Sachem:

A chemical cartridge for high-performance substructure search

Jakub Galgonek, Miroslav Kratochvíl and Jiří Vondrášek
Institute of Organic Chemistry and Biochemistry of the CAS

elixir CZECH REPUBLIC

Sachem

- means “the chief of an Indian tribe”
- is implemented as a PostgreSQL extension
- allows substructure search in large datasets
- is open-source
- outperforms other open-source cartridges
- also allows similarity search experimentally

elixir CZECH REPUBLIC

Fingerprint-based substructure search methods

- Step 1: screening
  - quickly screen out compounds that can be identified as unable to match the query
  - fingerprints are designed to capture important structural features of compounds – each fingerprint bit represents some structure feature
  - a feature present in a compound structure must also be present in all its structural extensions
- Step 2: substructure matching
  - instance of a relatively hard subgraph isomorphism problem
  - NP-complete

elixir CZECH REPUBLIC

Fingerprint-based substructure search methods

- currently used fingerprints
  - represented as binary vectors
  - typically contain hundreds or thousands bits
  - special part:
    - manually selected features
  - common part:
    - hashes of three-atom SMILES substructures
    - hashed cyclic subgraphs (up to 8 atoms) and hashed sub-trees (up to 7 atoms)
    - hashed linear paths (up to 7 atoms)

elixir CZECH REPUBLIC

Fingerprint-based substructure search methods

- currently used indexing methods:
  - GiST index
  - B-tree index
  - bitmap index

elixir CZECH REPUBLIC

Sachem

- Sachem/OrChem
  - performance-oriented reimplementation of OrChem
  - OrChem fingerprint
  - static bitmap index
- Sachem/Lucy
  - introduces our own fingerprint
  - full-text index

elixir CZECH REPUBLIC

Sachem/Lucy Fingerprint

- Defines bits of several categories:
  - each distinct atom is considered as a distinct fingerprint bit
  - all smallest rings of all compounds in the ChEBI is considered a substructural pattern for a fingerprint bit
  - all connected subgraphs of with a maximum of one ring and a limited number of bonds are considered a fingerprint bit
  - multiplicity of features was encoded by creating a new fingerprint bit for each power of 2 of the repetitions
- In PubChem:
  - 18.7 million distinct fingerprint bits
  - 860 non-zero bits per compound on average

elixir CZECH REPUBLIC

Sachem/Lucy Index

- identification of each fingerprint bit is encoded to a 6-byte word
- each compound stored in a database is expressed as a document containing words of its non-zero fingerprint bits
- documents are indexed by full-text index
- employs Apache Lucy engine library providing full-text search

elixir CZECH REPUBLIC

Query Fingerprint Bit-reduction Algorithm

- Step 1: discard redundant bits
  - if a occurrence of a bit implies a occurrence of other bits, than these bits are redundant in the query and can be discarded
  - has no effect on screening precision
- Step 2: discard less relevant bits
  - discard bits that are less relevant based on their statistical relevance for search
  - takes into the account that all query atoms have to be covered by several selected bits

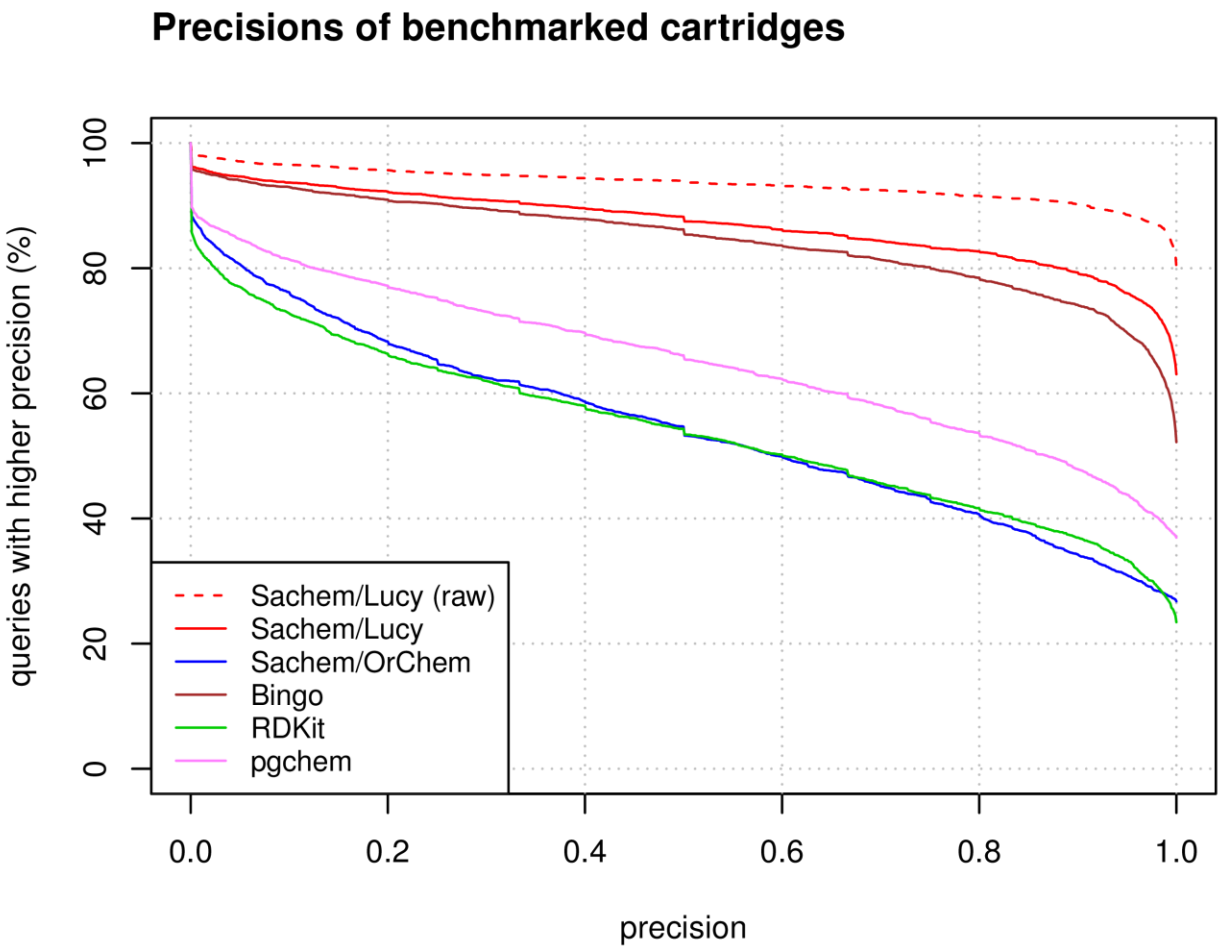
elixir CZECH REPUBLIC

Benchmark Datasets

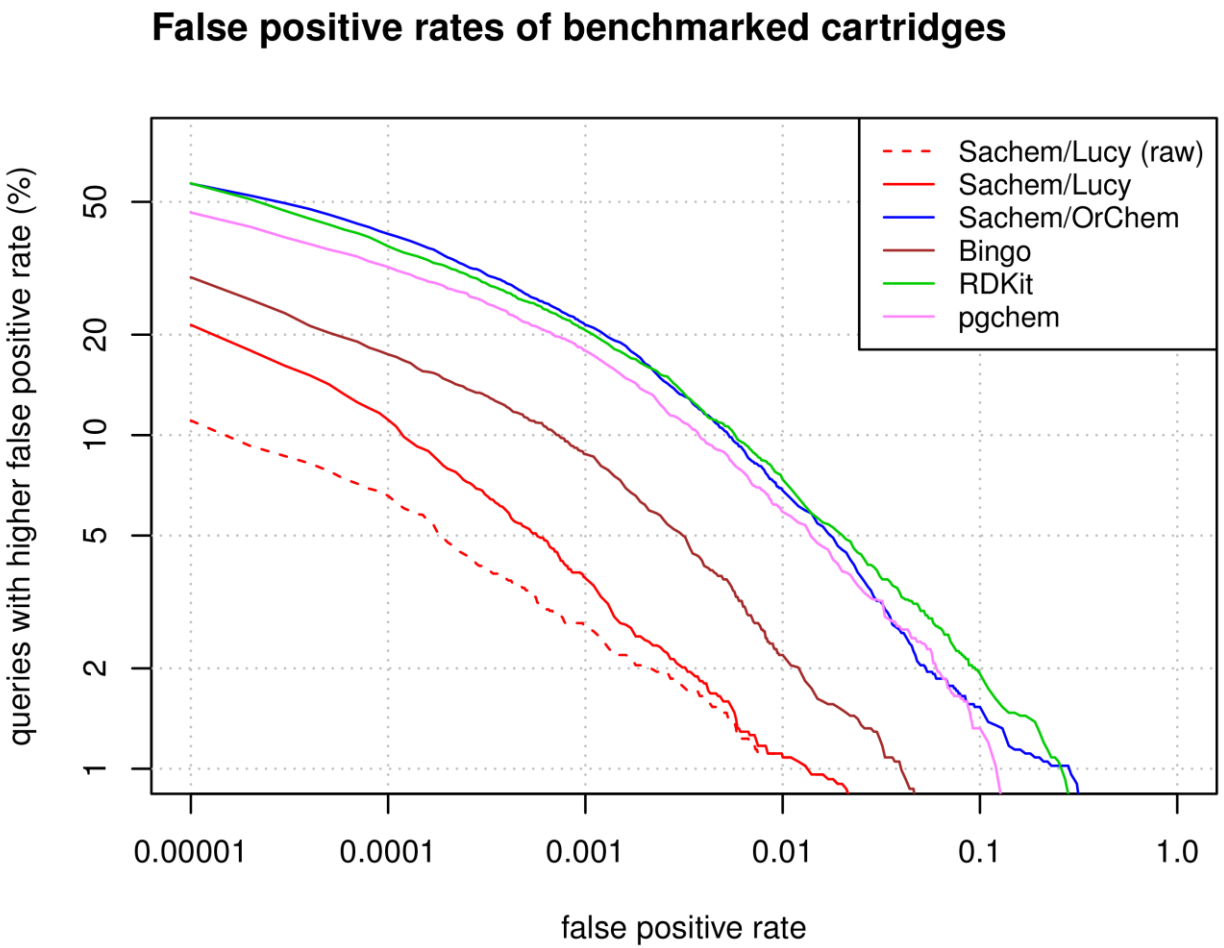
- Query Set:
  - 3329 queries from Substructure Query Collection (SQC)
- Datasets:
  - 94M: all compounds in the PubChem database
  - 10M: 10 million random compounds from PubChem
  - 1M: 1 million random compounds from 10M

elixir CZECH REPUBLIC

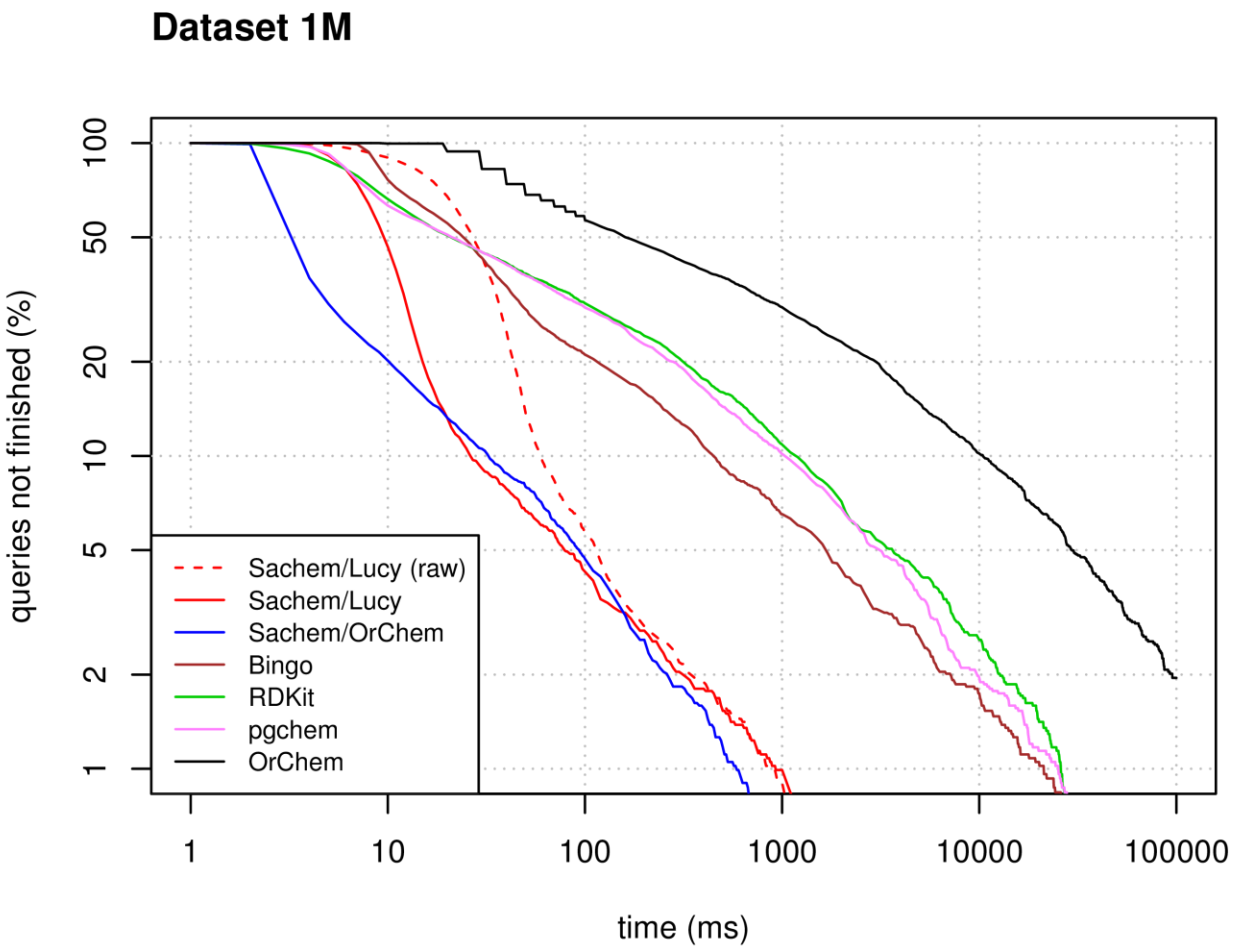
Screening Efficiency: Precision



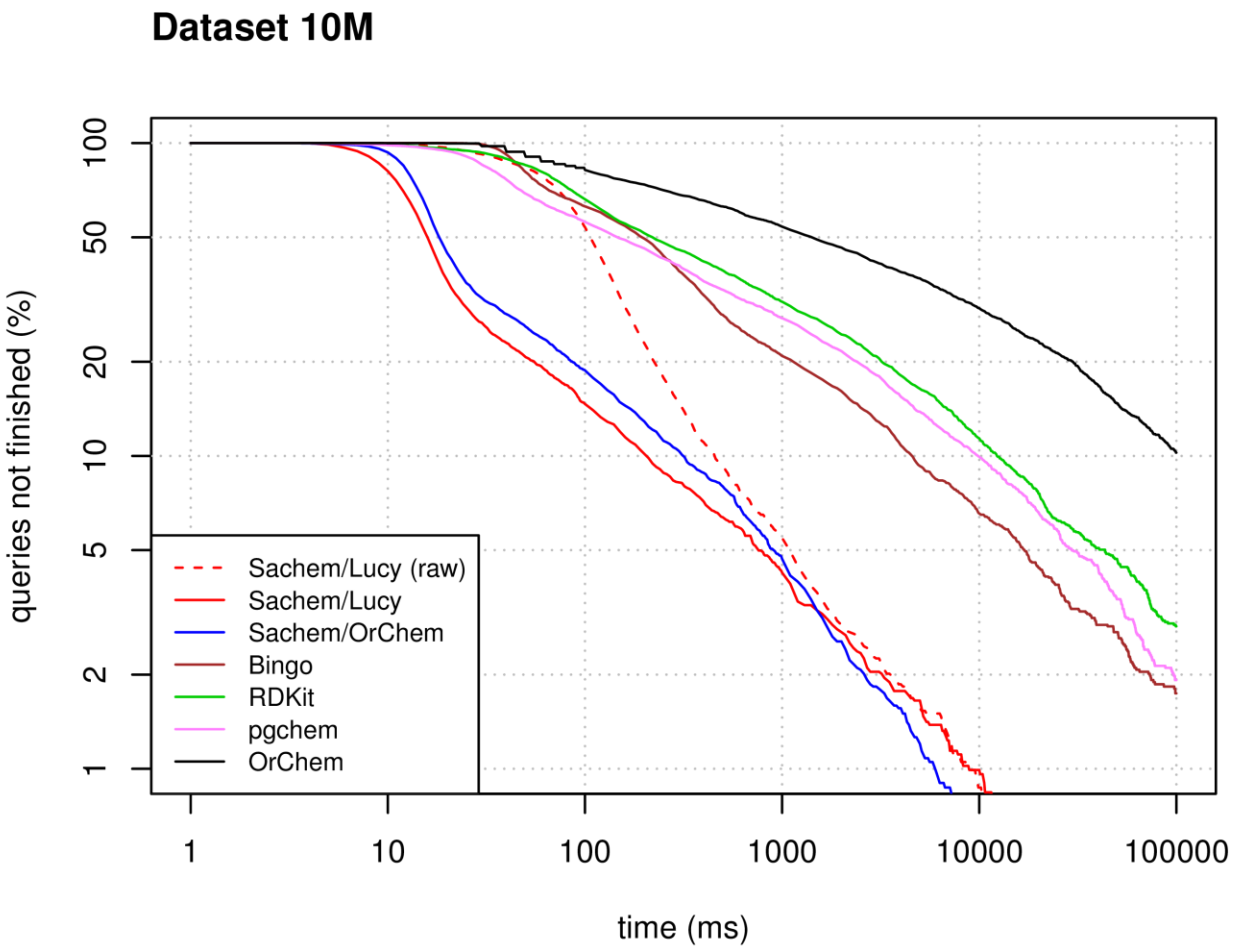
Screening Efficiency: False Positive Rate



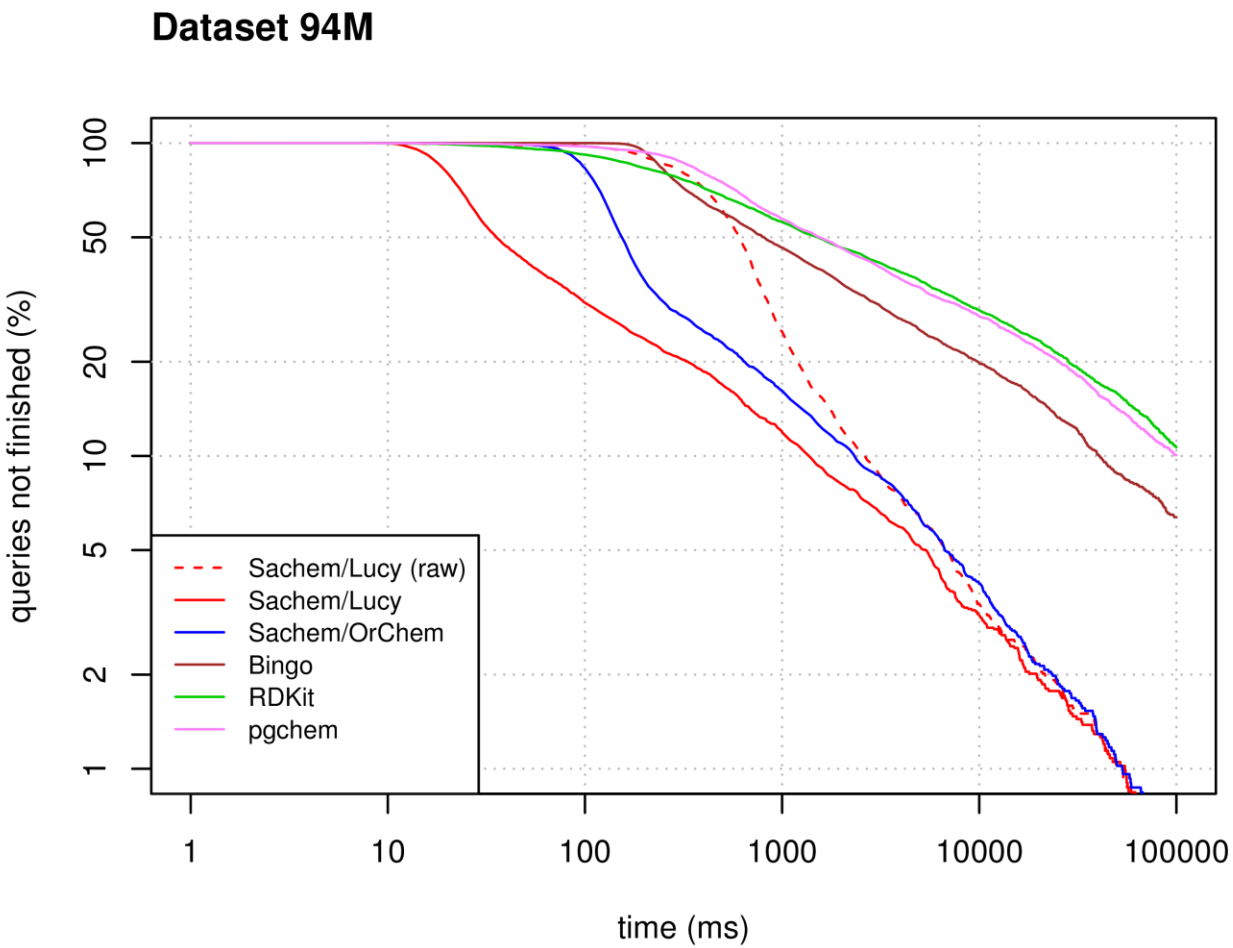
Overall performance comparison: 1M Dataset



Overall performance comparison: 10M Dataset



Overall performance comparison: 94M Dataset

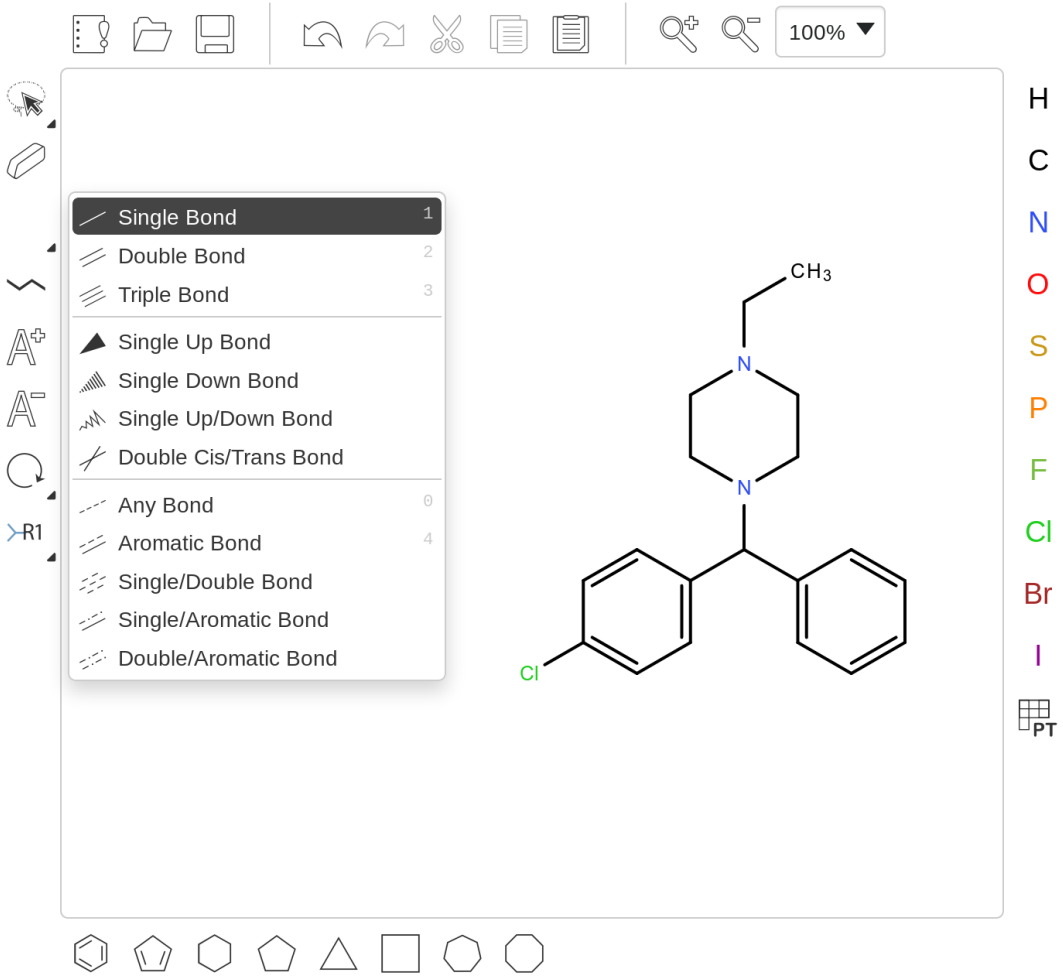


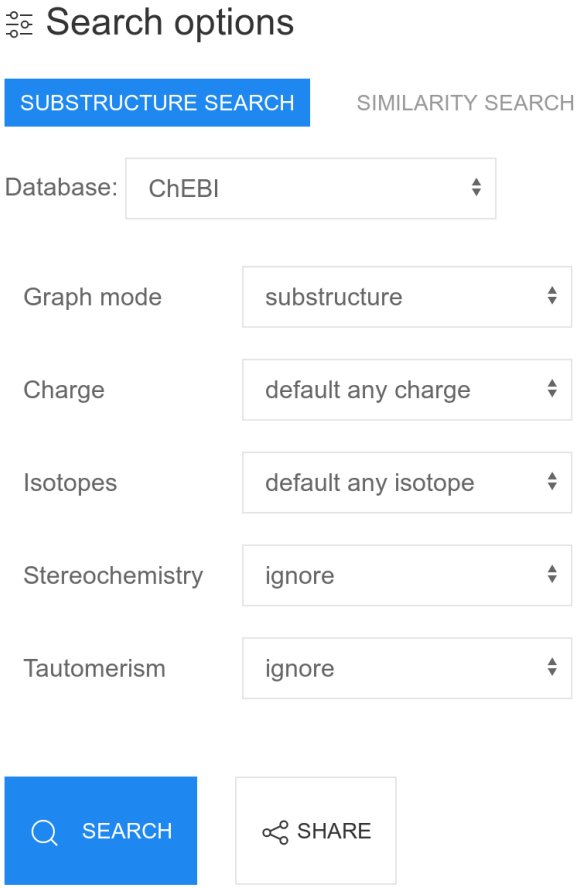
Sachem Web Interface

- beta version
- available at https://idsm.elixir-czech.cz/sachem/
- employs EPAM Ketcher to draw structures
- allows to use Sachem to search compounds in
  - Drugbank: ~ 9,000 compounds
  - ChEBI: ~ 98,000 compounds
  - ChEMBL: ~ 1.7 million compounds
  - PubChem: ~ 95 million compounds

elixir CZECH REPUBLIC

Sachem Web Interface





SPARQL Endpoint

- technical preview demo
- examples available at https://idsm.elixir-czech.cz/sparql/
- web application is based on YASGUI
- allows to integrate search with other SPARQL services
  - neXtProt
  - UniProt
  - ChEMBL

elixir CZECH REPUBLIC

SPARQL Interface: Example Query

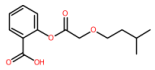
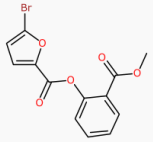
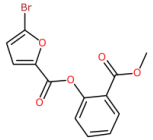
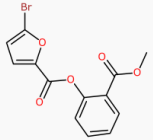
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX chembl: <http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX uniprot: <http://purl.uniprot.org/core/>
PREFIX sachem:<http://bioinfo.uochb.cas.cz/sparql-endpoint/sachem/>

SELECT ?COMPOUND ?UNIPROT ?ORGANISM_NAME WHERE
{
  SERVICE <https://www.ebi.ac.uk/rdf/services/sparql>
  {
    SERVICE sachem:chembl {
      ?COMPOUND sachem:substructureSearch [ sachem:query "CC(=O)Oc1ccccc1C(O)=O" ]
    }

    ?ACTIVITY rdf:type chembl:Activity; chembl:hasMolecule ?COMPOUND; chembl:hasAssay ?ASSAY.
    ?ASSAY chembl:hasTarget ?TARGET.
    ?TARGET chembl:hasTargetComponent ?COMPONENT.
    ?COMPONENT chembl:targetCmptXref ?UNIPROT.
    ?UNIPROT rdf:type chembl:UniprotRef.
  }

  ?UNIPROT uniprot:organism ?ORGANISM.
  ?ORGANISM uniprot:scientificName ?ORGANISM_NAME.
}
```

SPARQL Interface: Example Query

Table | Response | Pivot Table | Google Chart | Geo

Showing 1 to 50 of 1,024 entries (in 1.724 seconds)

Search: Show 50 entries

	COMPOUND	UNIPROT	ORGANISM_NAME
1	 http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL1598323	http://purl.uniprot.org/uniprot/P80244	Bacillus subtilis (strain 168)
2	 http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL1377610	http://purl.uniprot.org/uniprot/P9WHJ3	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)
3	 http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL1377610	http://purl.uniprot.org/uniprot/P9WMR3	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)
4	 http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL1377610	http://purl.uniprot.org/uniprot/P9WMR3	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)

elixir CZECH REPUBLIC

SPARQL Interface: Example Query

Table | Response | Pivot Table | Google Chart | Geo

Table

Available Variables

COMPOUND UNIPROT

Cells

Count

Columns

Rows

ORGANISM_NAME

ORGANISM_NAME	Totals
Homo sapiens	769
Rattus norvegicus	82
Ovis aries	79
Bos taurus	24
Mus musculus	12
Escherichia coli	8
Oryctolagus cuniculus	8
Cavia porcellus	8
Escherichia coli (strain K12)	5
Staphylococcus aureus	4
Enterobacter cloacae	4
Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)	3
Bacillus licheniformis	2
Geobacillus stearothermophilus	2
Photinus pyralis	2
Human immunodeficiency virus 1	2
Equus caballus	2
Plasmodium falciparum (isolate 3D7)	2

elixir CZECH REPUBLIC

Thank you

Thank you for your attention!

elixir CZECH REPUBLIC