
Towards automated mining of chemical structures in Chinese Patents

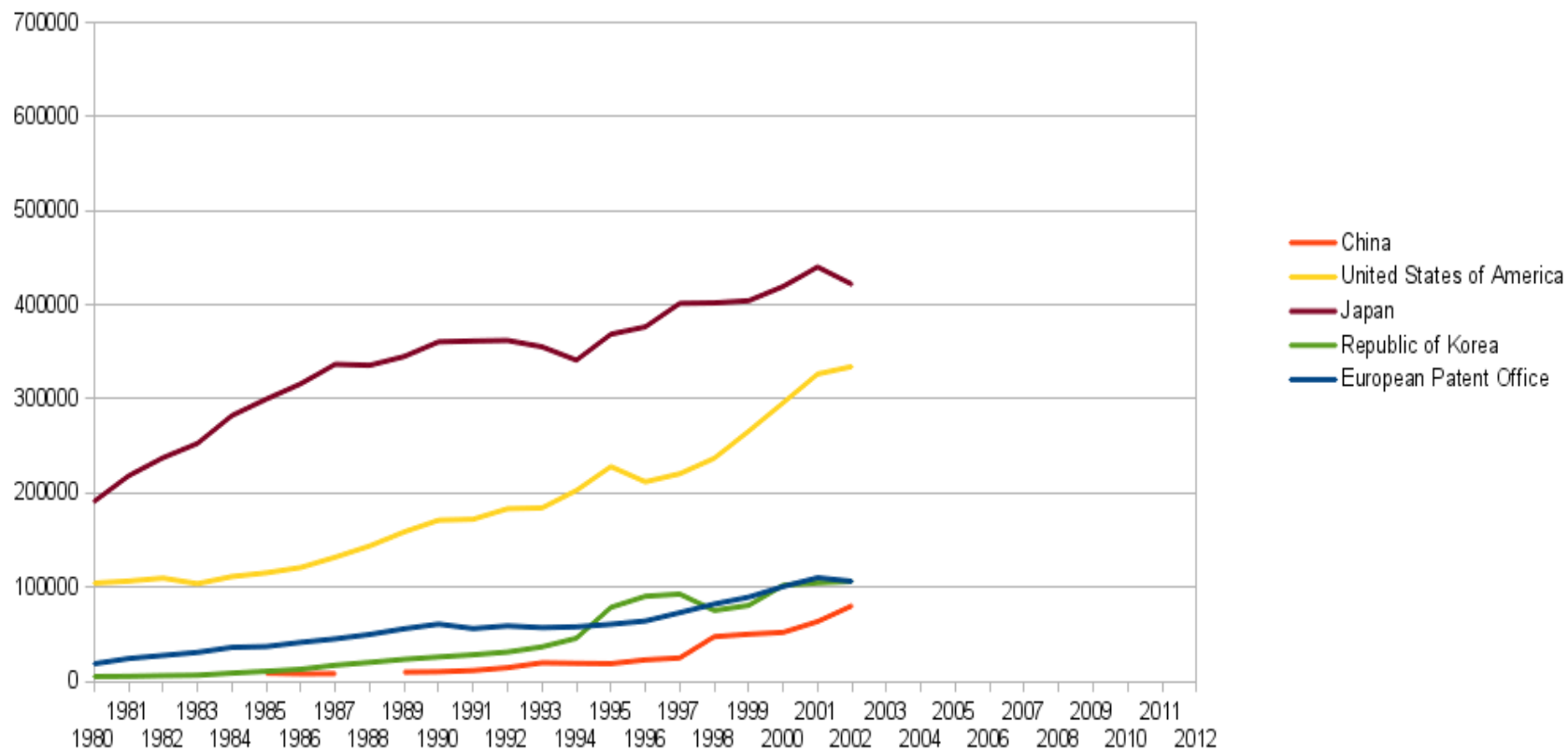
Daniel Bonniot de Ruisselet
ChemAxon

ICIC 2013, Vienna
October 16th 2013



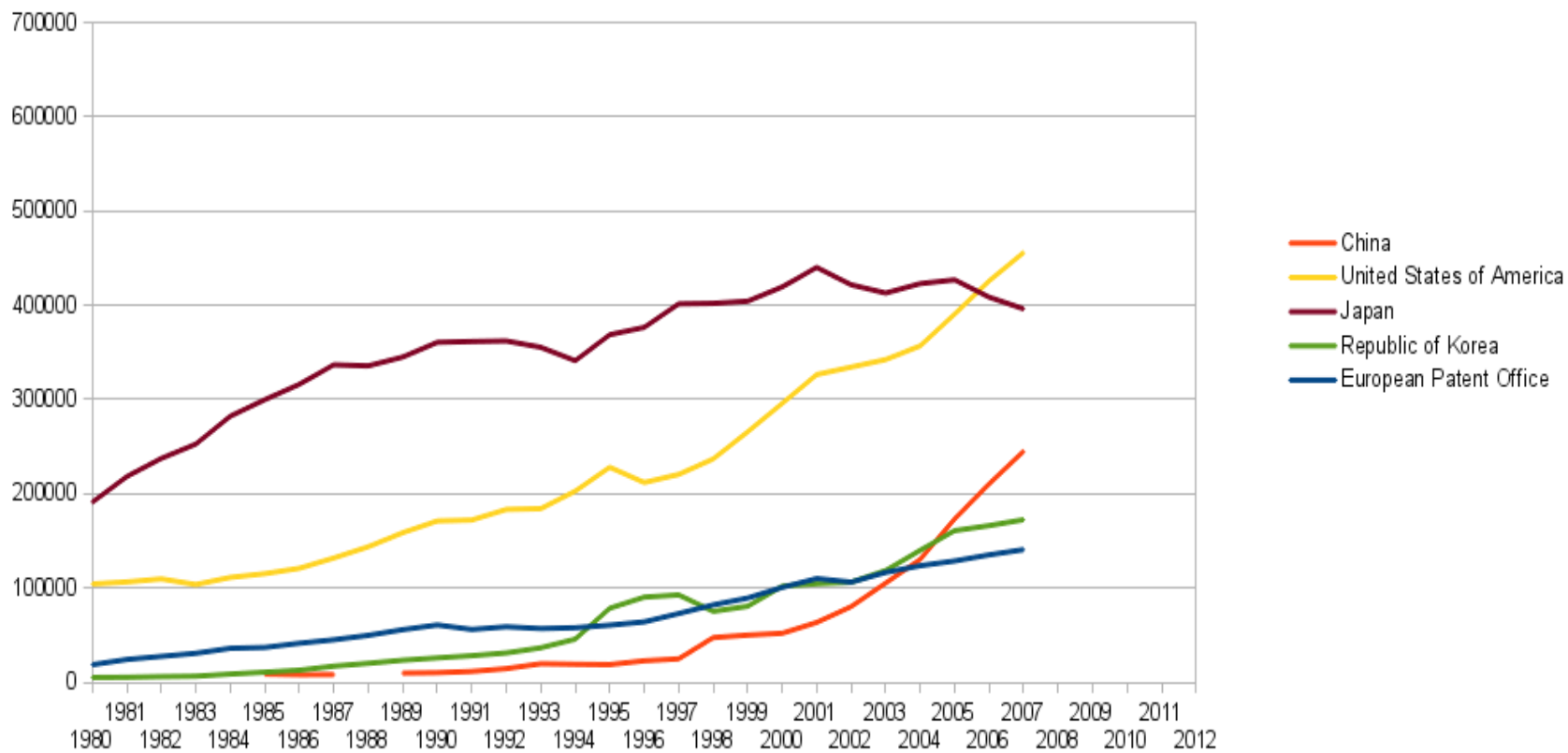
Patent filings at IP5 Offices 1980-2002

(source: WIPO)



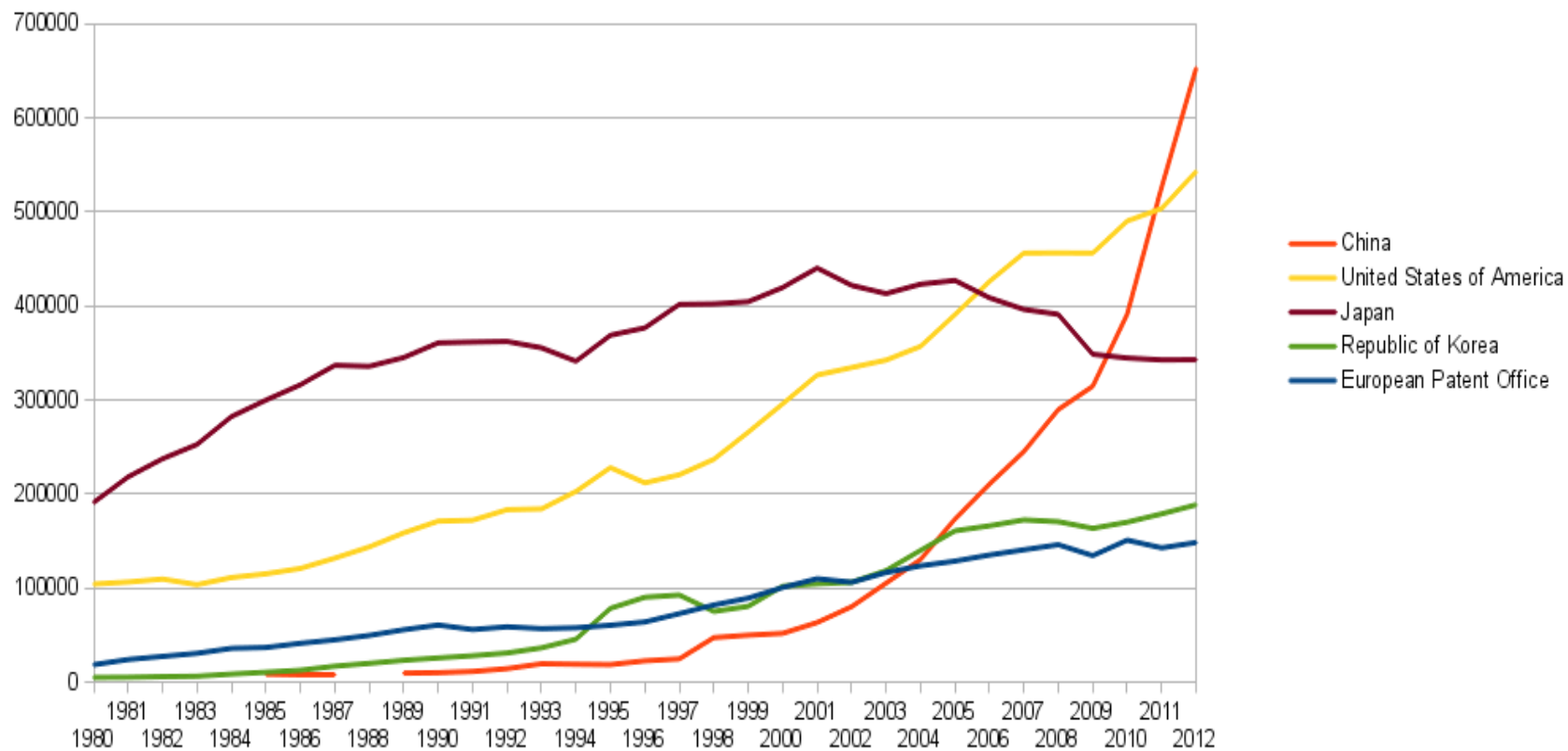
Patent filings at IP5 Offices 1980-2007

(source: WIPO)



Patent filings at IP5 Offices 1980-2012

(sources: WIPO, IP5)



Why Chinese patents matter

- Volume, exploding...
- Increasingly innovative
- Potential infringement, lawsuits
 - Apple (2008, 2012, 2013), Schneider Electric, Samsung, ...
- Hard to access because of language

Why chemical mining matters

- Find interesting patent(s) using text search
 - Each patent can contain 100s of chemical names
 - Convert them automatically to structures
 - Enables chemical calculations
- Find interesting patent(s) using chemical structure search
 - Requires building a chemical database index
- Track structures accross multiple patents
 - Including multiple languages
 - Searching for prior art, infringement, ...
 - Chemical similarity search
- ...

Chinese patents matter

&

chemical mining matters

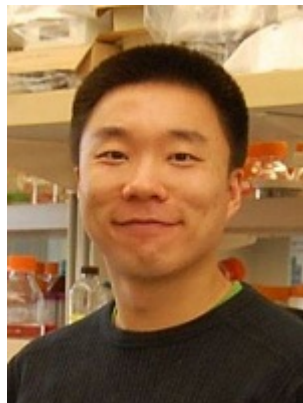
→

**Chemical mining of chinese
patents matters**

ChemAxon?

- Cheminformatics, since 1998
- All of the top 15 global pharmas are customers
- Chemical database: indexing and searching
- English Name to Structure
- Document to Structure
- Missing piece: Chinese Name to Structure

Chinese Name to Structure



邓巍 (Wei Deng, a.k.a. David)

Builds on english name to structure

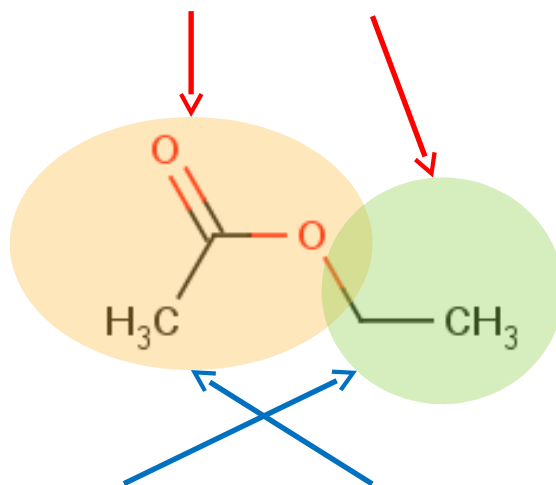
Specific dictionaries

Changes in algorithms...

The Challenges

1. Chinese texts have no spaces
2. Ester & Salt

乙酸乙酯



Ethyl Acetate

The Challenges

3. English: name alterations

丁烷 → buta + ane → butane

4. Chinese: many Characters have different meanings

盐 = salt

酸 = acid

盐酸 = hydrochloric acid

OCR Error correction

3-(笨基) 丙酸

笨 = benzene
笨基 = phenyl
丙酸 = proprionic acid

The screenshot shows the Google Translate interface. The source text is "3-(笨基)丙酸" (3-(phenyl)propionic acid) in Chinese. The interface is set to translate from Chinese to English. The target text area is empty, indicating a translation error. The interface includes the Google logo, a "SIGN IN" button, and various controls for language selection and translation.

Google

SIGN IN

Translate

From: Chinese To: English Translate

Chinese English Spanish Detect language

English Chinese (Simplified) Chinese (Traditional)

3-(笨基)丙酸

拼

ChemAxon

OCR Error correction

3-(笨基) 丙酸

苯 = benzene
苯基 = phenyl
丙酸 = propionic acid

The screenshot shows the Google Translate interface. The source text is "3-(笨基)丙酸" (Chinese). The target text is "3 - (stupid yl) propionic acid" (English). The interface includes the Google logo, a "SIGN IN" button, and translation controls. The source language is set to "Chinese" and the target language is "English". The translation is displayed in a box with a close button (X) and a microphone icon. The target text is displayed in a box with a star icon, a list icon, a speaker icon, a chat icon, and a checkmark icon.

Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 {5-[2-(4- 正辛基苯基) 乙基]-2 , 2- 二甲基 -1 , 3- 二氧六环 -5- 基 } 氨基甲酸叔丁酯是合成芬戈莫德及其衍生物的重要中间体。

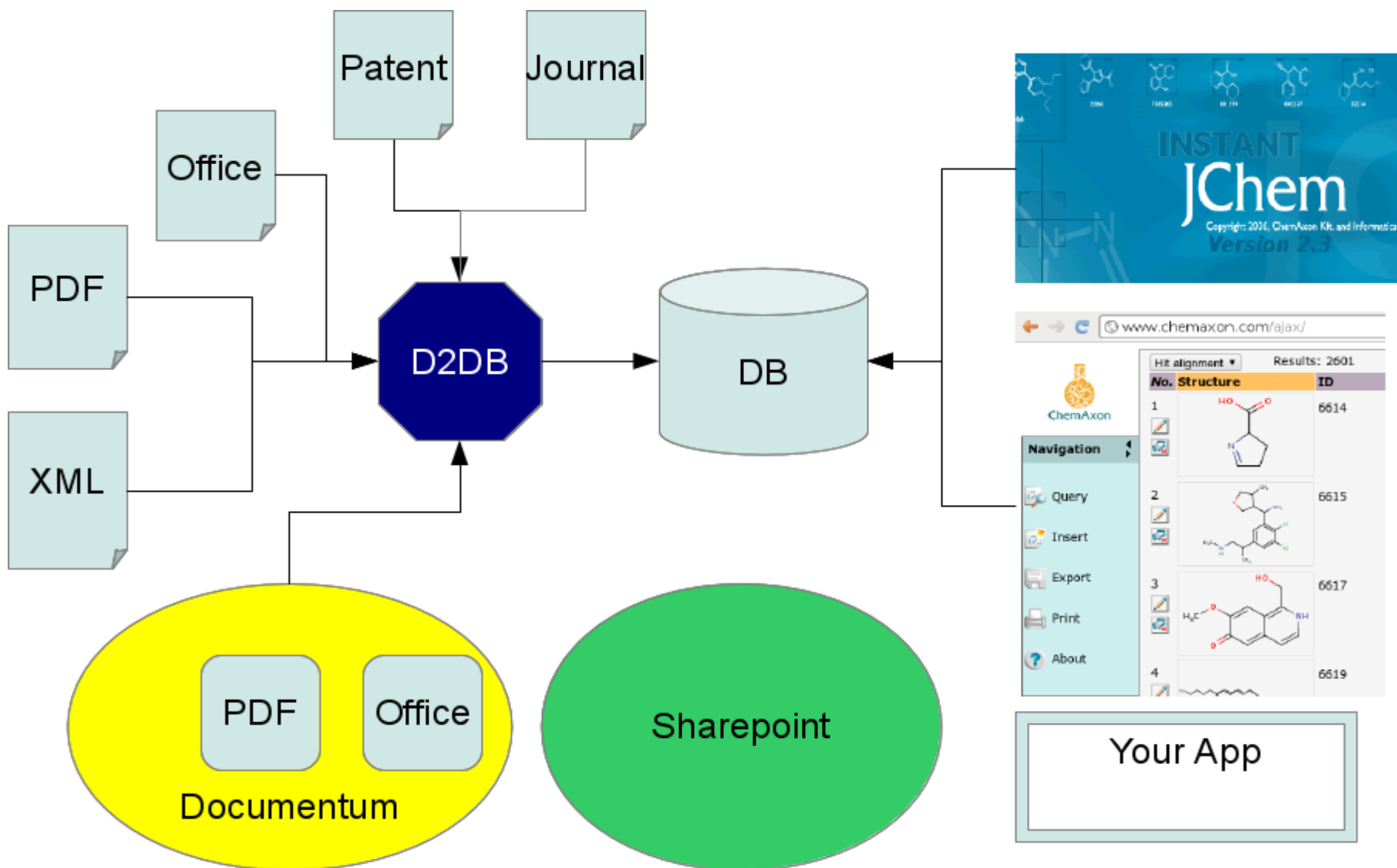
Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 {5-[2-(4-正辛基苯基)乙基]-2, 2-二甲基-1, 3-二氧六环-5-基} 氨基甲酸叔丁酯是合成芬戈莫德及其衍生物的重要中间体。

Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 {5-[2-(4-正辛基苯基)乙基]-2, 2-二甲基-1, 3-二氧六环-5-基} 氨基甲酸叔丁酯是合成芬戈莫德及其衍生物的重要中间体。
- XML Markup
 - Patent metadata
 - Encoding of characters
 - Tags (e.g. <p>)
- Document annotation

Document to Database



Document to Database

D2DB Search - Chromium

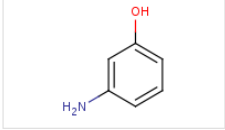
D2DB Search

<https://d2db.chemaxon.com/uspto/?jsessionid=D7A758FEF59A07D43FE5BD0A4C334434?0>

D2DB Search

Structure

Search type
SUBSTRUCTURE

Nc1ccc(O)cc1

Formula

Molecular Weight Range
 -

Options

Document Properties

Title

Assignees

Patent Id

Inventors

Submit Clear

Document to Database

D2DB Search - Chromium

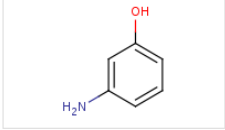
D2DB Search

https://d2db.chemaxon.com/uspto/?jsessionid=D7A758FEF59A07D43FE5BD0A4C334434?0

D2DB Search

Structure

Search type
SUBSTRUCTURE



Formula

Molecular Weight Range
 -

Options

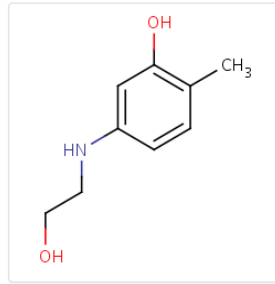
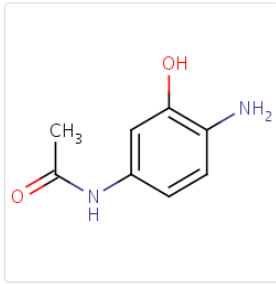
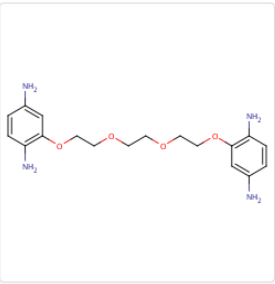
Document Properties

Title

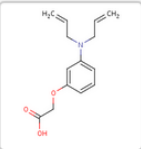
Assignees

Patent Id

Inventors



OH



Molecular Formula
C₁₄H₁₇NO₃

Molecular Weight
247.2897

SMILES
OC(=O)COC1=CC(=CC=C1)N(CC=C)CC=C

Documents

Patent Id	Title	Assignees	IPCR	Context
US-08088182-B2-20120103	Azo dyes, a process for the	Huntsman Interna	C09B 29/036	is prepared by addition of 5.3 g of 3-diallylamino
US-08088182-B2-20120103	Azo dyes, a process for the	Huntsman Interna	C09B 29/036 (c)	3-Diallylamino-phenoxyacetic acid methylester
US-08088182-B2-20120103	Azo dyes, a process for the	Huntsman Interna	C09B 29/036	of 4.6 g of sulfuric acid and 4.5 g of 3-diallylamino

Document to Database

D2DB Search - Chromium

https://d2db.chemaxon.com/uspto/?jsessionid=D7A758FEF59A07D43FE5BD0A4C33443470

D2DB Search

Structure

Search type
SUBSTRUCTURE

Formula

Molecular Weight Range

Options

Document Properties

Title
Search or select a value

Assignees
Search or select a value

Patent Id

Inventors

Submit Clear

Chemical structures shown in the search results include:

- A benzene ring with an amino group (H₂N) and a hydroxyl group (OH).
- A complex molecule with two benzene rings connected by a chain of ether linkages, with amino groups on the rings.
- A benzene ring with a hydroxyl group (OH) and an amino group (NH₂).
- A benzene ring with a hydroxyl group (OH) and a methyl group (CH₃).
- A benzene ring with a hydroxyl group (OH) and a methyl group (CH₃).
- A benzene ring with a hydroxyl group (OH) and a methyl group (CH₃).

Documents

Patent Id

US-08088182-B2-2012010

US-08088182-B2-2012010

US-08088182-B2-2012010

本发明活性化合物的制备和应用可见下述实施例。
-->制备实施例

实施例1

(方法(a))

将0.76g(2.9mmol)5-乙氧基-4-甲基-2-苯氧基咪唑-2,4-二氢-3H-1,2,4-三唑-3-酮溶于40ml乙腈中,在室温(约)拌下,以每次少量的方式与0.75g(3.2mmol)4-甲氧基咪唑基咪唑-3-磺酰胺和0.49g(3.2mmol)1,8-二氮杂二环[5.4-1-碳-7-烯(DBU)混合。将该反应混合物在室温搅拌12小时压浓缩。将残余物置于二氯甲烷中,依次用1N盐酸和水洗液酸干燥,并过滤。将滤液在水泵真空下浓缩,将残余物用煮,通过抽滤分离出所得结晶产物。

获得了0.70g(理论产率的60%)4-[[[(3-乙氧基-4,5-二氢-4-甲基-5-氧代-1H-1,2,4-三唑-1-基)咪唑基]磺酰基]苯-3-甲酸甲酯(别名5-乙氧基-4-甲基-2-[(4-甲氧基咪唑基-2-咪唑-3-基)磺酰基氨基咪唑基]-2,4-二氢-3H-1,2,4-三唑-3-酮点为163°C。

按照与实施例1类似的方法,并依据本发明制备方法的一般描述,还可以制备例如在下表1中列出的通式(I)化合物。

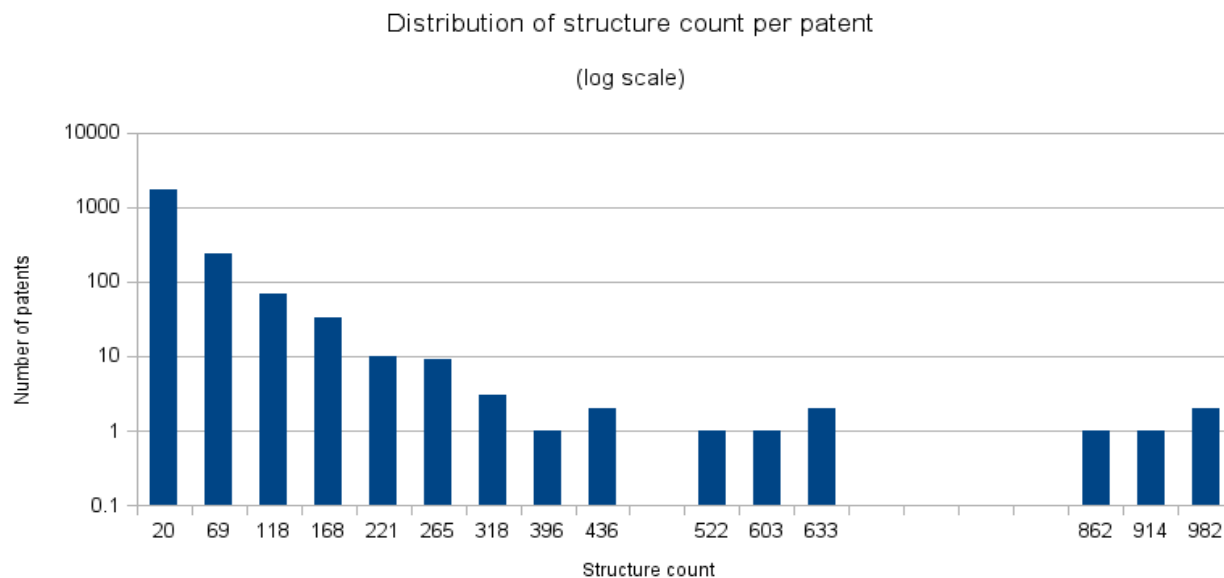
Chemical structure of the product: 5-乙氧基-4-甲基-2-苯氧基咪唑-2,4-二氢-3H-1,2,4-三唑-3-酮

Validation 1: Chinese name to structure

- Test set: 38,600 Chinese names + CAS number
- Contains unusual, incorrect, ambiguous names, radicals, inorganic salts,
- Conversion rate = 59 – 79 %
- Accuracy = 91%

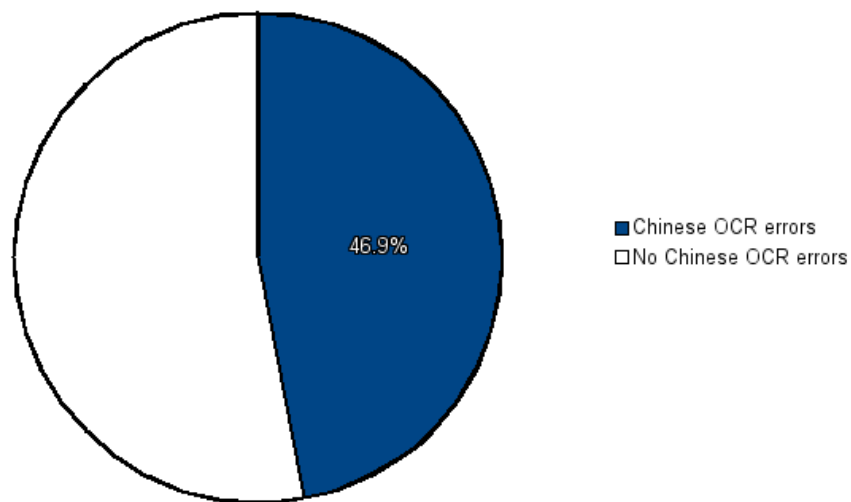
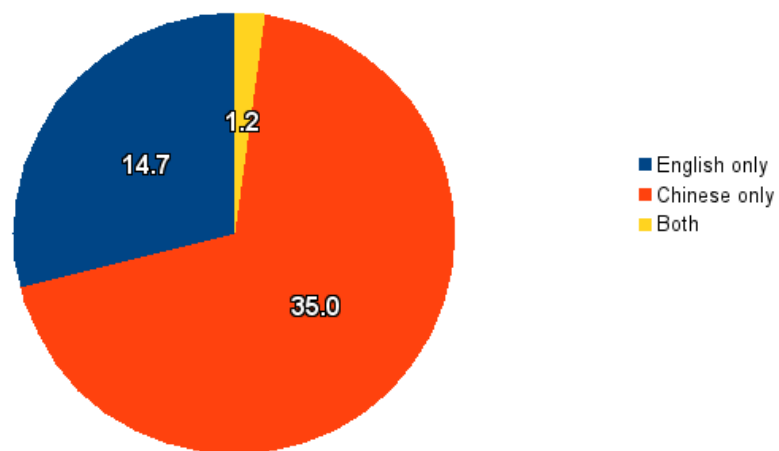
Validation 2: Chinese patents

- 54K chinese patents with automated english translation
- Filter: structures with at least 20 heavy atoms, and patents with at least 20 structures
- Remains: 2108 patents



Validation 2: Chinese patents

Origin of unique chemical structures



Conclusions

- Patent volume in chinese is booming
- It is important to mine & monitor it
- Automated solutions are needed, but hard
- General purpose auto translation is not enough
- Chinese N2S already gives better results
- ChemAxon can build solutions for specific workflows
- More collaboration with patent providers is needed to keep improving quality and solutions

谢谢！

Extra information

谢谢！

Automatic OCR Error Correction

(2R)-2-**rn**ethylsulfany**y**1-3-hydr**0**xybutanedi**0**ate



(2R)-2-methylsulfanyl-3-hydroxybutanedioate

Ar-benzyl-**Ar**-[3-(**1**H-tetrazol-5-yl)phenyl]propanamide





N-benzyl-*N*-[3-(1H-tetrazol-5-yl)phenyl]propanamide

我们**日**前**止**在研究开**友**中文化**字**名称的 OCR **自**动**纠**错**工**力**能**



我们目前正在研究开**发**中文化**学**名称的 OCR **自**动**纠**错**功**能

From Document to Structures

(19)  (11)  EP 2 377 850 A1

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication: 19.10.2011 Bulletin 2011/42 (51) Int. Cl.: C07D 239/02 (2006.01) C07D 235/02 (2006.01)
 C07D 263/02 (2006.01) C07D 265/02 (2006.01)
 C07D 401/12 (2006.01) C07D 403/12 (2006.01)
 C07D 407/12 (2006.01) A61K 31/04 (2006.01)

(21) Application number: 10158292.2 (22) Date of filing: 30.03.2010

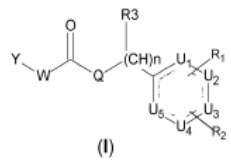
(74) Representative: Minoja, Fabrizio
 Bianchetti, Bracco Minoja S.r.l.
 Via Pirena, 63
 20129 Milano (IT)

(71) Applicant: Pharmeste S.r.l.
 44100 Ferrara (IT)

(72) Inventors:
 • NAPOLETANO, Mauro
 20127, MILANO (IT)

(54) TRPV1 vanilloid receptor antagonists with a bicyclic portion

(57) The invention discloses compounds of formula



(I)










wherein Y is selected from a group of formula

EP 2 377 850 A1

Printed by Jouve, 70027 MARSEILLE (FR) (Cont. next page)

EP2377850A1_5.10_nb401.mrv - MarvinView 5.10.0

File Edit View Table Structure Tools Help

#	structure	type	confidence	page	document	text
					backup\0Current\0 Id\20120313_Che	
470		common		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	MeOH
471		common		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	sodium sulfate
472		common		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	DMSO
473		systematic		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	1-(4-chloro-2-(3-hydroxypropylidino-1-yl)benzyl)-3-(2,3-dihydro-2-oxo-1H-benz[e][1,2,4]oxadiazol-5-yl)propan-1-amine
474		systematic		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	4-chloro-2-(3-hydroxypropylidino-1-yl)benzylidene nitrile
475		systematic		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	2-fluoro-4-chlorobenzonitrile
476		common		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	DMSO
477		systematic		19	C:\Users\David\Dropbox\ChemAxon_backup\0Current\0 Id\20120313_Che	1-(2-(aminomethyl)-5-chlorophenyl)propylidino-3-ol

ChemAxon's "Document to Structure"

- Extract chemical information from documents
 - Names: powered by the Naming Technology
 - Also import SMILES, InChI, CAS number ...
 - Images: OSRA, ...
 - Works with **scanned non-searchable PDF**
 - Returns structures and their **location** in the document

ChemAxon's "Document to Structure"

- Supported formats:
 - MS Office document: doc, docx, ppt, pptx, xls, xlsx, odt ...
 - Embedded structure objects (ChemDraw, Symyx, Marvin, ...)
 - PDF, text, XML, HTML

ChemAxon's "Document to Database"

- Data in DB:
 - Structures
 - Source (name, smiles, embedded, ...) and location
 - Documents, Authors, Metadata...
- Questions:
 - What structures appear in a specific document?
 - What documents contain a structure/substructure/...?
 - What documents written since 2010 in location X contain substructure S?
 - ...