ChemAxon

Jozsef David

CHEMICAL SEARCH WITH NEO4J AND KNOWLEDGE EXTRACTION FROM SCIENTIFIC PAPERS

ChemAxon

1998

Foundation
Owner-CEO: Ferenc Csizmadia
Product Owner: Jozsef David

2019

160 employees worldwide
30+ products

ChemAxon

HQ - Budapest (Hungary)

Development, Consultancy

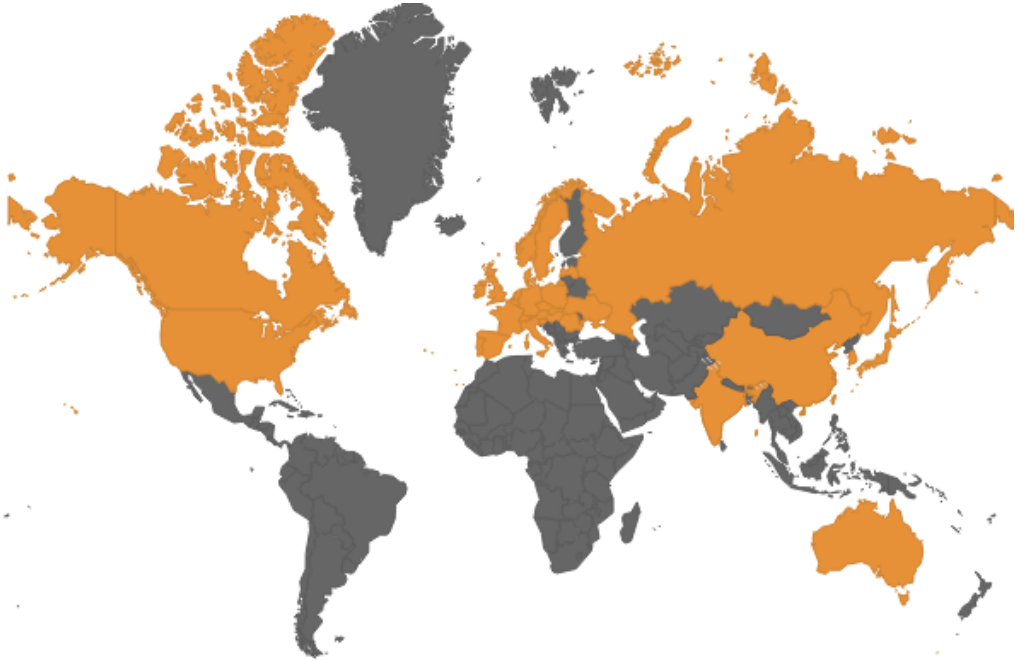
Prague (Czech Republic)

Sales, Consultancy and Application Scientists

Boston and San Diego (United States)

ChemAxon

ChemAxon around the Globe

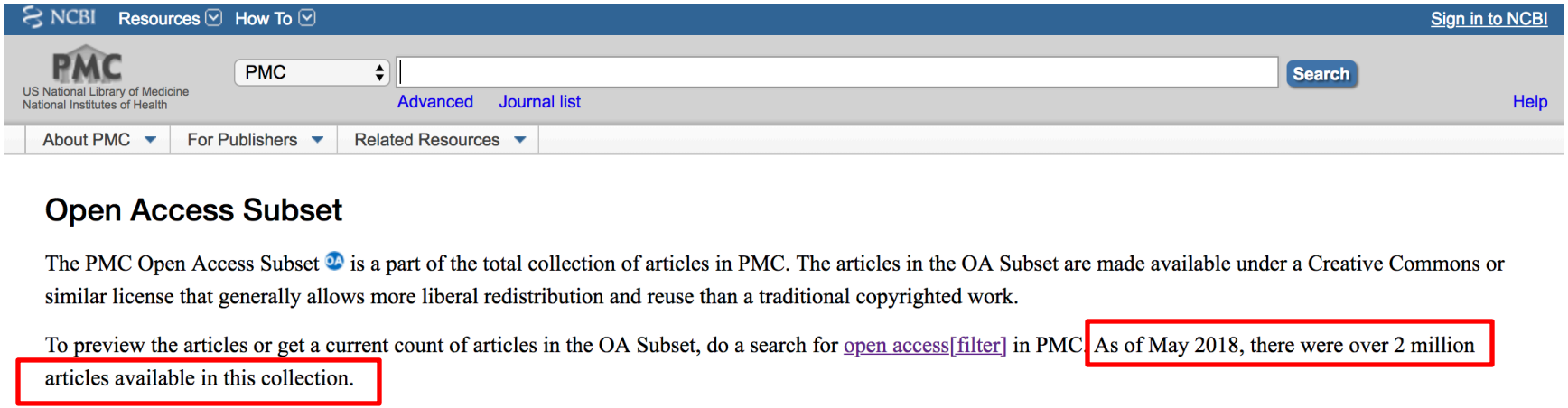


Distributors	Spain, India, China, Korea, Japan, Singapore, Australia
Clients	500+
Partners	50+
Users	300 000+

ChemAxon

THE PROJECT

Input data set



ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/

ChemAxon

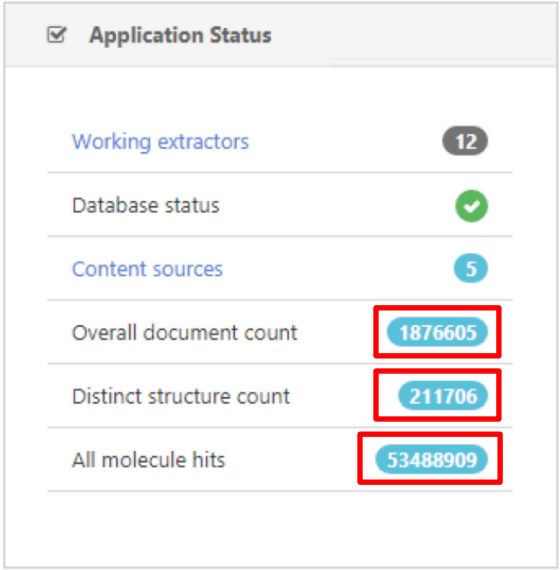
Data set

INPUT

- Documents from ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/
- size: **~30 Gb** (compressed)
- format: **xml**
- Document count: **~1.9 million**

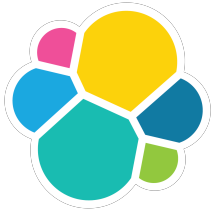
OUTPUT

- Chemical database
  - Distinct molecules: **~211.000**
  - All occurrences: **~53 million**
  - MSSQL DB: 800 GB
- Graph DB: 10.5 GB
- Free text DB: 110 GB

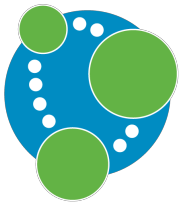


ChemAxon

Tools



elastic



neo4j

HARDWARE

PROVIDER	TYPE	MEMORY	CPU COUNT	DISK SIZE
Amazon EC2	M4.4xlarge	64 GB	16	1.5 TB

ChemAxon

CHEMICAL DATA EXTRACTION

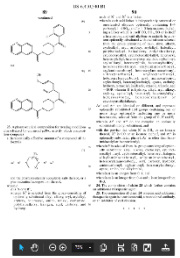
Chemical files and formats

OPTICAL CHARACTER RECOGNITION			MRV	PPTX	
RXN	SDF	OPTICAL STRUCTURE RECOGNITION			
CDXML	IUPAC	CAS	XLS	CDX	SKC
MARVIN OLE	ACCORD FOR EXCEL		DOC	MOL	
JCHEM FOR EXCEL	DOCX	PPT	ISIS OLE		
XLSX	JCHEM FOR SHAREPOINT		ISIS FOR EXCEL		
CORPORATE ID	SMILES	CHEMDRAW OLE			
RDF	CML	INCHI	SMARTS	CXSMILES	
CXSMARTS	XYZ	COMMON NAME	ONE		
SYSTEMATIC NAME	CHEMDRAW FOR EXCEL				

Corporate ID

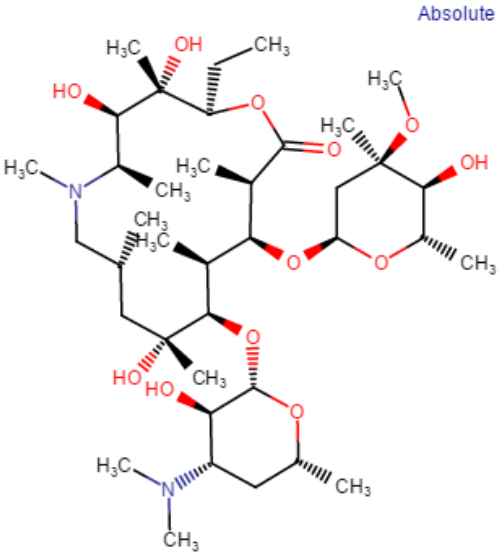
SMDL-00000420
SMDL-00000421
SMDL-00000422
SMDL-00000423
SMDL-00000424
...
SMDL-00116392
SMDL-00116393

OSR / OCR



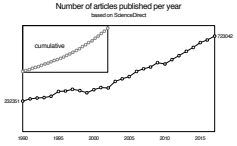
Nomenclature

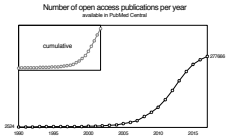
- Azadose
- Azithromycin
- Azithromycin Dihydrate
- Azithromycin Monohydrate
- Azithromycin Pfizer Brand
- Azitrocin
- Azythromycin
- Bayer Brand of Azithromycin Dihydrate
- Dihydrate, Azithromycin
- Funk Brand of Azithromycin Dihydrate
- Goxal
- Lesvi Brand of Azithromycin Dihydrate
- Mack Brand of Azithromycin Dihydrate
- Monohydrate, Azithromycin
- Pfizer Brand of Azithromycin
- Pfizer Brand of Azithromycin Dihydrate
- Pharmacia Brand of Azithromycin Dihydrate
- Sumamed
- Toraseptol
- Ultreon
- Vinzam
- Vita Brand of Azithromycin Dihydrate
- Zentavion
- Zithromax
- Zitromax

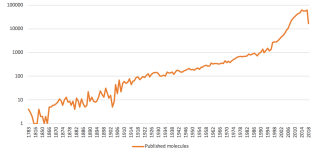


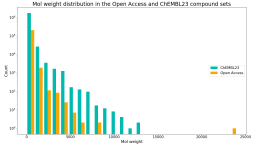
ChemAxon

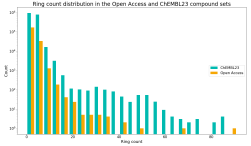
CONTENT ANALYSIS

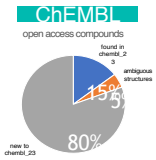

Number of articles published per year
based on ScienceDirect
cumulative
720042
252351
1980
1985
2000
2005
2010
2015

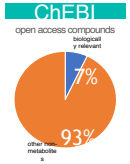

Number of open access publications per year
available in PubMed Central
cumulative
2024
1980
1985
2000
2005
2010
2015

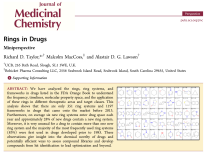

100000
10000
1000
100
10
1
Published molecules

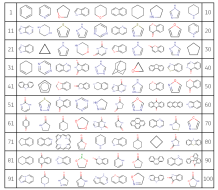

Mol weight distribution in the Open Access and ChEMBL23 compound sets
Count
ChEMBL23
Open Access
Mol weight

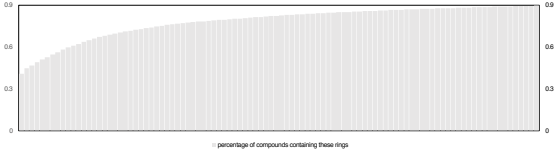

Ring count distribution in the Open Access and ChEMBL23 compound sets
Count
ChEMBL23
Open Access
Ring count

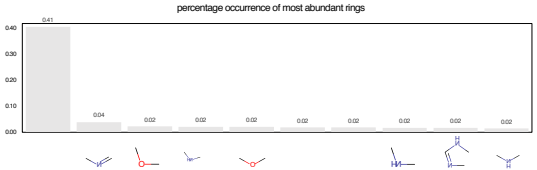

ChEMBL
open access compounds
found in chembl_23
15%
ambiguous structures
5%
80%
new to chembl_23

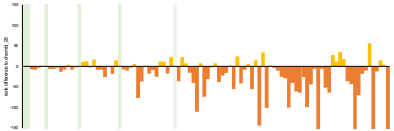

ChEBI
open access compounds
biologicall y relevant
7%
93%
other than metabolite s

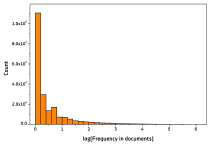

Journal of Medicinal Chemistry
Rings in Drugs

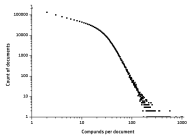


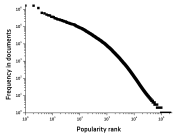

percentage of compounds containing these rings

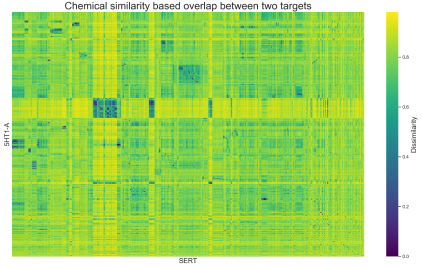

percentage occurrence of most abundant rings
0.41
0.04
0.02


diff. difference to chembl_23


Count
log(Frequency in documents)


Count of documents
Compounds per document


Frequency in documents
Popularity rank

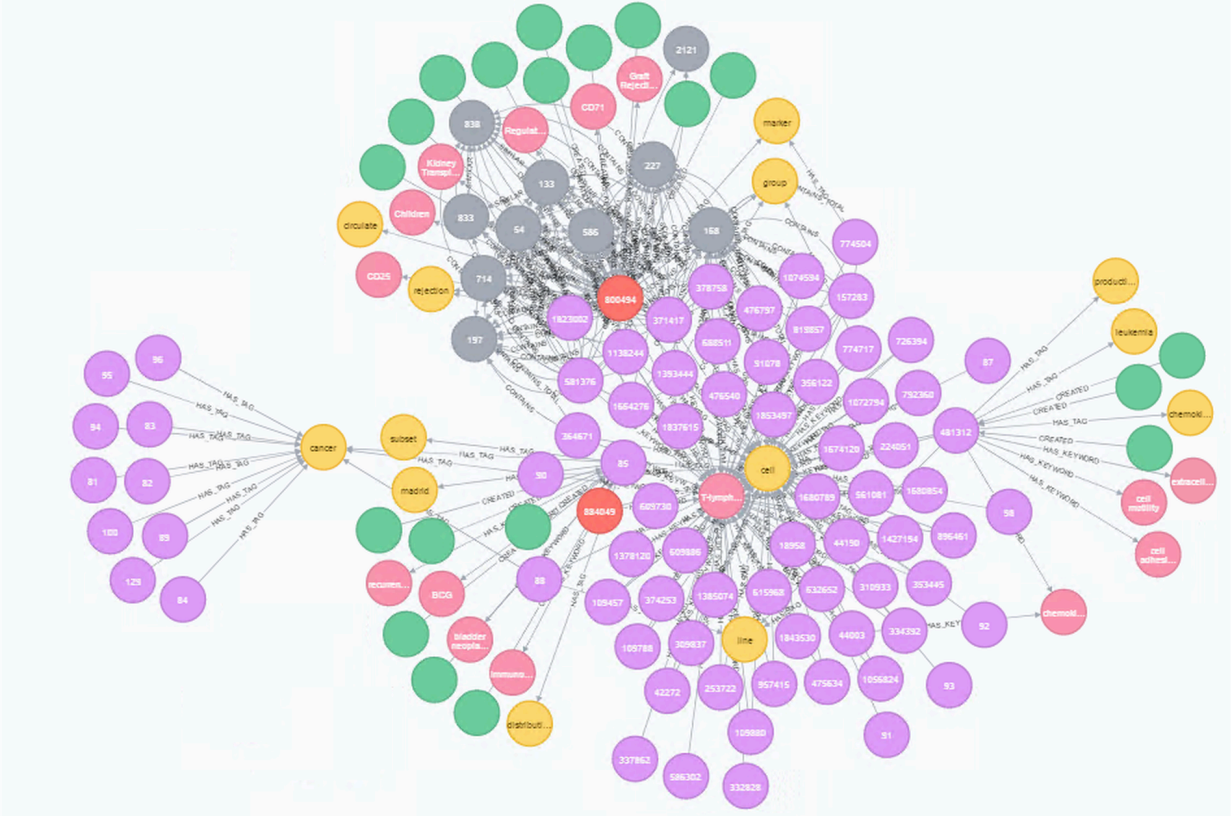

Chemical similarity based overlap between two targets
5HT1-A
SERT
Dissimilarity

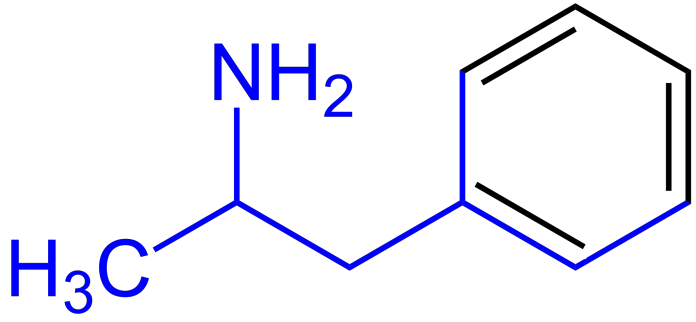
ChemAxon

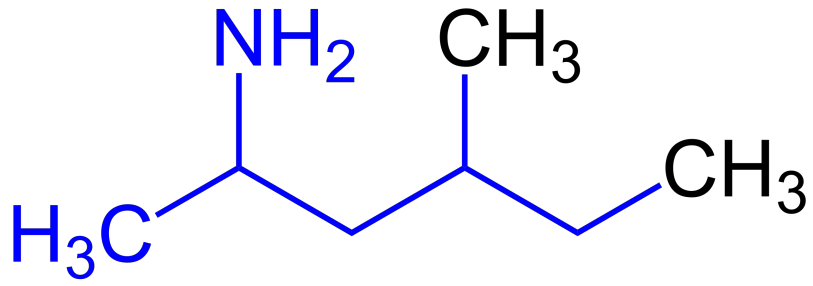
GRAPHS, FINALLY...

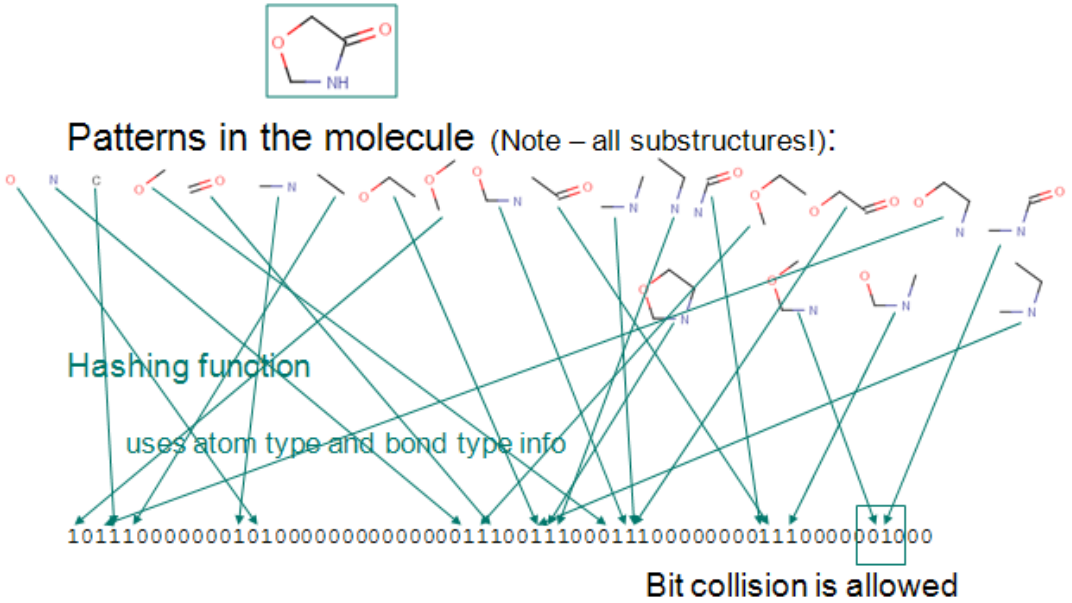
Graph path finding

DB size	10.42 GB
Nodes count	13 769 389
Relationships	124 442 817
Relationship types	Keys, tags, similarity, contains, references, created (by authors)







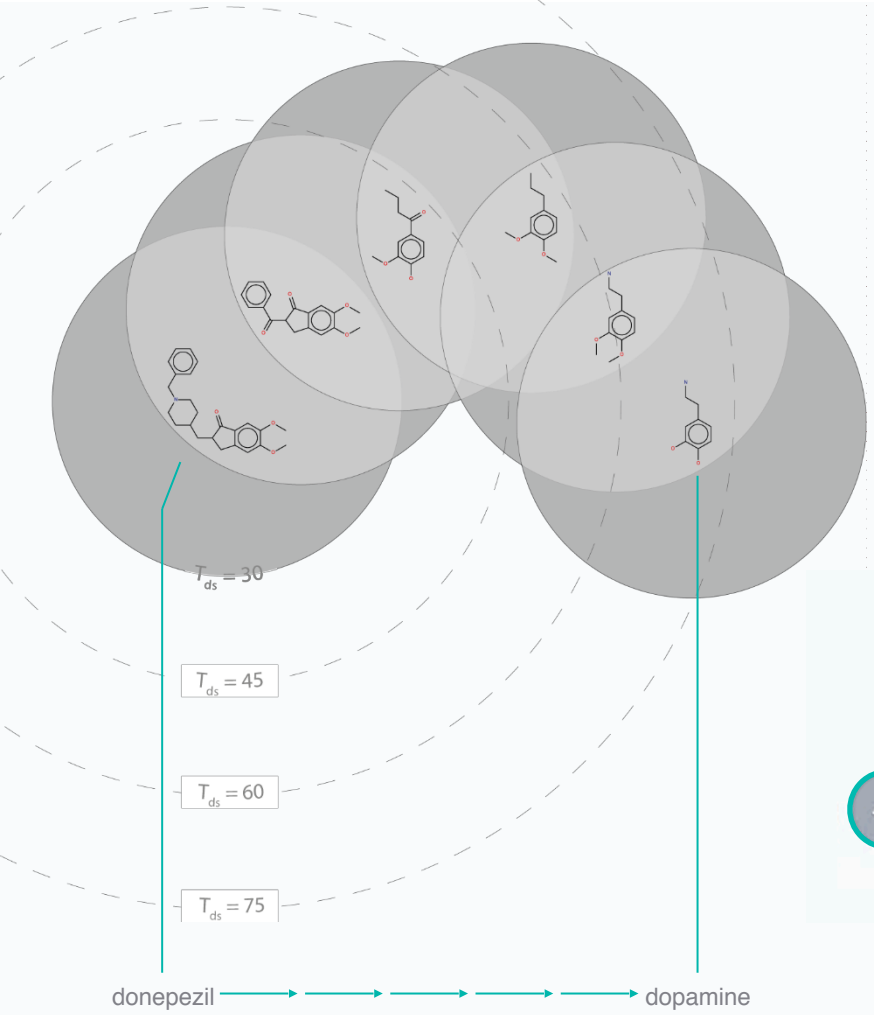


Chemical hashed fingerprint

$$T_s(X,Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

Tanimoto similarity

ChemAxon



GOAL

- Find a route from one chemical space to another hopping through molecules
- In maximum 5 steps

~8 sec