ChemAxon

Akos Tarcsay

BuZZ

#meetCXN

HOW FAST IS CHEMAXON RDBMS SEARCH?

2004- JOC - JChem Oracle Cartridge

2015- JPC-JChem PostgreSQL Cartridge

2019- CHR – Choral (Oracle Cartridge)

ChemAxon

STRUCTURE QUERIES

Duplicate, Substructure

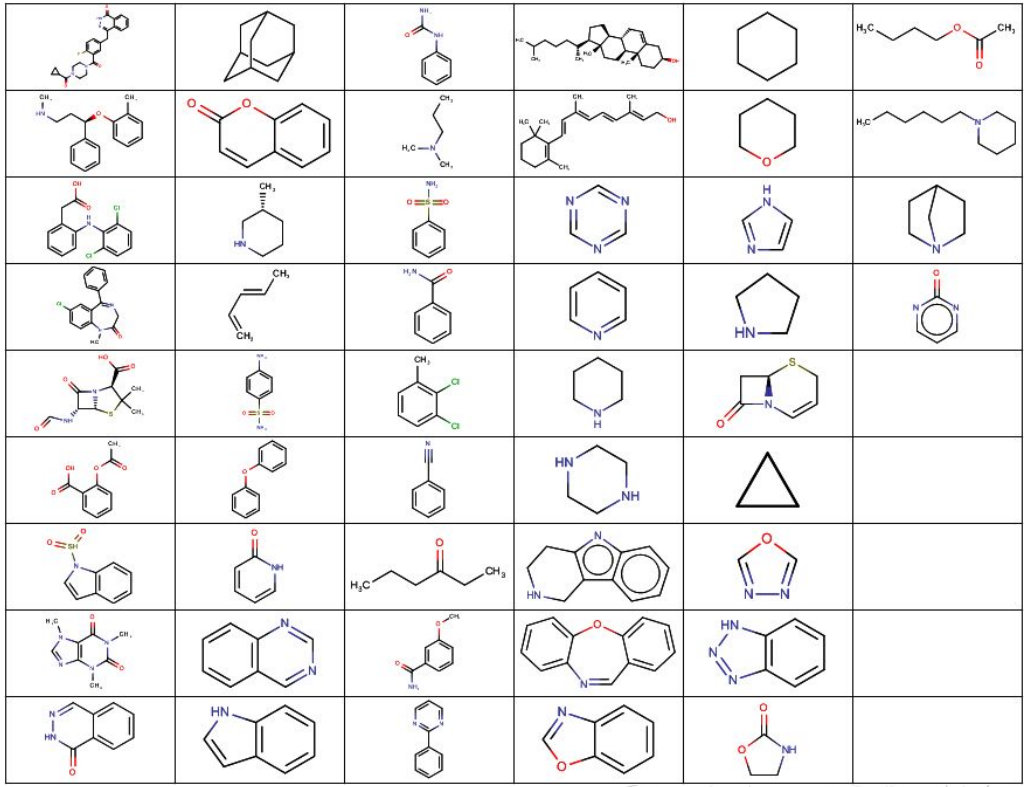
- 49 custom selected

(*Rings in drugs, Taylor RD, MacCoss M, Lawson AD. J. Med. Chem.*, **2014**, 57 (14), pp 5845–5859)

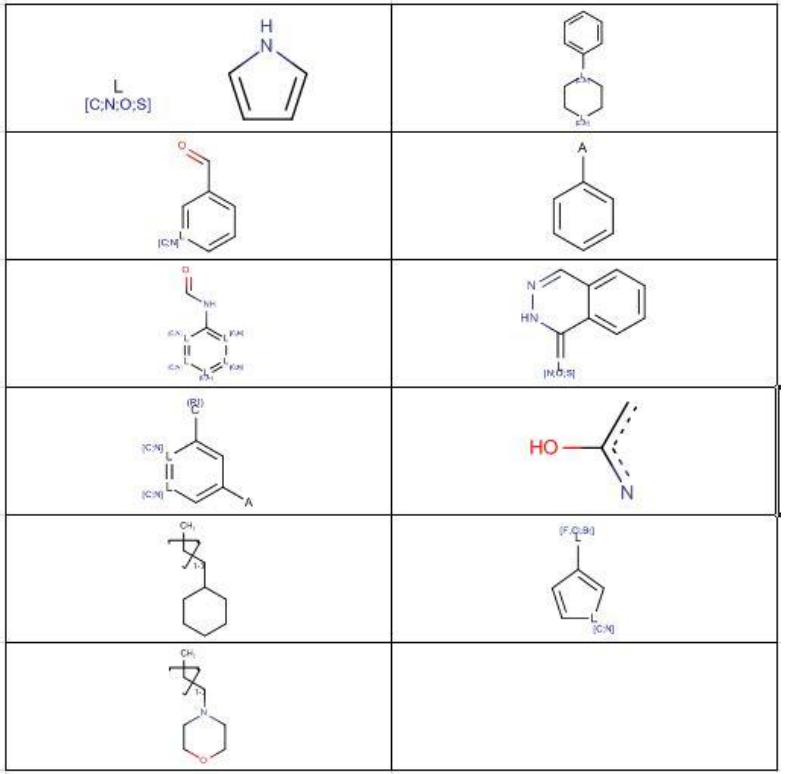
- 11 includes query features only for substructure search

S= 60

Similarity cutoff: 0.5



QUERY FEATURES

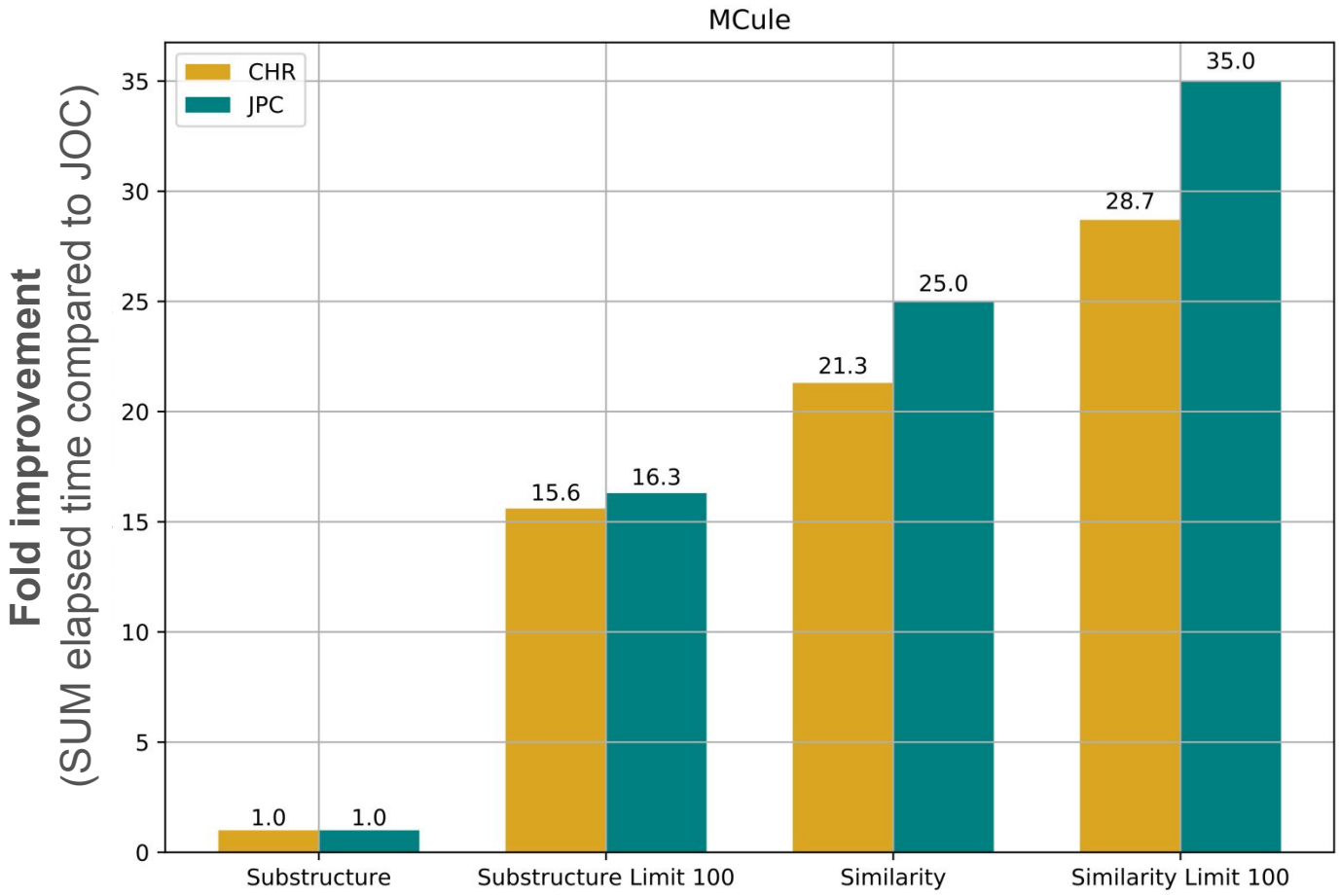


ChemAxon

SEARCH SPEED ASSESSMENT

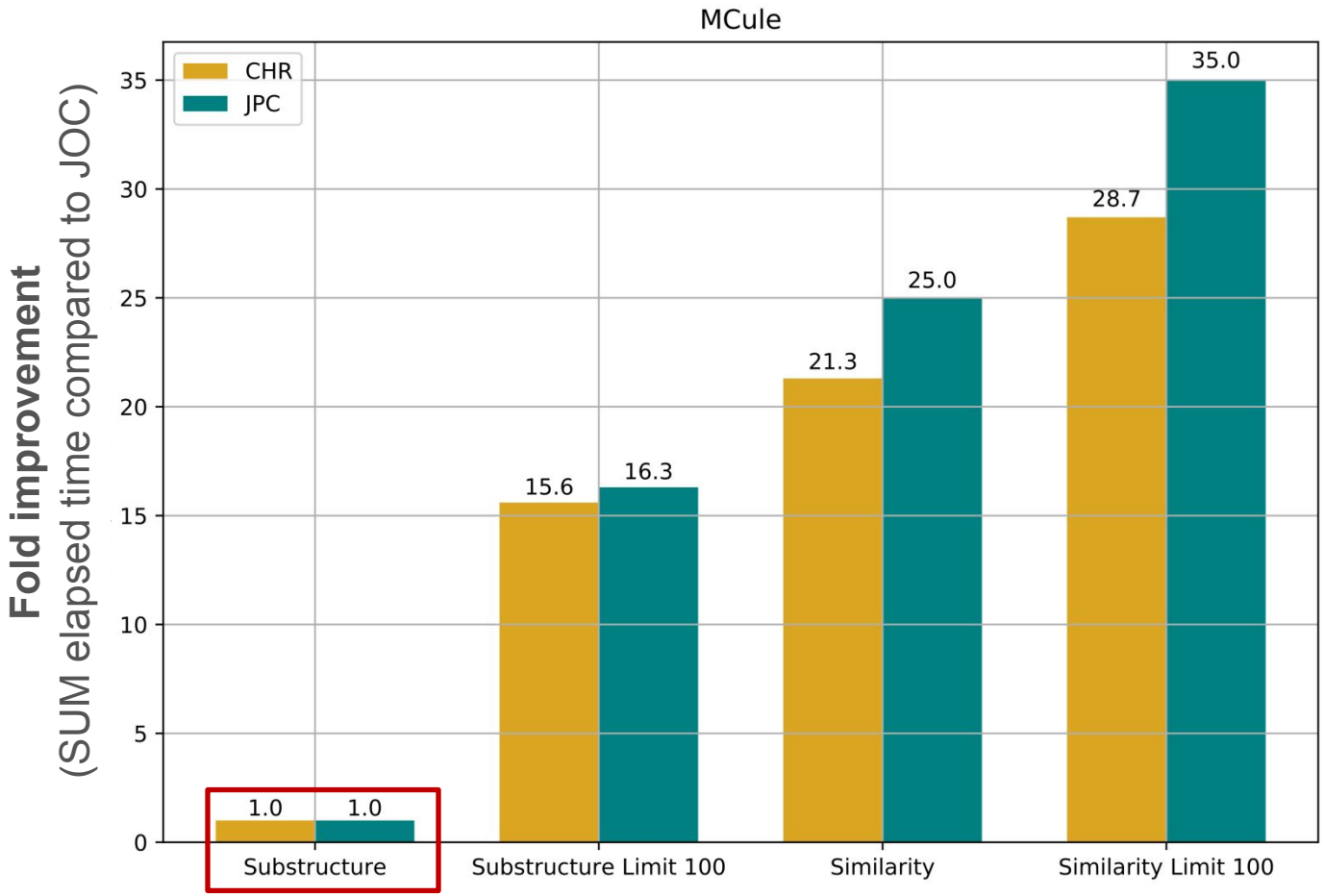
Large, simple data
42M structures, ID

ChemAxon



ChemAxon

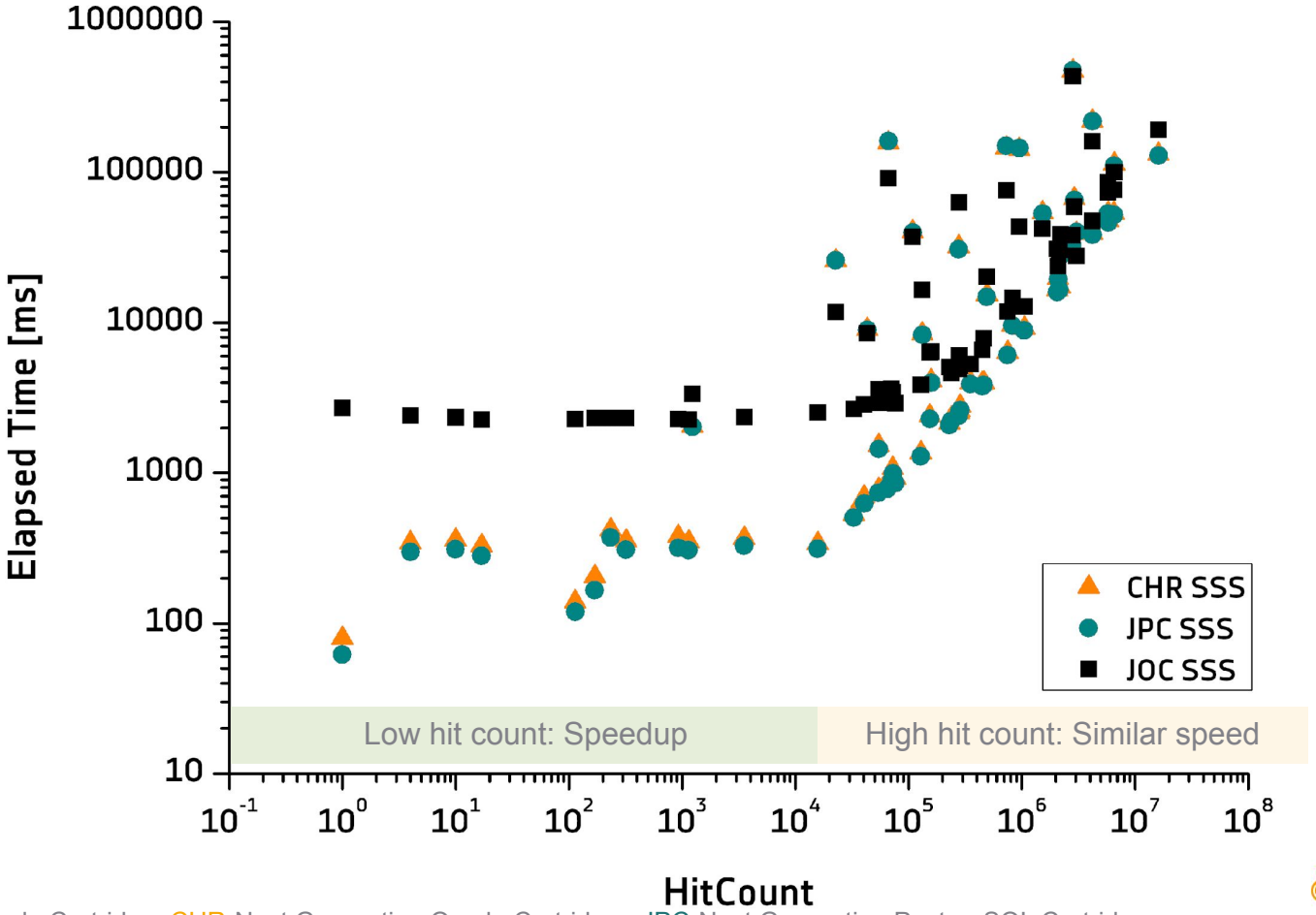
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge



JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon

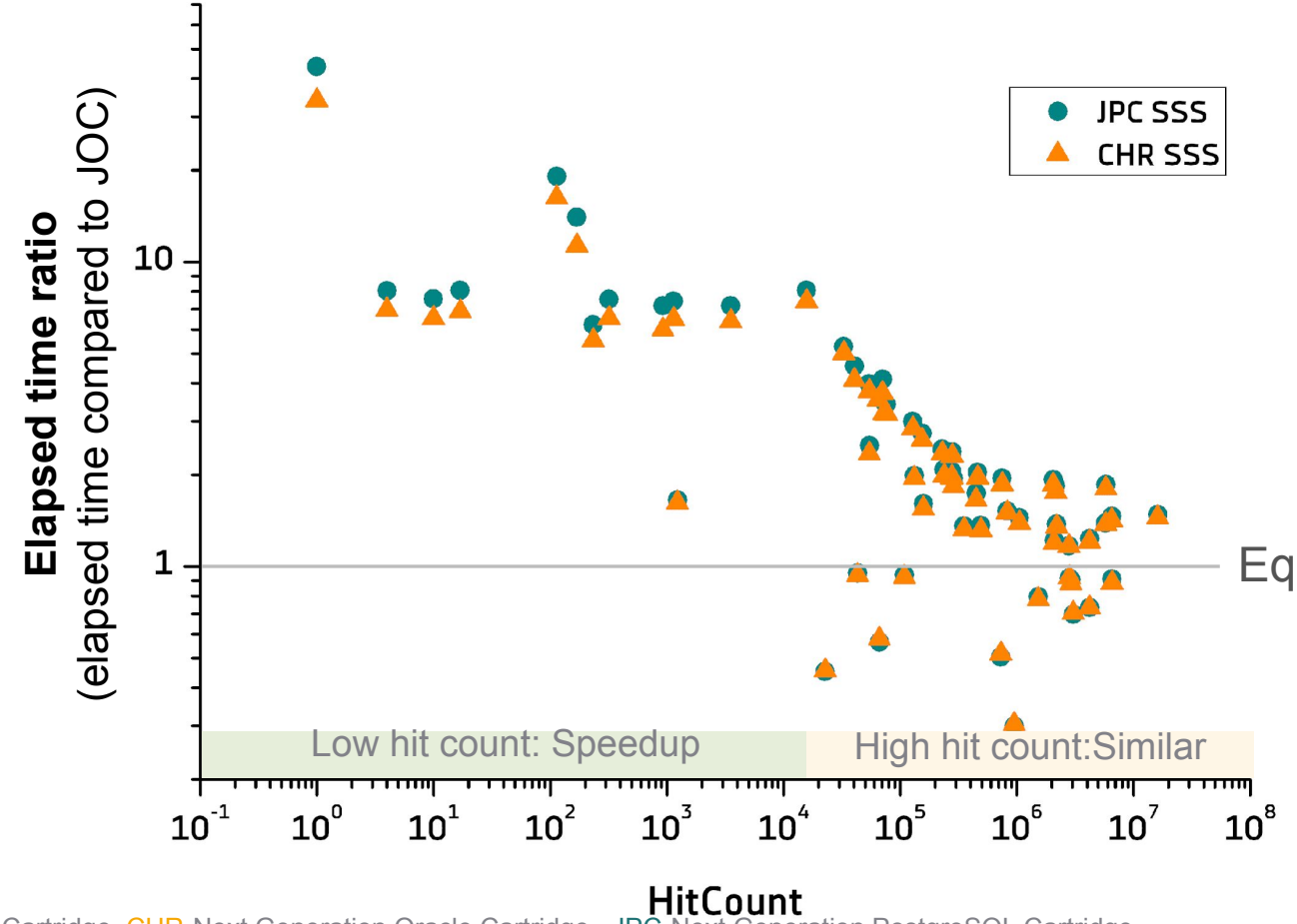
MCule Substructure search



JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon

MCule Substructure search



Median (ratio):
1.9 JPC
1.8 CHR

ChemAxon

JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

SEARCH SPEED ASSESSMENT

Large, simple data
42M structures, ID

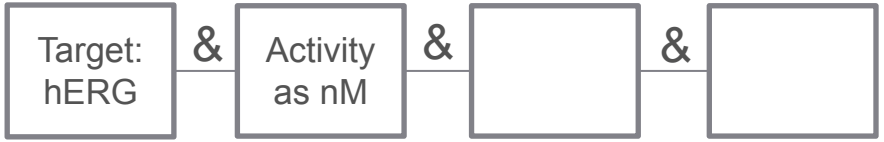
Significant performance improvement on limited substructure and similarity searches

ChemAxon

SEARCH SPEED ASSESSMENT

Real-life data:
~1.8M cpds, ~15M activities

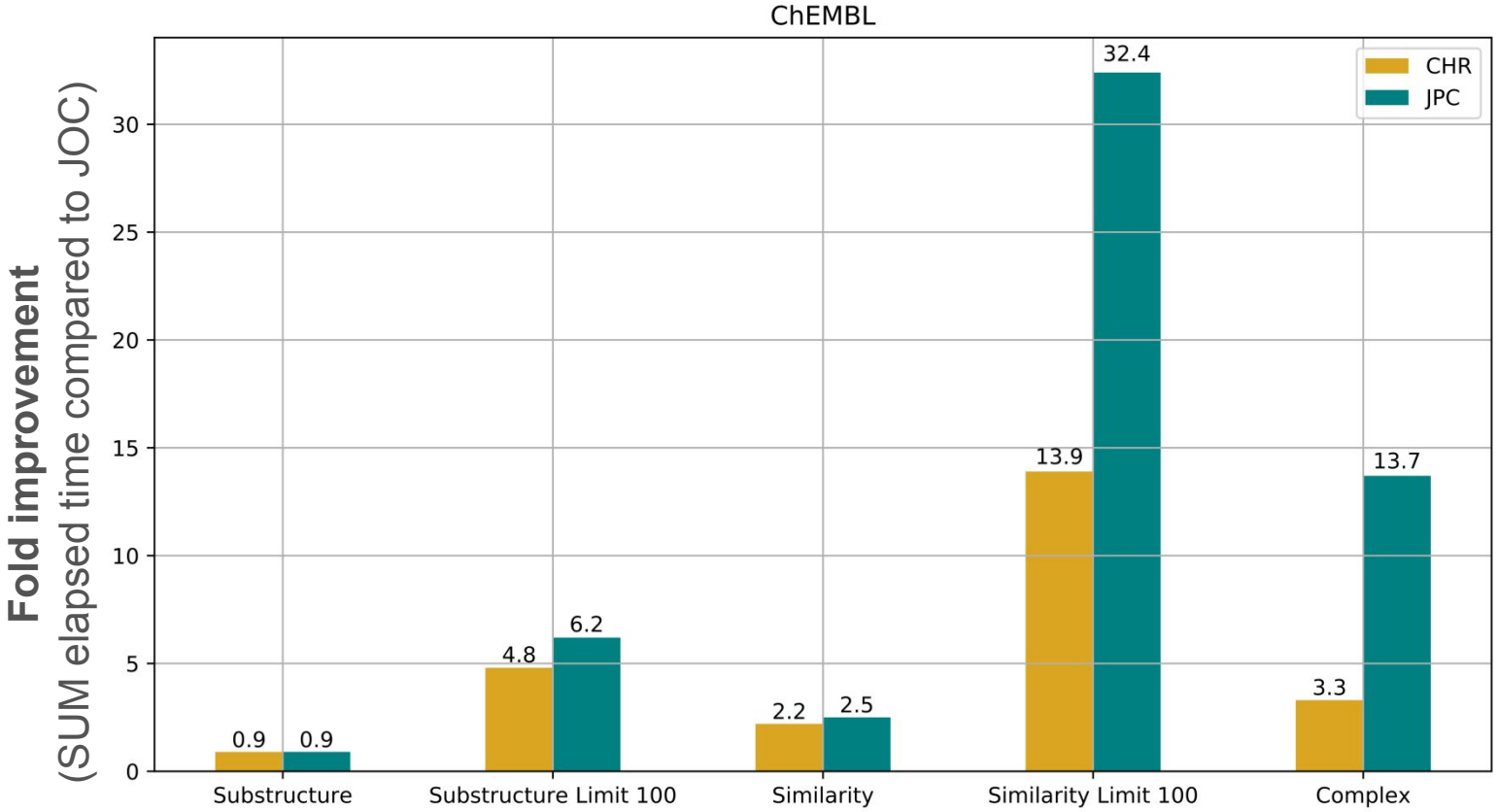
Combined query




```
SELECT count(distinct activities.molregno) FROM activities
JOIN assays ON activities.assay_id = assays.assay_id
JOIN target_dictionary ON assays.tid = target_dictionary.tid
JOIN compound_structures on activities.molregno = compound_structures.molregno
WHERE standard_units = 'nM'
AND standard_value IS NOT NULL
AND jc_compare(compound_structures.molfile, '[#6]-[#6](=O)-[#8]-[#6]-1=[#6]-[#6]=[#6]-[#6]=[#6]-1-[#6](-[#8])=O |c:6,8,t:4|', 't:s')=1
AND target_dictionary.chembl_id = 'CHEMBL1868'
AND jc_compare(compound_structures.molfile, 'c1ccncc1', 't:s')=0;
```

Query elements**: i) activity present (nM unit), ii) substructure match, iii) target name, iv) does not contain pyridine**
For different structures and target chembl_id, not structure is static

ChemAxon



3-10x performance improvement on complex queries

ChemAxon

SEARCH SPEED ASSESSMENT

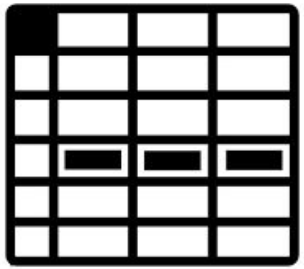
Real-life data:
~1.8M cpds, ~15M activities

Speed improvement on joined queries over complex data

ChemAxon

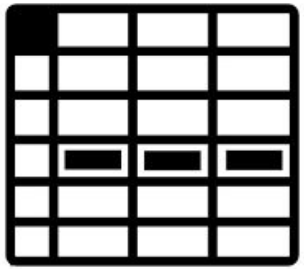
TestCase

10M MCule structures



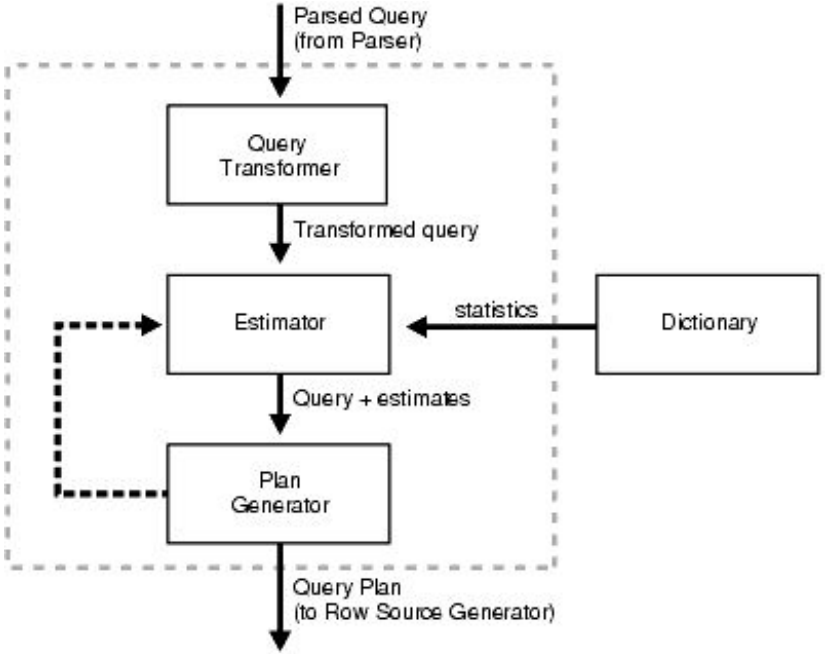
1:n

30M numberic values



ChemAxon

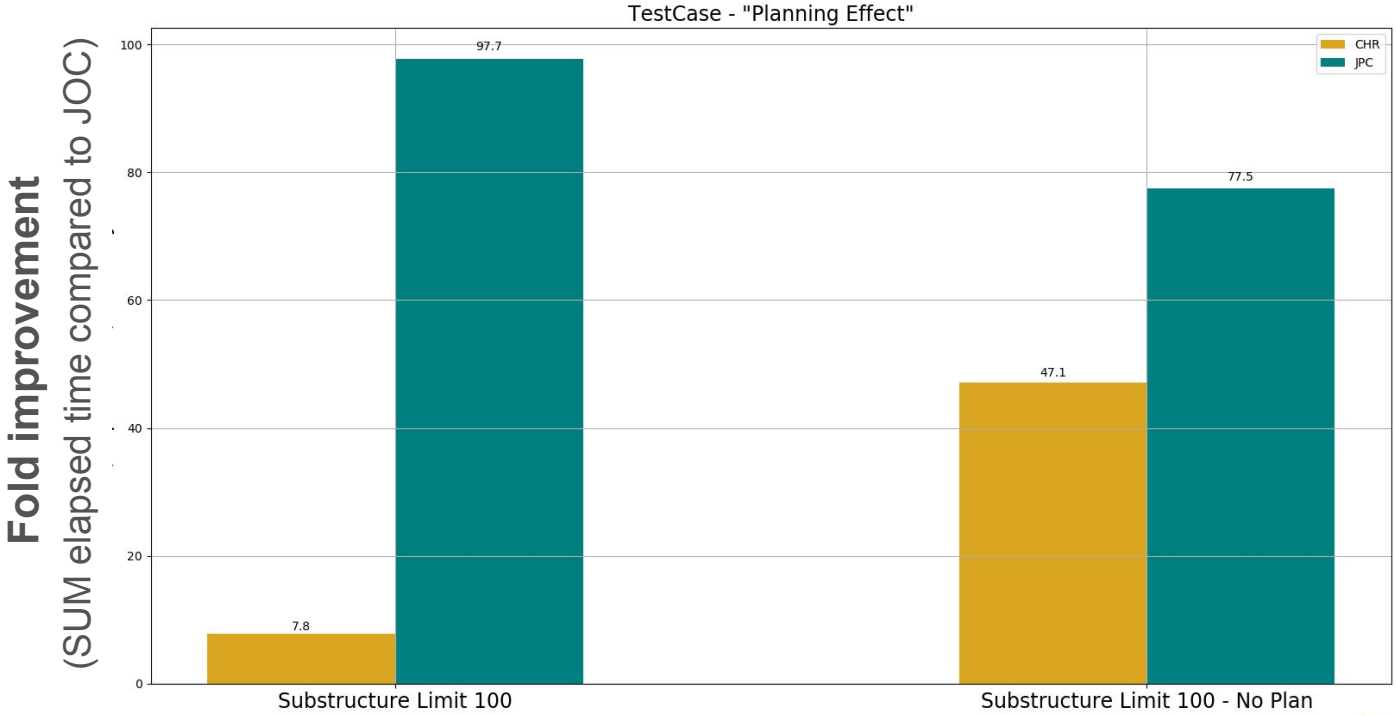
RDBMS INSIST ON PLANNING



https://docs.oracle.com/cd/E18283_01/server.112/e16638/optimops.htm

ChemAxon

WHY TO AVOID PLANNING IF ONLY SSS IS IN THE QUERY?



JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon

HOW TO AVOID PLANNING IF ONLY SSS IS IN THE QUERY?

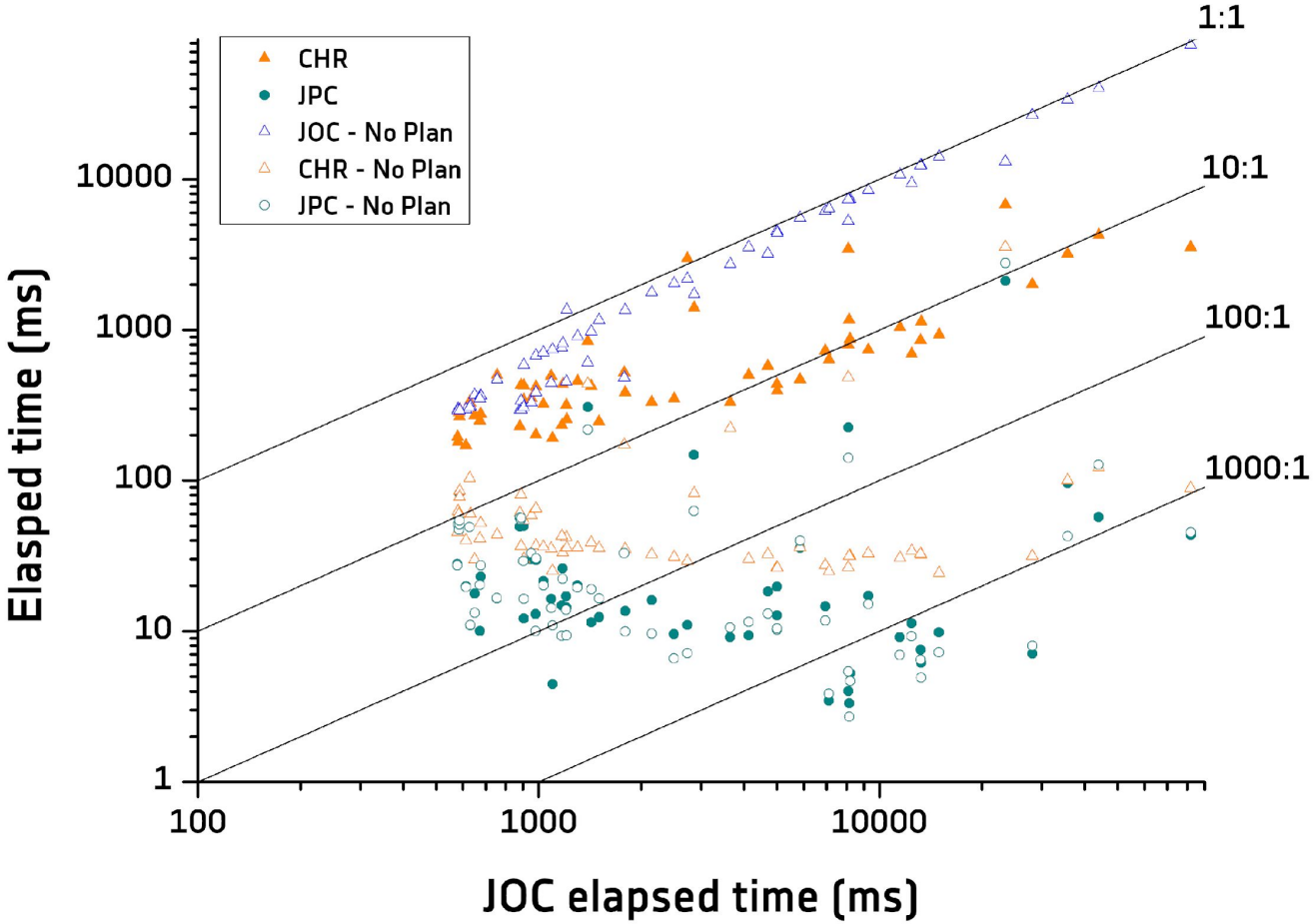
Create a dummy identity function x→x that only returns its argument:

```
create or replace function avoid_planning(v varchar2)
return varchar2 is
begin
  return v;
end;
/
```

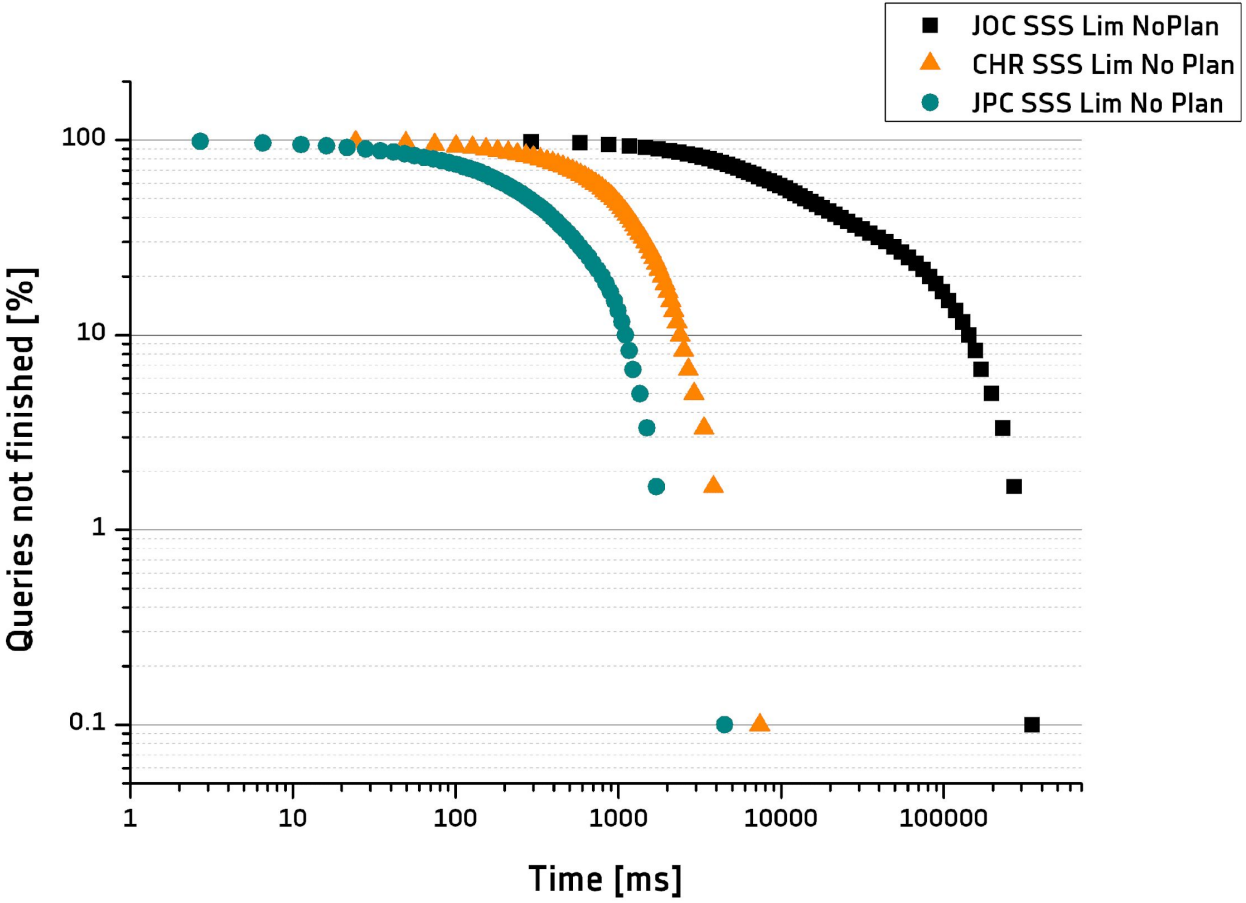
```
SELECT COUNT(*) FROM mcule WHERE
CHORAL_USER.SAMPLE_SEARCH(smi, avoid_planning('O=c1ncccn1'),
'SUBSTRUCTURE')=1;
```

ChemAxon

Limit 100 Substructure search – „Avoid planning" effect



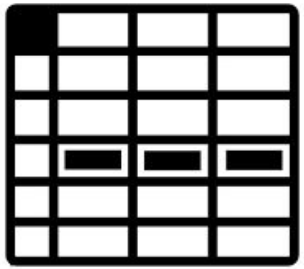
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge


JOC SSS Lim NoPlan
CHR SSS Lim No Plan
JPC SSS Lim No Plan
Queries not finished [%]
100
10
1
0.1
1
10
100
1000
10000
100000
Time [ms]

ChemAxon

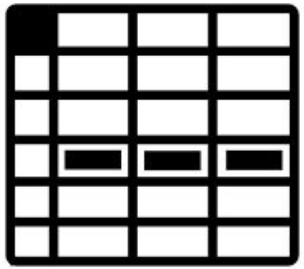
TestCase

10M MCule structures

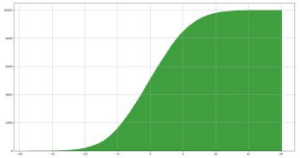


1:n

30M numberic values

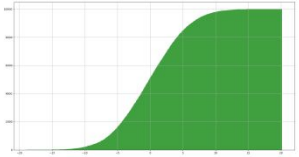


Control over selectivity



ChemAxon

Control over selectivity



Data Non selective ~10M

↓

Chemistry Access

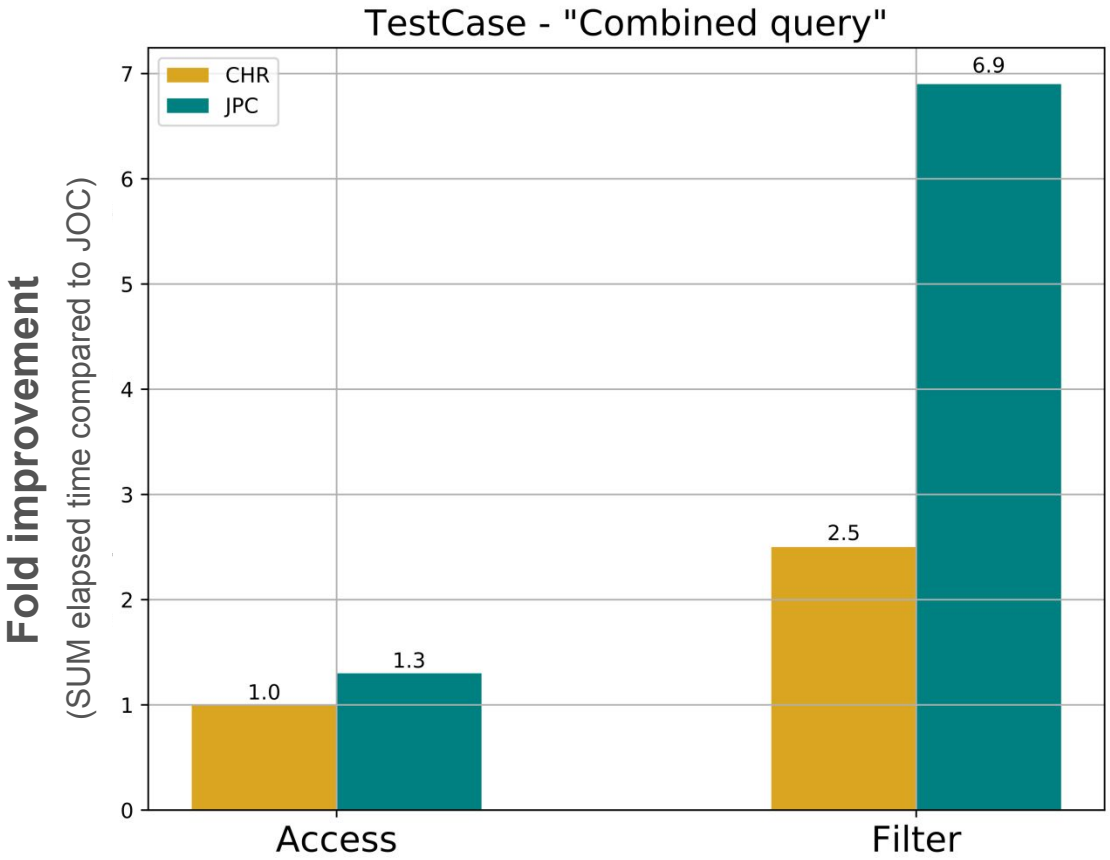
Data Selective ~1k

↓

Chemistry Filter

```
SELECT count(distinct data.id) FROM DATA
JOIN MCULE_10M ON DATA.id = MCULE_10M.id
WHERE DATA.value < ...
AND  jc_compare(MCULE_10M.mol, ...
```

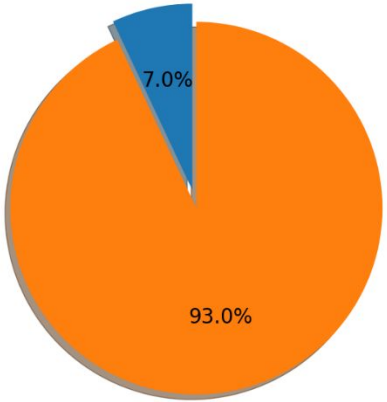
ChemAxon



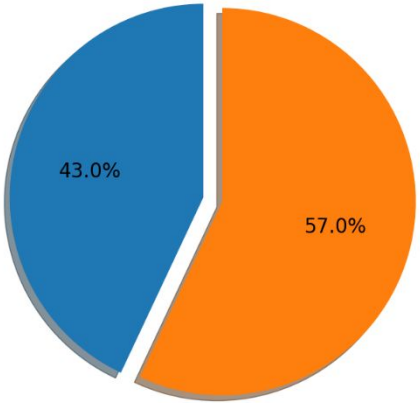
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon

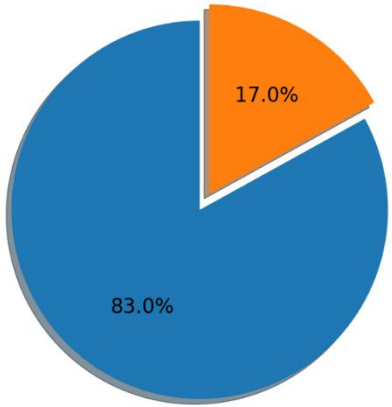
Does the engine obey our hypothesis?
Filter experiment



JOC



CHR

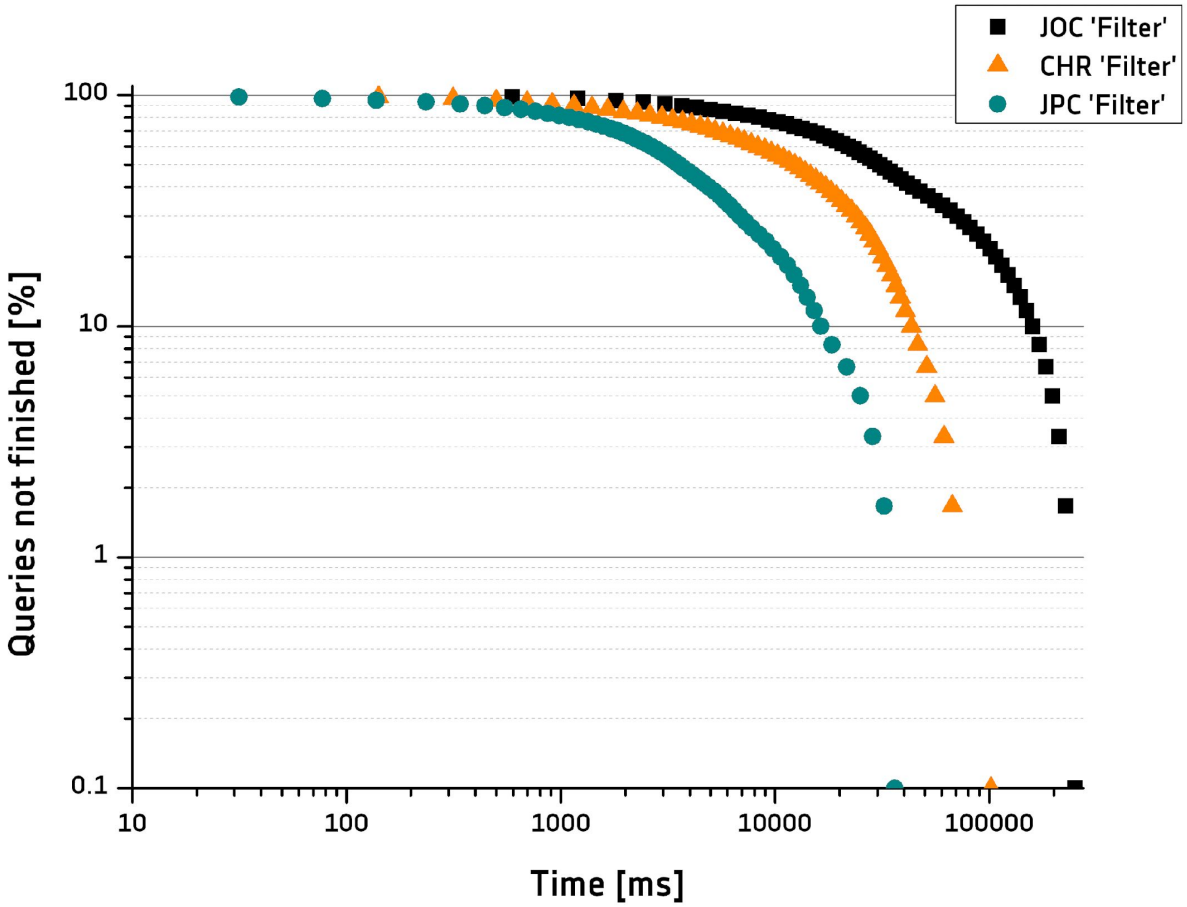


JPC

Access
Filter

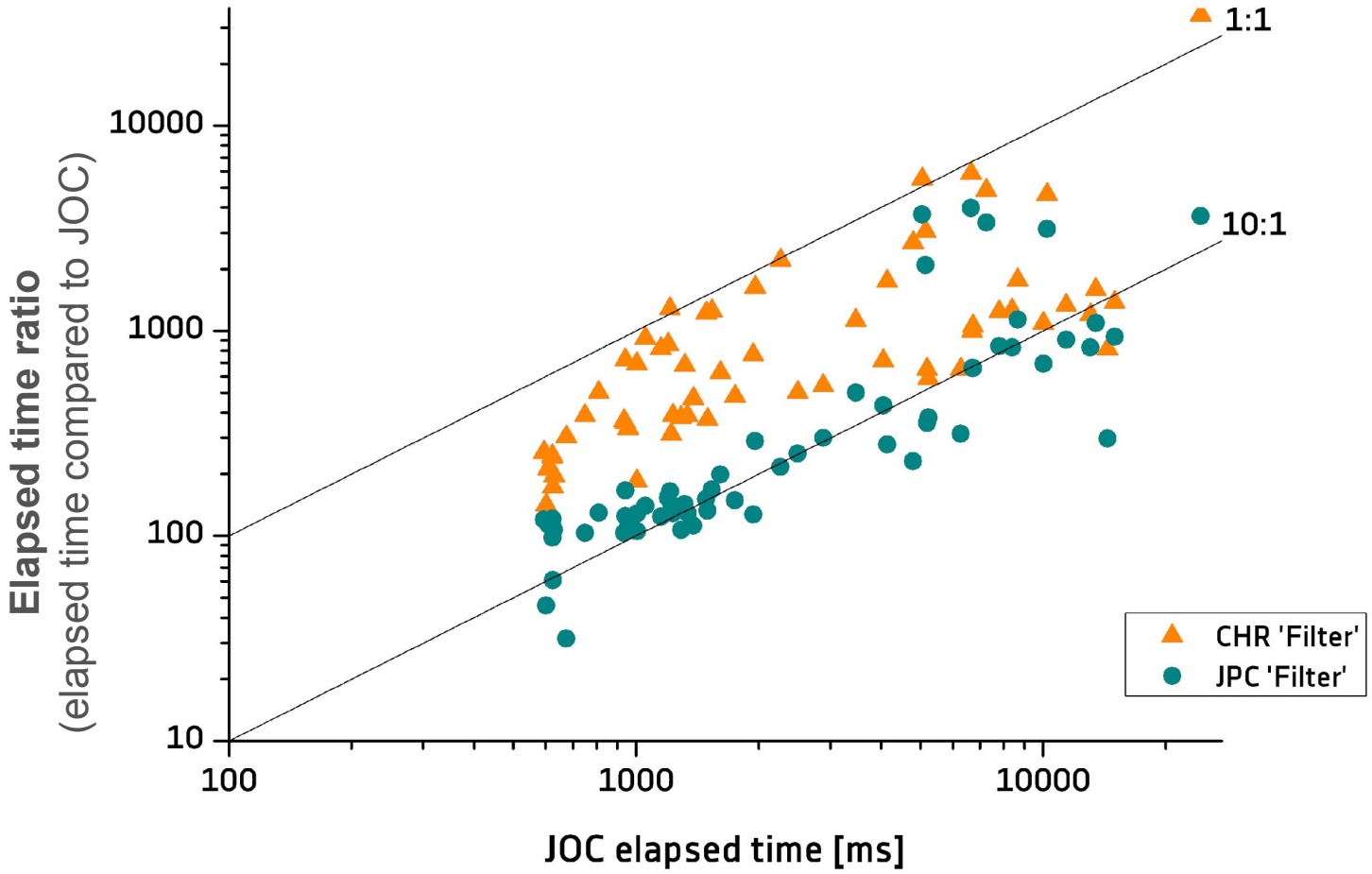
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon



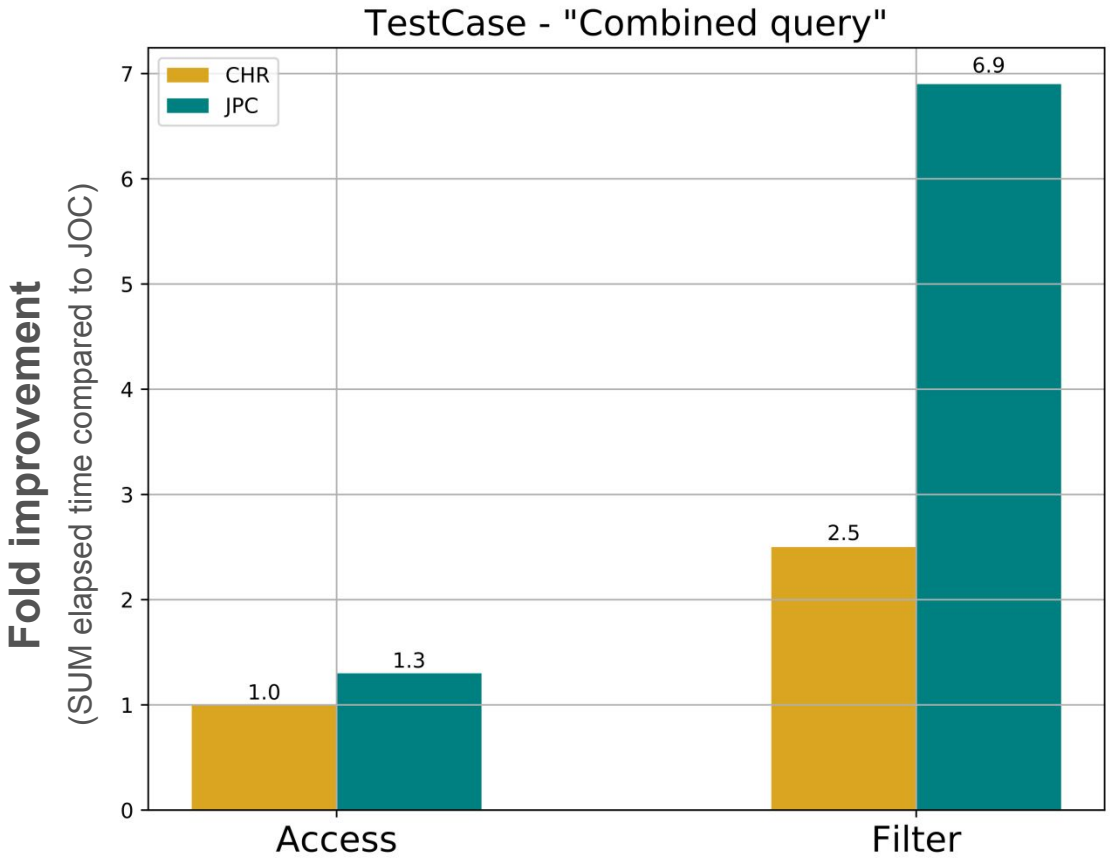
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon



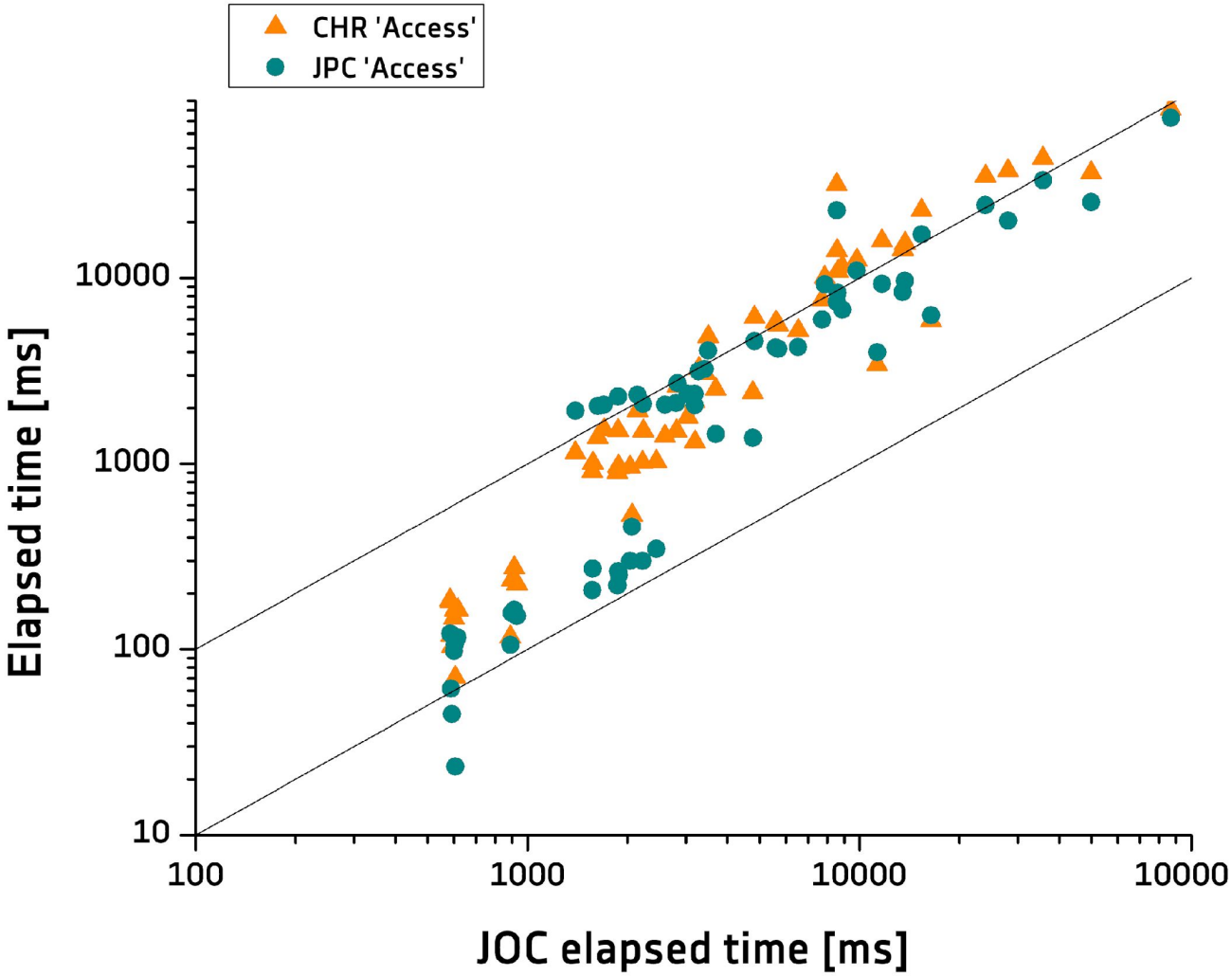
JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon

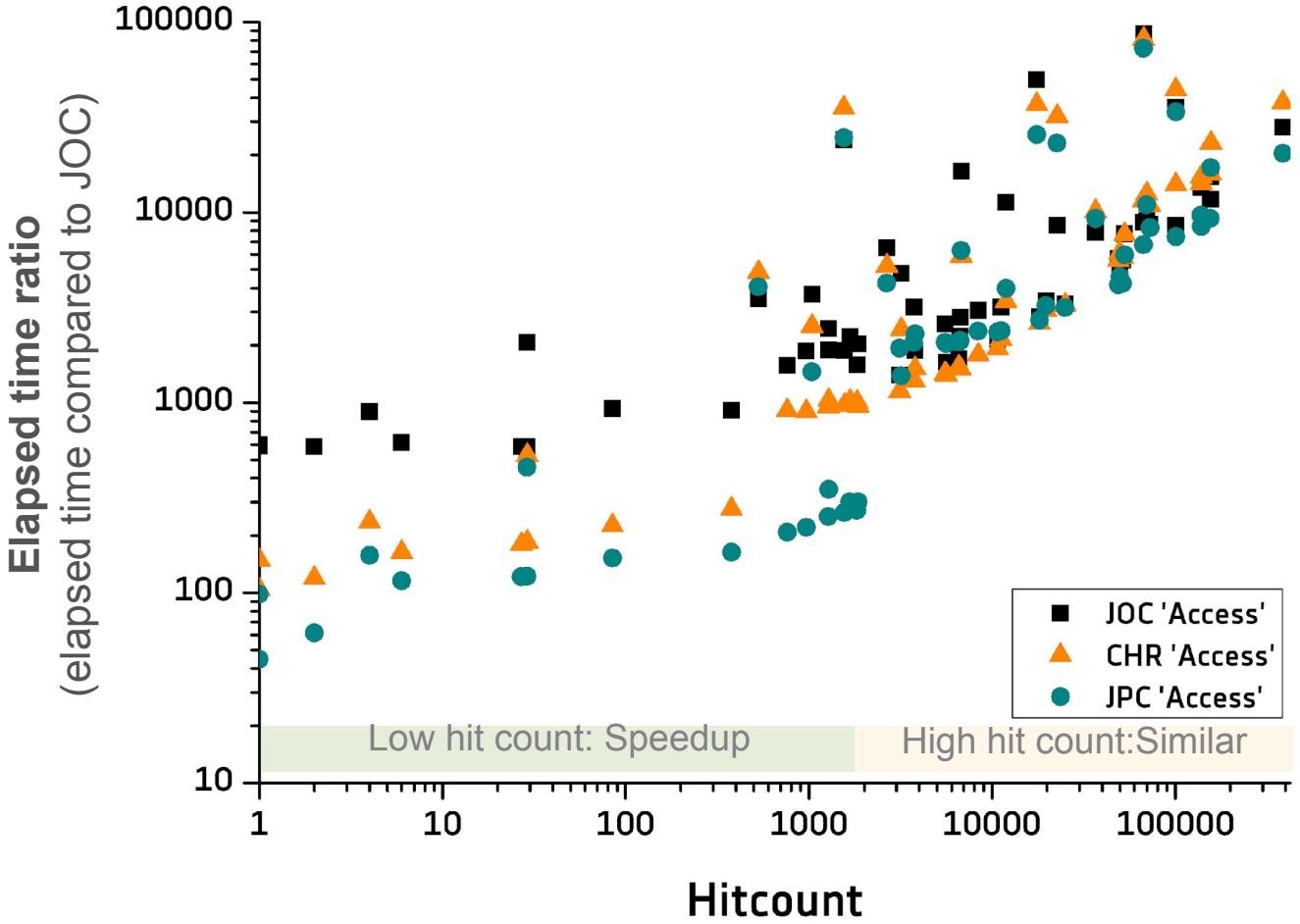


JOC -JChem Oracle Cartridge, CHR-Next Generation Oracle Cartridge, JPC-Next Generation PostgreSQL Cartridge

ChemAxon


CHR 'Access'
JPC 'Access'
Elapsed time [ms]
10
100
1000
10000
100
1000
10000
10000
JOC elapsed time [ms]

ChemAxon


Elapsed time ratio
(elapsed time compared to JOC)
100000
10000
1000
100
10
1
10
100
1000
10000
100000
Hitcount
JOC 'Access'
CHR 'Access'
JPC 'Access'
Low hit count: Speedup
High hit count:Similar

ChemAxon

Next generation cartrdiges (Choral, JPC):

- sorted SSS hits

- early hits, agile search

- large sets with low memory setup

Thank you

ChemAxon