The Value of Intellectual Substance Indexing

A Comparative Approach

Sebastian Brauch

STNext®

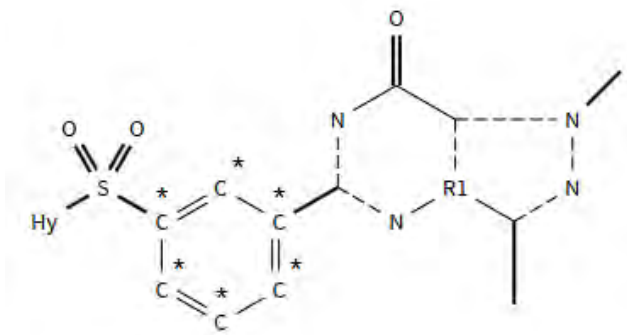
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS®
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Leading structure files under one roof

*Single **query structure** for all structure files*

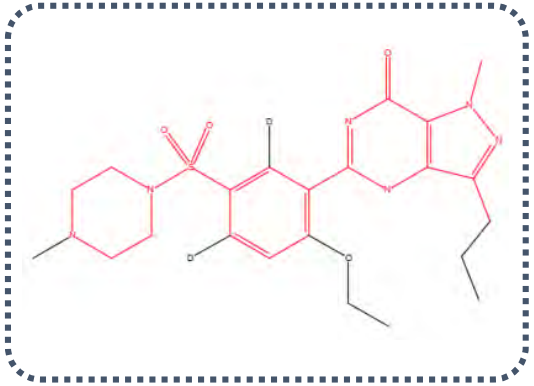
***Each file** has **unique content**!*



Query structure

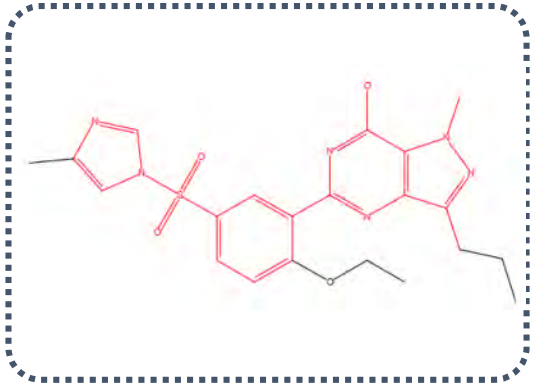
ReaxysFileSUB

~ 34 Mio. structures



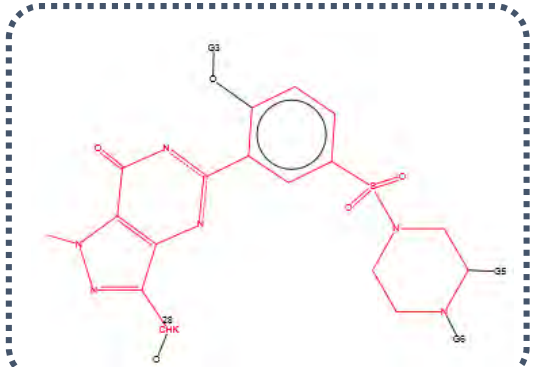
DCR

> 3.3 Mio. structures



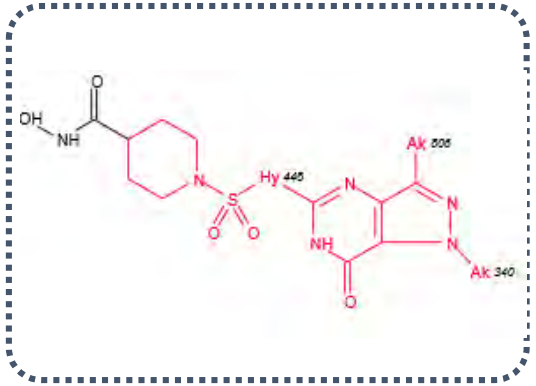
DWPIM

~ 2,2 Mio. structures



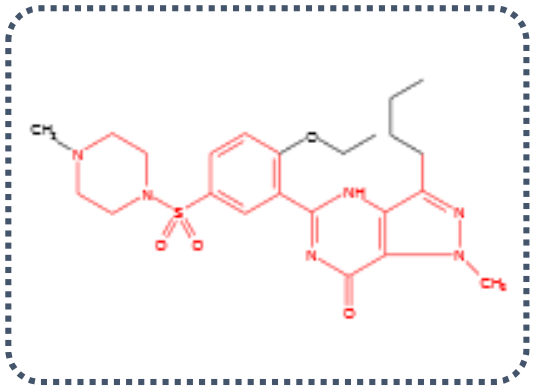
MARPAT

~ 1,2 Mio. structures



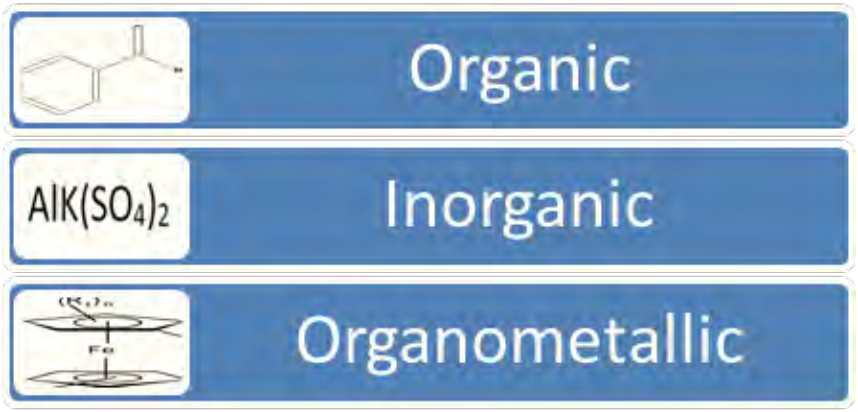
CAS Registry

> 139 Mio. structures



Now Available on STNext: Up-to date REAXYSFILE Databases

- REAXYSFILE**Sub**
  - **34 million** chemical substances (03/2021)
  - From 1771 to date



- REAXYSFILE**Bib**
  - **11.3 million** bibliographic references (03/2021)

    2,1 million patent records + **9,2 million** NPL records
  - Bibliographic documents from Beilstein and Gmelin handbooks of organic chemistry, chemistry journals and chemistry patents
  - Both databases are updated weekly

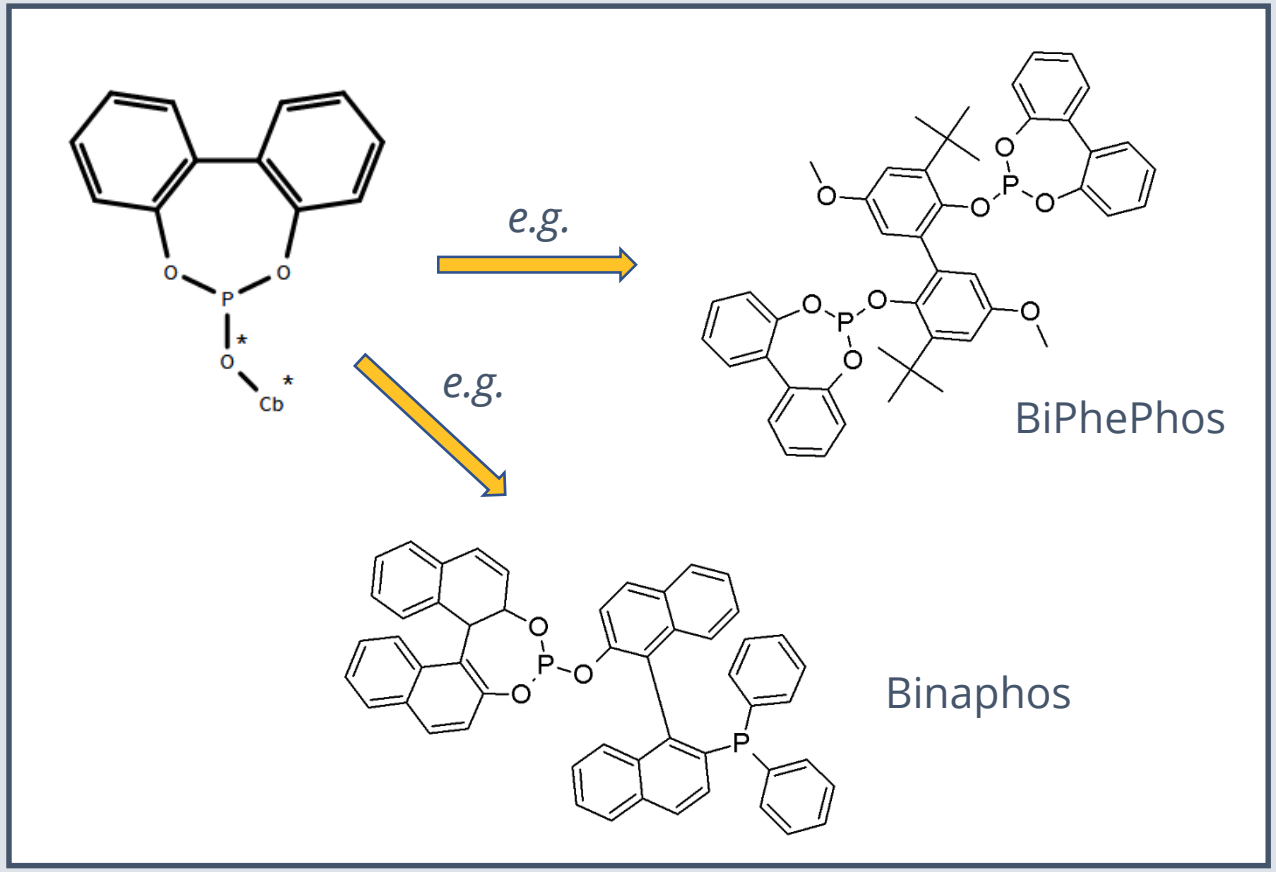
Eager to know more about Reaxysfile?
Watch these recorded webinars:

New REAXYSFILE Databases on STNext
Multi-file Structure Searching on STNext

!

Case Study: Phosphite Ligands in Catalysis

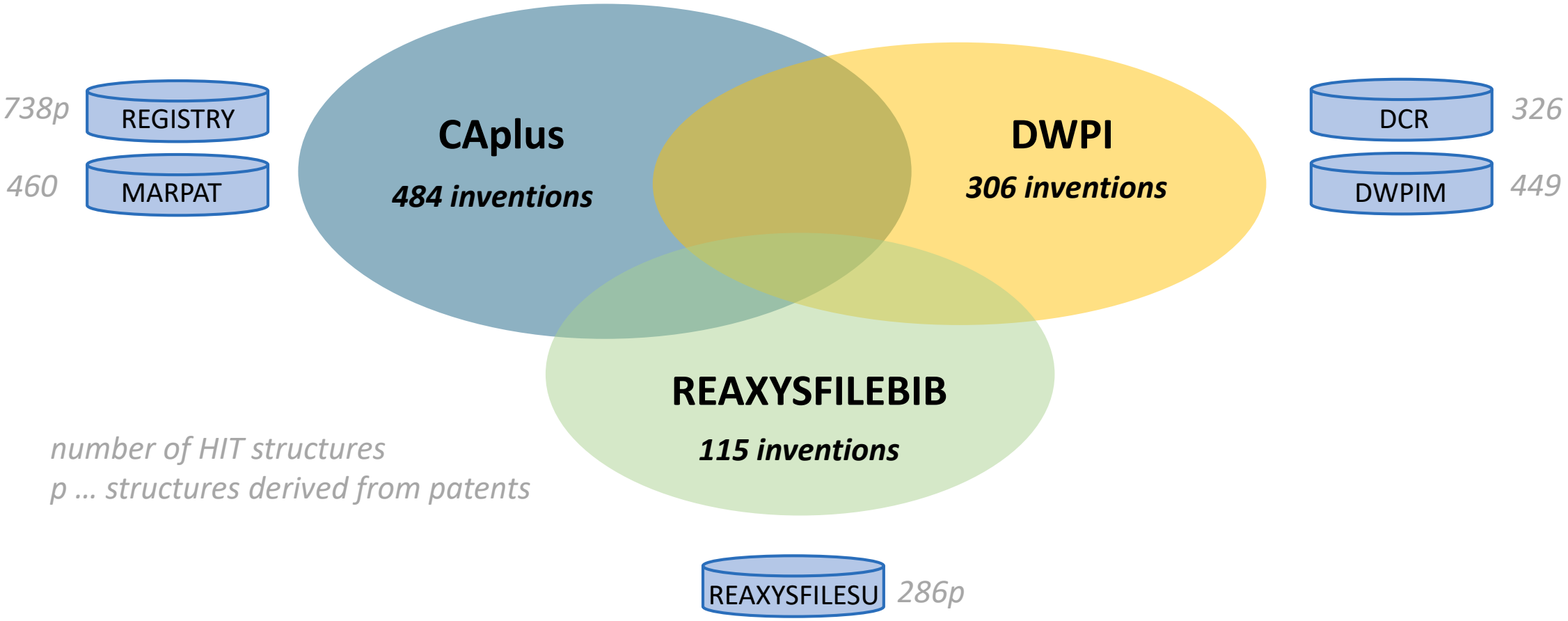
- phosphite ligands are used in metal complexes of e.g. Rh, Ru, Co, Ir in homogeneous catalysis
- applications include various hydroformylation reactions
- companies active in this field are Evonik, Union Carbide, Mitsubishi, BASF



Workflow:

- Multi-file structure search in STNext
- Crossover to bibliographic files, refine to patents
- Compare results between structure files on STNext
- Comparison to Patentscope

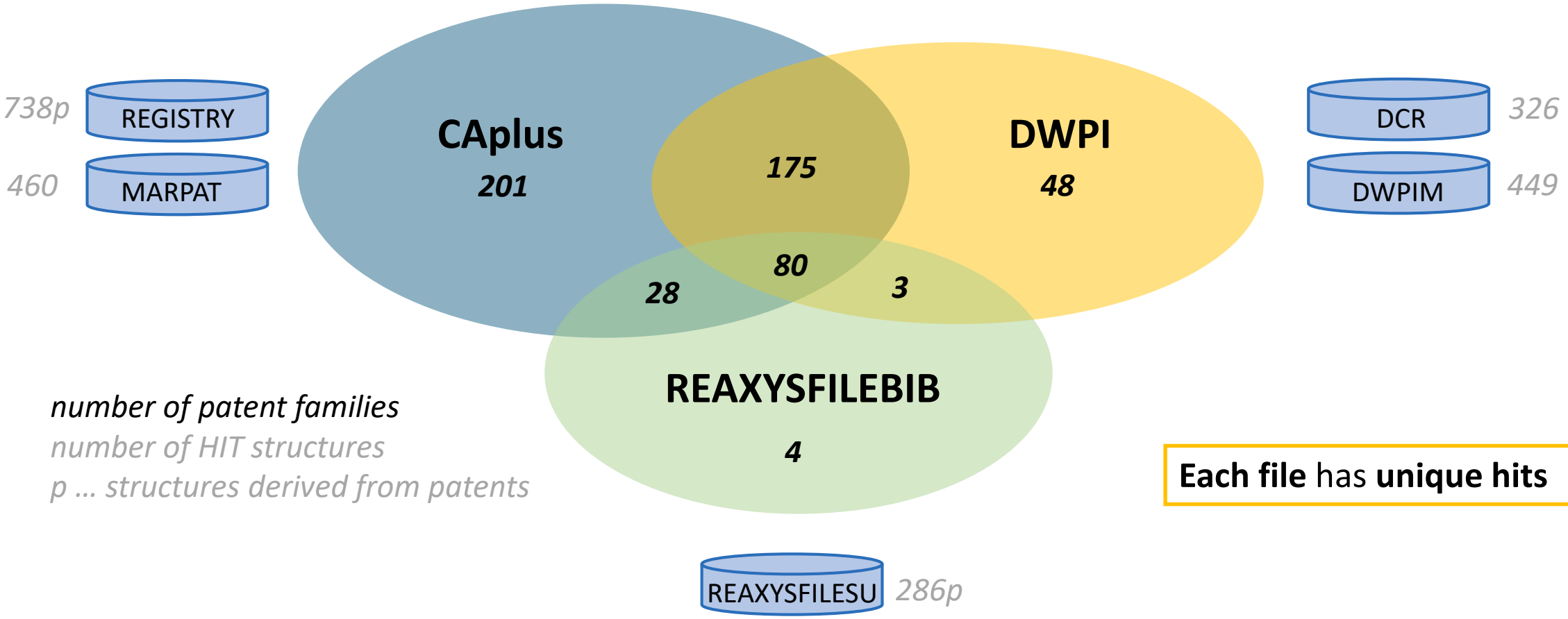
Case Study – Search Strategy and Analysis (1/2)



number of HIT structures
p … structures derived from patents

normalization via INPAFAMDB
date of search - 26th February 2021

Case Study – Search Strategy and Analysis (2/2)



number of patent families
number of HIT structures
p … structures derived from patents

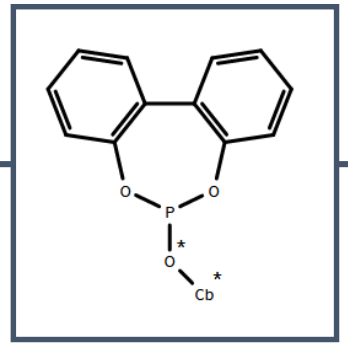
Each file has **unique hits** !

normalization via INPAFAMDB
date of search - 26th February 2021

Why are all structure files required for comprehensiveness?

- CAS, Clarivate and Elsevier apply different indexing guidelines to chemical patent publications
  - Different compounds are selected from the claims and the description for indexing
- Patent authority, document type and historical coverage varies between CAplus, DWPI and ReaxysFileBIB
- Timeliness of coverage and indexing
- For a particular invention CAS, Clarivate and Elsevier index the basic patent publication
  - The basic patent may vary between the database producers
  - The patent content depends on the family member

Comparison STNext vs. PATENTSCOPE



Hit structures

1198 CAplus

775 DWPI

286 REAXYSFILE

PATENTSCOPE *847*

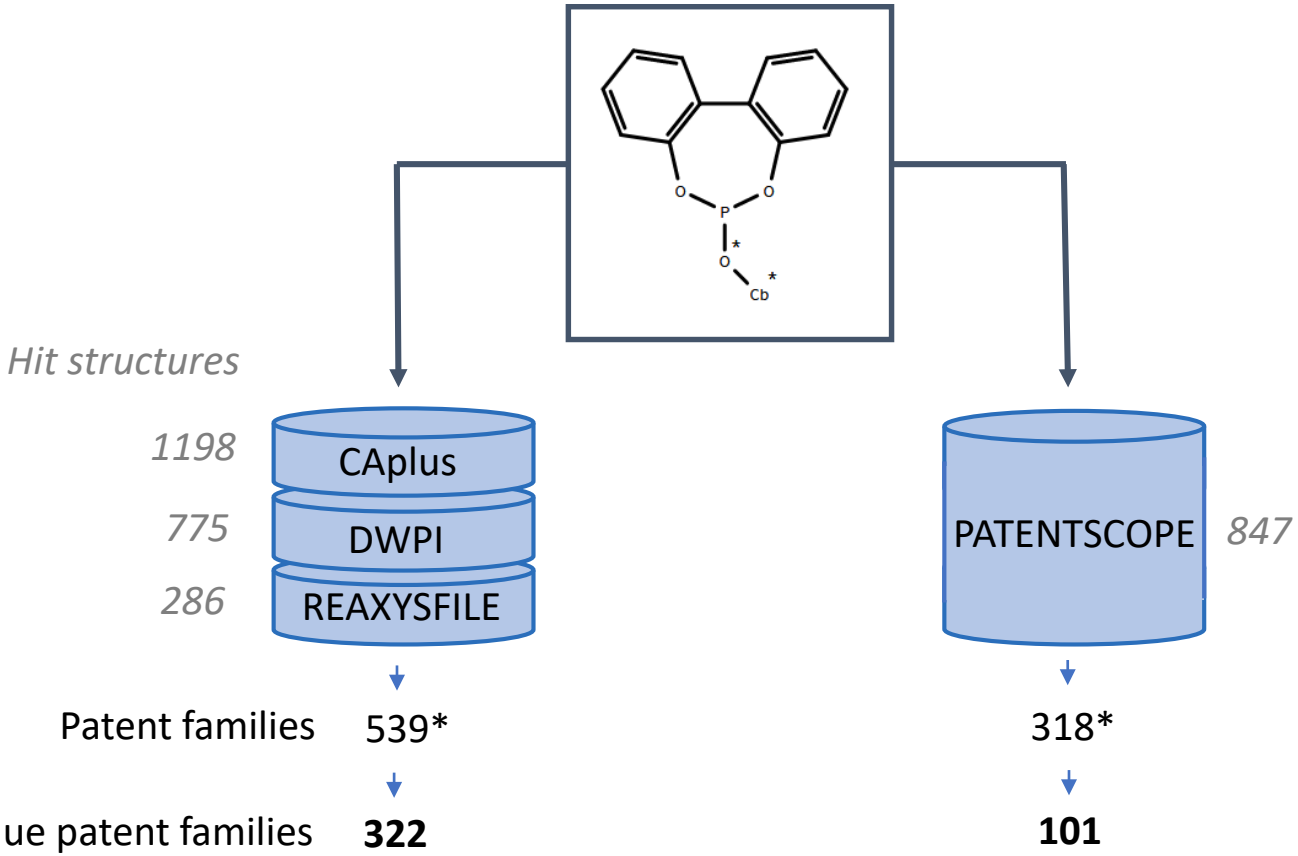
Patent families	539*	318*
Unique patent families	**322**	**101**

Unique patent families

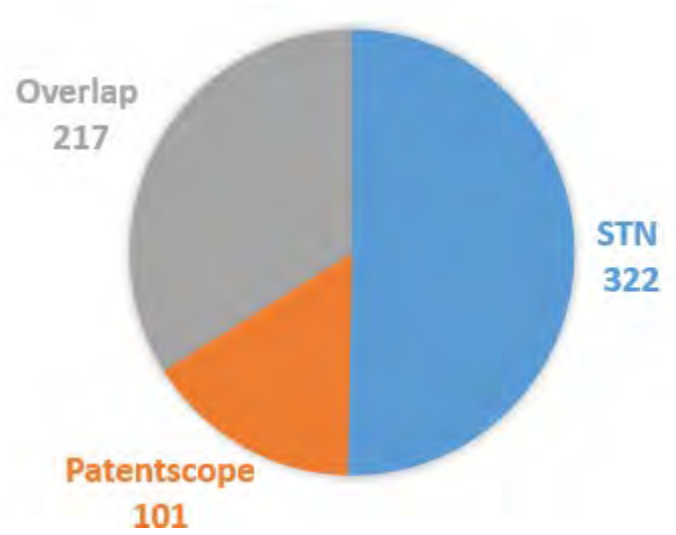
- **all** STN unique hits are highly relevant
- PATENTSCOPE retrieves only **40%** of all relevant answers
- **133** STN uniques are based on a Markush search
- **0 of 101!** *unique* PATENTSCOPE hits are relevant in the context of the search (**mostly background art, preparation of key intermediates, "laundry lists"**)

** Nomalization of results via INPADOCDB/INPAFAMDB*
date of search - 26th February 2021

Comparison STNext vs. PATENTSCOPE



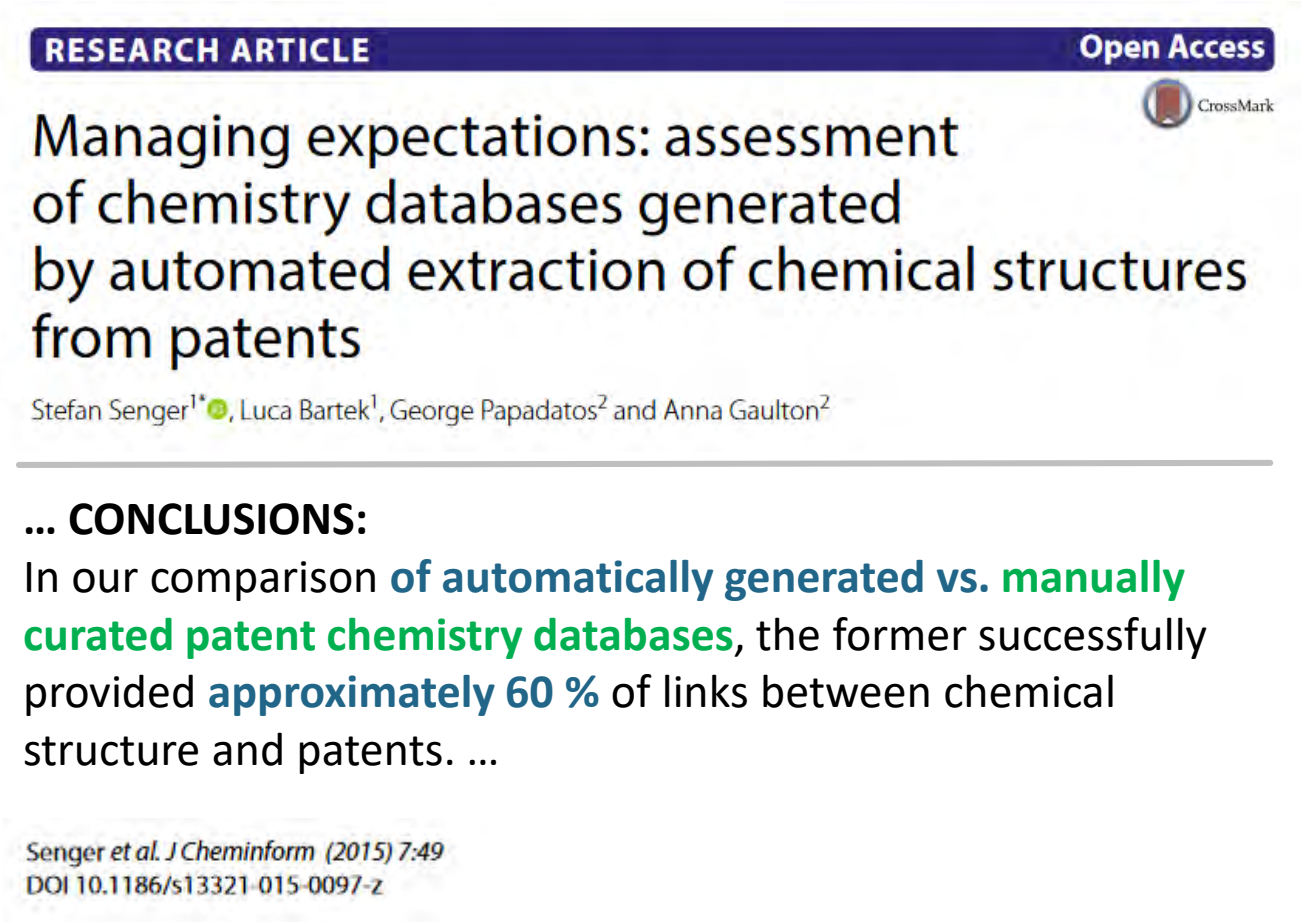
Unique patent families



- **all** STN unique hits are highly relevant
- PATENTSCOPE retrieves only **40%** of all relevant answers
- **133** STN uniques are based on a Markush search
- **0 of 101!** *unique* PATENTSCOPE hits are relevant in the context of the search (**mostly background art, preparation of key intermediates, "laundry lists"**)

** Nomalization of results via INPADOCDB/INPAFAMDB*
date of search - 26th February 2021

The value of intellectual indexing



RESEARCH ARTICLE Open Access

CrossMark

Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents

Stefan Senger[1*], Luca Bartek[1], George Papadatos[2] and Anna Gaulton[2]

… CONCLUSIONS:
In our comparison **of automatically generated vs. manually curated patent chemistry databases**, the former successfully provided **approximately 60 %** of links between chemical structure and patents. …

Senger *et al. J Cheminform (2015) 7:49*
DOI 10.1186/s13321-015-0097-z

Why do algorithms miss a relevant part of the disclosed substances?

- Ambiguous naming
- Markush representations
- No name – explanatory text or images rather than chemical names
- Stereochemistry issues
- Tautomer issues
- Multi-component substances
- Transliteration from non-latin languages

Key Takeaways

- Intellectually curated chemistry databases are superior to algorithm-curated databases
  - algorithms **miss relevant parts of the disclosed substances in a patent** (*e.g.* Markush, ambiguous naming, etc.)
  - algorithm-curated databases will **retrieve a lot of noise/artefacts**, *e.g.*, as no differentiation between intermediates, solvents, fragments, referenced compounds, etc., is possible
- All structure files available are required for comprehensiveness
  - Differences in indexing policies and coverage will retrieve unique hits in each intellectual curated database
- STN contains the industry-leading chemistry databases from CAS, Clarivate and Elsevier on a single platform making it the "gold standard" in all chemistry related searches

For any questions concerning structure searching contact the Helpdesk.

Or you can also assign structure searches to the **FIZ Search Service**.

Contact Us

CAS **help@cas.org**
www.cas.org

FIZ Karlsruhe **helpdesk@fiz-karlsruhe.de**
www.stn-international.de