Chemical Intelligence That Makes Hidden Knowledge Effortlessly Reachable

Anna Tomin, Ákos Tarcsay, Dóra Barna, Dávid Malatinszky, Gábor Imre, József Dávid

ChemAxon Kft. Záhony u.7 H-1031 Budapest

Our aim is to provide a method to easily access and explore the chemical space of large scientific knowledge bases stored in scientific articles, patents or reports. Chemistry is a unique field in this regard because chemical structures can be represented with various synonyms; moreover, navigating the knowledge base and the encapsulated chemical space requires special search methods like similarity or substructure searches.

Our study highlights computational approaches to turn chemistry related knowledge stored in all the Open Access articles easily accessible. Methods based on chemical similarity and graph databases are introduced to explore and analyze the content at various levels from a chemist's point of view.



fingerprint-based chemical similarity matrix and clustering by MadFast Similarity Search. [6]

amazon webservices

Amazon M4.4xlarge

64GB RAM

1.5TB disk size

16 CPU

Chemical space relevance

15%

found in chembl_23 ambiguous new to chembl_23 80%

Fig 2. Analyzed chemical space from open access articles compared to well-known public chemical data source.

- Chemically relevant space discussed in all Open Access scientific literature
- 76% of the ChEMBL drugs (Phase=4, Mw>0) are present [7]





The number of compounds per article provides a primary indication about the subject of the scientific article.

Our categorization reveals the logic:

- Low: not chemistry focused (amino acid, carbon dioxide, glass, peptide)
- Medium: SAR, medicinal chemistry, chemical biology, pharmacology, biotechnology journals
- High number: large-scale screening, QSAR/QSPR, method development, benchmark studies

"Five degrees of separation"



• Find a route from one chemical space to another hopping through Chemical similarity-based overlap between two targets

Chemical database:

• Distinct structures: ~211.000

All occurrences: ~53 million

• MSSQL DB: 800 GB

Graph database: 10.5 GB

Free text database: 110 GB



- molecules in 5 steps
- Chemical similarity-based steps in the graph
- Unique opportunity to discover new relationships and new ideas
- ~ 8 sec



Fig 4. Similarity path to possible new ideas via graph steps.

Fig 5. Chemical similarity matrix of compounds binding to serotonin transporter and Serotonin 1A receptor as target. Highlight shows identical compounds of the two targets.

- Overlap analysis of 600 x 1000 chemistry matrix was done by MadFast Similarity Search [5]
- Data cleaning and standardization was by Standardizer and Structure
- Red highlight shows identical compounds binding to both targets

Further exploration of a results set

Authors		Scatter plot			XYSSI
lected: Jun-ichi Kashiwakura / Adolf Ellinger / Alex KleinJan / Alistair Not	e / Alvaro A Cruz	* MISSING *			
Jun-ichi Kashiwakura 1 Adolf Ellinger 1 Alex KleinJan	Alistair Noble Alvaro A Cruz Amir M. Ghaemmaghami	allergen	• • • •		• • • •
Anna G. Mirer 1 Anne L. Astier 1 Anne-Marie Madore	Anthony H. Keeble 1 Anthony M. Szema 1 Antonio Di Marco 1	allergy 000000000000000000000000000000000000			
ntonio Nieto 1 Aruna Behera 1 Attilio Turchetta 1	Ayse Bilge Ozturk 1 Barbara Gepp 1 Barbara Kolarik 1 Beatriz Sastre 1	B cell O O O O O O O O O O O O O O O O O O	• • • •		000
ing K. Lam 1 Birgit Linhart 1 Blanca Estela del Rio-Nat	rro 1 Bobby Q Lanier 1 Bobby W. S. Li 1 Carl-Gustaf Bornehag 1	CD23 CD23 CD23 CD23 CD23 CD23 CD23 CD23	° ° ° ° ° ° °		• • • •
Carlos E Baena-Cagnani 1 Carolyn R Kaplan 1 Catherin	Laprise 1 Catherine de Lara 1 Cecilia M. Lindgren 1 Chang-Shin Park 1	-reteasing factor house dust mite jieke hypothesis	• • ••	0 00 00 0	• •
Keywords		IgE O O O IL-33 O O O O O O O O O O O O O O O O O O	• • • • • • •		•••••
lantad: asthma / InE	• • • •	nity IgE receptor	• • • •		0.00
allergy 5 asthma 3 IgE 2 group 2 innate lympho	cell 1 IL-33 1 report 1 allergen 1 allergic diseases 1	mouse model prevalence	• • • • •	· · · · · · · · ·	•
3 cell 1 CD23 1 basophil 1 Asthma 1 hi	amine-releasing factor 1 house dust mite 1 hygiene hypothesis 1	report	0	0 0 0	
nfection 1 low-affinity IgE receptor 1 mast cell 1	mouse model 1 prevalence 1 prevention 1 fungi 0 CONGA 0			2 5 10 0 5 16 18 19 20 10 20 50 10 10 10 10 10 10 10 10 10 10 10 10 10	Contraction of a state of the s
		N D + 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0	A K & A & B & A & A		4 10 No 1 10 40
			Call the to the service the	E Ma La LA HO LA LA DO LA LA CO LA	6 4.72
			HO HO HO HO Y SO JU CH	4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	No WHOLE. TIS
	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 IL-13 0 IL-18R 0 IL-25 0 Air pollution 0 IL-4 0 ITAM 0		10 10 10 10 10 19 33 10 1	2 20 0 1 1 2 20 20 20 20 20 20 20 20 20 20 20 20 2	Million F. 13
Contraction Con	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 IL-13 0 IL-18R 0 IL-25 0 Air pollution 0 IL-4 0 ITAM 0	Provide Contraction of the second sec	tore of to to to subject to to the subject to the	4 ⁶⁷ 4 ⁴⁰ 6 ⁶⁷ 6 ⁷ 7 4 ⁹ 4 ⁶⁷ 4 ³ 6 ⁷ 6 ¹⁷ 6	Antheol - 40
Contraction Cont	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0	Received and the second	S C C S JO HO N S JUDIN 4	4 ⁶ 4 ¹⁰ 6 ⁶ 5 ⁻⁵ 4 ⁰ 40 ⁵ 4 ³	No willook to
FccRip chain 0 Fel d 1 0 GSTP1 0 Gli1 0 Human mast cell 0 Hymenoptera venom 0 IL-10 0 Details of the selected document hits howing 5 filtered record (from total of 139)	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 IL-13 0 IL-18R 0 IL-25 0 Air pollution 0 IL-4 0 ITAM 0	A C R C R C R C R C R C R C R C R C R C	s Gree store too too subjects to a	4 ⁶ 4 ¹⁰ 1 ⁰ 1 ⁰ 1 ³ 4 ¹⁰ 4 ¹⁰ 1 ⁰ 1 ⁰ 1 ¹⁰	So will of t. 10
FeeRiß chain 0 Fet d 1 0 GSTP1 0 Gill 0 Human mast cell 0 Hymenoptera venom 0 IL-10 0	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 IL-13 0 IL-18R 0 IL-25 0 Air pollution 0 IL-4 0 ITAM 0	Calos	Tags	د ^{ور} باره ومتر ² به به در مرم ² من	No lecules
Contract of the selected document hits B Image: Selected document hits	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 II-13 0 II-18R 0 II-25 0 Air pollution 0 II-4 0 ITAM 0 *	Calos Contraction of the Calos	Tags Cell mast fceri induce antigen	Left Loft de lot 1 + Herol Lodon Paul Lo Herold de lot aver Herold de lot 1 + Herol Lodon Paul Lo Herold de lot aver Herold de lot aver and the lot aver a	Molecules
Contraction Contempotention as remained of the contempotent	HIF-1 0 Helicobacter pylori 0 Histamine H2 receptor antagonists 0 IAM 0		Image: Second part of the second product of the second pr	Left Hoto do to to to to the hotor ho	Molecules
Energination of the selected document hits Procease of the selected document hits Details of the selected document hits howing 5 filtered record (from total of 139) Title Image: Image	HE-1 Helicobacter pytori Histamine H2 receptor antegonists I - 13 Authors Toshiaki Kawakami Jun-ichi Kashiwakura Yuko Kawakami or Catherine de Lara Alistair Noble C Ruby Pawankar Carlos E Baena-Cagnani Jean Bousquet G Walter Canonica Alvaro A Cruz Mich	ael A Kaliner Bobby Q Lanier Karen Henley	Image: Section of the section of th	Left Hoto for the Hotor for	
Contraction of the selected document hits Petails of the selected document hits	HiF-1 Heticobacter pytori Histamine H2 receptor antagonists L-13 L-18R L-25 Air polution L-4 TAM TAM Authors Toshiaki Kawakami Jun-ichi Kashiwakura Yuko Kawakami Toshiaki Kawakami Jun-ichi Kashiwakura Yuko Kawakami Toshiaki Kawakami Jun-ichi Kashiwakura Yuko Kawakami or Catherine de Lara Alistair Noble Ruby Pawankar Carlos E Baena-Cagnani Jean Bousquet G Walter Canonica Alvaro A Cruz Mich Bobby W. S. Li Ralph Stadhouders Marjolein J. W. de Bruijn Melanie Lukkes Dior M. J. M. Beeren Mirjam Kool Wilfred F. J. van IJoken Tata Nageswara Rao Hans Jörg Fehling Rudi W. Hendriks	ael A Kaliner Bobby Q Lanier Karen Henley	Image: Section of the section of th	Left Hull product 1 - Hull CT robot Part of the state of the st	\mathbf{C} $\mathbf{Molecules}$

Conclusion & scientific summary

The knowledge, that is being produced and stored in the forms of reports, patents and scientific journal articles is expanding exponentially. Our use-case highlights the potential of novel technologies to pre-process, search and explore the information network enfolded in large document sets on the field of chemistry.

ChemLocator provided the framework to explore the hidden chemical and related knowledge of that large corpus. Chemical space was analyzed with calculation of fingerprint-based chemical similarity matrix and clustering by MadFast Similarity Search. In order to explore the scaffold diversity of this exclusive chemical space, the obtained set was fragmented to yield rings and ring systems. Hidden relationships were explored by combining text and chemical information in graph data model and related visualization.

Fig 6. Crossfilter based visualization approach to explore large search results.

- Search of a loosely specified query can yield many document results. Exploration by a crossfilter [8] based tool allows the user to further slice and dice the results set.
- This example allows real time crossfiltering by document authors, keywords and extracted molecules.

References

[1] https://chemaxon.com/products/chemlocator

- [2] ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/
- [3] https://chemaxon.com/products/chemical-structure-representation-toolkit
- [4] https://chemaxon.com/products/jchem-engines
- [5] https://neo4j.com/
- [6] https://chemaxon.com/products/madfast
- [7] https://www.ebi.ac.uk/chembl/
- [8] http://crossfilter.github.io/crossfilter/



atomin@chemaxon.com