

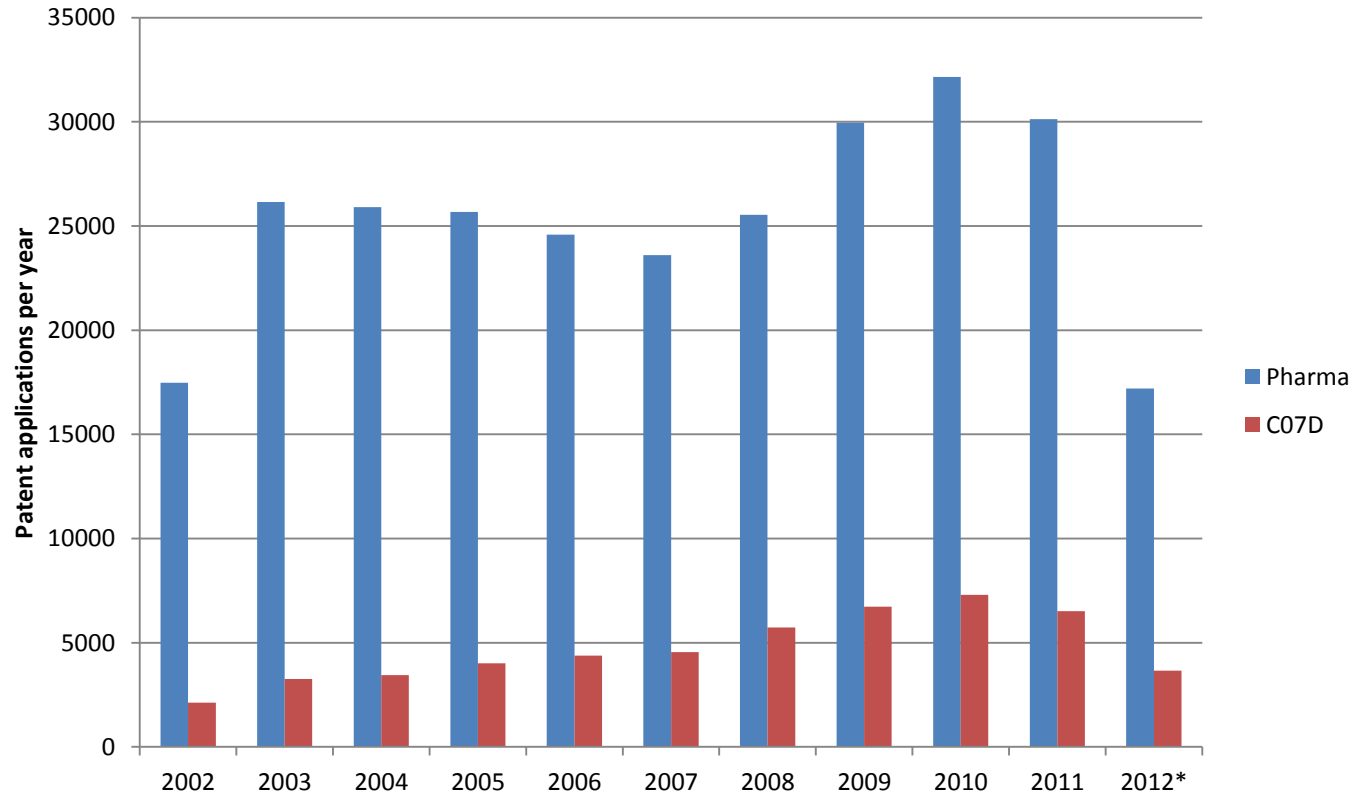


Chemical Text Mining for Current Awareness of Pharmaceutical Patents

Daniel Lowe and Roger Sayle
NextMove Software
Cambridge, UK



US PATENT APPLICATIONS BY YEAR



*2012 includes patent applications published on or before 9th August 2012.
"Pharma" is defined as IPC codes C07*, A61K, A61P and A01N.





USPTO BULK DOWNLOADS

The following USPTO [patent products](#) are available for free download.

Patent Grants

- [Patent Grant Multi-Page Images \(1790 – present\)](#)
- [Patent Grant Full Text with Embedded Images \(2001 – present\)](#)
- [Patent Grant Full Text \(1976 – present\)](#)
- [Patent Grant Bibliographic Data \(1976 – present\)](#)
- [Patent Grant OCR Text \(1920 – 1979\)](#)
- [Patent Grant Single-Page Images \(Oct 2010 – present\)](#)

Patent Application Publications

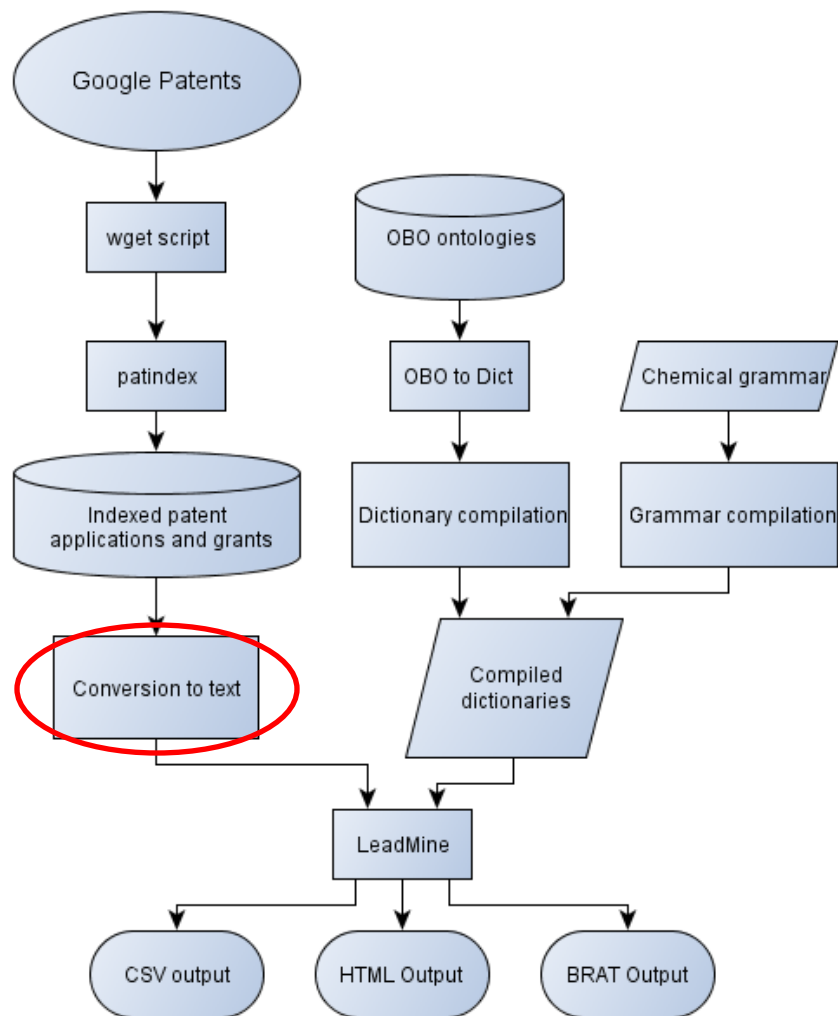
- [PAIR \(Patent Application Information Retrieval\) Data](#)
- [Patent Application Publication Multi-Page Images \(2001 – present\)](#)
- [Patent Application Publication Full Text with Embedded Images \(2001 – present\)](#)
- [Patent Application Publication Full Text \(2001 – present\)](#)
- [Patent Application Publication Bibliographic Data \(2001 – present\)](#)
- [Patent Application Single-Page Images \(Oct 2010 – present\)](#)

Additional Patent Data

- [Patent Assignment Text \(1980 – present\)](#)
- [Patent Maintenance Fee Events \(1981 – present\)](#)
- [Patent Classification Information \(current\)](#)
- [Patent IFW Petition Decisions](#)



WORKFLOW



GREEN BOOK 1976-2000

PAC EXAMPLE 30

PAC >6-(2-Chloro-ethoxy)-7-(2-methoxy-ethoxy)-quinazolin-4-yl]-(3-ethynyl-phenyl)-amine Hydrochloride

PAR The title product was prepared from

4-chloro-6-(2-chloro-ethoxy)-7-(2-methoxyethoxy)-quinazoline (399 mg, 1.26 mmol) and 3-ethynyl-aniline (147 mg, 1.26 mmol) as described for Example 29. (515 mg; 94%; M.P. 215.degree.-225.degree. C. (dec); LC-MS: 398 (MH.sup.+); anal. RP18-HPLC RT: 4.85 min.).



EXAMPLE 30

[6-(2-Chloro-ethoxy)-7-(2-methoxy-ethoxy)-quinazolin-4-yl]-(3-ethynyl-phenyl)-amine Hydrochloride

The title product was prepared from 4-chloro-6-(2-chloro-ethoxy)-7-(2-methoxyethoxy)-quinazoline (399 mg, 1.26 mmol) and 3-ethynyl-aniline (147 mg, 1.26 mmol) as described for Example 29. (515 mg; 94%; M.P. 215°-225° C. (dec); LC-MS: 398 (MH⁺); anal. RP18-HPLC RT: 4.85 min.).



SGML (2001 GRANTS)

- Uses different tags to Red Book documents
- May contain unclosed tags:

<PCIT>

<DOC><DNUM><PDAT>5154857</PDAT></DNUM>

<DATE><PDAT>19921000</PDAT></DATE></DOC>

<PARTY-US>

<NAM><SNM><STEXT><PDAT>Goto et al.</PDAT></STEXT></SNM></NAM>

</PARTY-US>

<PNC><PDAT>25229963</PDAT></PNC></PCIT><CITED-BY-EXAMINER>



RED BOOK (2002 – PRESENT)

<heading id="h-0082" level="1">Example 30</heading>

<heading id="h-0083" level="1">Preparation of (E)-2-amino-N-(3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)phenyl)acetamide</heading>

<p id="p-0883" num="1052"><chemistry id="CHEM-US-00235" num="00235">

</chemistry>

</p>

<p id="p-0884" num="1053">(E)-2-amino-N-(3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)phenyl)acetamide was prepared following the method used in Example 15.</p>

<p id="p-0885" num="1054">Step 1: Coupling of Wittig reagent 24 with 3-nitrobenzaldehyde gave 1-nitro-3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)benzene as a light yellow oil. Yield (0.639 g, 95%), isomer ratio 4:1 ratio trans:cis.</p>

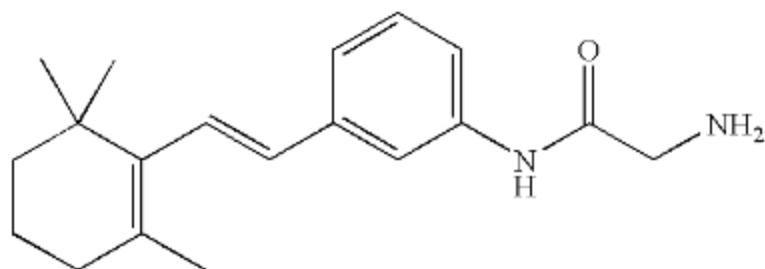
<p id="p-0886" num="1055">trans-isomer: ¹H NMR (300 MHz, CDCl₃)
δ 8.24 (t, J=1.9 Hz, 1H), 8.04 (m, 1H), 7.69 (d, J=7.7 Hz, 1H), 7.47 (t, J=8.0 Hz, 1H), 6.83 (dd, J=16.3, 0.85 Hz, 1H), 6.40 (d, J=16.3 Hz, 1H), 2.06 (t, J=6.2 Hz, 2H), 1.81 (s, 3H), 1.65 (m, 2H), 1.52 (m, 2H), 1.08 (s, 6H);</p>



HTML GENERATED FROM RED BOOK

Example 30

Preparation of (E)-2-amino-N-(3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)phenyl)acetamide



(E)-2-amino-N-(3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)phenyl)acetamide was prepared following the method used in Example 15.

Step 1: Coupling of Wittig reagent 24 with 3-nitrobenzaldehyde gave 1-nitro-3-(2-(2,6,6-trimethylcyclohex-1-enyl)vinyl)benzene as a light yellow oil. Yield (0.639 g, 95%), isomer ratio 4:1 ratio trans:cis.

trans-isomer: $^1\text{H NMR}$ (300 MHz, CDCl_3) δ 8.24 (t, $J=1.9$ Hz, 1H), 8.04 (m, 1H), 7.69 (d, $J=7.7$ Hz, 1H), 7.47 (t, $J=8.0$ Hz, 1H), 6.83 (dd, $J=16.3, 0.85$ Hz, 1H), 6.40 (d, $J=16.3$ Hz, 1H), 2.06 (t, $J=6.2$ Hz, 2H), 1.81 (s, 3H), 1.65 (m, 2H), 1.52 (m, 2H), 1.08 (s, 6H);



BENEFITS OF CLEAN INPUT

- Patent feed text:

Cis-2,3,6,7,12,12a-hexahydro-2-benzyl6-(4-methoxyphenyl)-pyrazino[2',1':6,1]pyrido[3,4-b]indole-1,4-dione

New line in
middle of name!

- Extracted from USPTO source:

Cis-2,3,6,7,12,12a-hexahydro-2-benzyl6-(4-methoxyphenyl)-pyrazino[2',1':6,1]pyrido[3,4-b]indole-1,4-dione

- LeadMine entity:

Cis-2,3,6,7,12,12a-hexahydro-2-benzyl-6-(4-methoxyphenyl)-pyrazino[2',1':6,1]pyrido[3,4-b]indole-1,4-dione













LEADMINE 2.0

- Dictionary and grammar based general entity recogniser
- Tokenization determined by the terms to be recognised



DEFAULT DICTIONARIES

Dictionary		Example	Size
Molecule		benzoic acid	Infinite
Dictionary		ranitidine	11,201
Registry #		GW-409544	Large but finite
CAS Number		7732-18-5	Large but finite
Element		gold	185
Fragment		phenyl	Infinite
Atom Fragment		chloro	11
Polymer		polystyrene	74
Generic		alkane	362
Noise		formal	16



LEADMINE 2.0

- Dictionaries to be used are configurable e.g. protein targets, genes, diseases, reaction names etc.
- Matching speed is independent of dictionary size
- Any dictionary can be used with spelling correction to match lexically close entities



LEADMINE 2.0 CONFIGURATION

#A company registry number for a compound

[dictionary]

location CFDictR.cfx

entityType R ←

htmlColor #90b0ff

caseSensitive false

useSpellingCorrection false

Multiple dictionaries
can map to the same
entity type

#A molecule e.g. 2-methylpyridine

[dictionary]

location CFDictM.cfx

entityType M

htmlColor violet

enforceBracketing true

caseSensitive false

useSpellingCorrection true

minimumCorrectedEntityLength 9

maxCorrectionDistance 1

Spelling correction can
be adjusted on a per
dictionary basis



BUILDING DICTIONARIES

- Uses Daciuk/Mihov's algorithm to allow building dictionaries with millions of entities in linear time
- Extremely large dictionaries are often smaller when compiled than the original input
- 54 million synonyms from PubChem can be compiled to a dictionary slightly less than 1gb in 17 minutes and 20 seconds!

Daciuk, J.; Mihov, S.; Watson, B. W.; Watson, R. E. Incremental construction of minimal acyclic finite-state automata.

Computational linguistics **2000**, 26, 3–16.



FOREIGN LANGUAGE SUPPORT

- Chinese and Japanese chemical names may be rapidly converted to English as a pre-processing step

“Translating IUPAC-like chemical nomenclature to and from simplified Chinese” 9:10 am, Wednesday, Global Opportunities in Chemical Information



SAMPLE OUTPUT (HTML)

To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0°C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol). The reaction mixture was then warmed to room temperature and stirred overnight. After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate. The drying agent was filtered off and the filtrate concentrated under reduced pressure. The residue obtained was purified by flash chromatography (silica gel, 0-5% methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).



SAMPLE OUTPUT (CSV)

```
"in",E,"M",1,0,"COC1=CC=C(CN2CCC(CC2)O)C=C1","1-(4-methoxybenzyl)piperidin-4-ol",  
"in",E,"M",1,0,"BrCC1=CC=C(C=C1)OC","1-(bromomethyl)-4-methoxybenzene",  
"in",E,"M",1,0,"OC1CCNCC1","4-hydroxypiperidine",  
"in",E,"G",1,0,, "brine",  
"in",E,"M",1,0,"CN(C=O)C","dimethylformamide",  
"in",E,"M",1,0,"C(C)(=O)OCC","ethyl acetate",  
"in",E,"M",1,0,"CO","methanol",  
"in",E,"M",1,0,"C(Cl)Cl","methylene chloride",  
"in",E,"G",1,0,, "silica gel",  
"in",E,"M",1,0,"S(=O)(=O)([O-])[O-].[Na+].[Na+]", "sodium sulfate",  
"in",E,"M",1,0,"C(C)N(CC)CC","triethylamine",  
"in",E,"N",1,0,"O","water",
```



BRAT (BRAT RAPID ANNOTATION TOOL)

T1	M 25 44	4-hydroxypiperidine
T2	M 78 95	dimethylformamide
T3	M 121 153	1-(bromomethyl)-4-methoxybenzene
T4	M 178 191	triethylamine
T5	M 404 417	ethyl acetate
T6	N 439 444	water
T7	G 457 462	brine
T8	M 495 509	sodium sulfate
T9	G 658 668	silica gel
T10	M 675 683	methanol
T11	M 684 702	methylene chloride
T12	M 714 747	1-(4-methoxybenzyl)piperidin-4-ol



BRAT (BRAT RAPID ANNOTATION TOOL)

- 1 To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0°C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol).
- 2 The reaction mixture was then warmed to room temperature and stirred overnight.
- 3 After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate.
- 4 The drying agent was filtered off and the filtrate concentrated under reduced pressure.
- 5 The residue obtained was purified by flash chromatography (silica gel, 0-5% methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).



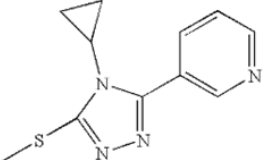
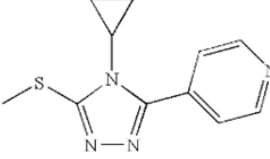
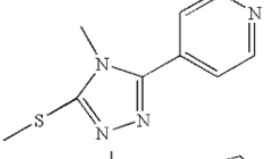
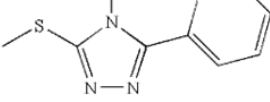
PATFETCH

Patent number:

e.g. 6356863 or 2007129372

Output type:Raw HTML UTF-8 Escaped ASCII **Leadmine:**Off HTML CSV Corrected

The following examples were synthesized in a manner analogous to that for 3-[4-methyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]pyridine

Structure	Name	Analytical data	Example No.
	3-[4-cyclopropyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]pyridine	LC-MS ($M^+ + 1$): 233	835
	4-(4-Cyclopropyl-5-methylsulfanyl-4H-1,2,4-triazol-3-yl)-pyridine	$^1\text{H NMR}$: 8.77 (d, 2 H), 7.75 (m, 2 H), 3.23 (m, 1 H), 2.82 (s, 3 H), 1.17 (m, 2 H), 0.80 (m, 2 H).	836
	4-[4-methyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]-pyridine	$^1\text{H NMR}$: (DMSO-D6): 2.7 (s, 3 H) 3.6 (s, 3 H) 7.7 (m, 2 H) 8.8 (d, 2 H)	837
	3-(4-fluorophenyl)-4-methyl-5-(methylthio)-4H-1,2,4-triazole	Used directly in the next step towards 3-(4-fluorophenyl)-4-methyl-5-(methylsulfonyl)-4H-1,2,4-triazole.	838

Example 839**4-methyl-3-(methylthio)-5-(trifluoromethyl)-4H-1,2,4-triazole**

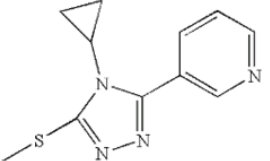
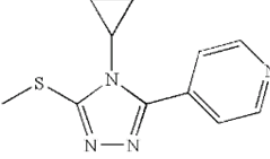
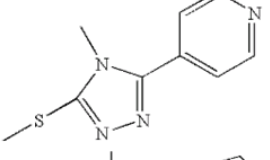
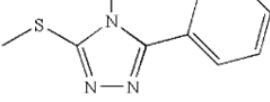
PATFETCH

Patent number:

e.g. 6356863 or 2007129372

Output type:Raw HTML UTF-8 Escaped ASCII **Leadmine:**Off HTML CSV Corrected

The following examples were synthesized in a manner analogous to that for 3-[4-methyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]pyridine

Structure	Name	Analytical data	Example No.
	3-[4-cyclopropyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]pyridine	LC-MS ($M^+ + 1$): 233	835
	4-(4-cyclopropyl-5-methylsulfanyl-4H-1,2,4-triazol-3-yl)pyridine	$^1\text{H NMR}$: 8.77 (d, 2 H), 7.75 (m, 2 H), 3.23 (m, 1 H), 2.82 (s, 3 H), 1.17 (m, 2 H), 0.80 (m, 2 H).	836
	4-[4-methyl-5-(methylthio)-4H-1,2,4-triazol-3-yl]pyridine	$^1\text{H NMR}$: (DMSO- D_6): 2.7 (s, 3 H) 3.6 (s, 3 H) 7.7 (m, 2 H) 8.8 (d, 2 H)	837
	3-(4-fluorophenyl)-4-methyl-5-(methylthio)-4H-1,2,4-triazole	Used directly in the next step towards 3-(4-fluorophenyl)-4-methyl-5-(methylsulfonyl)-4H-1,2,4-triazole.	838

Example 839**4-methyl-3-(methylthio)-5-(trifluoromethyl)-4H-1,2,4-triazole**

PATFETCH-CONT.

- Recognises common USPTO grant/application number variants e.g. 6356863/US
6356863/006356863/US 6,356,863 B1
- Allows all USPTO patent grant/applications to be accessed as text or html from simple URLs e.g. patfetch/patents/6356863.html



"MACROSCOPIC" ANALYSIS

- Having all patents available allows for analysis that spans the entire corpus rather than being limited to a single patent
- Example use cases
 - Identifying the key compounds in a patent
 - Finding the first instance of a molecule in the patent literature
 - Identify patents containing novel chemistry



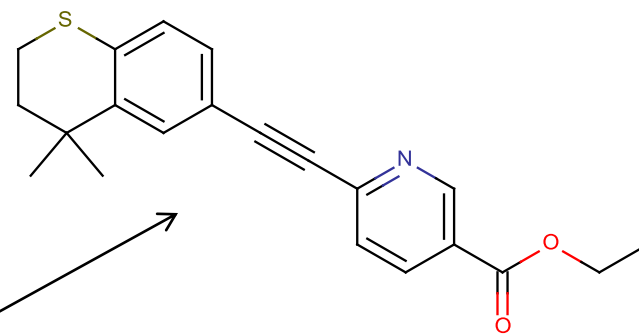
FILTERING IRRELEVANT PATENTS

- Most irrelevant patents can be excluded by IPC codes. These are assigned by the USPTO to classify each patent.
- Typical pharmaceutical IPC codes
 - C07 (Organic Chemistry)
 - A61K (Preparations for medical, dental or toilet purposes)
 - A61P (Specific therapeutic activity of chemical compounds or medicinal preparations)
 - A01N (Preservation of bodies of humans or animals or plants or parts thereof)



FINDING THE FIRST MENTION OF A COMPOUND

- Trivial names of compound often won't be present in the first patent synthesising a compound
- Brand name Fabior, approved May 11, 2012
- Generic name Tazarotene
- First mentioned in US05023341



Hz, 2.2 Hz), 7.59 (1H, d, J~7.8 Hz), 7.66 (1H, d, J~2.2 Hz), 8.30 (1H, dd, J~7.8 Hz, 2.3 Hz), 9.2

Alternative synthesis: Ethyl 6-[(4,4-dimethylthiochroman-6-yl)ethynyl]nicotinate (Compound 99) v

A solution of 15.4 g (76.2 mmol) of 4,4-dimethyl-6-ethynyl-thiochroman (Compound 1) and 14.0

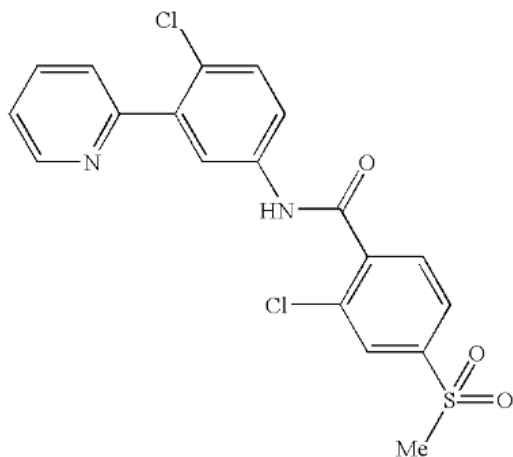


FINDING THE FIRST MENTION OF A COMPOUND

- Brand name Erivedge, approved January 30, 2012
- Generic name Vismodegib
- First mentioned in US20060063779A1

Example 37

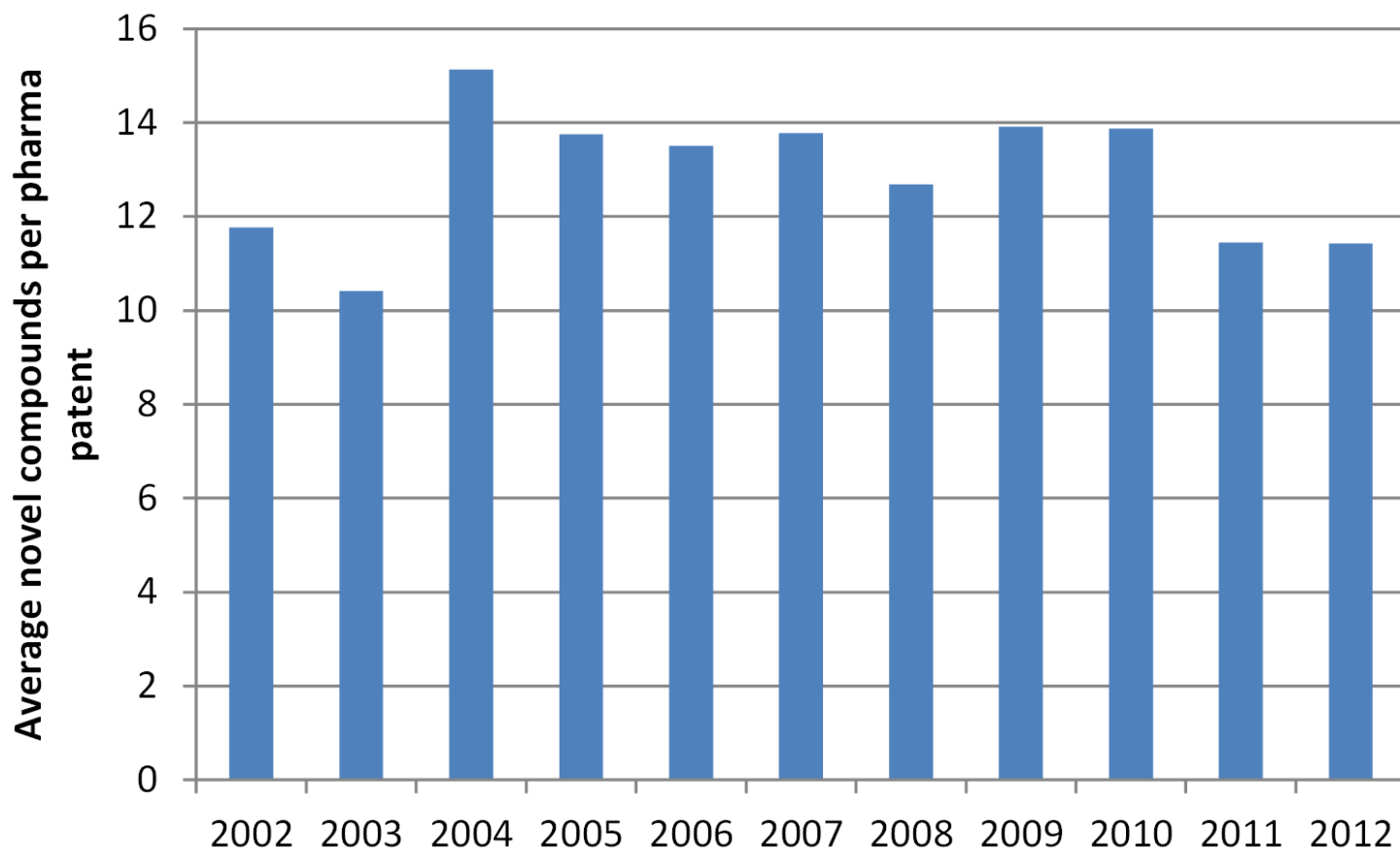
2-chloro-N-(4-chloro-3-(pyridin-2-yl)phenyl)-4-(methylsulfonyl)benzamide



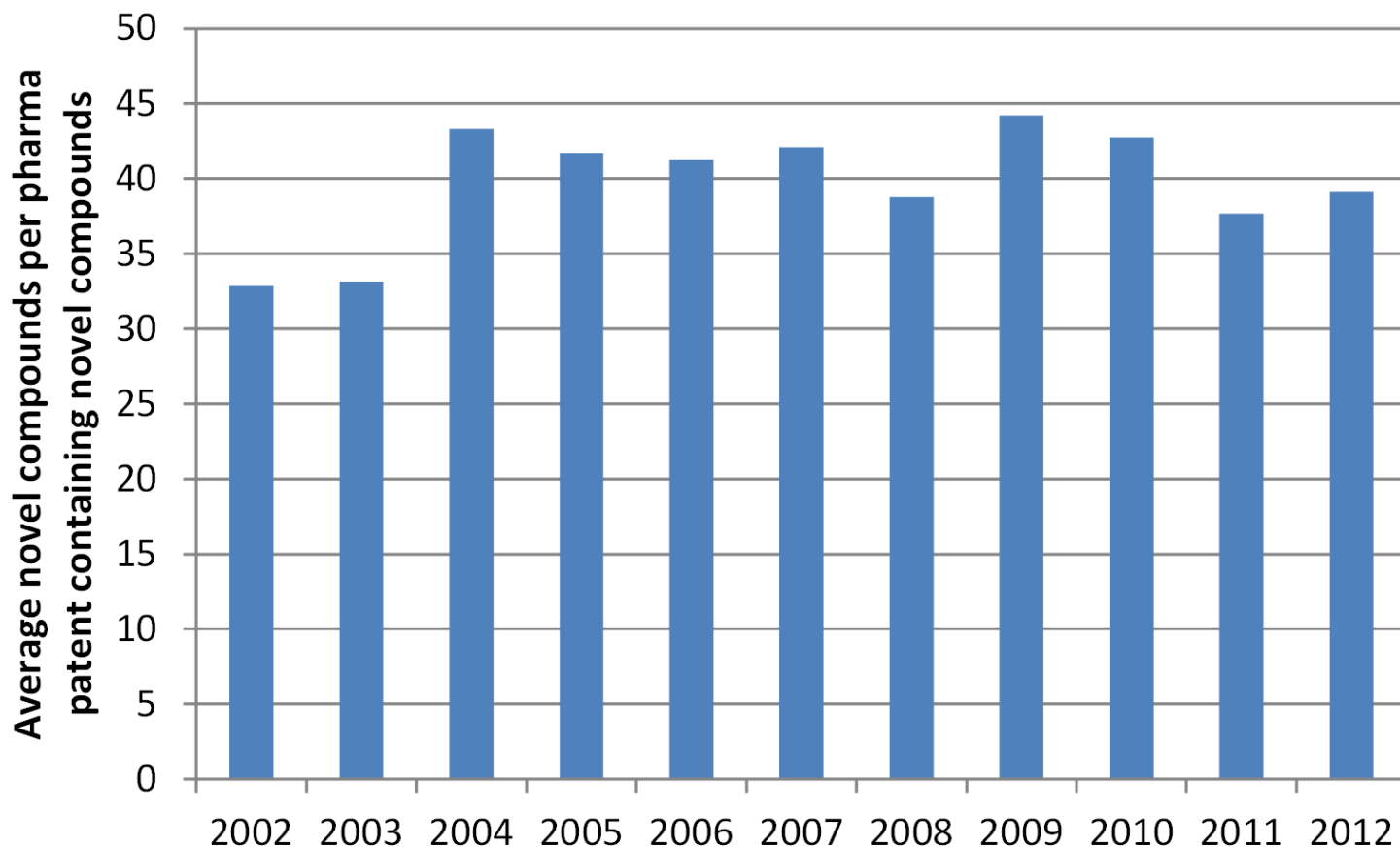
Procedure G was used to couple 4-chloro-3-(pyridin-2-yl)aniline (50 mg) and 2-chloro-4-methylsulfonylbenzoic acid to proc



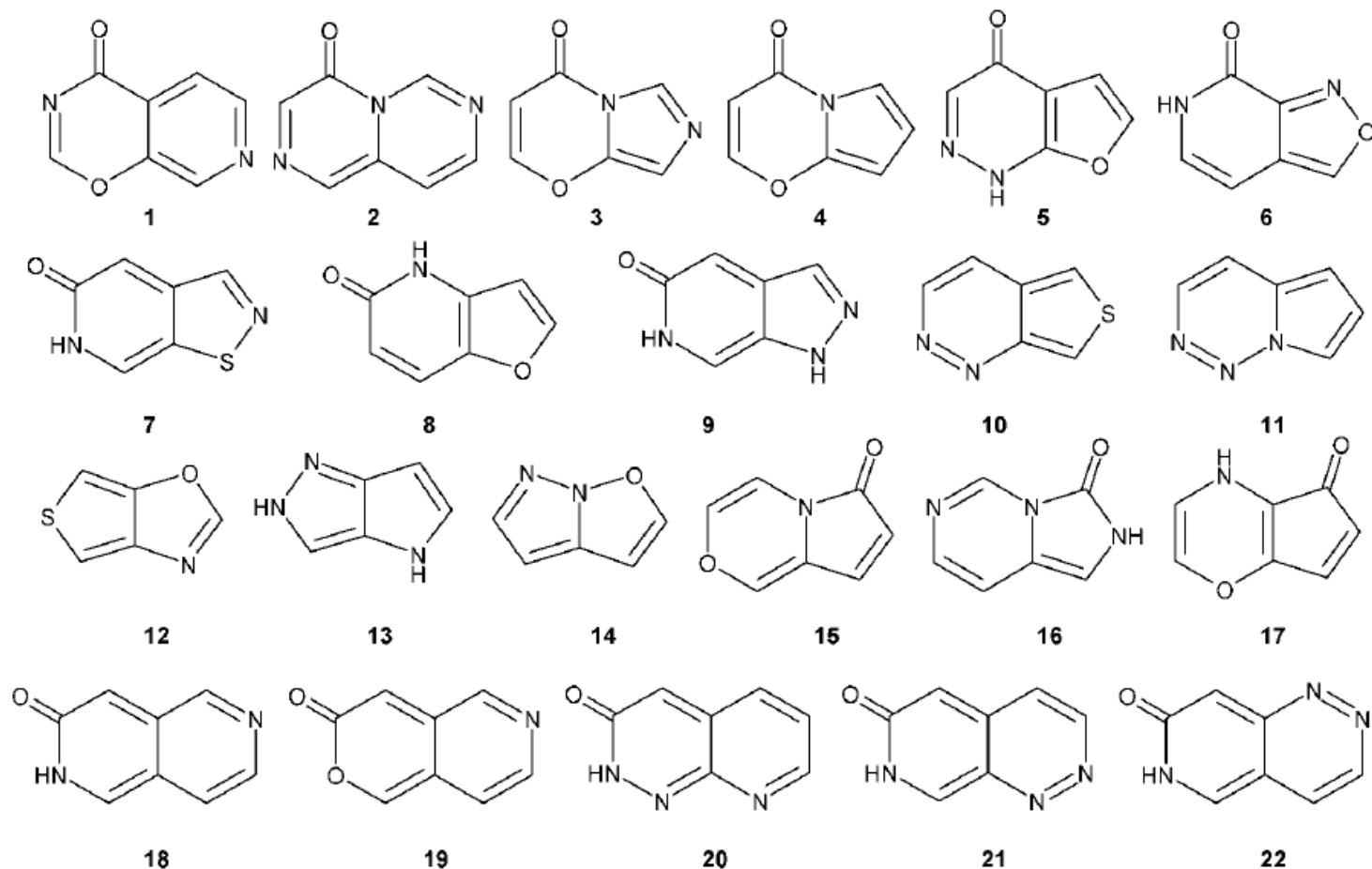
NOVEL COMPOUNDS PER PATENT



NOVEL COMPOUNDS PER PATENT



AWARENESS OF NOVEL SCAFFOLDS

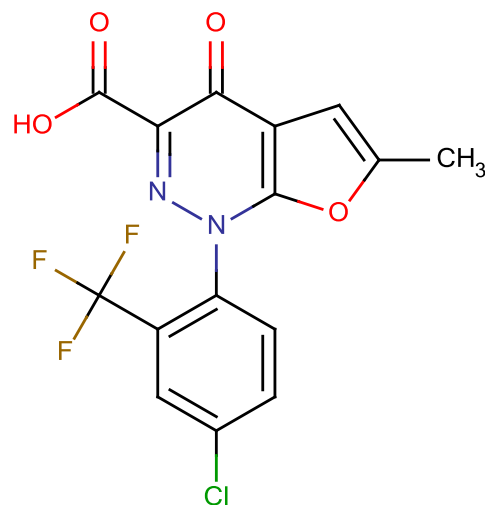
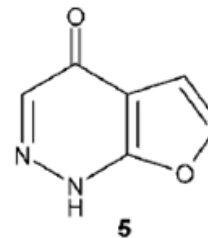


Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *Journal of medicinal chemistry* **2009**, *52*, 2952–2963.

ACS National Meeting, Philadelphia, USA 19th August 2012



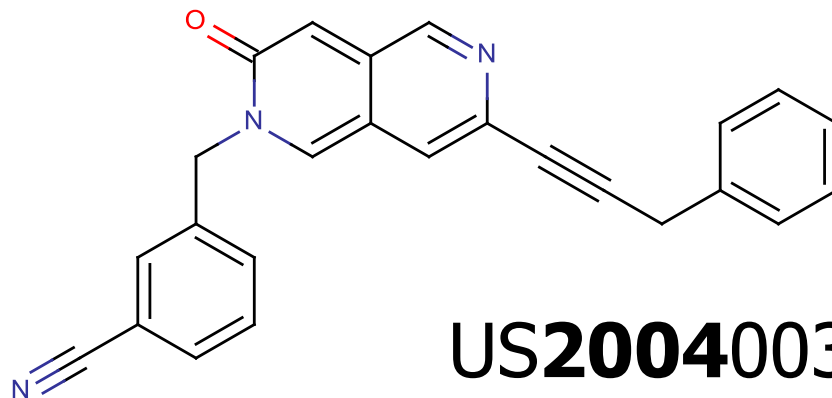
AWARENESS OF NOVEL SCAFFOLDS



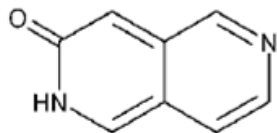
US46041346, 1983



AWARENESS OF NOVEL SCAFFOLDS



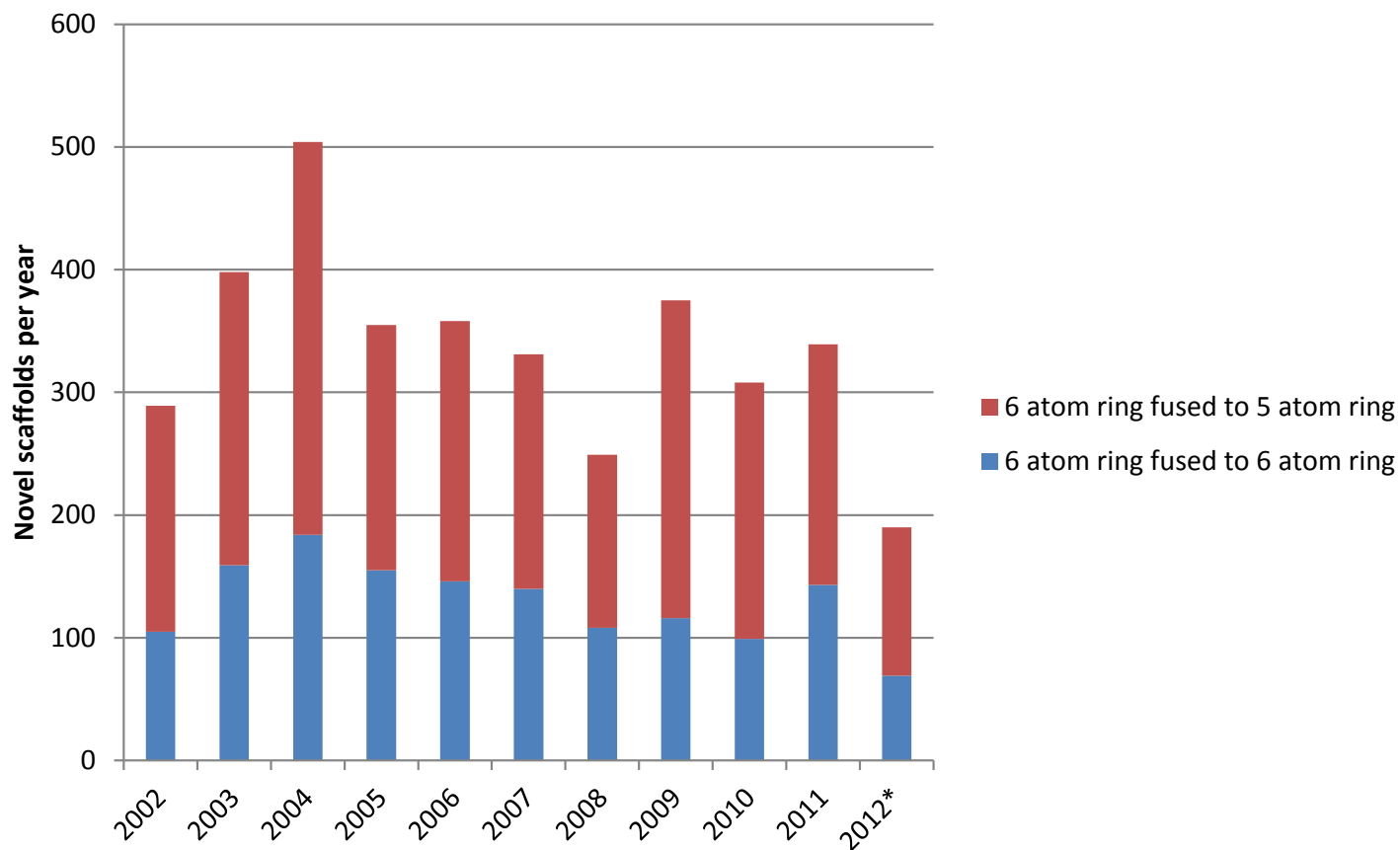
US20040038959A1



18



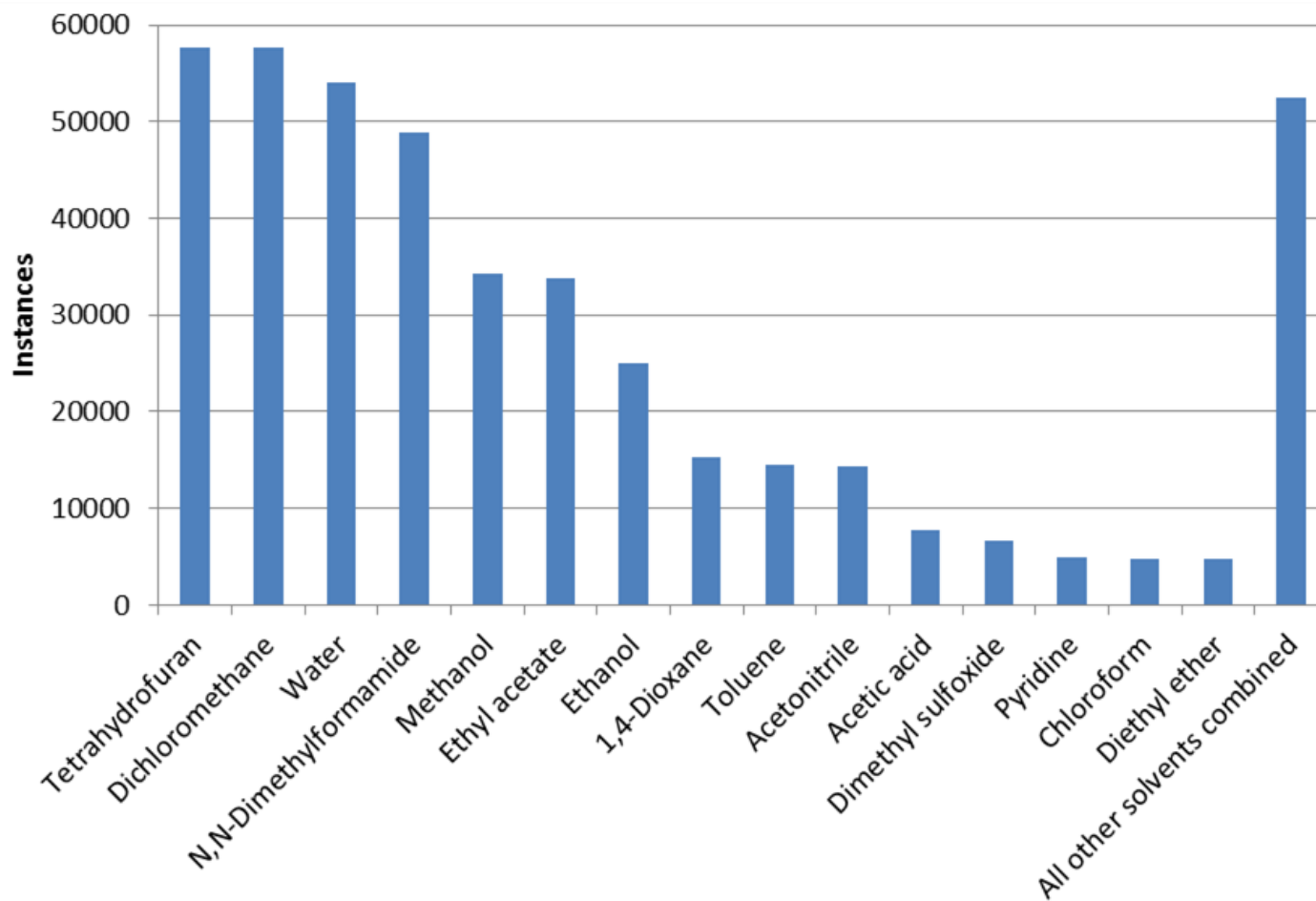
RATE OF NOVEL SCAFFOLD DISCOVERY



*year not yet complete



SOLVENT OCCURRENCES



Extracted from reactions in 2008-2011 USPTO patent applications



CONCLUSIONS

- Getting clean text from patents is an important starting point
- LeadMine offers a highly configurable environment for performing entity extraction
- Comprehensive coverage of the patent literature can assist in identifying the interesting aspects of a new patent



ACKNOWLEDGEMENTS

- Sorel Muresan and Paul Hongxing Xie, AstraZeneca.
- Nicko Goncheroff, SureChem/Digital Science.
- Colin Batchelor, Royal Society of Chemistry.
- Peter Loew and Heinz Saller, InfoChem.
- Pat Walters, Vertex Pharmaceuticals.

- Thank you for your time.



ACS National Meeting, Philadelphia, USA 19th August 2012



SIMPLE JAVA API

```
ExtractEngine engine = new ExtractEngine();
EntityCollector collector = engine.processString("text
to analyse");
List<Entity> foundEntities = collector.getEntities();
for (Entity entity : entities) {
    entity.getText();
    entity.getEntityType();
    entity.getBeg();
    entity.getEnd();
}
```

