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.

Chemical database:
- Distinct structures: **~211.000**
- All occurrences: **~53 million**
- MSSQL DB: 800 GB

Graph database: 10.5 GB
Free text database: 110 GB

Amazon M4.4xlarge
64GB RAM
16 CPU
1.5TB disk size

1. Extraction of chemical data on this large corpus with ChemAxon technology
- Content extraction from nearly 1.9M articles [1] with **ChemLocator.** [2]
- Structure standardization and data correction and validation was done using ChemAxon's **Standardizer** and **Structure Checker**. [3]

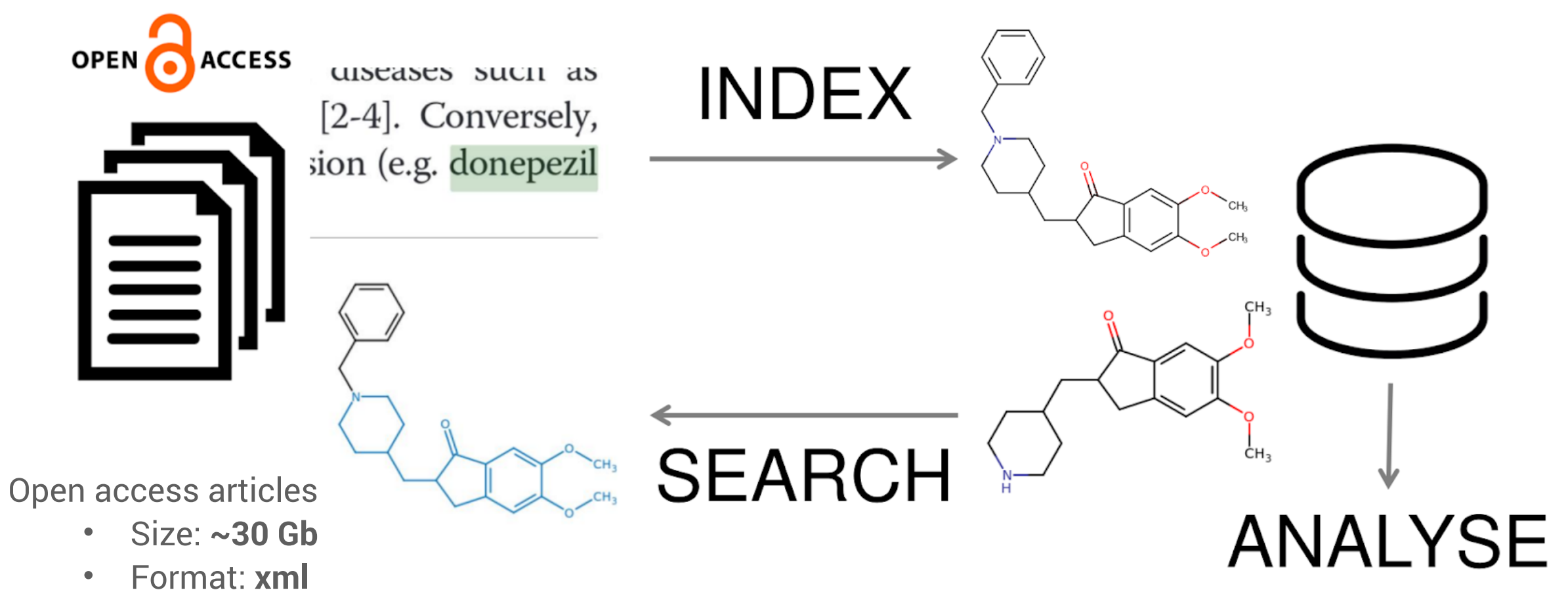


Fig 1. Large-scale conversion of chemical text content to chemical objects with **ChemLocator**.

2. Automated preparation of databases to store and organize chemical and relevant data
- Chemical content storage in **JChem Base**. [4]
- Hidden relationships were explored by combining text and chemical information in **graph data model** and related visualization. [5]

3. Analysis and exploration of the collected chemical space
- Chemical space was analyzed with calculation of fingerprint-based chemical similarity matrix and clustering by **MadFast Similarity Search**. [6]

Chemical space relevance

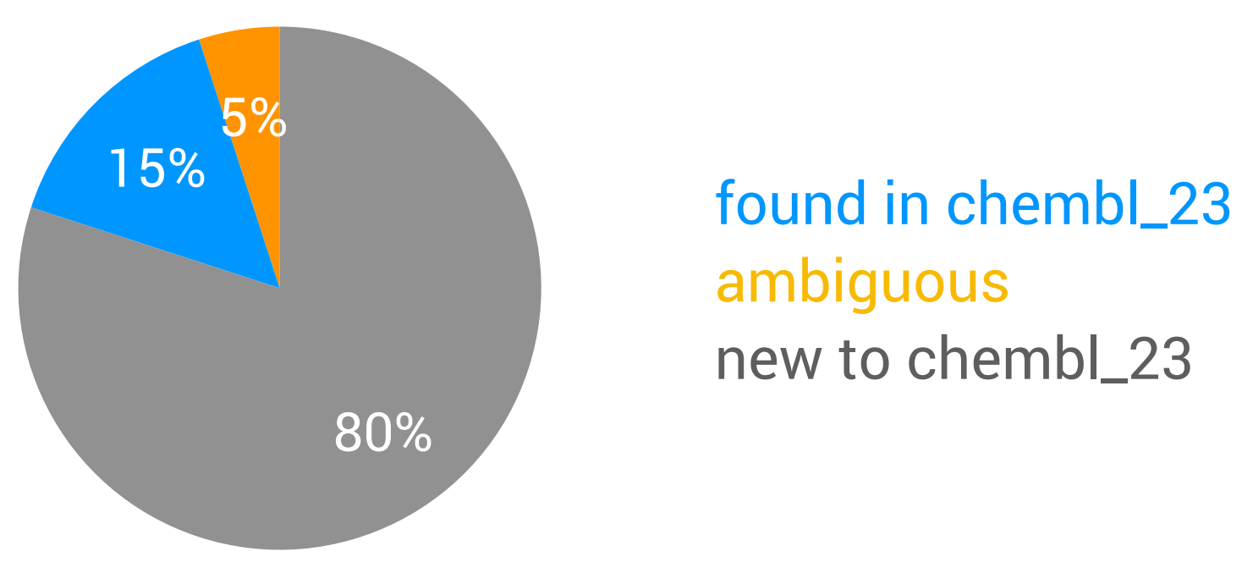


Fig 2. Analyzed chemical space from open access articles compared to well-known public chemical data source.

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

How many structures are extracted from documents?

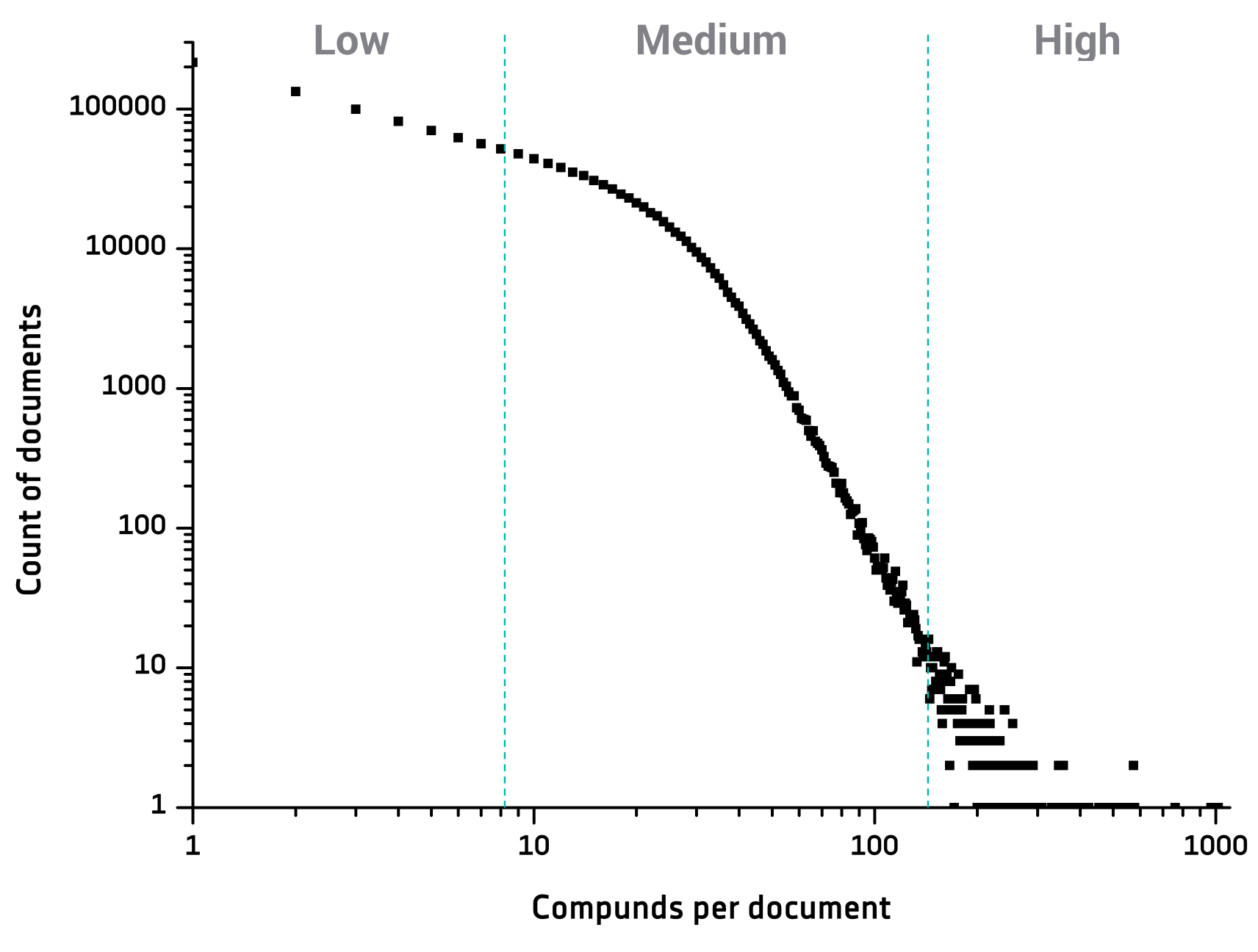


Fig 3. Number of compounds extracted from individual documents.

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, biotechnology journals
- **High number:** large-scale screening, QSAR/QSPR, method development, benchmark studies

"Five degrees of separation"

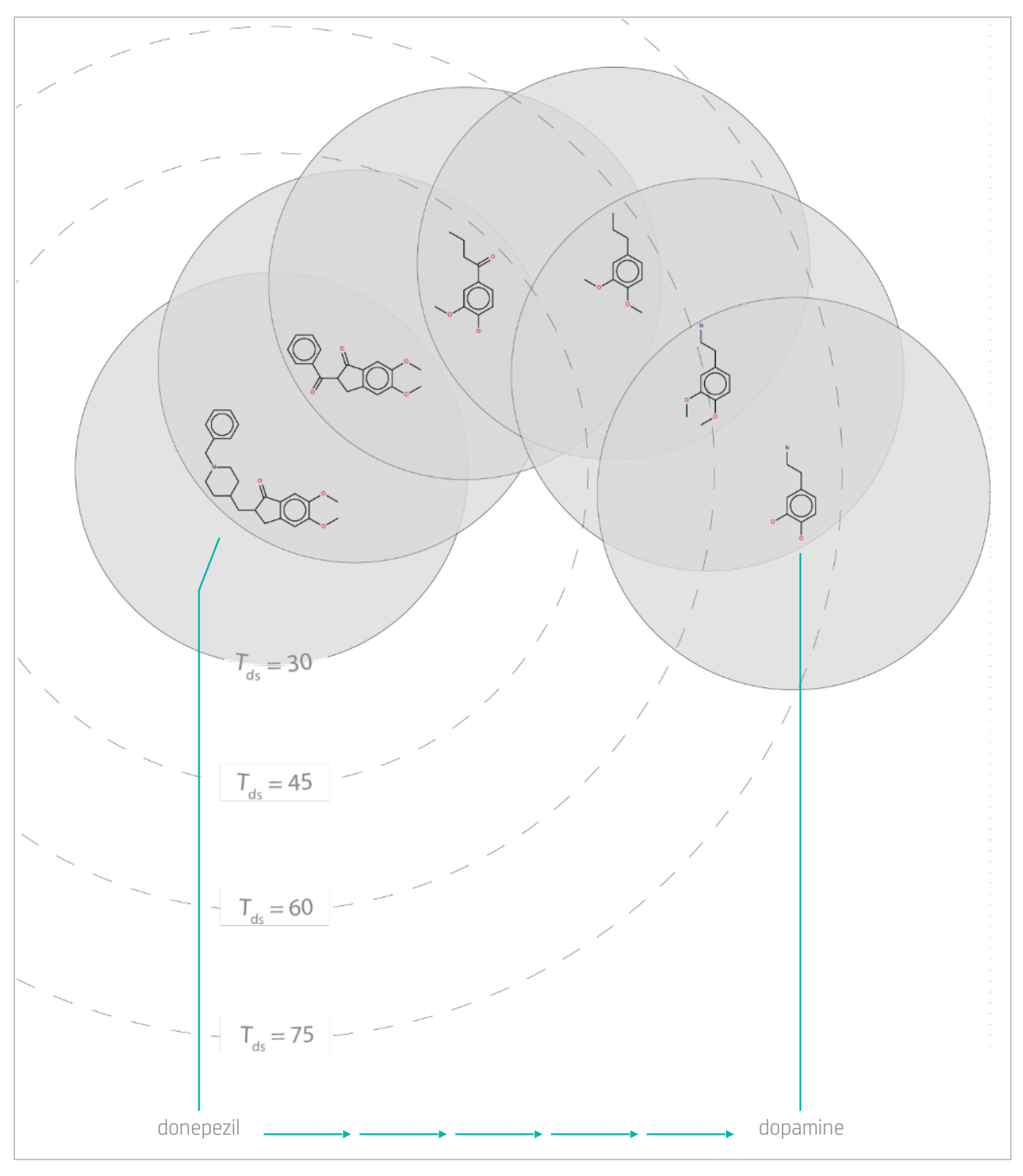
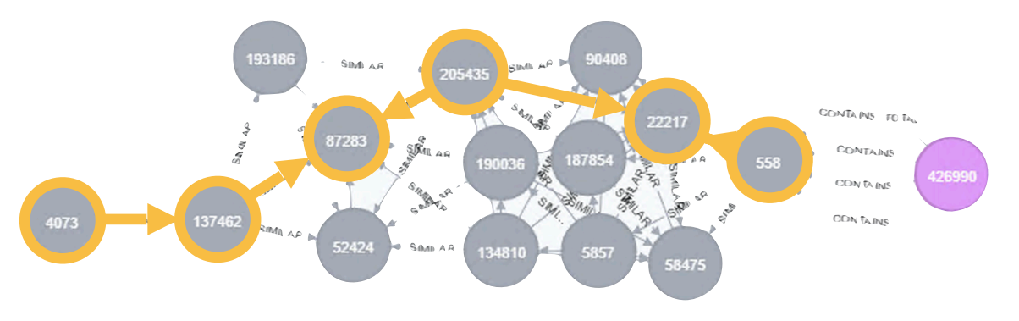


Fig 4. Similarity path to possible new ideas via graph steps.

- Find a route from one chemical space to another hopping through molecules in 5 steps
- Chemical similarity-based steps in the graph
- Unique opportunity to discover new relationships and new ideas
- ~ 8 sec



Chemical similarity-based overlap between two targets

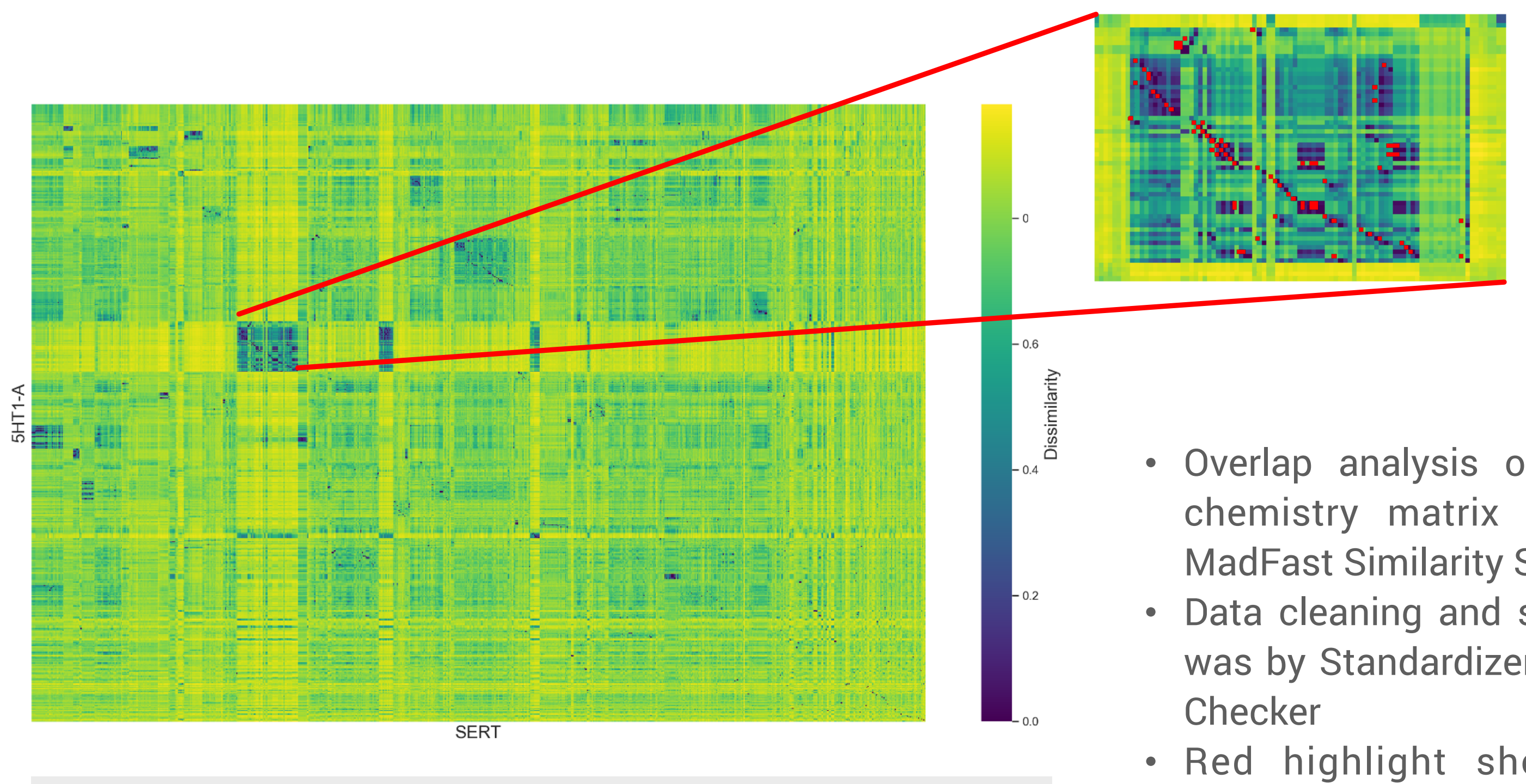
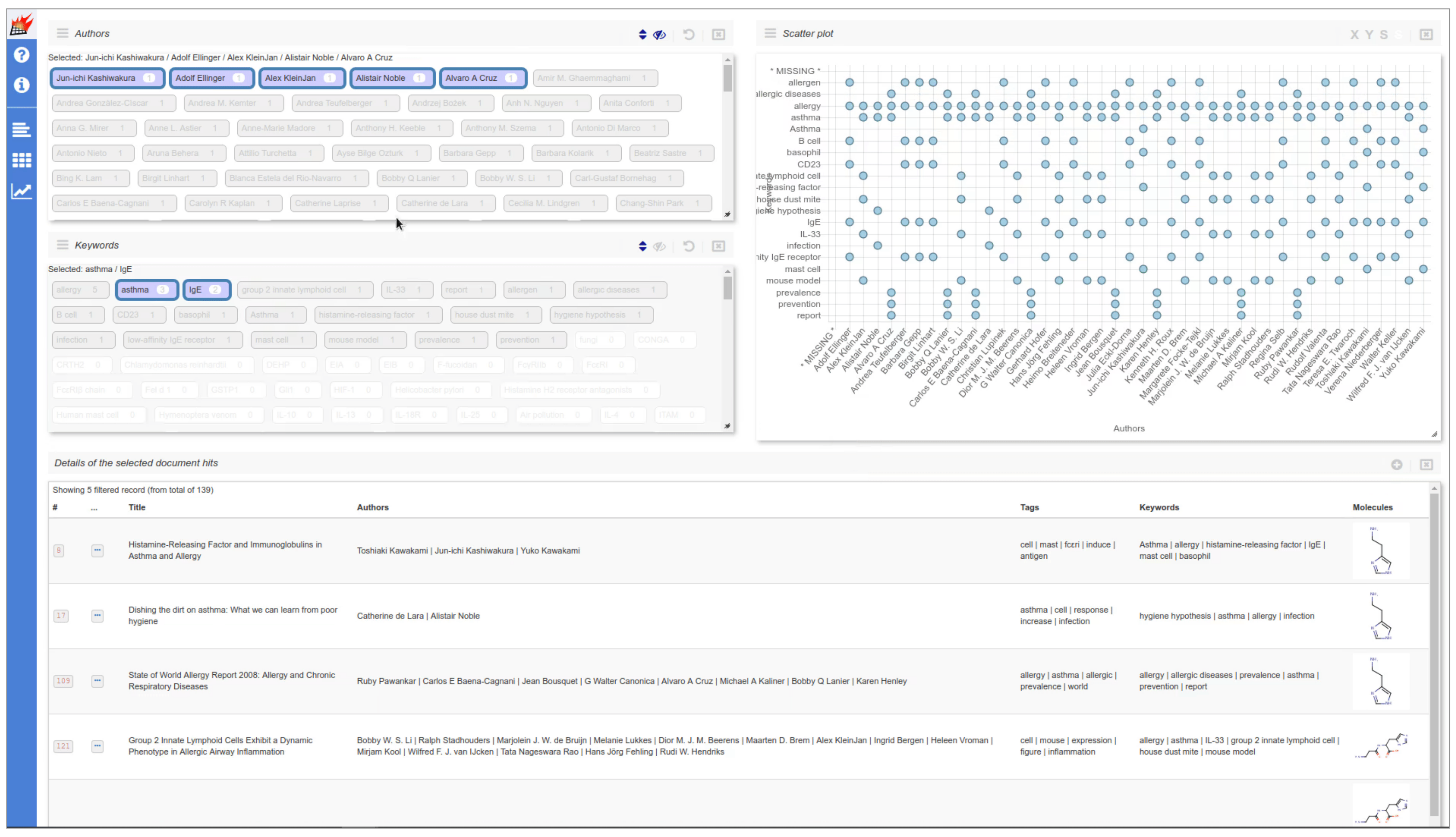


Fig 5. Chemical similarity matrix of compounds binding to serotonin transporter and Serotonin 1A receptor as target. Highlight shows identical compounds of the two targets.

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

Further exploration of a results set



Showing 5 filtered record (from total of 139)

#	...	Title	Authors	Tags	Keywords	Molecules
8	···	Histamine-Releasing Factor and Immunoglobulins in Asthma and Allergy	Toshiaki Kawakami | Jun-ichi Kashiwakura | Yuko Kawakami	cell | mast | fcεri | induce | antigen	Asthma | allergy | histamine-releasing factor | IgE | mast cell | basophil	
17	···	Dishing the dirt on asthma: What we can learn from poor hygiene	Catherine de Lara | Alistair Noble	asthma | cell | response | increase | infection	hygiene hypothesis | asthma | allergy | infection	
109	···	State of World Allergy Report 2008: Allergy and Chronic Respiratory Diseases	Ruby Pawankar | Carlos E Baena-Cagnani | Jean Bousquet | G Walter Canonica | Alvaro A Cruz | Michael A Kaliner | Bobby Q Lanier | Karen Henley	allergy | asthma | allergic | prevalence | world	allergy | allergic diseases | prevalence | asthma | prevention | report	
121	···	Group 2 Innate Lymphoid Cells Exhibit a Dynamic Phenotype in Allergic Airway Inflammation	Bobby W. S. Li | Ralph Stadhouders | Marjolein J. W. de Bruijn | Melanie Lukkes | Dior M. J. M. Beerens | Maarten D. Brem | Alex KleinJan | Ingrid Bergen | Heleen Vroman | Mirjam Kool | Wilfred F. J. van IJcken | Tata Nageswara Rao | Hans Jörg Fehling | Rudi W. Hendriks	cell | mouse | expression | figure | inflammation	allergy | asthma | IL-33 | group 2 innate lymphoid cell | house dust mite | mouse model	

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.

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.

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

ChemAxon