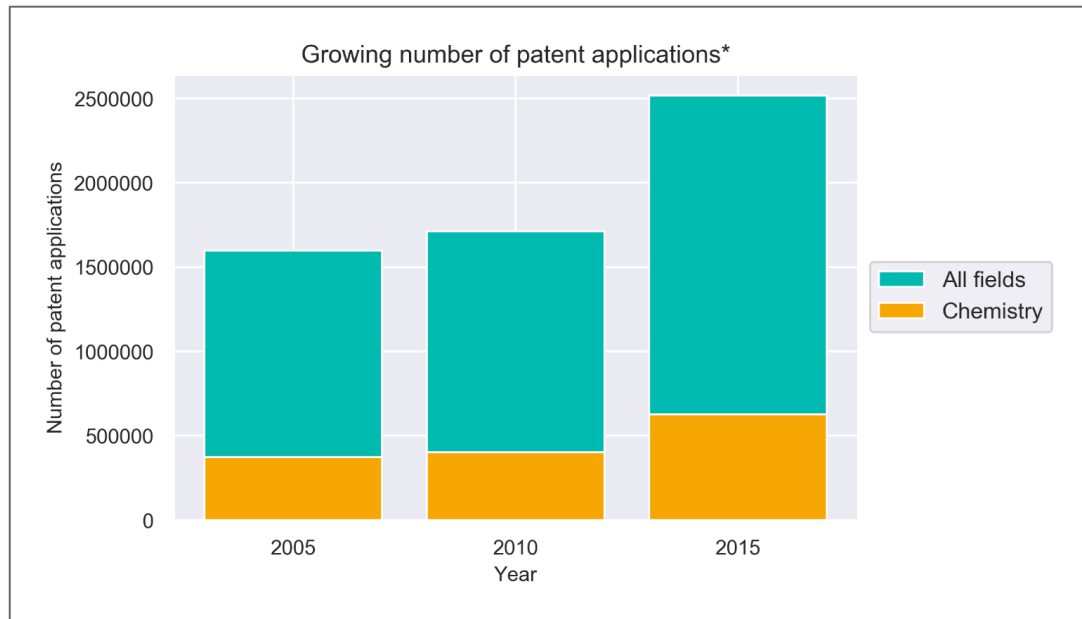
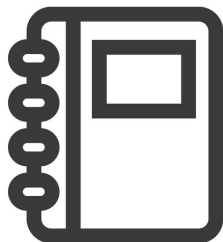


How to reach the hidden world of chemistry in patents

How to reach the hidden world of chemistry in documents

THE PROBLEM

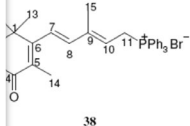


*World Intellectual Property Indicators 2017 Patents,
https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2017-chapter2.pdf



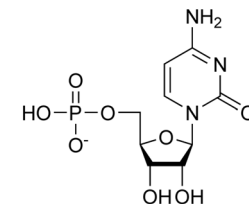
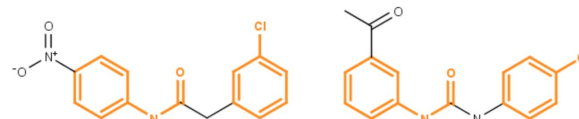
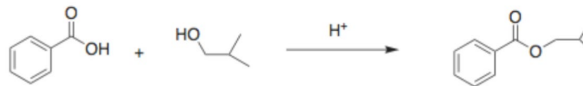
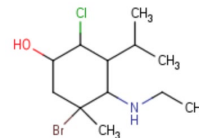
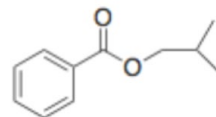
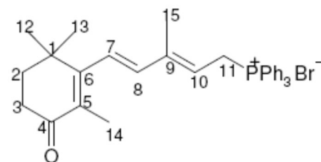
4-Oxo-β-ionylidene-ethyl-triphenylphosphonium bromide (38):

To a solution of PPh₃HBr (10.3 g, 28.6 mmol) in anhydrous methanol (100 mL) was added 36 (6.7 g, 28.6 mmol) in methanol (100 mL). After stirring for 75 h at room temp., the solvent was evaporated and 38 (10.57 g, 18.9 mmol, 66 % lit.³⁰ yield; 84 %) was obtained as a yellow solid. (m.p. 50 °C). The NMR spectra correspond to the



v⁺-Br⁻.

cm⁻¹ (w), 1655 (s, C=O), 1588 (w), 1480 (s), 1107 (s), 756



List of Schemes

Model Fischer Esterification with 1 and 2

1 Background

Our research project requires a target ester, whose starting material requires 15 synthetic steps. We decided to optimize the conditions of this Fischer esterification by preparing a closely related analogue 3. The starting materials in this model reaction is the readily available benzoic acid (1) and *i*-PrOH (2).



United States Patent

[19]

[11] Patent Number: 5,859,006

[43] Date of Patent: Jan. 12, 1999

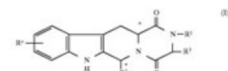
ABSTRACT

Primary Examiner—Makand J. Shah
Assistant Examiner—Tamara T. Ngo
Attorney, Agent, or Firm—Marshall, O'Toole, Gerstein,
Murray & Borun

Wash.

ABSTRACT

[57] A compound of formula (I)



and salts and solvates thereof, in which:
R¹ represents hydrogen, halogen or C₁₋₆ alkyl;
R² represents hydrogen, C₁₋₆ alkyl, C₂₋₆ alkenyl,
C₂₋₆ alkynyl, haloC₁₋₆ alkyl, C₂₋₆ cycloalkyl,
C₂₋₆ cycloalkyl(C₁₋₆ alkyl), aryl(C₁₋₆ alkyl) or
heteroaryl(C₁₋₆ alkyl); R³ represents an optionally sub-
stituted monocyclic aromatic ring selected from
benzene, thiophene, furan and pyridine or an optionally
substituted bicyclic ring.

9401000
ADVN 43/42
CYTD 471.00
SQ: 514/292



United States Patent Office

3,150,632
Patented Dec. 8, 1966

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37



UNSTRUCTURED → STRUCTURED



ChemLocator



ChemLocator

Aspirin, also known as acetylsalicylic acid (ASA), is a medication used to treat pain, fever, or inflammation.[4] Specific inflammatory conditions in which aspirin is used include Kawasaki disease, pericarditis, and rheumatic fever.[4] Aspirin given ...



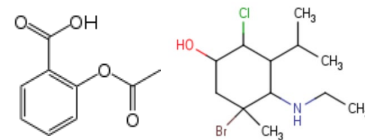
pptx	ppt	one	cdx	mrv	mol
pdf	html	doc	docx	xls	xlsx
iupac	cdxml	inChi	cas	smiles	rxn
rdf	email	aspx	xml	sdf	etc..

Free text indexing
ChemAxon Naming +

Semantic indexing



Chemical NER*
OCR
OSR
Metadata extraction
Tagging
Keywords
Biological NER
etc.



*NER: Named Entity Recognition



24,000 patent documents

- From the **USPTO** patent database
- Medicinal chemistry patents from 2015



771,000 unique molecules

- exact molecules: 740,000
- fragments: 23,000
- generic structures: 8,800

3.1 million molecules in total

An Experiment

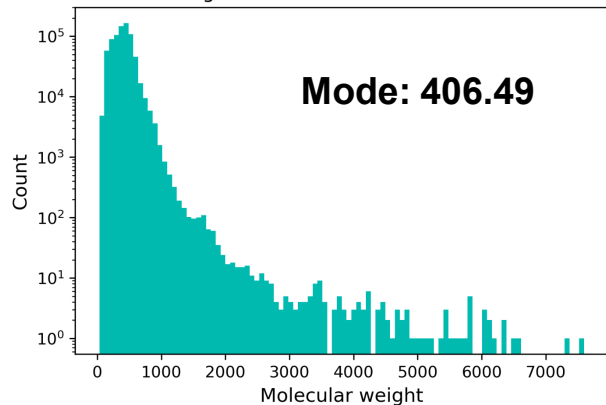


Processing

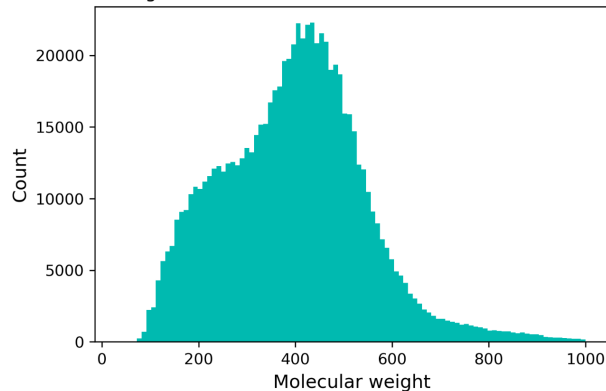
- Indexing time: 7.08 days
- Chemical database: 13.5 GB
- Elastic database: 25 GB

Chemical space - Basic properties

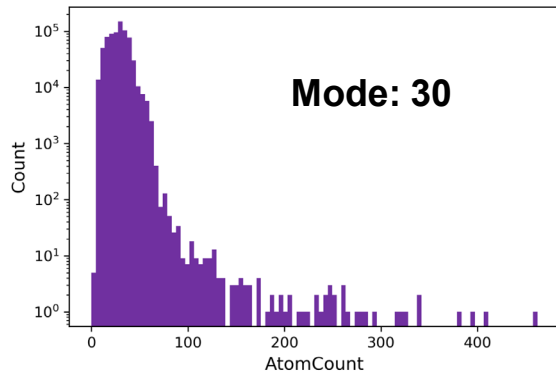
Molecular weight distribution of the extracted molecules



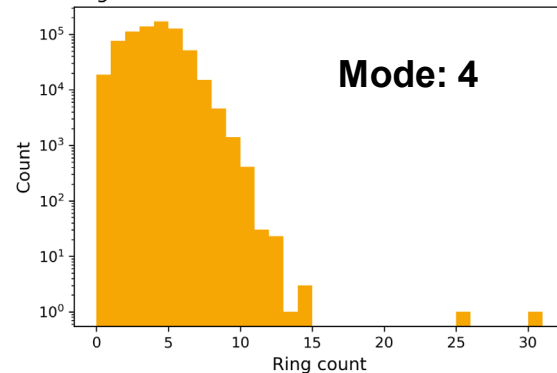
Molecular weight distribution of the extracted molecules for MW < 1000



AtomCount distribution of the extracted molecules

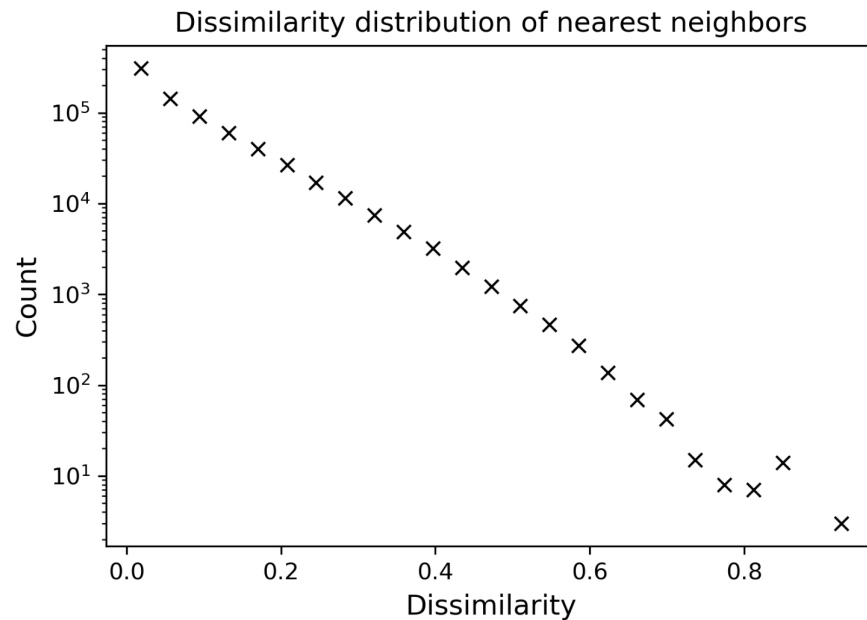
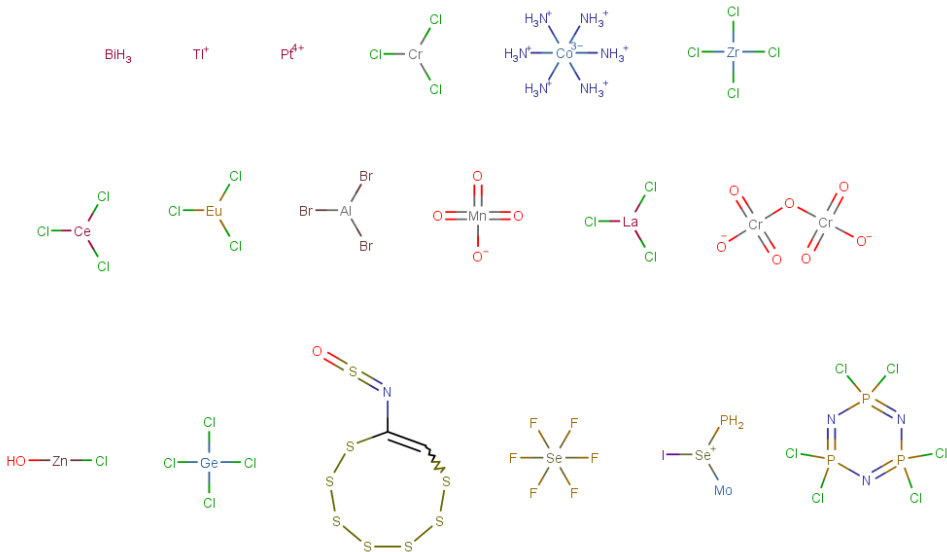


RingCount distribution of the extracted molecules

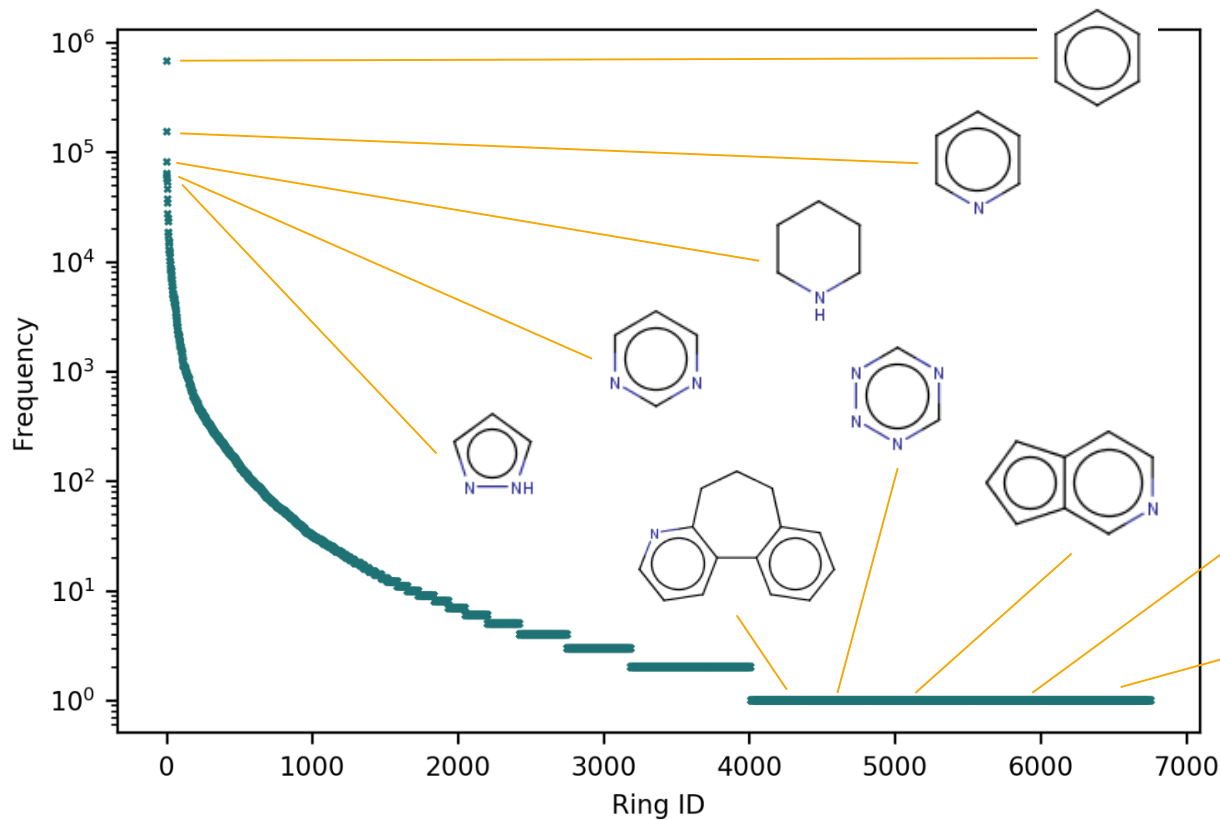


Chemical space - Structural similarity

- 50% of the molecules had its most similar pair within 0.05 dissimilarity;
- Molecules that are very different from their nearest neighbors:



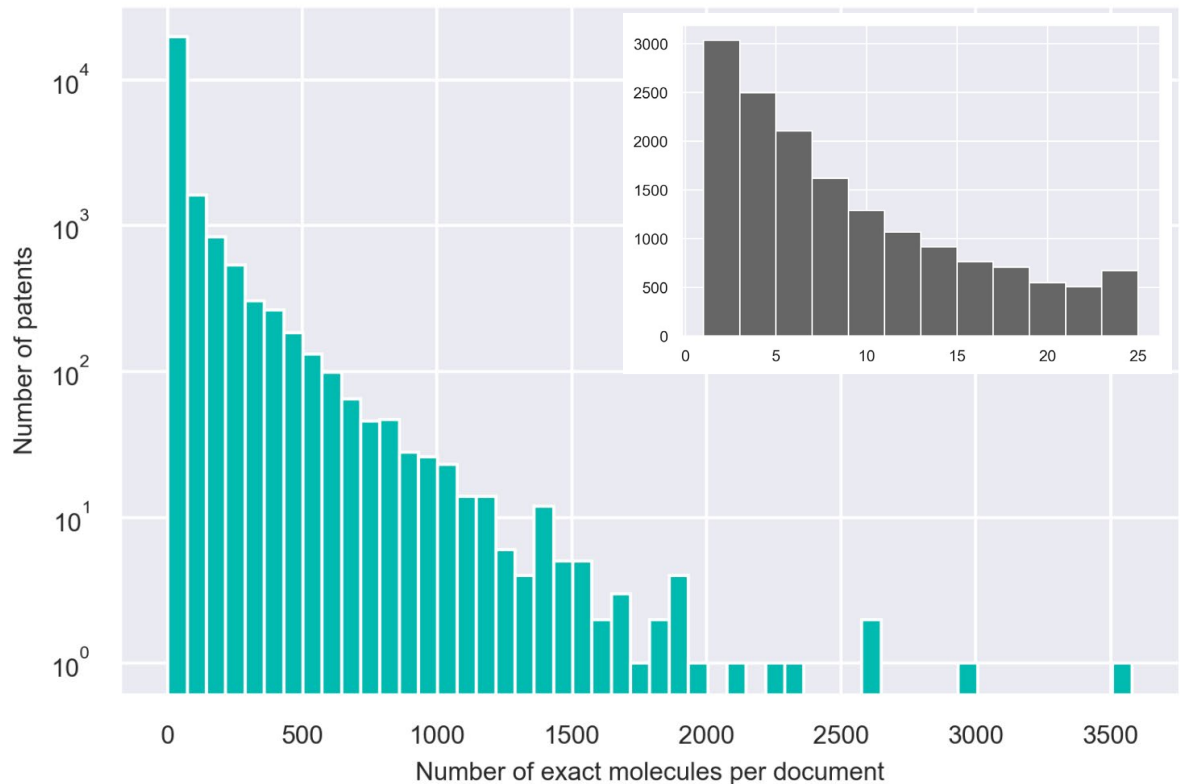
Chemical space - Typical ring systems



- 6756 different ring systems;
- 2745 rings occur only once;

Chemical structures per document

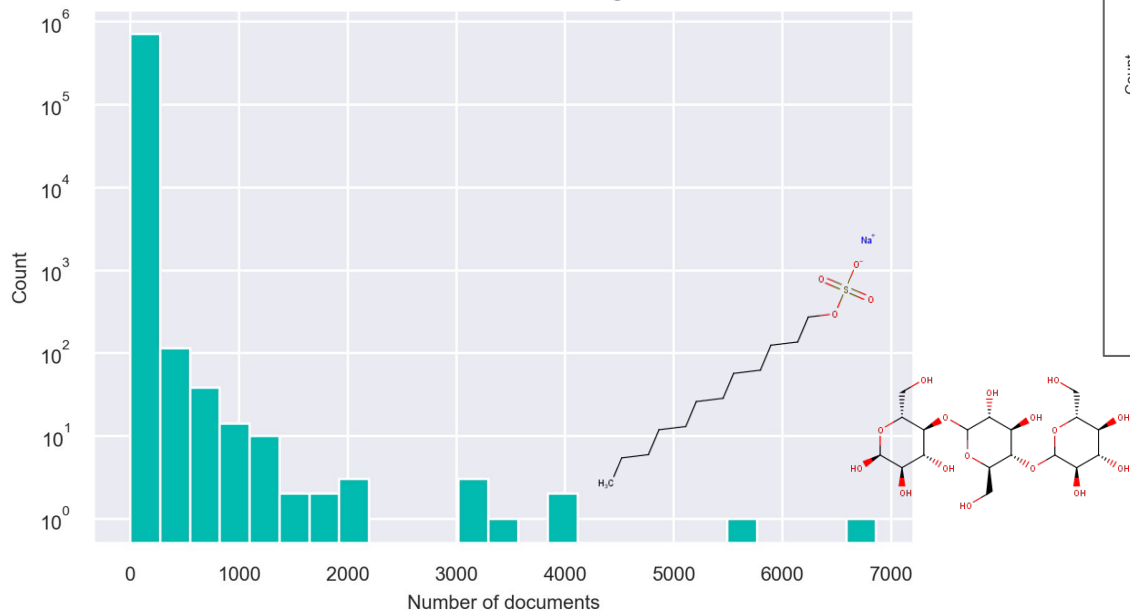
Distribution of number of exact structures per patent document



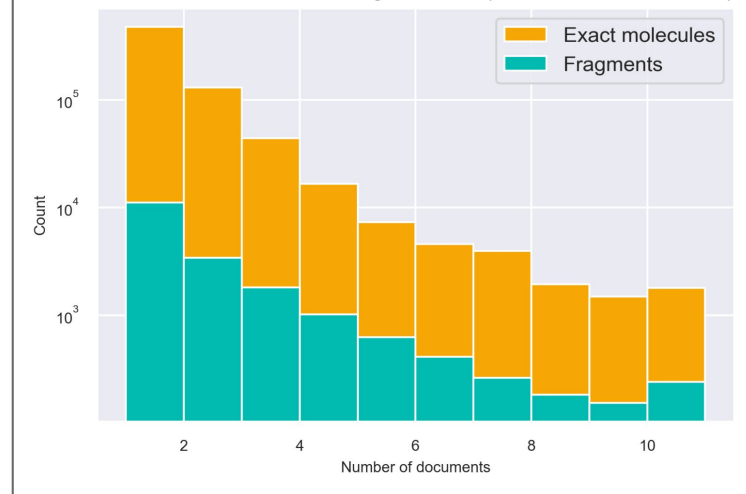
- 1548 patents with a single exact molecule;
- Topics:
 - pharmaceutical preparation
 - pharmaceutical composition
 - method of treatment
- 25 documents with more than 1500 exact molecules;
- kinase inhibitors, H4 receptor inhibitors, glucokinase activators...

Frequency of chemical structures

Number of documents mentioning a molecule



Number of documents mentioning a molecule (document number below 12)



Chemical space overlap of two targets 1.

ChemLocator v2.19.0221.1

Query History Document

Document details

XML US09884052-20180206

Size 96KB
Created Feb 11, 2019, 1:19:20 PM
Modified Feb 5, 2019, 10:47:27 PM
Last crawled Feb 11, 2019, 2:21:40 PM

Tags

alkyl wherein compound Edit tags

Taxonomy hits

PROTYP4GENE: acetylcholinesterase

Non-hit taxonomy terms

- PROTYP4GENE: cholinesterase
- CHEMBL_PC: Hydrolase
- PROTYP4GENE: acetylcholinesterase
- GO_MF: acetylcholinesterase activity
- GO_BP: acetylcholine catabolic process
- GO_BP: negative regulation of synaptic transmission, cholinergic

Actions

- View all structure hits (7)
- Export all structures (7)
- Copy link

Downloads

- View document map
- Download original document
- Add document to Downloads list

Downloads list is empty

XML US09884052-20180206

US 09884052...

D:\TestFiles\Patents\Chemistry\2018\ipg180206\US09884052-20180206.XML

XML US09186345-20151117

US 09186345...

D:\TestFiles\Patents\Chemistry\2015\ipg151117\US09186345-20151117.XML

XML US09750746-20170905

US 09750746...

D:\TestFiles\Patents\Chemistry\2017\ipg170905\US09750746-20170905.XML

XML US10137164-20181127

US 10137164...

D:\TestFiles\Patents\Chemistry\2018\ipg181127\US10137164-20181127.XML

XML US09084800-20150721

US 09084800...

D:\TestFiles\Patents\Chemistry\2015\ipg150721\US09084800-20150721.XML

XML US09346818-20160524

US 09346818...

D:\TestFiles\Patents\Chemistry\2016\ipg160524\US09346818-20160524.XML

ChemLocator v2.19.0221.1

Query History Document

Document details

XML US09867798-20180116

Size 186KB
Created Feb 11, 2019, 1:19:10 PM
Modified Feb 5, 2019, 10:43:44 PM
Last crawled Feb 11, 2019, 2:21:40 PM

Tags

embodiment ladostigil disease Edit tags

Taxonomy hits

PROTYP4GENE: acetylcholinesterase

Non-hit taxonomy terms

- MESHIND: Inflammation
- MEDDRA: Inflammation
- CHEMBL_PC: Unclassified protein
- MESHIND: Ischemia
- PROTYP4GENE: cholinesterase
- CHEMBL_PC: Hydrolase
- PROTYP4GENE: acetylcholinesterase
- GO_MF: acetylcholinesterase activity
- GO_BP: acetylcholine catabolic process

Actions

- View all structure hits (15)
- Export all structures (15)
- Copy link

Downloads

- View document map
- Download original document
- Add document to Downloads list

Downloads list is empty

XML US09795646-20171024

US 09795646...

D:\TestFiles\Patents\Chemistry\2017\ipg171024\US09795646-20171024.XML

XML US09867798-20180116

US 09867798...

D:\TestFiles\Patents\Chemistry\2018\ipg180116\US09867798-20180116.XML

XML US09227937-20160105

US 09227937...

D:\TestFiles\Patents\Chemistry\2016\ipg160105\US09227937-20160105.XML

XML US09676997-20170613

US 09676997...

D:\TestFiles\Patents\Chemistry\2017\ipg170613\US09676997-20170613.XML

XML US09757338-20170912

US 09757338...

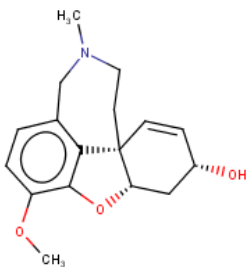
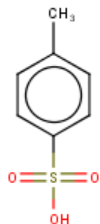
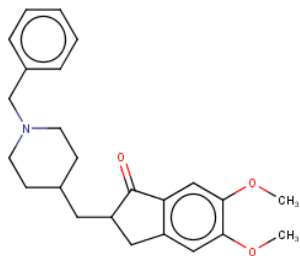
D:\TestFiles\Patents\Chemistry\2017\ipg170912\US09757338-20170912.XML

XML US09132135-20150915

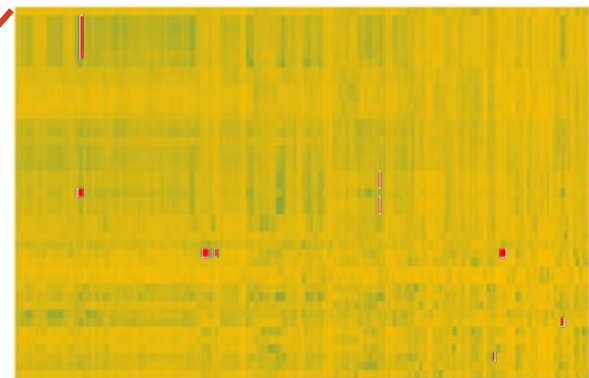
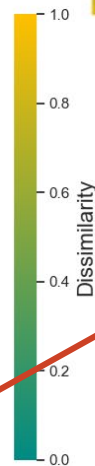
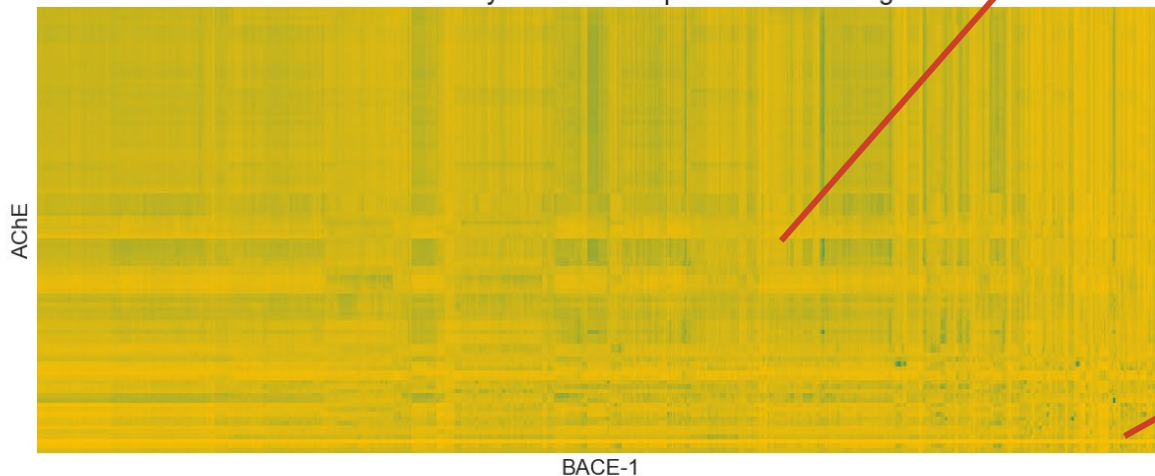
US 09132135...

D:\TestFiles\Patents\Chemistry\2015\ipg150915\US09132135-20150915.XML

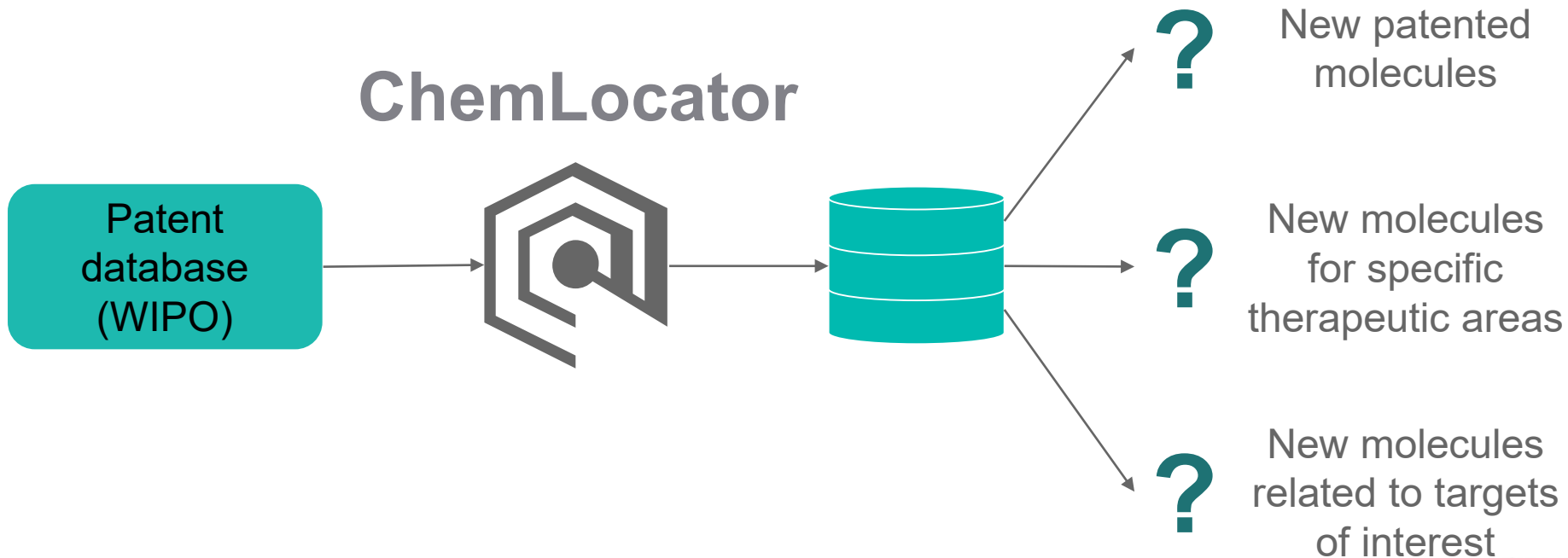
Chemical space overlap of two targets 2.



Chemical similarity based overlap between two targets



Up-to-date information automatically





THANK YOU!