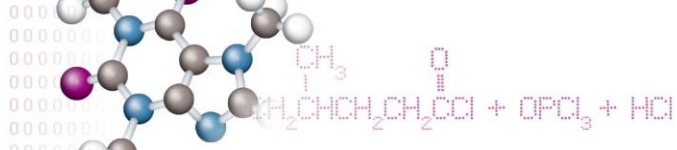




-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



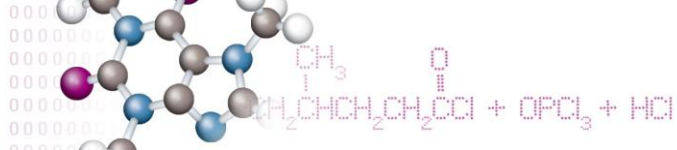
Challenges in Next Generation Scientific and Patent Information Mining

Josef Eiblmaier (InfoChem), Hans Kraut (InfoChem), Larisa Isenko (InfoChem), Heinz Saller (InfoChem),
Peter Loew (InfoChem)

ICIC 2011 Barcelona, October 23 – 26



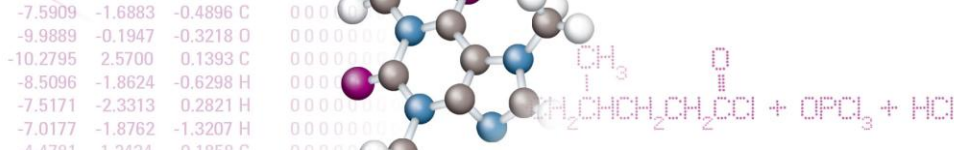
-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



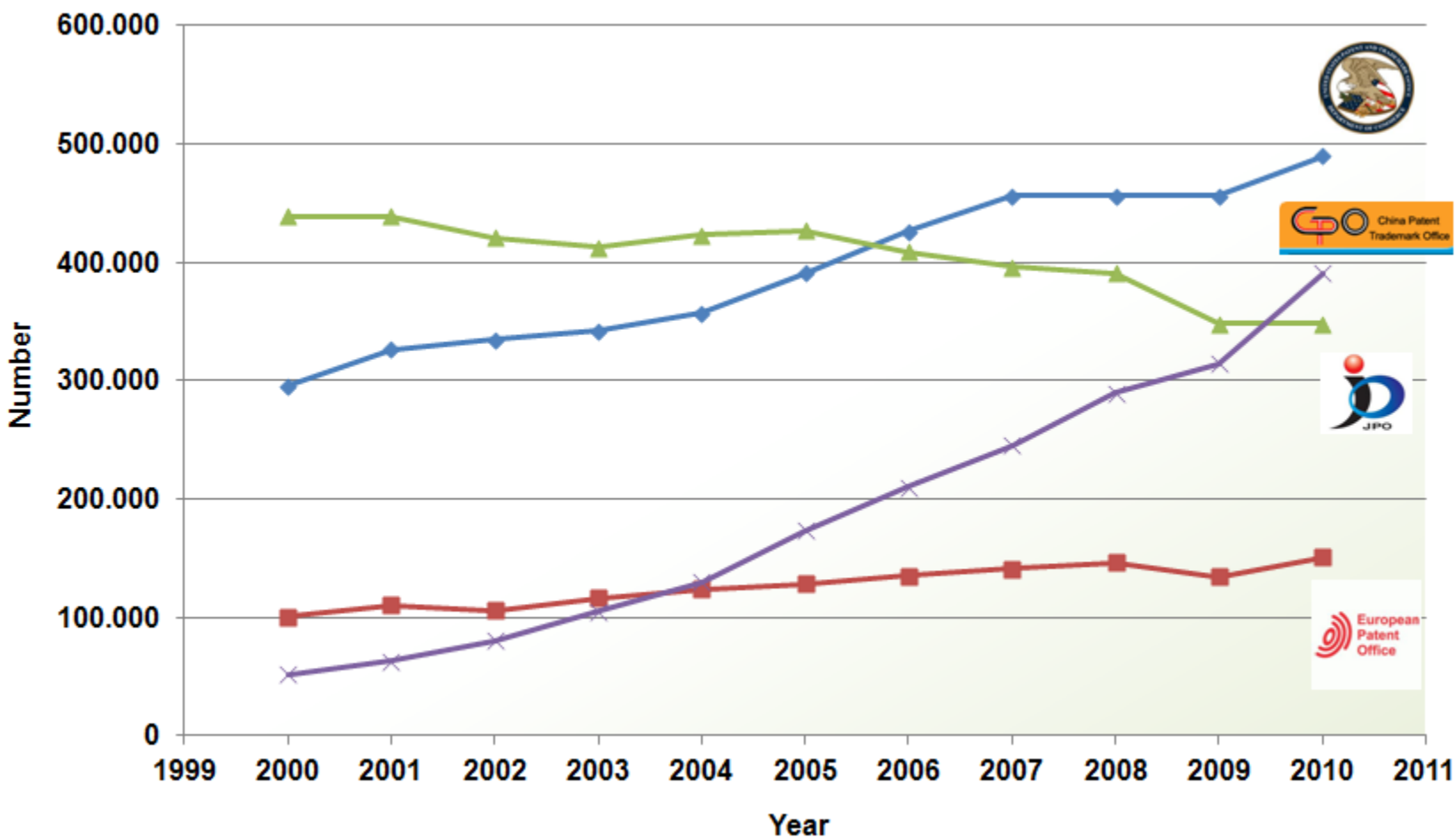
Outline

- » Patents, a buried treasure?
- » Chemical NER
- » CDX work-up
- » Image recognition
- » Combining the techniques:
ChemProspector





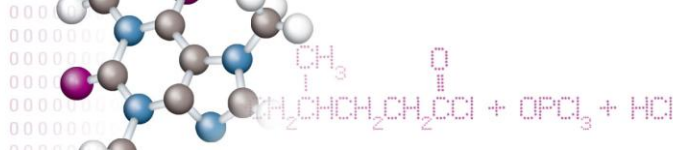
No. Applications Filed (Selected Patent Offices)



Source: EPO (http://www.epo.org/searching/asian/trends_de.html), USPTO (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm)



-7.5909	-1.1333	-0.4896	C	0 0 0 0
-9.9889	-0.1947	-0.3218	O	0 0 0 0
-10.2795	2.5700	0.1393	C	0 0 0 0
-8.5096	-1.8624	-0.6298	H	0 0 0 0
-7.5171	-2.3313	0.2821	H	0 0 0 0
-7.0177	-1.8762	-1.3207	H	0 0 0 0

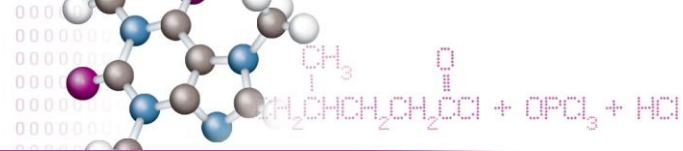


Patents, a Buried Treasure?

- » Scientific knowledge contained in patents comprises about 70 – 80 percent of all scientific knowledge
- » In most cases, a patent is the first publication in the research process
- » Important for competitor monitoring, technology assessment, R&D portfolio management
- » Access to external technological knowledge



-7.5909 -1.1313 -0.4896 C
 -9.9889 -0.1947 -0.3218 O
 -10.2795 2.5700 0.1393 C
 -8.5096 -1.8624 -0.6298 H
 -7.5171 -2.3313 0.2821 H
 -7.0177 -1.8762 -1.3207 H



Chemical Pharmaceutical Patents: a Buried Treasure?

Chemical structures (images)

Chemical names/fragments (text)

Markush structures (text, images, CDX)

Chemical structures (CDX files)

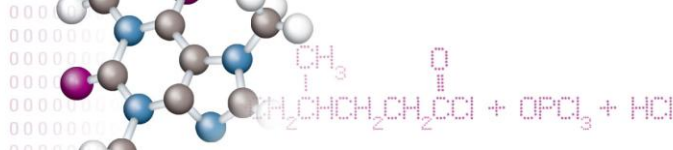
Reactions (images, CDX files)

Structure	Name	Met IC ₅₀ (µM)	¹ H-NMR	MS (M)
(a)	5-Bromo-3-(2,6-dichlorobenzoyloxy)-pyridin-2-ylamine	5.3	(400 MHz, DMSO-d ₆) 8.7, 8.2 (m, 1H), 7.56 (m, 2H), 7.46 (m, 2H), 5.80 (s, 2H), 5.22 (s, 2H)	341
(b)			(400 MHz, DMSO-d ₆) 8.7, 5.6 (d, 1H)	
(c)				
(d)				
(e)				





-7.5909 -1.1943 -0.4896 C
-9.9889 -0.1947 -0.3218 O
-10.2795 2.5700 0.1393 C
-8.5096 -1.8624 -0.6298 H
-7.5171 -2.3313 0.2821 H
-7.0177 -1.8762 -1.3207 H



Chemical Named Entity Extraction: the Compulsory Exercise?

» Named entity extraction (ICANNOTATOR)

- Dictionary based
- Morphology based
- Context based

» Name to structure conversion

» Tagging of source documents

- Highlighting
- Enrichment with additional attributes
- Linking (deep linking, cross document)

Related Issue Journal CHEMISTRY AND MATERIALS SCIENCE

View Related Documents

Journal Article Metabolism of ricinoleate by Neurospora crassa M. Goodrich-Tanrikulu, A. E. Stafford, J. T. Lin and T. A. McKeon Applied Microbiology Volume 46, Number 1, January 1989

Download PDF

Journal Article Conversion of p

Chemistry Term Highlight n Generic S Complete S Fragment Substance

Permissions & Reprints EXPORT CITATION ABOUT

of these compounds with 3,2,6-pyrimidin-4(3H)-one

no 4,3-dipyridin-6-one

pyridin-5-one

Challenges NER: OCR'ed Documents

Scanned source document:

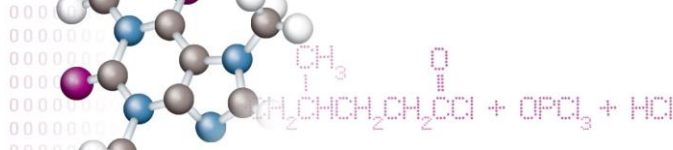
The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (**17**) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (**18**) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{(1)}-C_{(2)}-C_{(3)}-S_{(1)}$ is 175.0° . The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound **18** are

OCR results:

The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (**17**) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (**18**) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{(1)}-C_{(2)}-C_{(3)}-S_{(1)}$ is 175.0° . The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound **18** are



-7.5909	-1.8883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



Challenges NER: 'Digital Born' PDF's

PDF document:

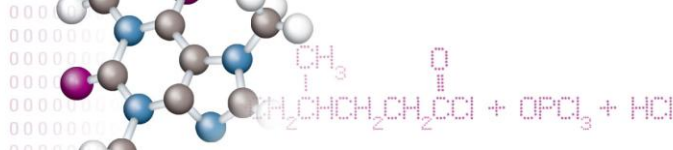
[7 α ,8 α ,3',4']-N'-(Phenyl)succinimido-6,14-*endo*-etheno-6,7,8,14-tetrahydrothebaine (VI) was synthesized using the method described in [16];

PdfToText results:

[7**a**,8**a**,3**c**,4**c**]-N**c**-(Phenyl)succinimido-6,14-endo-etheno....



-7.5909	-1.1933	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



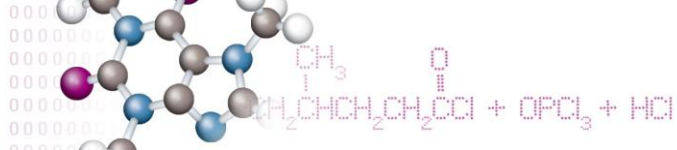
CDX Workup: Nothing is as it Seems!

- » For various sources CS ChemDraw files are available on large scale
 - MRW's (e.g. Science of Synthesis)
 - Journals (e.g. Thieme, Wiley, Springer)
 - Patents (e.g. USPTO: 2000 - present)
- » The contained structures/reactions can be worked-up automatically (*ICSchemeProcessor*)
- » But be careful: nothing is as it seems!



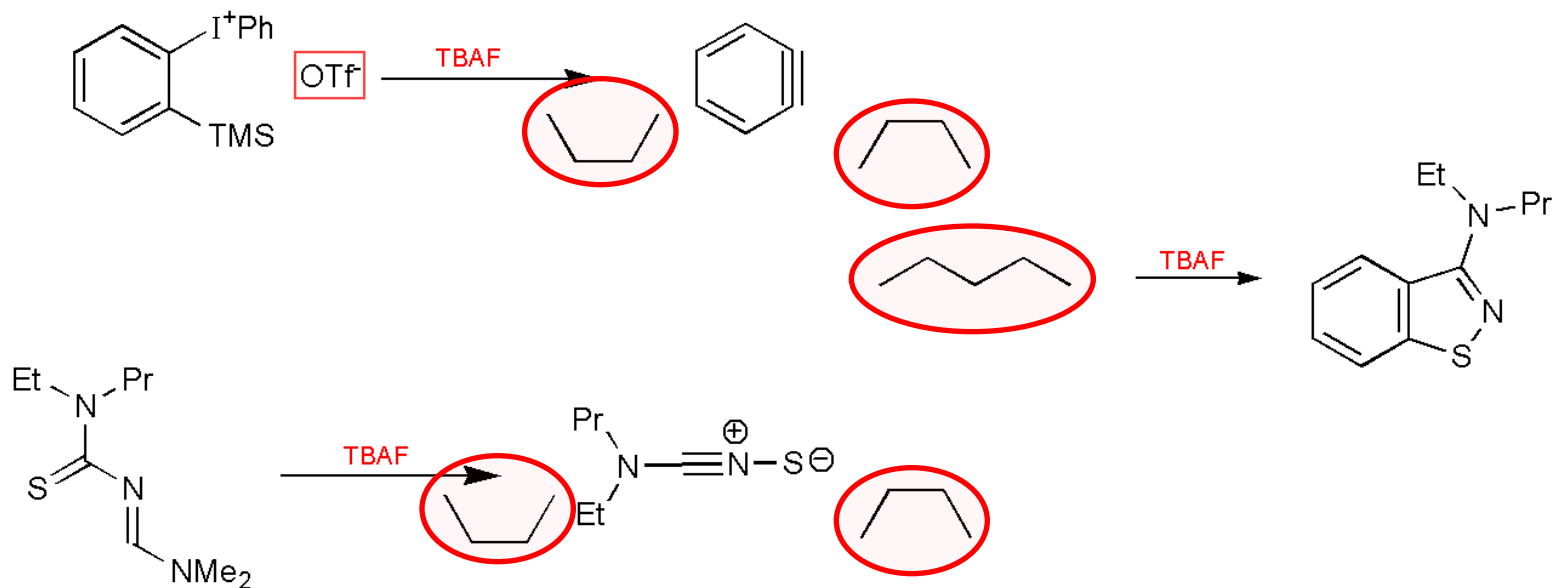


-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



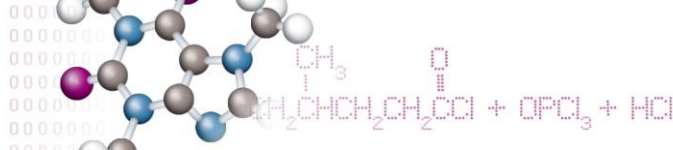
CDX Work-up: Nothing is as it Seems!

» Reaction arrows / forked arrows / brackets





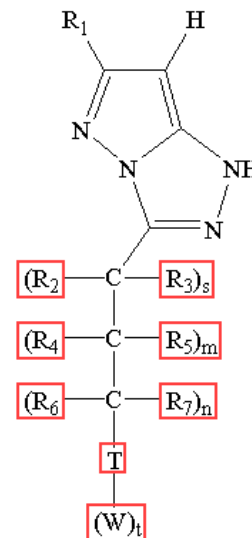
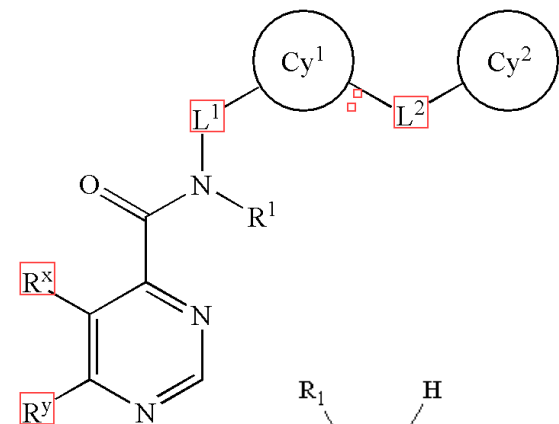
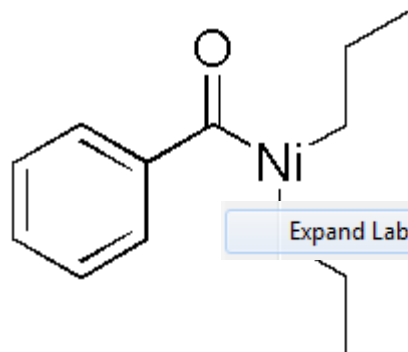
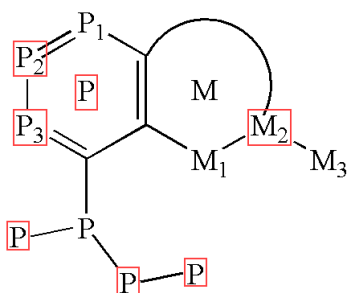
-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



CDX Work-up: Nothing is as it Seems!

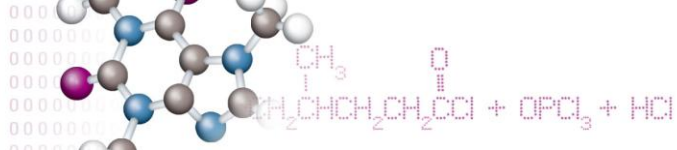
» Unresolvable labels

- Labels not defined
- Element symbols used as R-group labels
- Ambiguous fragment labels (e.g. molecular formula)



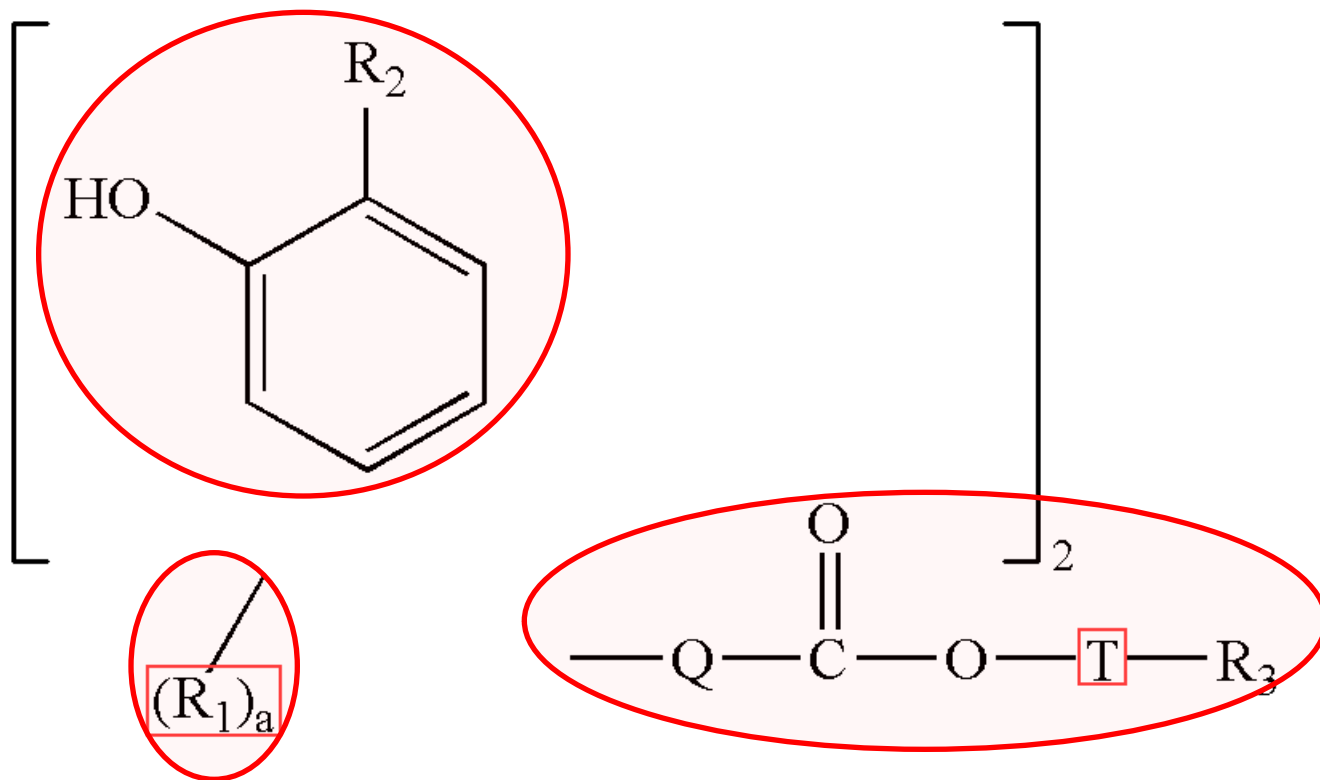


-7.5909	-1.3883	-0.4896	C	0 0 0 0 0 0
-9.9889	-0.1947	-0.3218	O	0 0 0 0 0 0
-10.2795	2.5700	0.1393	C	0 0 0 0 0 0
-8.5096	-1.8624	-0.6298	H	0 0 0 0 0 0
-7.5171	-2.3313	0.2821	H	0 0 0 0 0 0
-7.0177	-1.8762	-1.3207	H	0 0 0 0 0 0



CDX Work-up: Nothing is as it seems!

» Variable points of attachment



CDX Work-up: Nothing is as it Seems!

» R-group enumeration

- in the CDX file
- unstructured in the text
- structured in the text as table (XML)

```
<sdoc name="07EBF-DOC-6" key="S0105100700300000006">
```

```
<pmme>5.1.16.6.5 Method 5: Application of Acylgermanes as Amide Precursors</pmme>
```

```
<sosc name="07EBF-SCH-6.1" key="S010510070030000000601">
```

```
<posc>Scheme 5 Synthesis of Amides from Acylgermanes[6]</posc>
```

```
<pisc><file name="S_102EBF005" key="S01051007003000000060101"><dsup>16,17</dsup></file></pisc>
```

```
<tbl>
<thdr>
<tr><tcol>R</tcol><ctfs>1</ctfs></tcol><tcol>R</tcol><ctfs>2</ctfs></tcol><tcol>R</tcol><ctfs>3</ctfs></tcol><tcol>Yield
(%)</tcol><tcol>Ref</tcol></tr>
</thdr>
```

```
<tr><tcol><ctfi>t</ctfi>-Bu</tcol><tcol>Bn</tcol><tcol>Bn</tcol><tcol>79</tcol><tcol><cbib>[6]</cbib></tcol></tr>
```

```
<tr><tcol>Ph</tcol><tcol>Bn</tcol><tcol>Bn</tcol><tcol>90</tcol><tcol><cbib>[6]</cbib></tcol></tr>
```

```
</tbl></sosc>
```

```
</sdoc>
</sman>
```

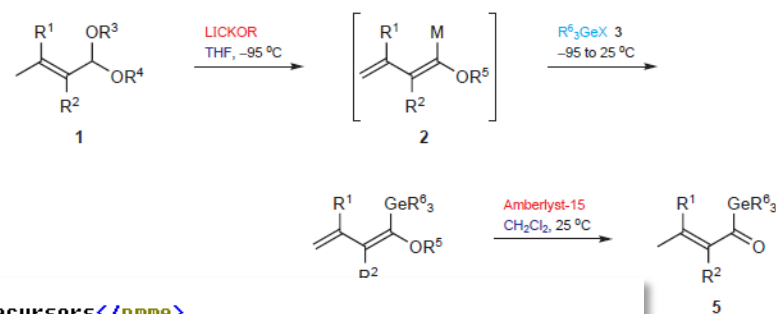
5.1.16.6.1

Method 1:

Synthesis from Germlyl Enol Ethers by Hydrolysis

The reaction of α,β -unsaturated *O,O*-acetals **1** with the butyllithium/potassium *tert*-butoxide superbase (LICKOR) gives 1-alkoxy-1-metallobuta-1,3-diene intermediates **2** via 1,4-elimination. The sequential trapping of these organometallic species with trialkyl(halo)germanes **3** affords 1-alkoxy-1-germylbuta-1,3-dienes **4** in moderate to excellent yields. Hydrolysis in the presence of Amberlyst-15 resin then furnishes α,β -unsaturated acylgermanes **5** (Scheme 1).^[3]

Scheme 1 Acylgermanes by Hydrolysis of Germlyl Enol Ethers^[3]

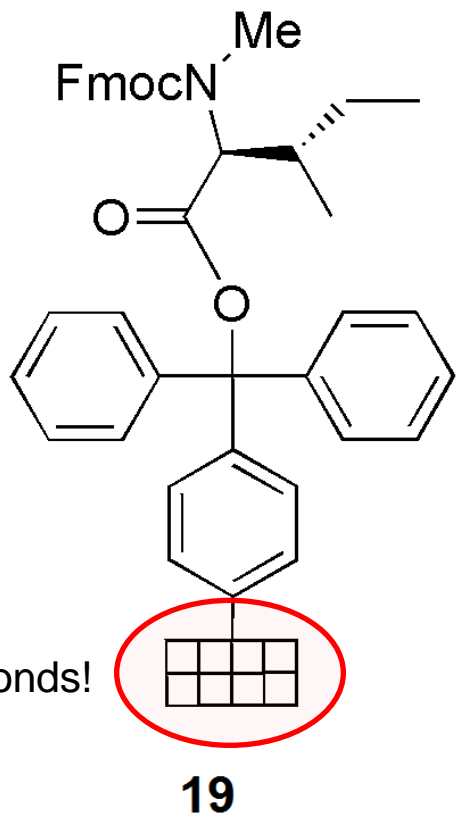


	Yield ^a (%) of 5	Ref
	5	[3]
	n.r.	[3]
	5	[3]
	7	[3]
	n.r.	[3]
	n.r.	[3]
	n.r.	[3]

H H CH₂CMe₂CH₂ CH₂CMe₂CH₂OH Bu 90 n.r.

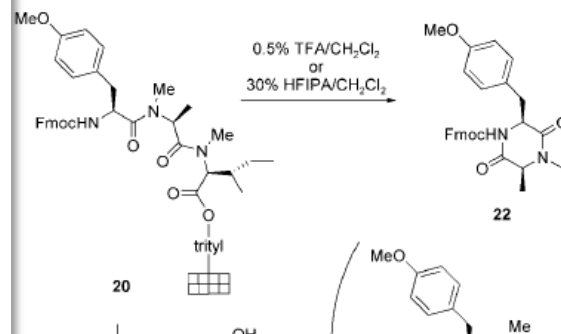
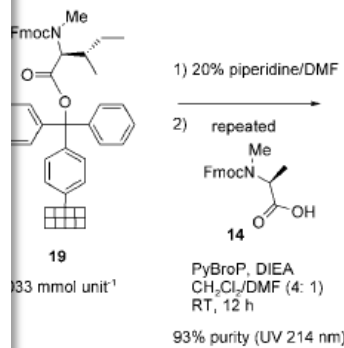
CDX Work-up: Nothing is as it Seems!

» 'Non chemical' objects



C-C bonds!

Total Synthesis of (-)-Apratox



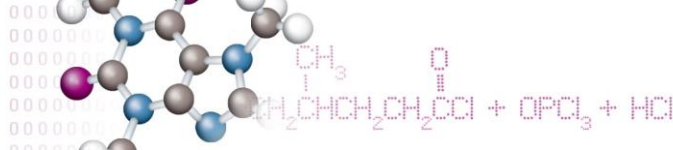
trichloride SynPhase Lanterns, **12** preformed ally available trityl alcohol SynPhase Lantern (lantern) with AcCl, was performed by triethylamine (DIEA)/CH₂Cl₂ (Scheme 2).^[12] of **13** from the polymer-supported **19** with Cl₂, the loading amount of **13** was determined to be 33 mmol per lantern. After removal of the lantern with 20% piperidine/DMF, acylation with

synthesized by suitable for the fore, we investi

Coupling of poly-line-containing previously reported PyAOP^[20] was complete the reaction the Fmoc group precursor **11** w



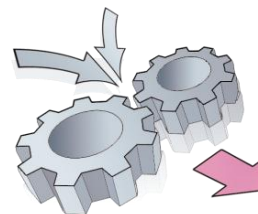
-7.5909	-1.133	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



CDX Work-up: Solution/Approach ICSchemeProcessor

» Algorithmic detection of features

- Repeating groups
- Aliases, R-groups within CDX files



» Guidelines for authors and typesetters

- Syntax definitions for tables, R-groups etc.
- Syntax rules for captions
- Reaction arrangement, forked arrows
- Location and colour rules for reaction conditions
(reactants, catalysts, solvents)



Image Recognition: the Supreme Challenge

» In many cases only raster images are available

» Multiple step process:

- page segmentation
- image classification
- vectorization, OCR, reconstruction

» Part of ChemProspector: *ICImg2Struct*

- InfoChem proprietary development, started mid 2010

The collage shows the following elements:

- Patent Snippet:** A portion of a patent document (EP 1 298 129 A2) containing the text: "Description BACKGROUND OF THE INVENTION [0001] The present invention relates to a process for producing a 2,2-bis(fluoro)zopyrin-4-carboxylic acid...".
- Chemical Structures:** A chemical structure of a substituted pyridine ring with a COOH group and a CH2F group, and another structure showing a transformation with labels R1, COOH, and CH2.
- Grid:** A grid of numerical data, likely representing a feature vector or classification matrix, with columns labeled 'SWISSCHEM' and 'ZS 1'.



-7.5909	-1.1833	-0.4896	C	0 0 0
-9.9889	-0.1947	-0.3218	O	0 0 0 0 0
-10.2795	2.5700	0.1393	C	0 0 0 0 0 0
-8.5096	-1.8624	-0.6298	H	0 0 0 0 0
-7.5171	-2.3313	0.2821	H	0 0 0 0 0 0
-7.0177	-1.8762	-1.3207	H	0 0 0 0 0 0
-1.4794	-1.2124	-0.1058	C	0 0 0 0 0 0

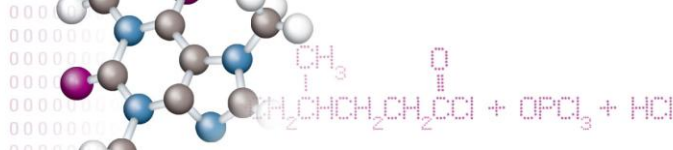


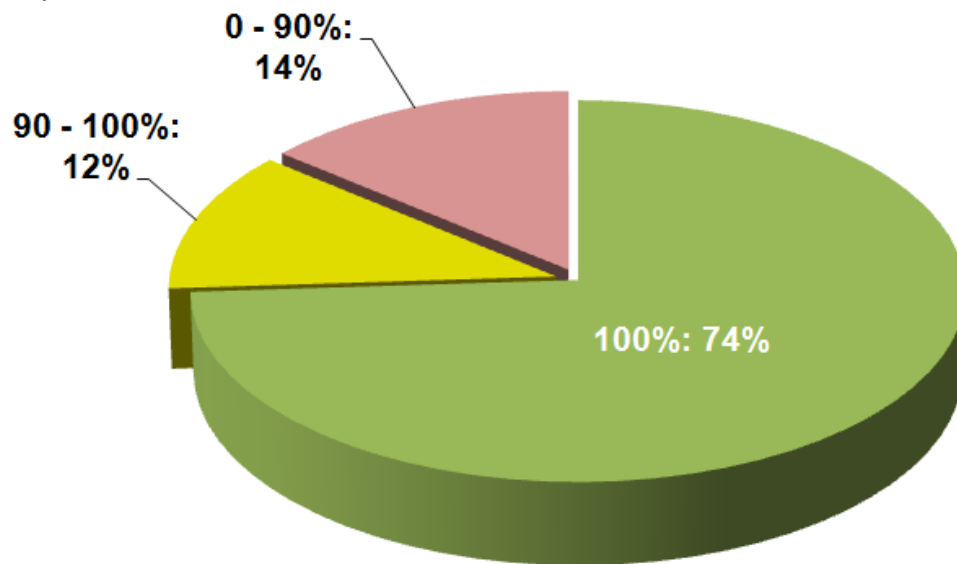
Image Recognition ICImg2Struct: Status and Evaluation

» Used benchmark set: OSRA validation set (USPTO)

- 5,735 chemical structures (images and associated MOL files)
(<http://cactus.nci.nih.gov/osra/uspto-validation.zip>)

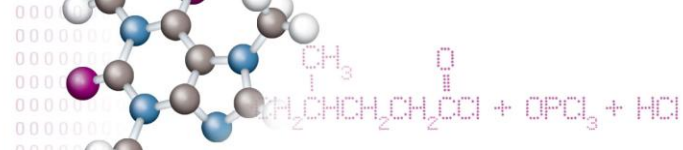
» Degree of match: InfoChem SimilarityMCD

- 100% correct: 4,234 structures (74%)
- 90 - 100%: 722 structures (12%)
- 0 - 90%: 779 structures (14%)





-7.5909 -1.1433 -0.4896 C 0 0 0 0 0 0
-9.9889 -0.1947 -0.3218 O 0 0 0 0 0 0
-10.2795 2.5700 0.1393 C 0 0 0 0 0 0
-8.5096 -1.8624 -0.6298 H 0 0 0 0 0 0
-7.5171 -2.3313 0.2821 H 0 0 0 0 0 0
-7.0177 -1.8762 -1.3207 H 0 0 0 0 0 0



Original Image	ICImg2Struct	SimilarityMCD
		100%
		100%
		96.3%



-7.5909	-1.1933	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
1.4791	1.8131	0.1052	C

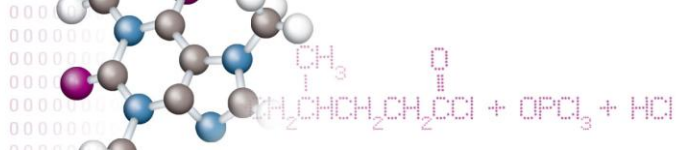
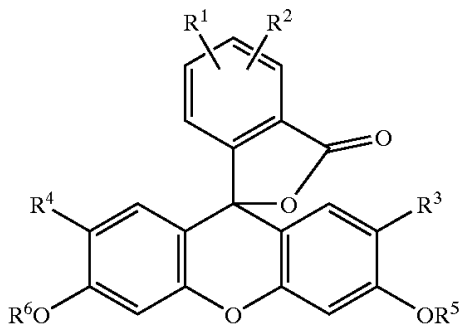
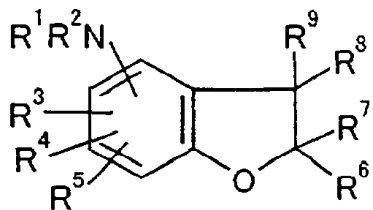
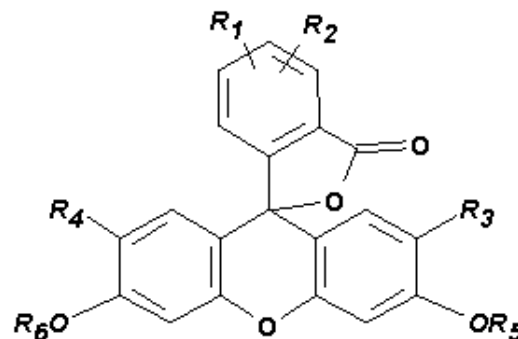


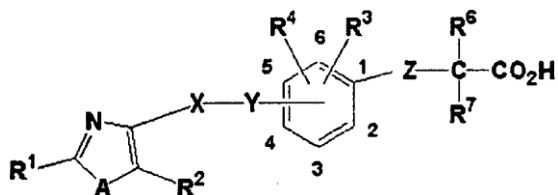
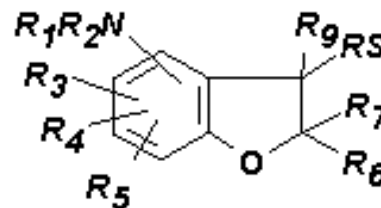
Image Recognition: Variable Points of Attachment



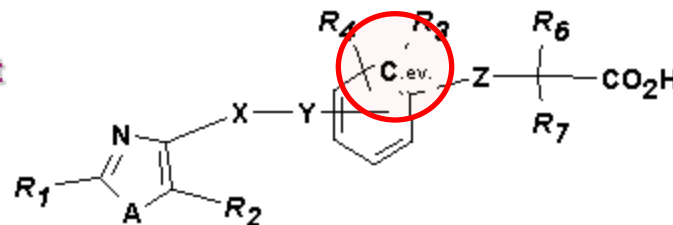
ICImg2Struct



ICImg2Struct



ICImg2Struct





-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
-1.4791	1.2121	0.1059	C

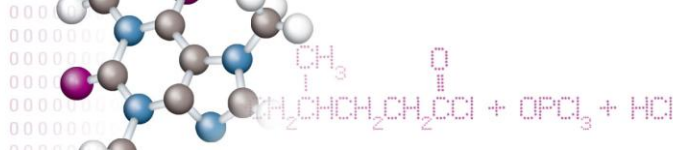
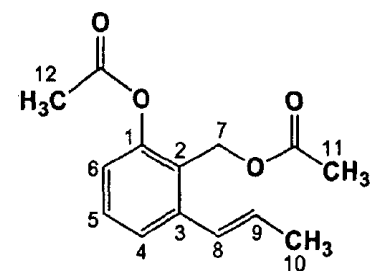
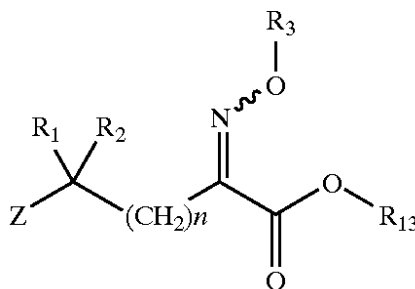
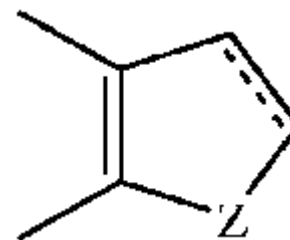
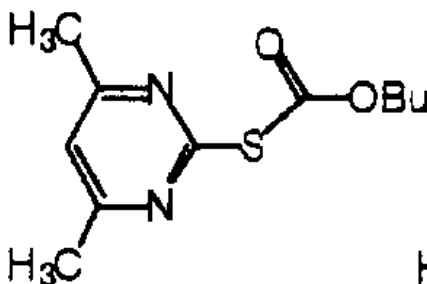
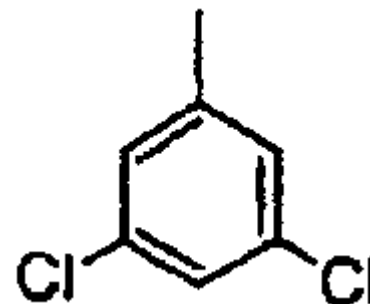
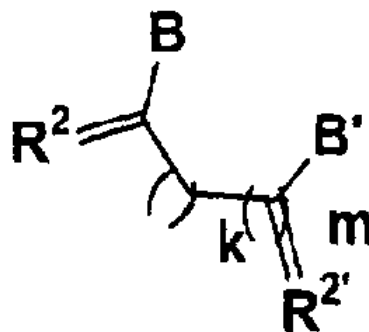


Image Recognition: More Challenges

- » Wavy bonds
- » Atom numbers
- » Brackets
- » Chlorine, Iodine
- » Circles
- » Fused characters
- » Charges
- » Crossing bonds
- » Variable bonds
- » ...



(111)





-7.5909	-1.1813	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
-1.4791	1.2121	0.1059	C

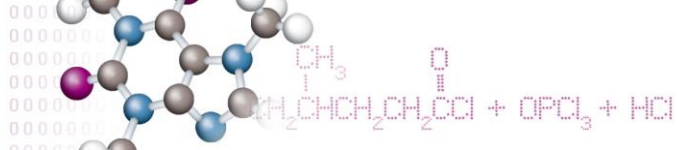
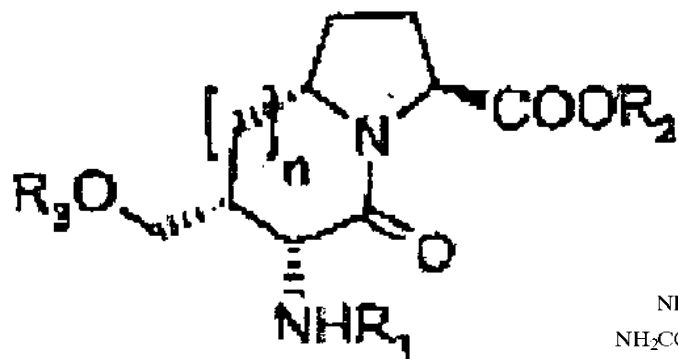
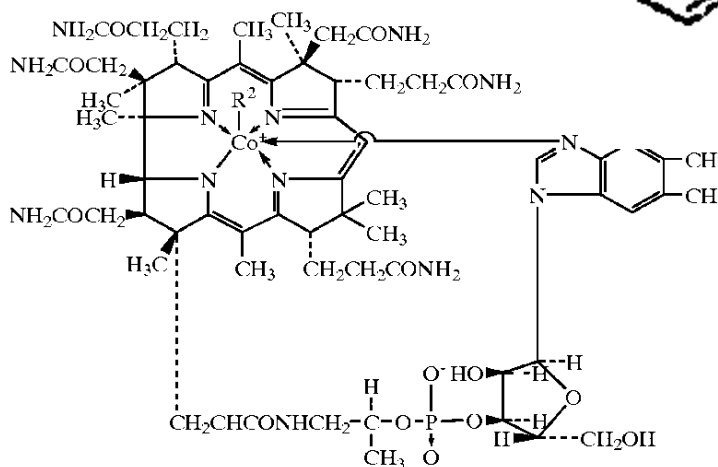
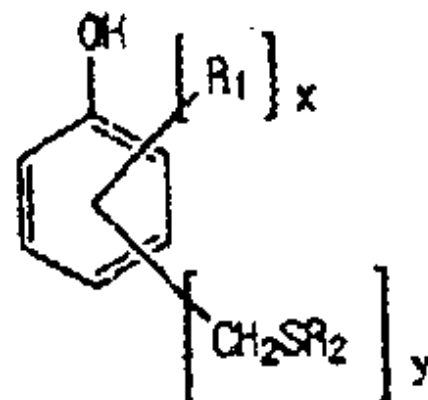
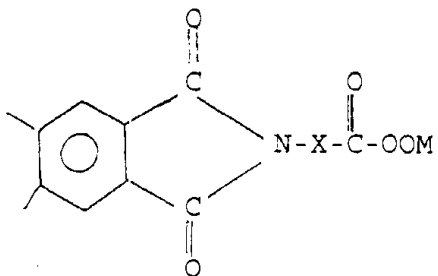
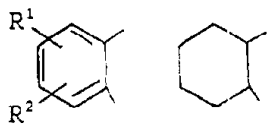


Image Recognition: Even More Challenges

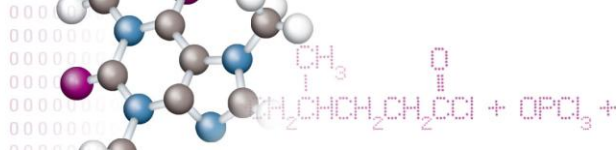


(Iz)



- [0021] R²=CN: Cyanocobalamin (I)
- [0022] R²=OH: Hydroxocobalamin (II)
- [0023] R²=CH₃: Methylcobalamin (V)

-7.5909	-1.1333	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



Combining the Techniques: ChemProspector



» Main emphasis:

*‘The automatic extraction of Markush Structures
from patent documents’*

- » Research SME-project within the THESEUS research program
- » THESEUS: ‘New Technologies for the Internet of Services’
- » Duration THESEUS: five years (2007 - 2011)
- » Duration ChemProspector: July 2009 – end of 2011



Supported by:

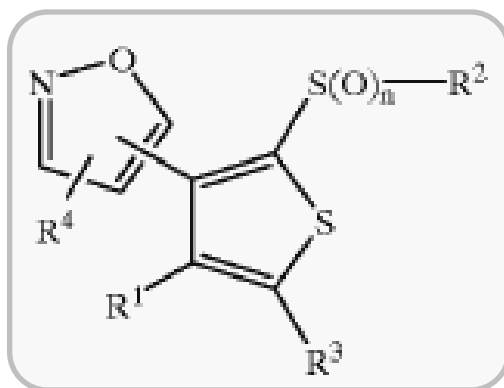


Federal Ministry
of Economics
and Technology

on the basis of a decision
by the German Bundestag

Main Challenges

More specifically, the present invention is directed to a substituted isoxazolylthiophene compound represented by the formula (I)



wherein R^1 and R^2 individually represent an alkyl group of 1–5 carbon atoms, R^3 represents a cyano group or a group $CONR^5R^6$ (in which R^5 and R^6 individually represent a hydrogen atom or an alkyl group of 1–10 carbon atoms), R^4 represents an alkyl group of 1–5 carbon atoms or a phenyl group, and n is an integer of 0–2, or a salt thereof.



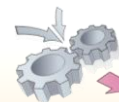
Image classifier



Chemical recognition



ICANNOTATOR



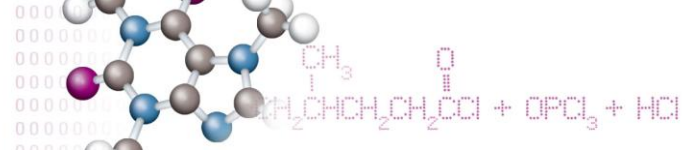
Markush-Parser



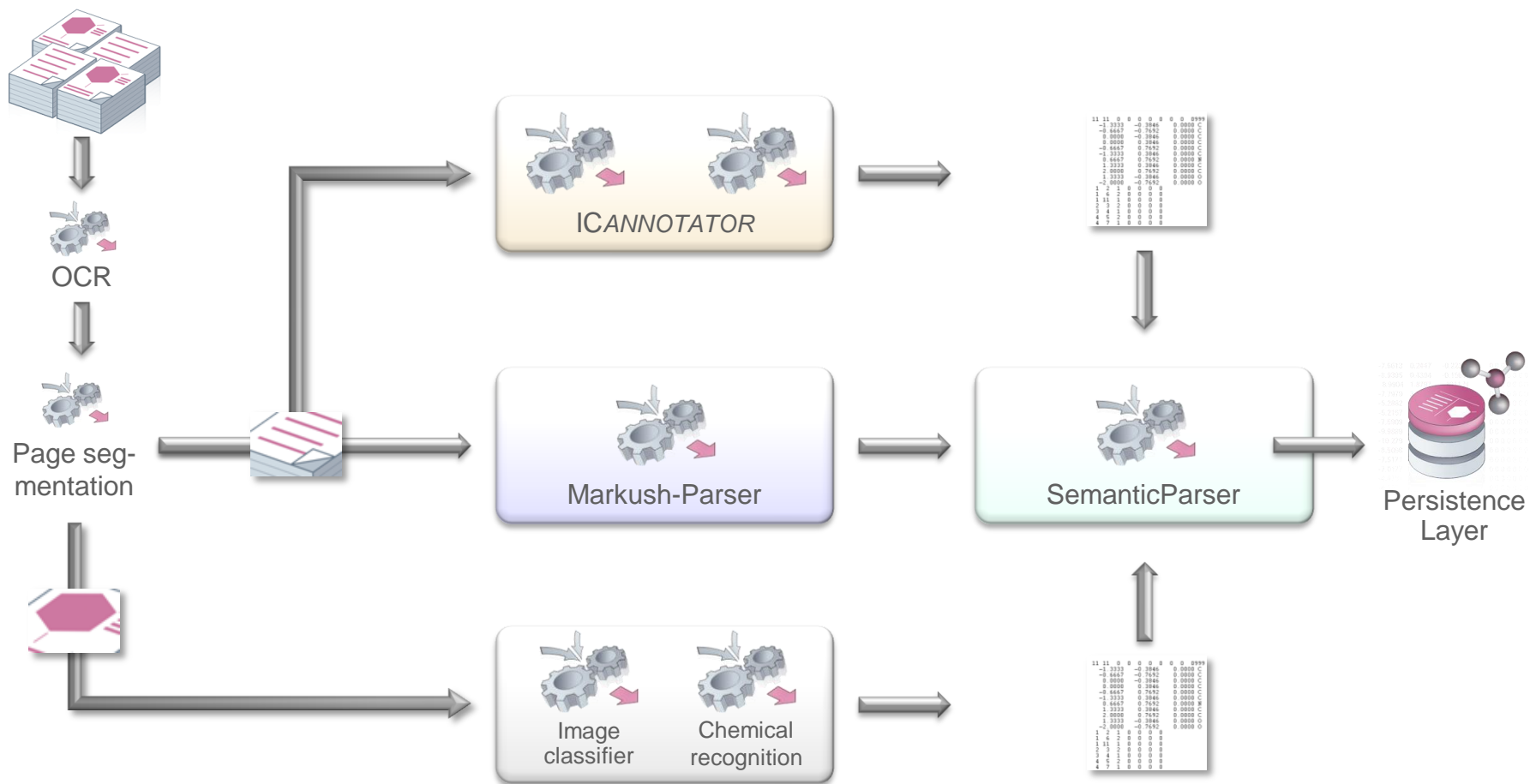
SemanticParser



-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
-1.4794	1.2424	0.1459	C



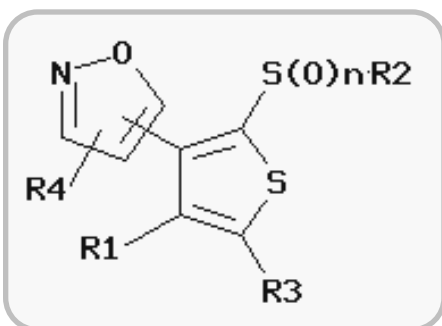
ChemProspector: Approach



Results: Example

start markush description

US06355663-20020312-C00001



end markush description

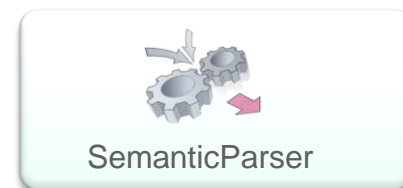
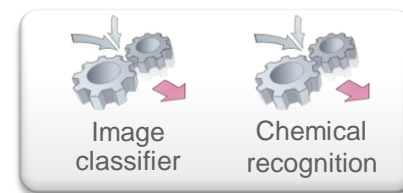
R1 = `alkyl[Size(1-5)]`
 R2 = `alkyl[Size(1-5)]`

R3 = `'cyano'` OR `'CONR5R6'`

R5 = `'H'` OR `alkyl[Size(1-10)]`
 R6 = `'H'` OR `alkyl[Size(1-10)]`

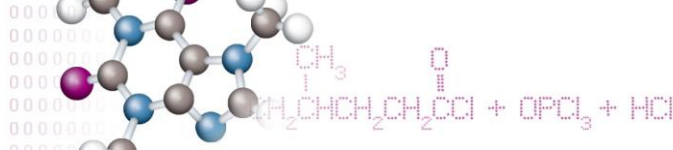
R4 = `alkyl[Size(1-5)]` OR `'phenyl'`

n = 0-2



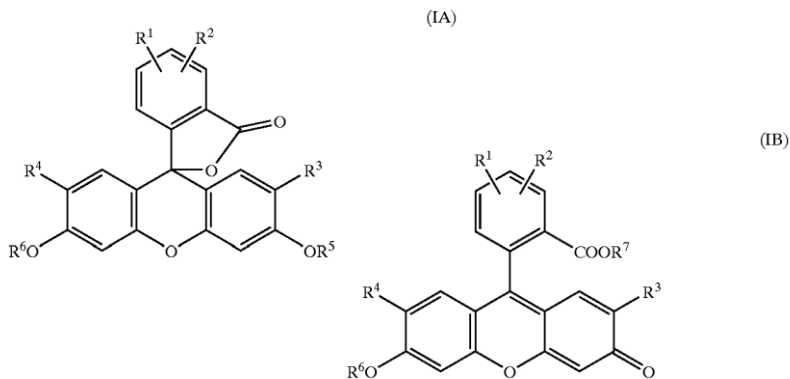


-7.5909	-1.143	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H

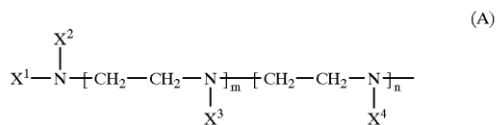


Results: US2003162298 A1

1. A compound represented by general formula (IA) or (IB) or a salt thereof:



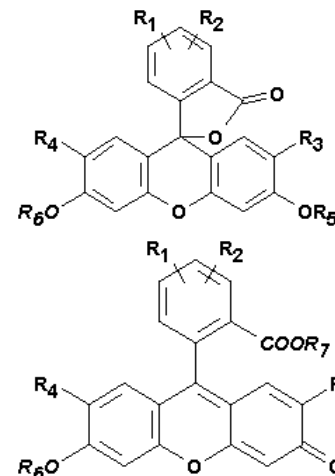
wherein R¹ and R² independently represent a hydrogen atom or a group represented by formula (A):



wherein X¹, X², X³, and X⁴ independently represent a hydrogen atom, an alkyl group, a 2-pyridylmethyl group, or a protective group for an amino group, and m and n independently represent 0 or 1, provided that R¹ and R² do not simultaneously represent hydrogen atoms; R³ and R⁴ independently represent a hydrogen atom or a halogen atom; R⁵ and R⁶ independently represent a hydrogen atom, an alkylcarbonyl group, or an alkylcarbonyloxymethyl group, and R⁷ represents a hydrogen atom or an alkyl group.

start markush description

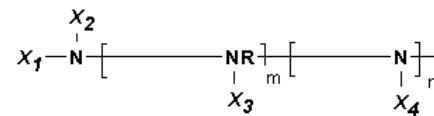
US20030162298A1-20030828-C00018



end markush description

R1 = "H" OR "STRUCT_IN_FORMULA(A)"
R2 = "H" OR "STRUCT_IN_FORMULA(A)"

US20030162298A1-20030828-C00019



X1 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X2 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X3 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X4 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"

m = 0-1

n = 0-1

IF R1 AND R2 NOT SIMULTANEOUSLY = "H"

R3 = "H" OR "Hal"

R4 = "H" OR "Hal"

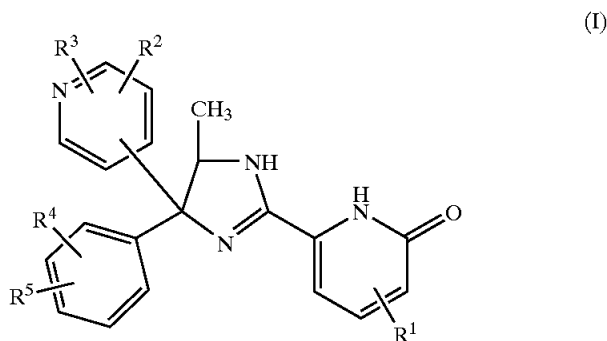
R5 = "H" OR "alkylcarbonyl" OR "alkylcarbonyloxymethyl"

R6 = "H" OR "alkylcarbonyl" OR "alkylcarbonyloxymethyl"

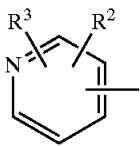
R7 = "H" OR "alkyl"

Results: US 20050154025 A1

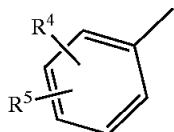
1. A compound of the formula (I):



wherein R¹ is hydrogen, halogen, cyano, lower alkyl, halo-lower alkyl, hydroxy, lower alkoxy or aralkyloxy; R² and R³ are each independently hydrogen, halogen or halo-lower alkyl; and R⁴ and R⁵ are each independently hydrogen or halogen, provided that when R¹ is hydrogen, a group of the formula:

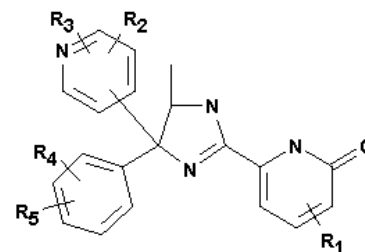


and a group of the formula:



start markush description

US20050154025A1-20050714-C00031



end markush description

R1 = "H" OR "Hal" OR "cyano" OR "alkyl[Size(1-6)]" OR "alkyl" OR "hydroxy" OR "alkoxy[Size(1-6)]" OR "aralkyloxy"

R2 = "H" OR "Hal" OR "alkyl"

R3 = "H" OR "Hal" OR "alkyl"

R2 AND R3 INDEPENDENTLY

R4 = "H" OR "Hal"

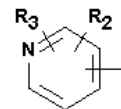
R5 = "H" OR "Hal"

R4 AND R5 INDEPENDENTLY

IF R1 = "H" OR "Struct_E" OR "Struct_F"

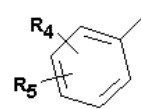
Struct_E

US20050154025A1-20050714-C00032



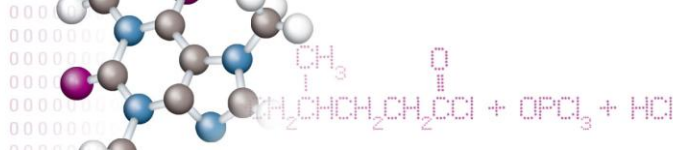
Struct_F

US20050154025A1-20050714-C00033





-7.5909	-1.3883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
1.4794	1.2124	0.1059	F



Acknowledgements

» The InfoChem Team

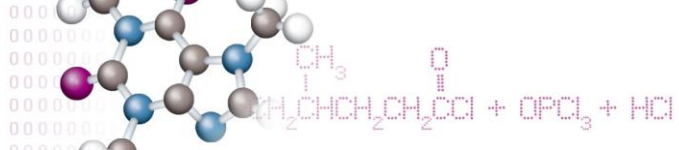


» The German Federal Ministry of Economy and Technology (BMWi)





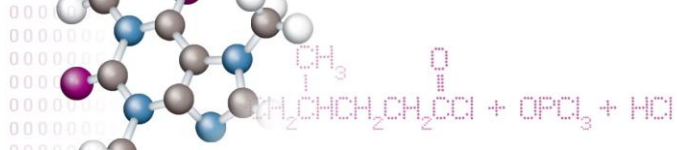
-7.5909	-1.8883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H



Thank you!



-7.5909	-1.8883	-0.4896	C
-9.9889	-0.1947	-0.3218	O
-10.2795	2.5700	0.1393	C
-8.5096	-1.8624	-0.6298	H
-7.5171	-2.3313	0.2821	H
-7.0177	-1.8762	-1.3207	H
1.4791	1.8121	0.1050	C



Questions?