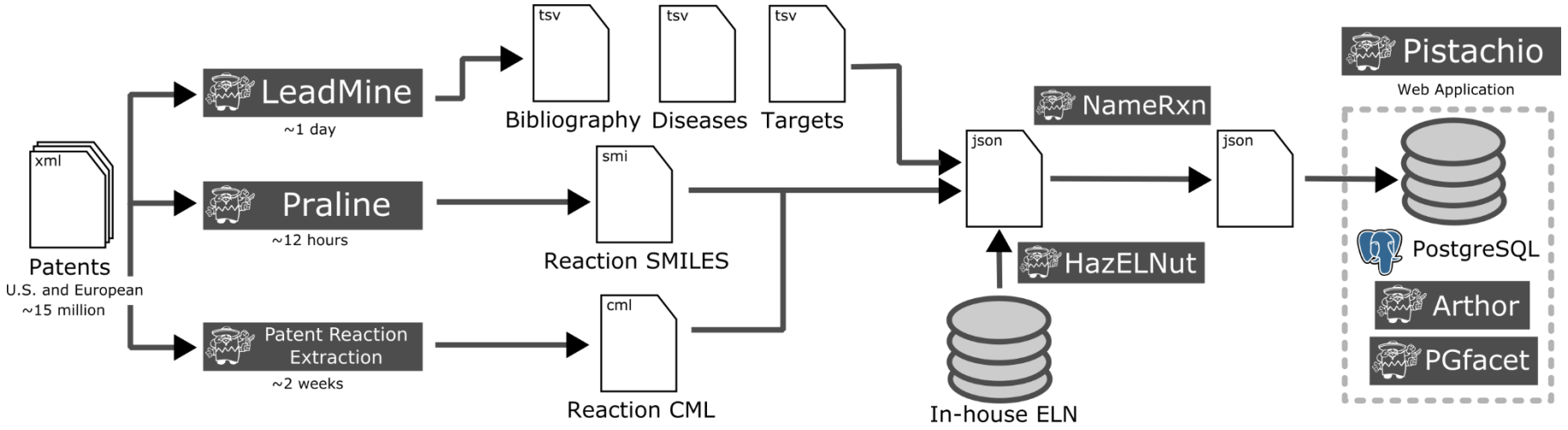
The reactions are available as reaction SMILES or CML from here, as 7zip archives. The CML representation includes quantities and yields where these were found. A documentation zip provides further information on the format of the data. This data is made available under CC-Zero i.e. without copyright.



REACTION DATABASE LANDSCAPE

- Manually curated databases
  - Elsevier Reaxys
  - CAS SciFinder
  - InfoChem SPRESI
- Machine curated databases
  - USPTO
  - Pistachio
- Evaluation Metrics
  - Availability/Price
  - Coverage/Size/Frequency
  - Quality/Annotation

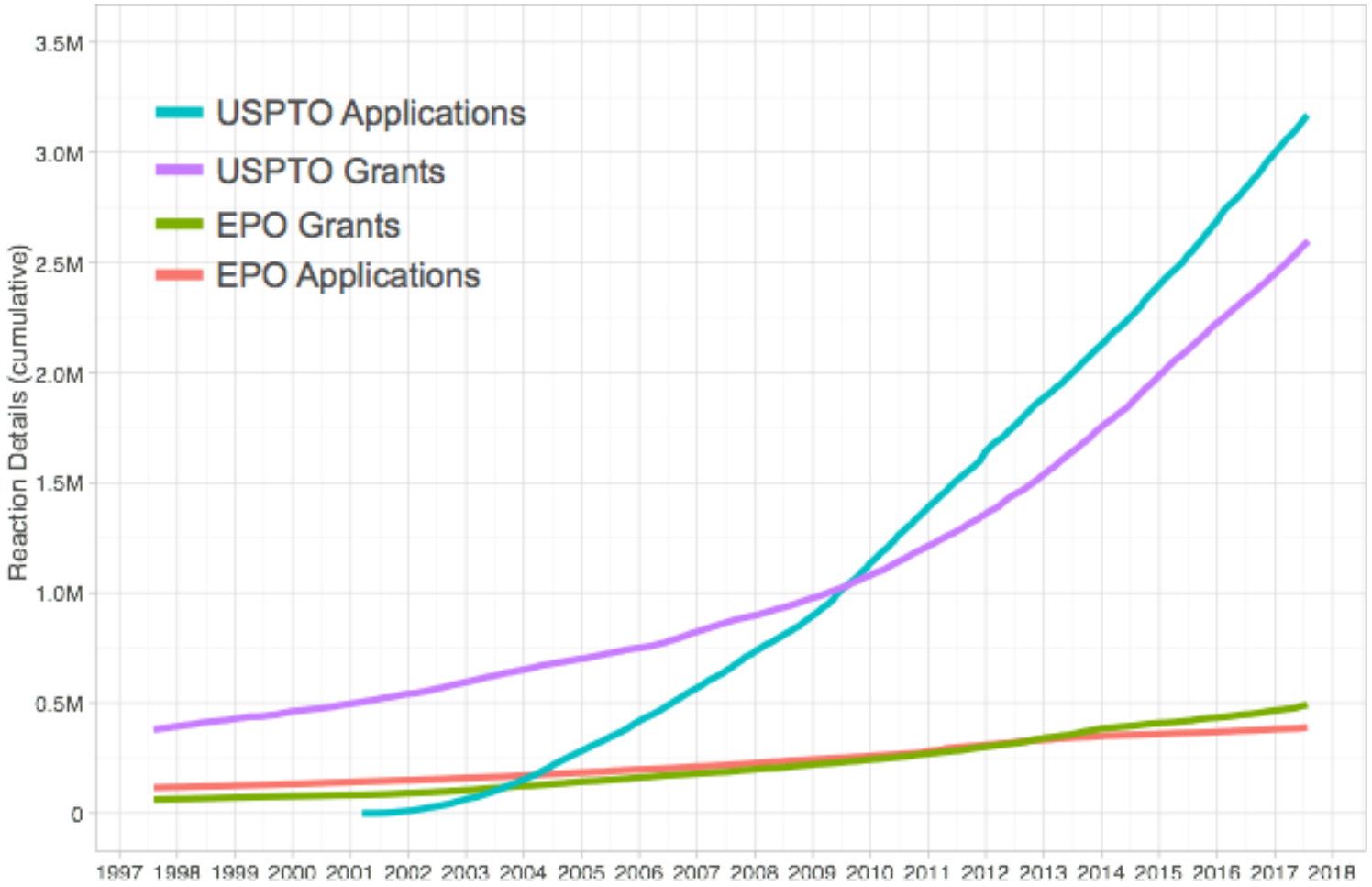
PISTACHIO WORKFLOW



CURRENT PISTACHIO STATISTICS

• U.S. Application Text	2,902,949	2020-02-27
• U.S. Grant Text	2,662,420	2020-02-25
• EPO Application Text	553,143	2020-02-19
• EPO Grant Text	866,530	2020-02-19
• U.S. Application Sketches	1,419,737	2020-02-27
• U.S. Grant Sketches	915,226	2020-02-25
• Total	9,320,005	
• Unique (parents)	2,945,919	

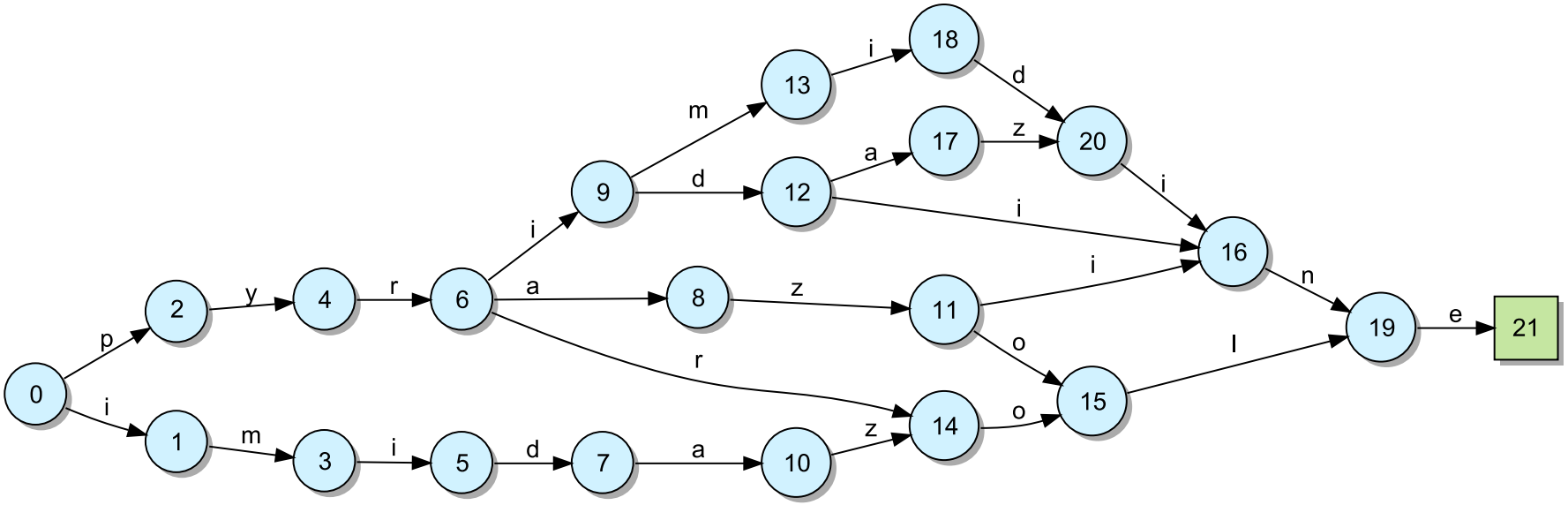
TOTAL REACTIONS OVER TIME



5.3TB SOURCE "BIG DATA"

- /patents/applications
  - 6,085,900 2001-2020 2.2TB
- /patents/grants
  - 7,305,481 1976-2020 1.5TB
- /patents/ep
  - 1,871,242 2013-2020 1.1TB [1978-]
- /patents/epo
  - 3,516,326 1978-2013 492GB €20K

EXAMPLE ENTITY DICTIONARY AS DAG



- Nitrogen containing heterocycles as minimal DFA:
  - Pyrrole, Pyrazole, Imidazole, Pyrdine, Pyridazine, Pyrimidine, Pyrazine
- CaffeineFix supports (very large) user dictionaries.

SPELLING CORRECTION

di-terf-butyl (4S)-/V-(fert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

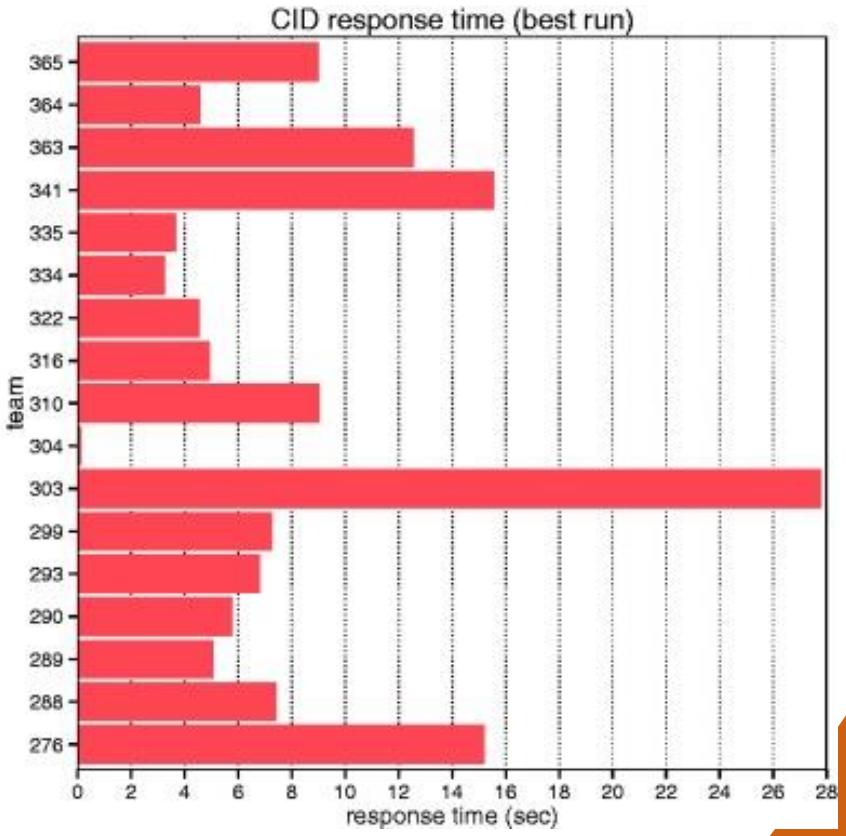
CaffeineFix corrected to:

di-tert-butyl (4S)-N-(tert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate

RULE-BASE TEXT-MINING SPEED

BioCreAtIvE V challenge evaluating text-mining and extraction systems.

Web service **response time** to annotate an **abstract** evaluated for CDR task.



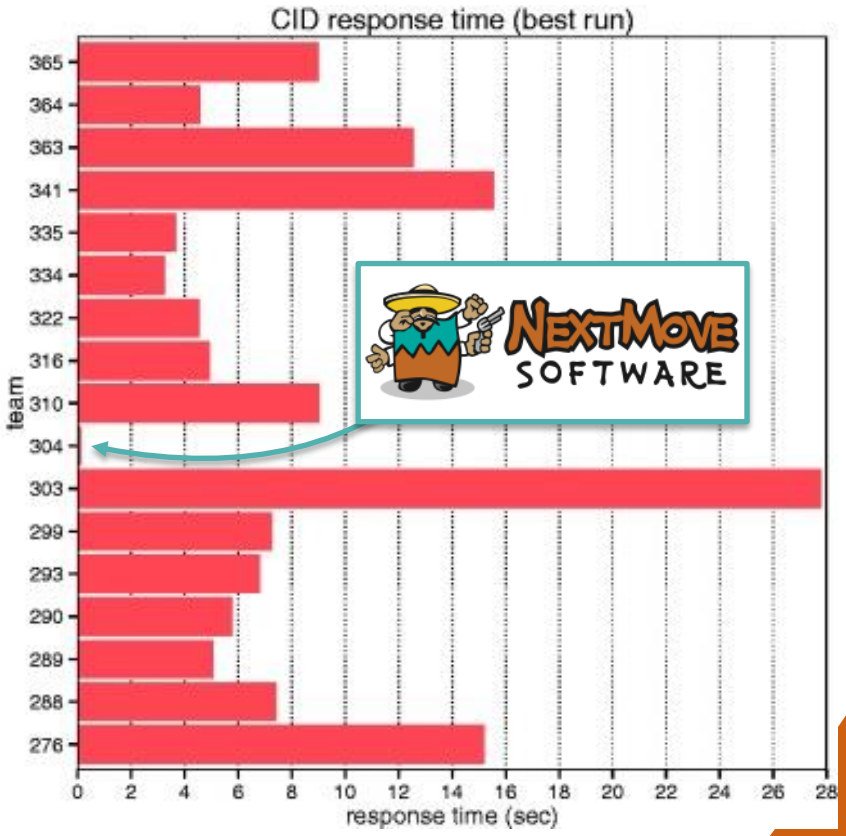
Chih-Hsuan Wei et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford). 2016; 2016: baw032. PMC4799720

RULE-BASE TEXT-MINING SPEED

BioCreAtIvE V challenge evaluating text-mining and extraction systems.

Web service **response time** to annotate an **abstract** evaluated for CDR task.

Efficient rule-based text-mining provides provenance for annotations and can mine entire back-archive of US patents in **~24 hours** on a single machine.



Chih-Hsuan Wei et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford). 2016; 2016: baw032. PMC4799720

ADVANCED ENTITY RECOGNITION

- Name2Structure improvements.
- Dictionaries and ontologies.
- Molecular formulae and line formulae
  - K_2CO_3, $PdCl_2(PPh_3)_2$, $Pd(P(o\text{-}tolyl)_3)_2$, $PdCl_2(dppf)\text{-}CH_2Cl_2$
- Inorganics, organometallics and salts.
  - vanadium oxychloride, cupric chloride.
- Mixtures/Formulations.
  - 5% 2M methanolic ammonia/DCM, 10% H2O2 in water
- Apparatus.
  - A 1 L three necked round bottomed flask

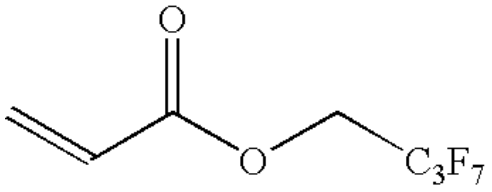
CHEMDRAW SKETCH PROCESSING

Re-interpretation of ChemDraw sketches

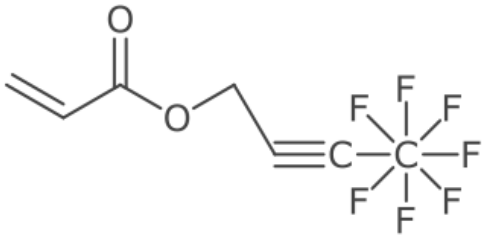
1. Correct systematic errors
2. Extract extra semantics (structure variation, reaction schemes)
3. Categorise output (is this something we can't interpret)

US 2004/101442 C00025

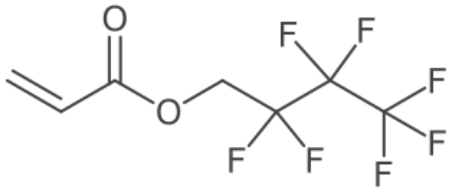
Original Sketch



Default Interpretation (USPTO molfile)

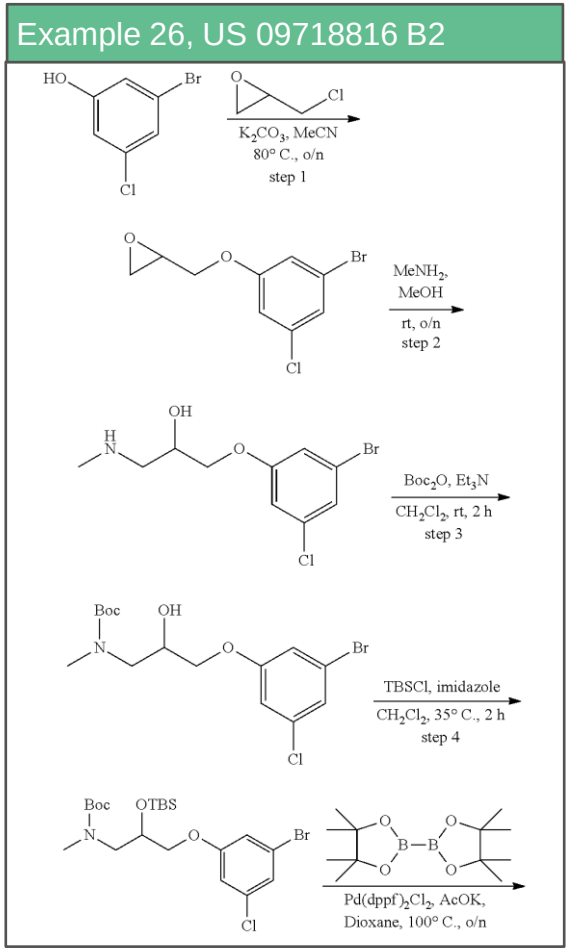


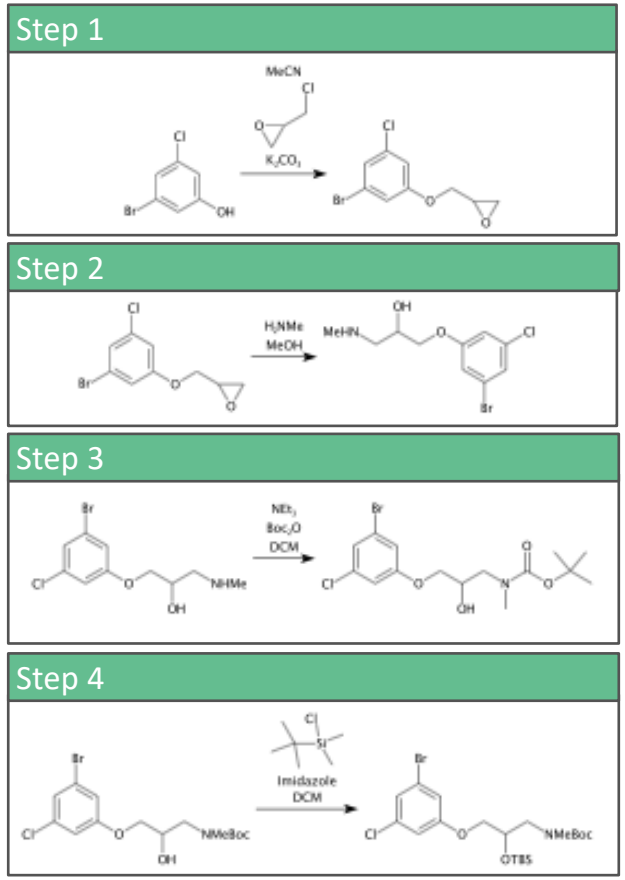
Our Interpretation



John May, *et al.* Sketchy Sketches: Hiding Chemistry in Plain Sight. Seventh Joint Sheffield Conference on Cheminformatics. 2016

REACTION SCHEME SKETCHES





etc..

John May, *et al.* Sketchy Sketches: Hiding Chemistry in Plain Sight. Seventh Joint Sheffield Conference on Cheminformatics. 2016

RESOLVING IDENTIFIERS

- Need to “name space” identifiers
  - “Compound 1”, “Reference compound 1”, “Example 1”
  - But “Compound 1” = “cmpd 1” = “cpd. #1”
- Identifier may be defined multiple times e.g. as a sketch and chemical name

RESOLVING IDENTIFIERS
(TEXT-MINING)

EXAMPLE 3

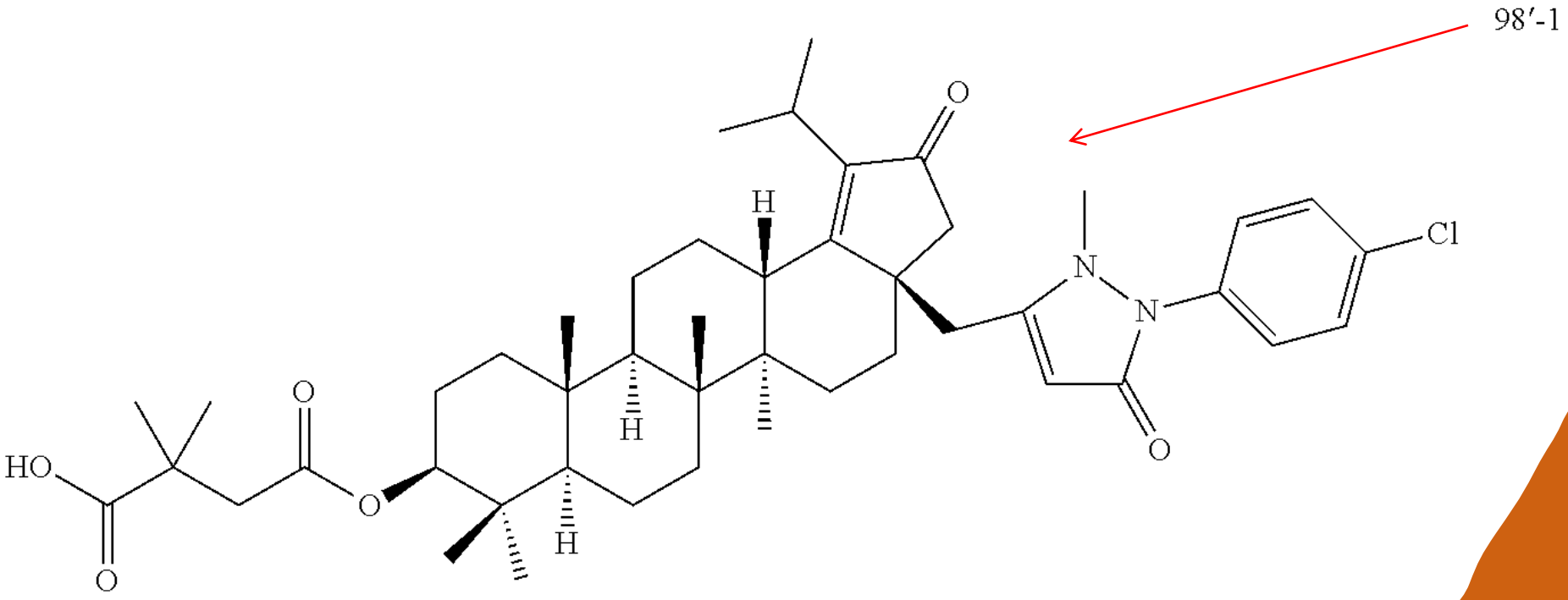
1,2-Dihydro-2-(4-trifluoromethylphenyl)-5-(3-trifluoromethyl-2-pyridinyl)-3H-indazol-3-one

This compound was prepared as described for Example 1 replacing 2-azido-4-bromobenzoic acid with 2-azido-5-bromobenzoic acid in Step 2. ^{1}H NMR (360 MHz, DMSO): 7.52 (1H, d, J8.0 Hz), 7.65-7.75 (1H, m), 7.76 (1H, dd, J 1.2 and 8.0 Hz), 7.83 (1H, s), 7.91 (2H, d, J 8.0 Hz), 8.18 (2H, d, J 8.0 Hz), 8.34 (1H, d, J8.0 Hz), 8.9 (1H, d, J 8.0 Hz), 11.1 (1H, s).

EXAMPLE 4

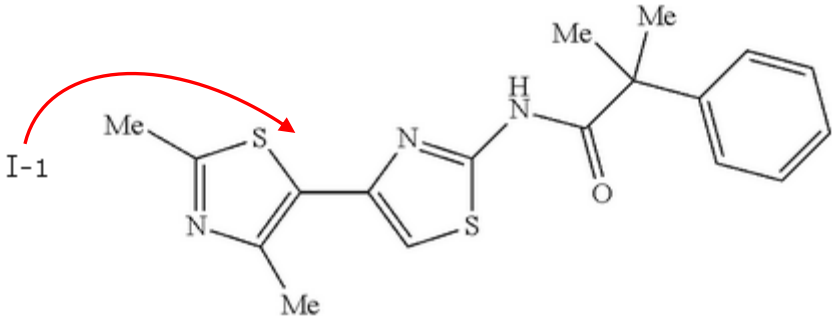
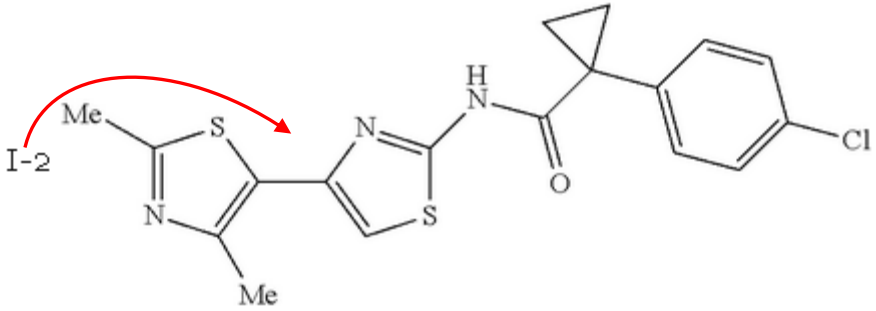
1,2-Dihydro-6-(2-methoxyphenyl)-2-(4-trifluoromethylphenyl)-3H-indazol-3-one

RESOLVING IDENTIFIERS
(SKETCHES)



RESOLVING IDENTIFIERS
(TABLES)

TABLE 1

No.	Chamical Structure
I-1	
I-2	

R GROUP TABLES

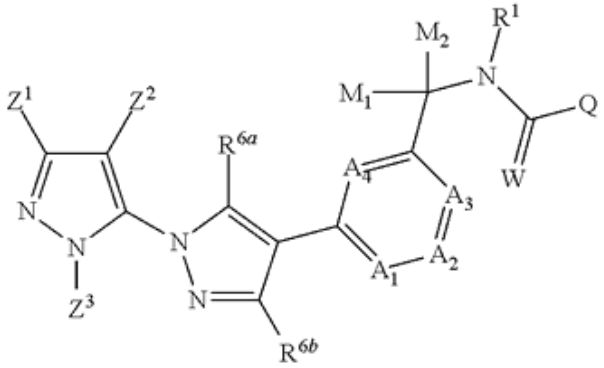
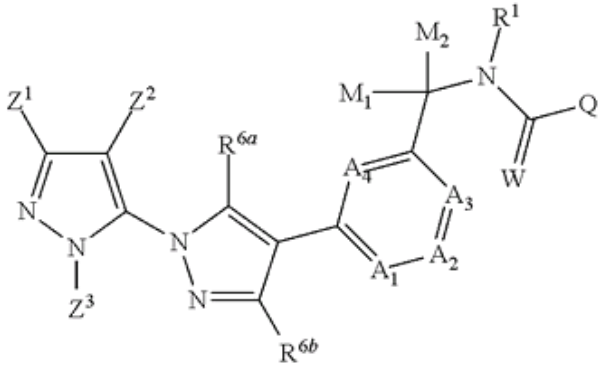


TABLE 2

Ex. No.	Z^1	Z^2	Z^3	R^1	R^{6a}	R^{6b}	A_4	A_3	A_2	A_1	M_1	M_2	W	Q	logP a)	Mass a)
Ic-1	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	C—H	C—H	H	H	O	ethyl	3.59	496.2
Ic-2	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	ethyl	3.72	514.1
Ic-3	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	1,3-thiazol-yl	3.93	569.0
Ic-4	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	CH_3	3.40	500.1
Ic-5	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	2,2,2-trifluoroethyl	3.93	568.0
Ic-6	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	trifluoromethyl	4.24	554.0
Ic-7	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	thiophen-2-ylmethyl	4.12	582.0
Ic-8	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	N	C—H	H	H	O	CH_3	2.35	483.0

US 2016/0002208 A1

R GROUP TABLES



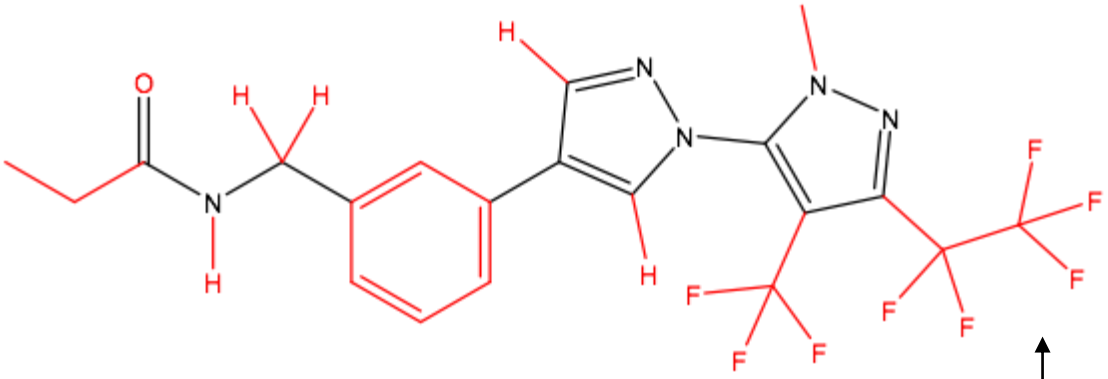


TABLE 2

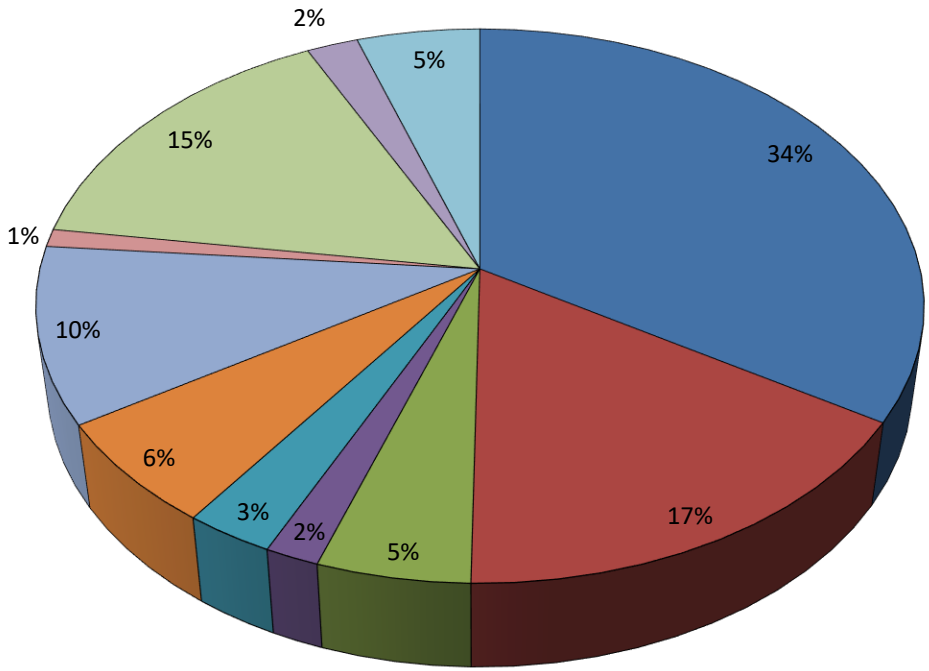
Ex. No.	Z^1	Z^2	Z^3	R^1	R^{6a}	R^{6b}	A_4	A_3	A_2	A_1	M_1	M_2	W	Q	logP a)	Mass a)
Ic-1	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	C—H	C—H	H	H	O	ethyl	3.59	496.2
Ic-2	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	ethyl	3.72	514.1
Ic-3	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	1,3-thiazol-yl	3.93	569.0
Ic-4	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	CH_3	3.40	500.1
Ic-5	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	2,2,2-trifluoroethyl	3.93	568.0
Ic-6	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	trifluoromethyl	4.24	554.0
Ic-7	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	thiophen-2-ylmethyl	4.12	582.0
Ic-8	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	N	C—H	H	H	O	CH_3	2.35	483.0

US 2016/0002208 A1

ADDITIONAL ANNOTATION

- Company Ontology
  - Ciba-Giegy Corp. = Ciba-Giegy → Novartis
- Calculated Yields
  - Density of $POCl_3$ is 1.64 g/cm^3.
- Reaction Steps/Recipes (ISA-88)
  - Add, Synthesize, Wait, Degass, Yield, Wash, Reflux, Irradiate, Stir, Extract, Precipitate, Mill, Remove, Filter, Partition, Sample, Heat, Concentrate, Dry, Quench, Cool, Transfer, Purify, Dissolve…

CATEGORIZATION OF REACTIONS



- Heteroatom alkylation and arylation
- Acylation and related processes
- C-C bond formations
- Heterocycle formation
- Protections
- Deprotections
- Reductions
- Oxidations
- Functional group conversion
- Functional group addition
- Resolution

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.

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 ~1150 named reactions/types.
  - e.g. 3.5.3 Negishi coupling (RXNO:0000088)
- These require ~2490 SMIRKS-like transformations.

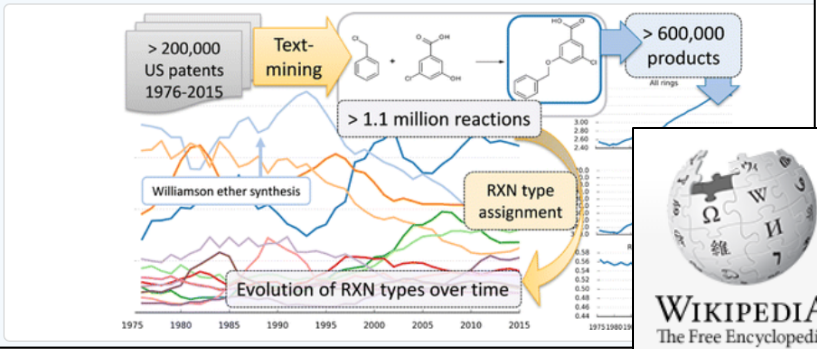
Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter

Nadine Schneider*†, Daniel M. Lowe§, Roger A. Sayle§, Michael A. Tarselli‡, and Gregory A. Landrum†
† Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland
‡ Novartis Institutes for BioMedical Research, 186 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States
§ NextMove Software Ltd., Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, U.K.

J. Med. Chem., **2016**, *59* (9), pp 4385–4402
DOI: 10.1021/acs.jmedchem.6b00153
Publication Date (Web): March 30, 2016

*E-mail: nadine-1.schneider@novartis.com.
This article is part of the Computational Methods for Medicinal Chemistry special issue.

Abstract



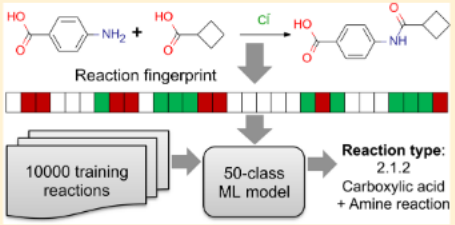
JOURNAL OF CHEMICAL INFORMATION AND MODELING
Article
pubs.acs.org/jcim

Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity

Nadine Schneider,† Daniel M. Lowe,‡ Roger A. Sayle,† and Gregory A. Landrum*,†
†Novartis Institutes for BioMedical Research, Novartis Campus, 4002 Basel, Switzerland
‡NextMove Software, Ltd., Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, United Kingdom

Supporting Information

ABSTRACT: Fingerprint methods applied to molecules have proven to be useful for similarity determination and as inputs to machine-learning models. Here, we present the development of a new fingerprint for chemical reactions and validate its usefulness in building machine-learning models and in similarity assessment. Our final fingerprint is constructed as the difference of the atom-pair fingerprints of products and reactants and includes agents via calculated physicochemical properties. We validated the fingerprints on a large data set of reactions text-mined from granted United States patents from the last 40 years that have been classified using a substructure-based expert system. We applied machine learning to build a 50-class predictive model for reaction-type classification that correctly predicts 97% of the reactions in an external test set. Impressive accuracies were also observed when applying the classifier to reactions from an in-house electronic laboratory notebook. The performance of the novel fingerprint for assessing reaction similarity was evaluated by a cluster analysis that recovered 48 out of 50 of the reaction classes with a median F-score of 0.63 for the clusters. The data sets used for training and primary validation as well as all python scripts required to reproduce the analysis are provided in the Supporting Information.



BiomolecularGraphics4All Talk Sandbox Preferences Beta Watchlist Contributions Log out

Article Talk Read Edit View history More Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Negishi coupling

From Wikipedia, the free encyclopedia

The **Negishi coupling** is a widely employed transition metal catalyzed cross-coupling reaction. The reaction couples organic halides or triflates with organozinc compounds, forming carbon-carbon bonds (c-c) in the process. A palladium (0) species is generally utilized as the metal catalyst, though nickel is sometimes used:[1][2]

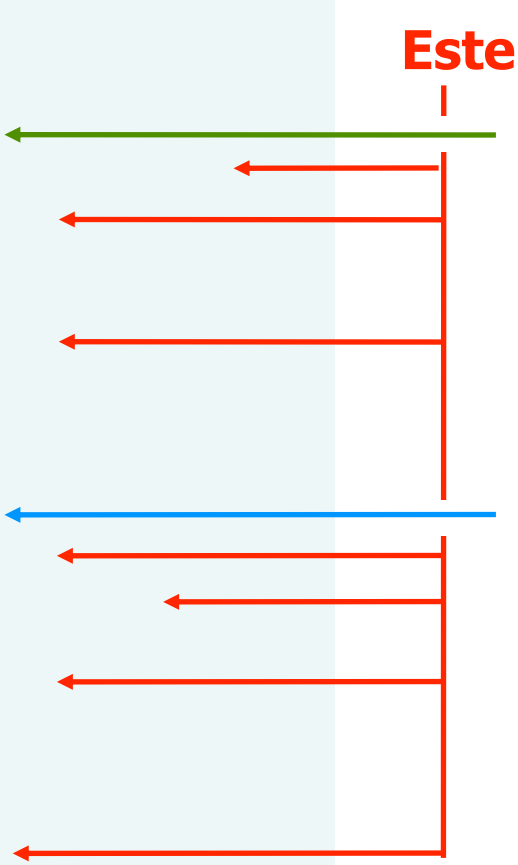
Negishi coupling	
Named after	Ei-ichi Negishi
Reaction type	Coupling reaction
Identifiers	
Organic Chemistry Portal	negishi-coupling
RSC ontology ID	RXNO:0000088

CONCEPTS AND RXNO

1 Heteroatom alkylation and arylation
.7 O-substitution
.1 Chan-Lam ether coupling
.2 Diazomethane esterification
.3 Ethyl esterification
.4 Hydroxy to methoxy
.5 Hydroxy to triflyloxy
.6 Methyl esterification
.n
2 Acylation and related processes
.6 O-acylation to ester
.1 Ester Schotten-Baumann
.2 Esterification (generic)
.3 Fischer-Speier esterification
.4 Baeyer-Villiger oxidation
.5 Yamaguchi esterification
.6 Hydroxy to imidazolecarbonyloxy
.7 Imidazolecarbonyl to ester
.8 Hydroxy to acetoxy
.9 Steglich esterification
.n

CONCEPTS AND RXNO

1 Heteroatom alkylation and arylation
.7 O-substitution
.1 Chan-Lam ether coupling
.2 Diazomethane esterification
.3 Ethyl esterification
.4 Hydroxy to methoxy
.5 Hydroxy to triflyloxy
.6 Methyl esterification
.n
2 Acylation and related processes
.6 O-acylation to ester
.1 Ester Schotten-Baumann
.2 Esterification (generic)
.3 Fischer-Speier esterification
.4 Baeyer-Villiger oxidation
.5 Yamaguchi esterification
.6 Hydroxy to imidazolecarbonyloxy
.7 Imidazolecarbonyl to ester
.8 Hydroxy to acetoxy
.9 Steglich esterification
.n



Esterification (7)

Chan-Lam coupling (3)

Schotten-Baumann Reaction (9)

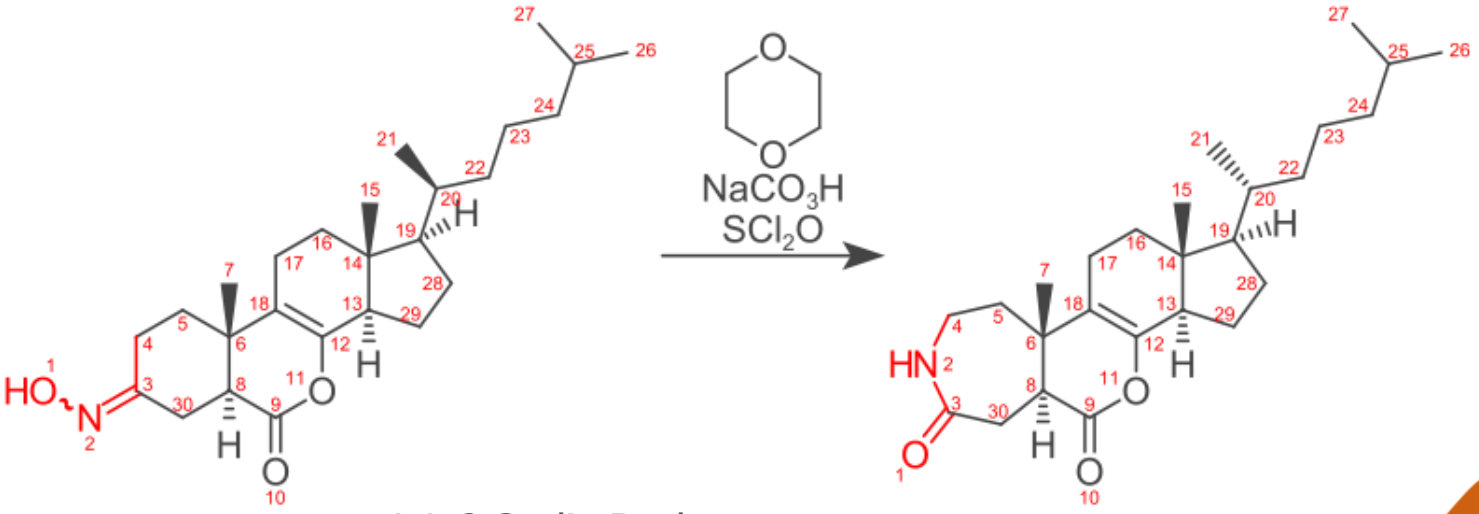
RXNO: http://github.com/rsc-ontologies/rxno

NAMERXN

Assigns reactions to **1150+** reaction categories using transformations

Can guarantee **perfect** Atom-Atom Mapping

- Atom-Atom Mapping is an output **not** an input
- MCS mappers struggle with rearrangements:



4.1.6 Cyclic Beckmann rearrangement

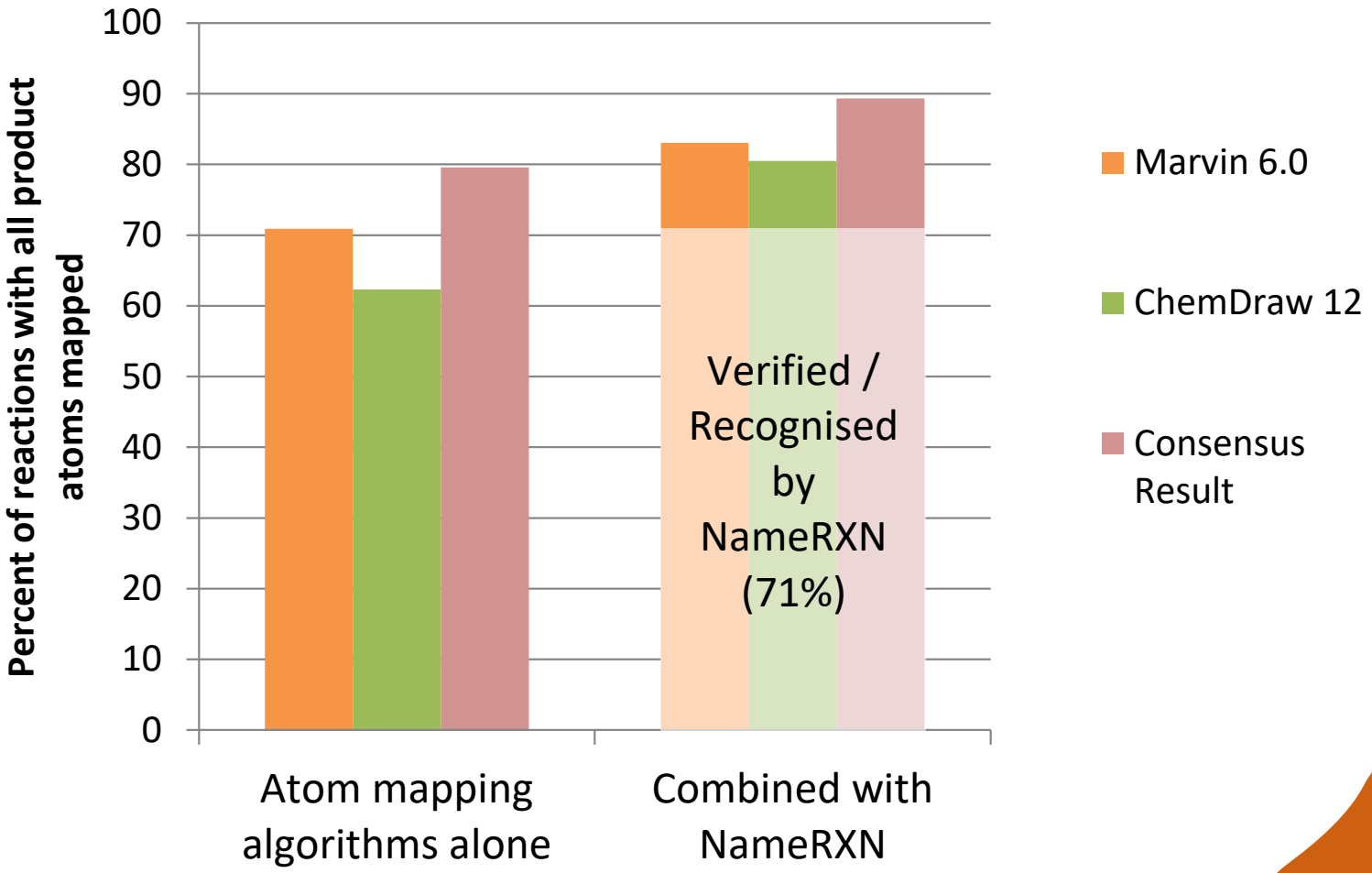
EXAMPLE SMARTS/SMIRKS

```
# NOZAKI_HIYAMA_KISHI_REACTION
[#6v4+0;X4,X3:1][BrD1h0+0:2].[Ni].[Cr].[OD1h
0+0:3]=[CD2h1v4+0:4]>>[#6:1][C:4]-[Oh1:3]

# PAAL_KNORR_THIOPHENE_SYNTHESIS
[OD1h0+0:1]=[CX3v4+0:2][CX4v4+0:3]([H])[CX4v
4+0:4]([H])[CX3v4+0:5]=[OD1h0+0:6]>>[S:1]1[C
:2]=[C:3][C:4]=[C:5]1
```

- Writing SMIRKS is both an art and a science.

ATOM MAPPING + CLASSIFICATION



10 MOST POPULAR REACTIONS

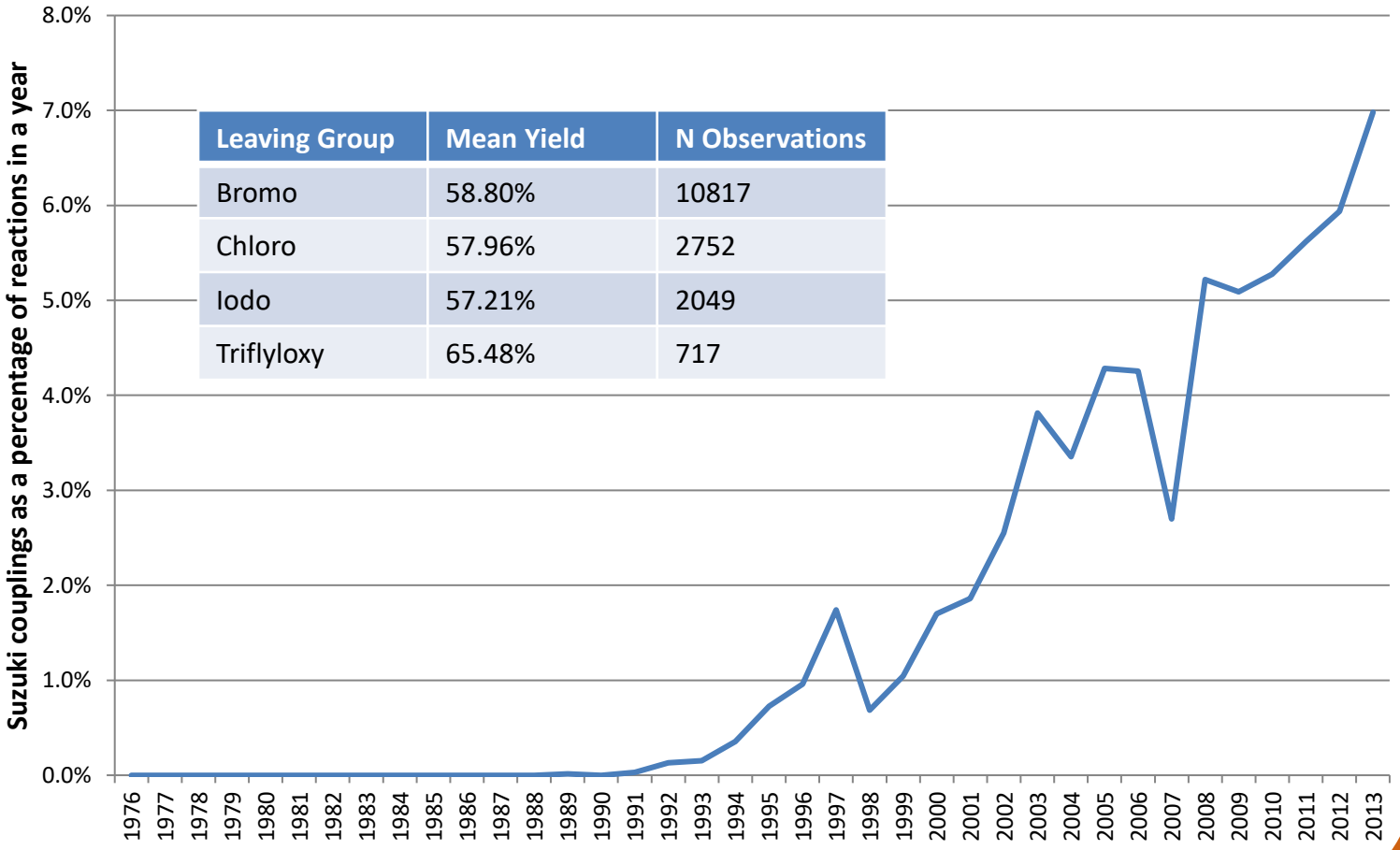
ID	Name	Count
2.1.2	Carboxylic acid + amine	26,040
1.3.1	Buchwald-Hartwig amination	22,048
3.1	Suzuki coupling	16,508
1.7.6	Williamson ether synthesis	15,665
2.1.1	Amide Schotten-Baumann	11,016
7.1	Nitro to amino	10,234
6.1.1	N-Boc deprotection	9,821
6.2.2	CO2H-Me deprotection	9,487
6.2.1	CO2H-Et deprotection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

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

TRENDS IN REACTION TYPES

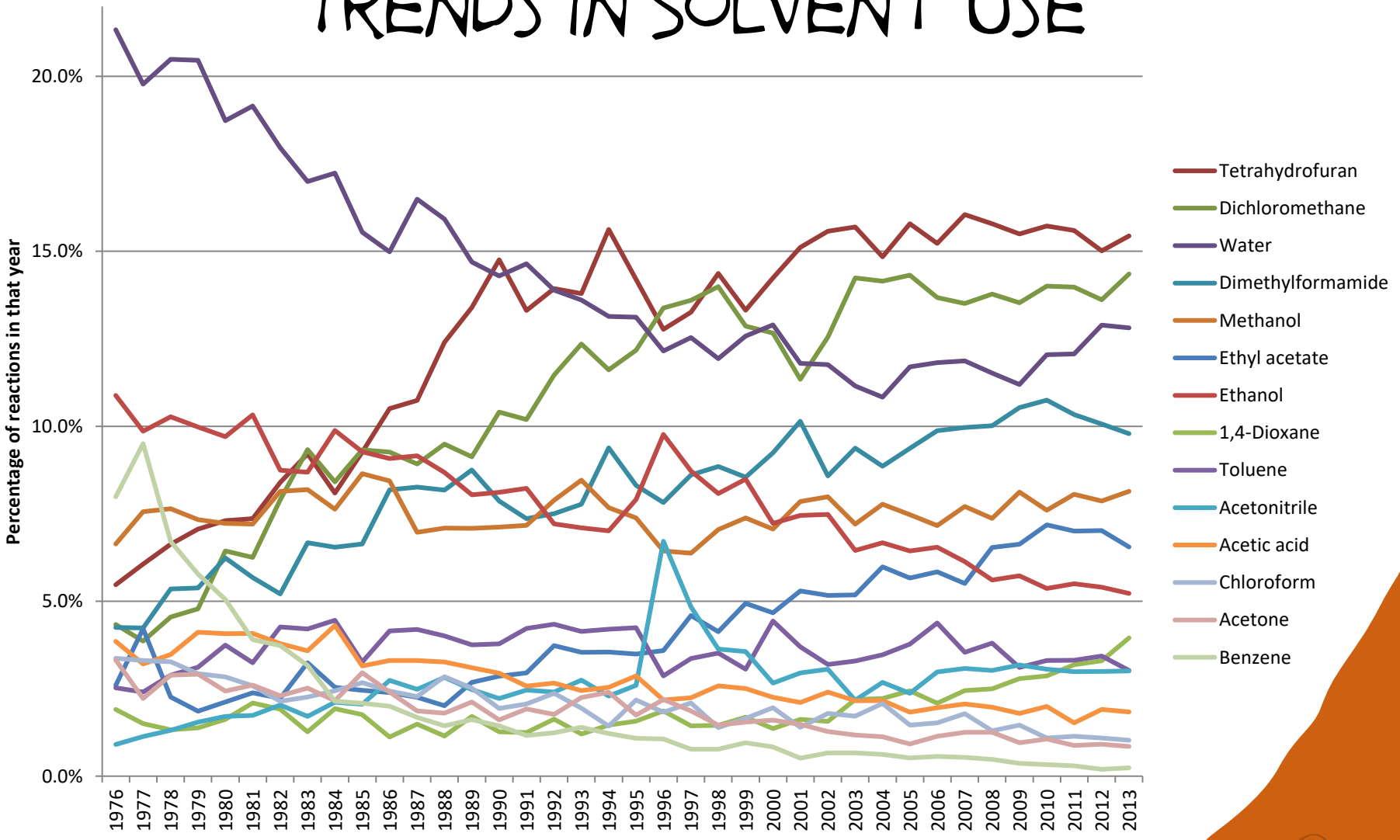
Leaving Group	Mean Yield	N Observations
Bromo	58.80%	10817
Chloro	57.96%	2752
Iodo	57.21%	2049
Triflyloxy	65.48%	717



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 SOLVENT USE



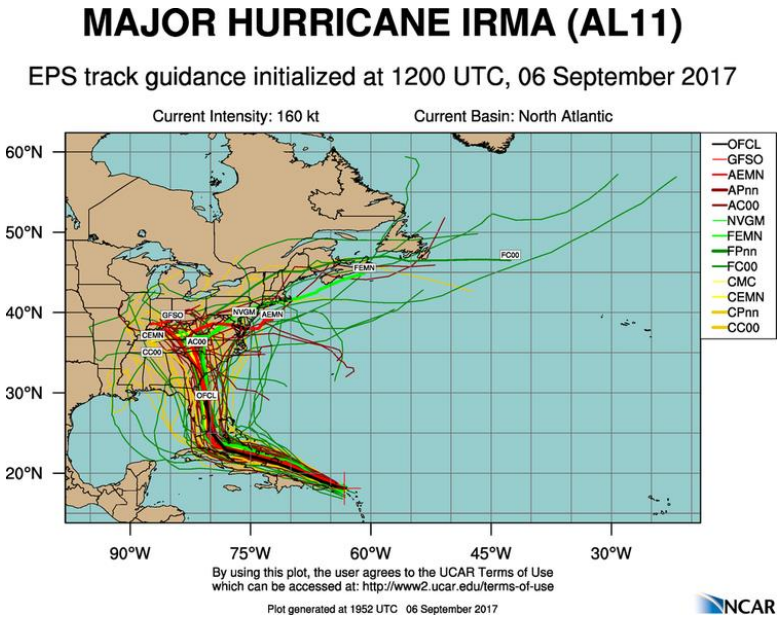
ARE SOLVENTS GETTING GREENER?

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%

RARE NAMED REACTIONS

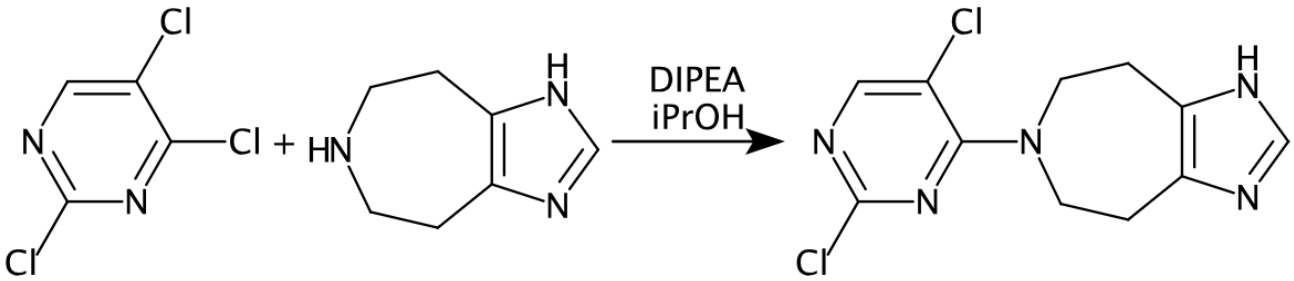
- Adams decarboxylation
- Angeli-Rimini reaction
- Aza-Baylis-Hillman reaction
- Boyer reaction
- Buchwald-Fischer indole synthesis
- Castro-Stephens coupling
- **Chapman rearrangement**
- Chugaev elimination
- Cook-Heilbron thiazole synthesis
- Fischer-Hepp rearrangement
- Gasman indole synthesis
- Fukuyama indole synthesis
- Imine Hosomi-Sakurai reaction
- Koch reaction
- Leuckart reaction
- Liebeskind-Srogl coupling
- Lossen rearrangement
- Ponzio reaction
- Prins reaction
- Reimer-Tiemann carboxylation

ANALYSIS VS. PREDICTION



THE CHALLENGE OF REGIOSELECTIVITY

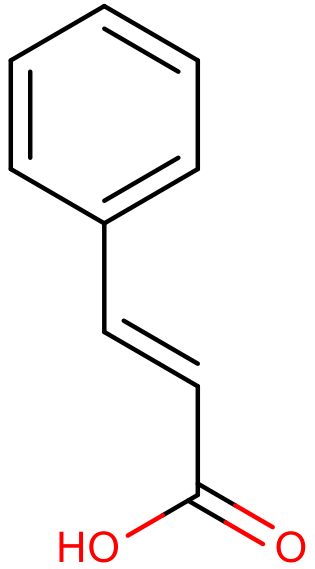
- A tricky benchmark is reactions of 2,4,5-trichloropyrimidine



- The nature of pyrimidine makes the chloro at the 4-position more reactive than the 2 position which is more reactive than the 5 position.
- Simple quantum mechanical have difficulty discerning this order.

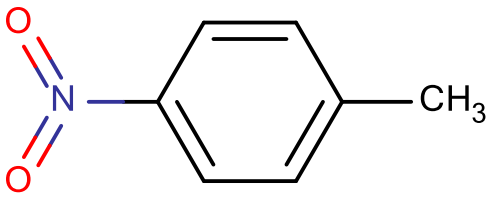
APPLICATION TO PLANNING 1

- Cinnamic Acid ($PhCHCHCO_2$)
  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)



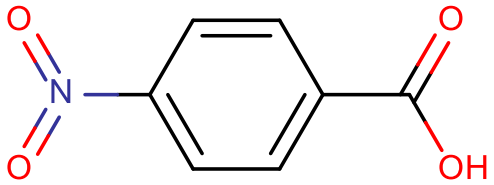
APPLICATION TO PLANNING 2

- p-Nitrotoluene



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

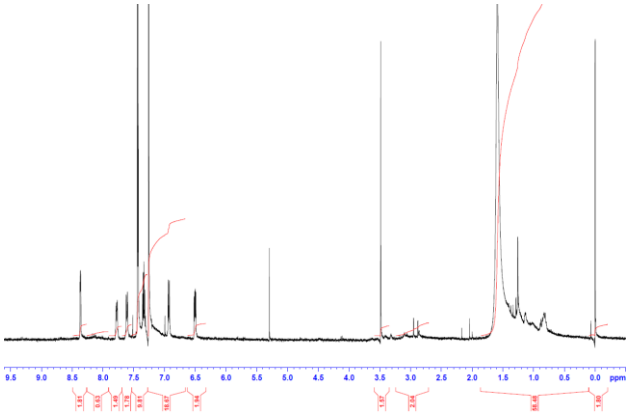
- p-Nitrobenzoic acid



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

EXPERIMENTAL VALIDATION

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



ACKNOWLEDGEMENTS

- NextMove Software
- NextMove Alumni
  - Daniel Lowe
  - Noel O'Boyle

- Thank you for you time.
- Questions?
- Thoughts?

- AbbVie
- AstraZeneca
- Bristol-Myers Squibb
- **Eli Lilly**
- GlaxoSmithKline
- Hoffmann-La Roche
- **IBM Research Zurich**
- IKTOS
- Merck
- **MIT**
- **Novartis**
- PendingAI
- Relay Therapeutics
- Royal Society of Chemistry
- Syngenta
- Vernalis
- Vertex Pharmaceuticals
- WuXi

NAMERXN NAMING RESOURCES

- L. Kurti and B. Czako, “Strategic Applications of Named Reactions in Organic Synthesis”, Elsevier, 2005.
- A. Hassner and I. Namboothiri, “Organic Synthesis based on Name Reactions”, Elsevier, 3rd Edition 2011.
- The Merck Index, 15th Edition 2003. (13th Edition 2001).
- J.J. Li, “Name Reactions”, Springer, 5th Edition, 2019.
- github.com/rsc-ontologies/rxno
- Wikipedia