ChemAxon

Gábor Imre

MADFAST SIMILARITY SEARCH

Brief history

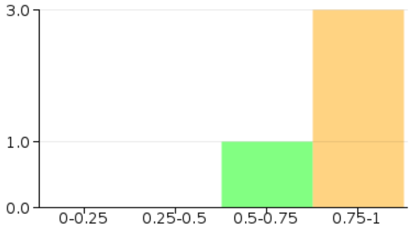
A similarity based overlap analysis of 5k query structures was needed to be executed against 12M targets - all under 1h.

Explored two approaches:

- Use clustering based heuristic to reduce set sizes.

- Optimize multy query similarity search implementation.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Q1	< .9	< .9	< .9	.9	< .9	< .9	< .9	< .9	< .9	< .9
Q2	< .9	< .9	< .9	< .9	.9	< .9	< .9	< .9	< .9	< .9
Q3	< .8	< .8	< .8	< .8	< .8	< .8	.8	< .8	< .8	< .8
Q4	< .7	< .7	< .7	.7	< .7	< .7	< .7	< .7	< .7	< .7



Who won?

Multi core machines are fast and can have huge memory:

- Exhaustive search won
- Original goal was reached with a few minutes execution time

The fundamentals of a faster than expected similarity search engine was born.
What could we do with it?

- Overlap analysis of large sets?
- Push the limits of similarity based clustering?
- Real time search?

REAL TIME SEARCH

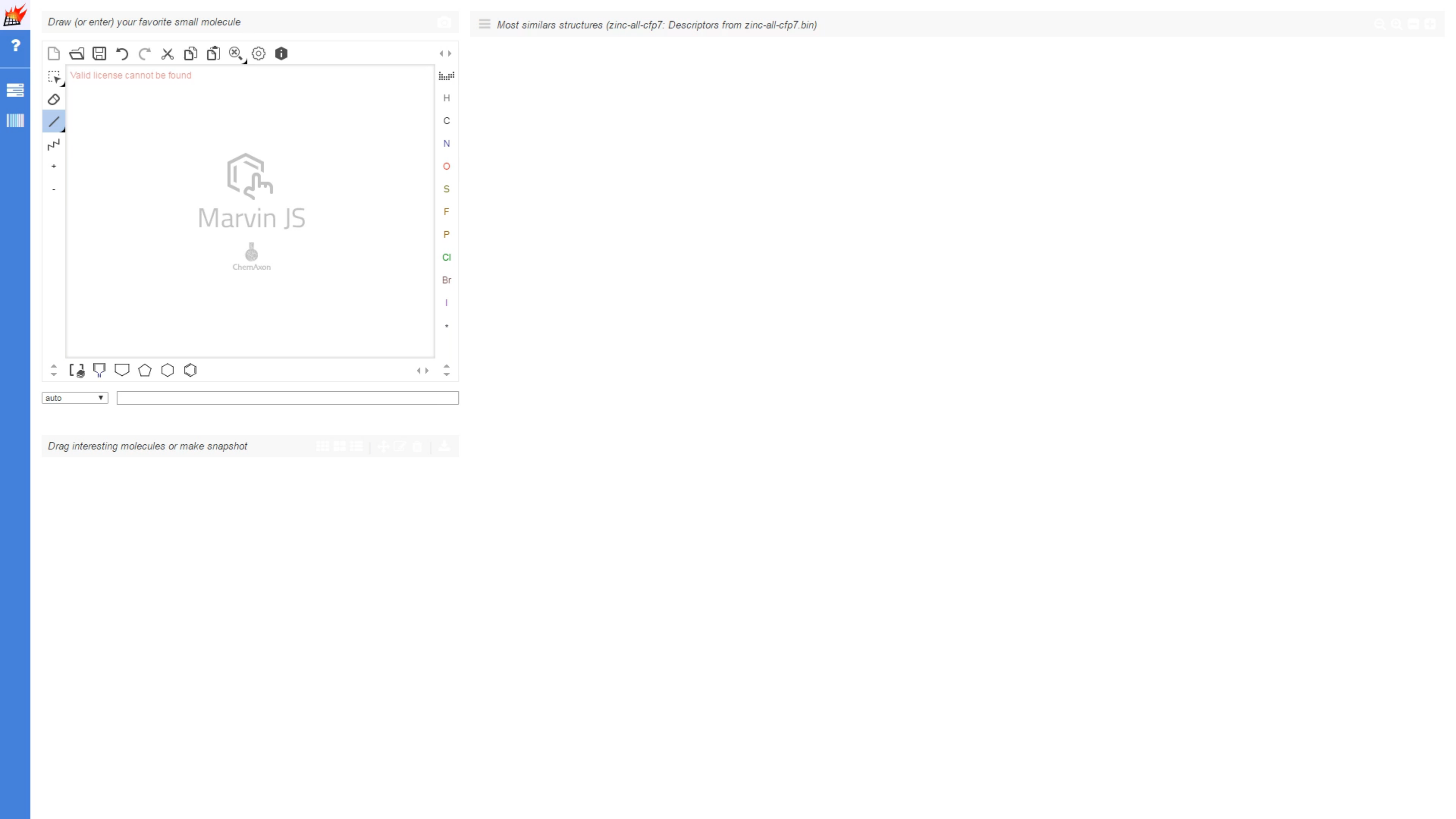
Demo

Demo

Lets see real time search on a large set:

- Search against the Zinc database containig 16M structures

- On an Amazon EC2 virtual machine (32vCPU class)



Draw (or enter) your favorite small molecule

Valid license cannot be found

Drag interesting molecules or make snapshot

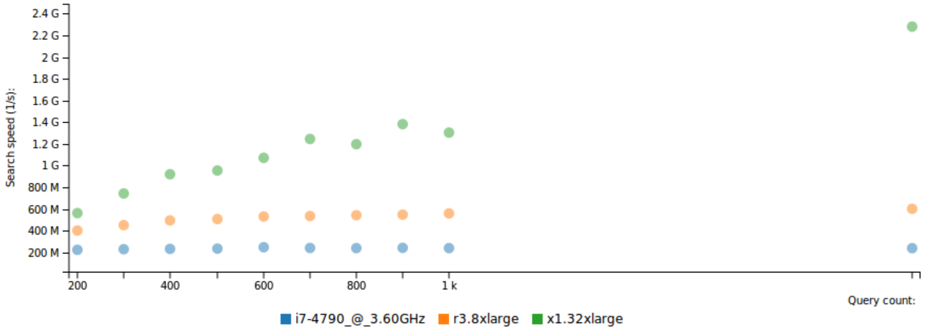
Most similars structures (zinc-all-cfp7: Descriptors from zinc-all-cfp7.bin)

Performance

Similarity search time was ~0.08 sec (80 ms) per query translating to ~5ns per query-target comparison. The most similar targets are shown as you type/draw.

Efficiency of multi query search:

- With >600 query batches >600M comparison/s sustained on a c3.8xlarge instance

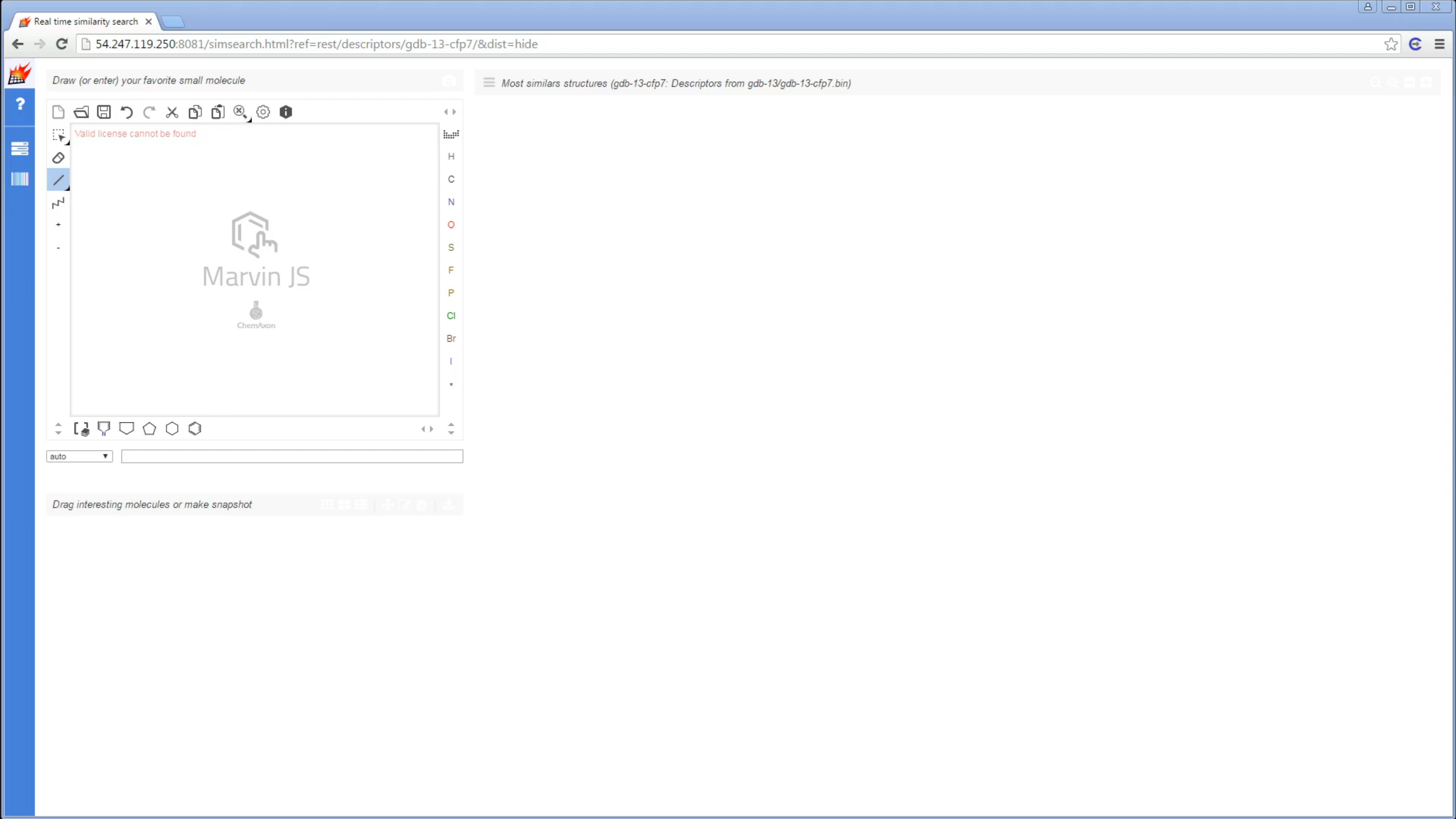


- Or 2.2G comparison/s on an x1.32xlarge instance - <8min run time for doing an 1M x 1M exhaustive search

Go larger

Searching against 16M targets is fast. What are the limits?

- Amazon EC2 provides r3.8xlarge instance with 32 vCPU and 244 GB memory
- GDB-13 is the largest publicly available small organic molecules database containing 977M structures. (*Small organic molecules enumerated up to 13 atoms of C, N, O, S and CL following simple chemical stability and synthetic feasibility rules.*)



Real time similarity search

54.247.119.250:8081/simsearch.html?ref=rest/descriptors/gdb-13-cfp7/&dist=hide

Draw (or enter) your favorite small molecule

Valid license cannot be found

Most similars structures (gdb-13-cfp7: Descriptors from gdb-13/gdb-13-cfp7.bin)

Drag interesting molecules or make snapshot

Further notes

- Nearly 1B structures were the limits for the r3.8xlarge instance type.

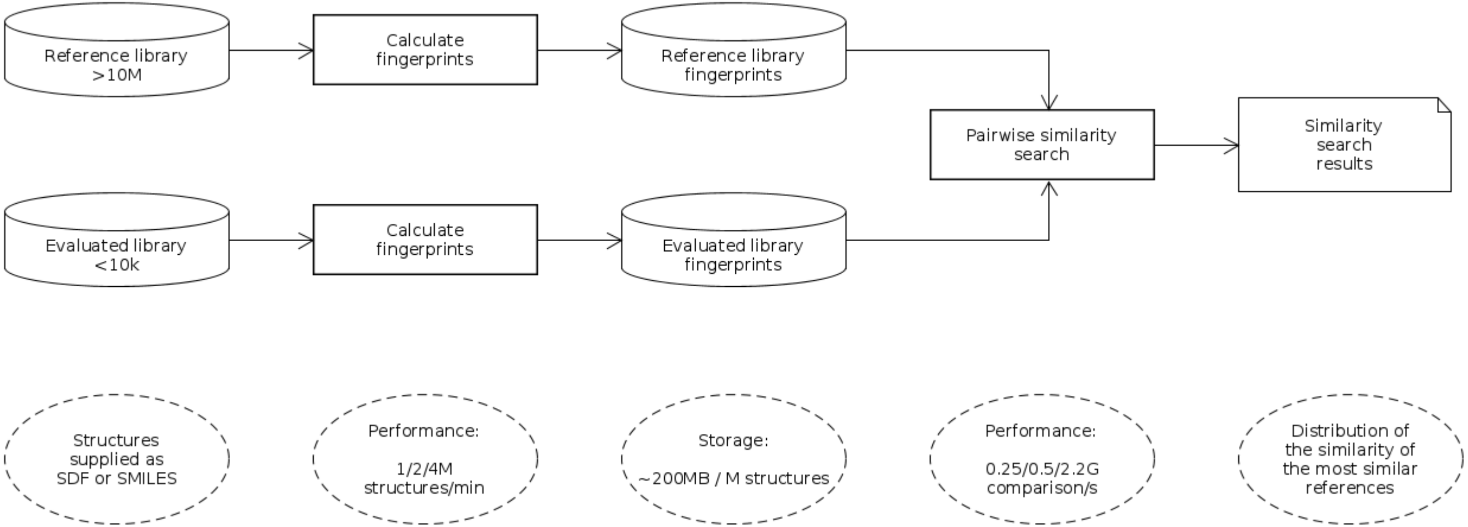
- The new x1.32xlarge contains nearly 2TB RAM and 128vCPU, for ~13$/h

- So even ~8B structures could be handled using a single machine

OTHER USE CASES

Beyond real time similarity search

Similarity based overlap analysis



Reference library >10M

Calculate fingerprints

Reference library fingerprints

Evaluated library <10k

Calculate fingerprints

Evaluated library fingerprints

Pairwise similarity search

Similarity search results

Structures supplied as SDF or SMILES

Performance:

1/2/4M structures/min

Storage:

~200MB / M structures

Performance:

0.25/0.5/2.2G comparison/s

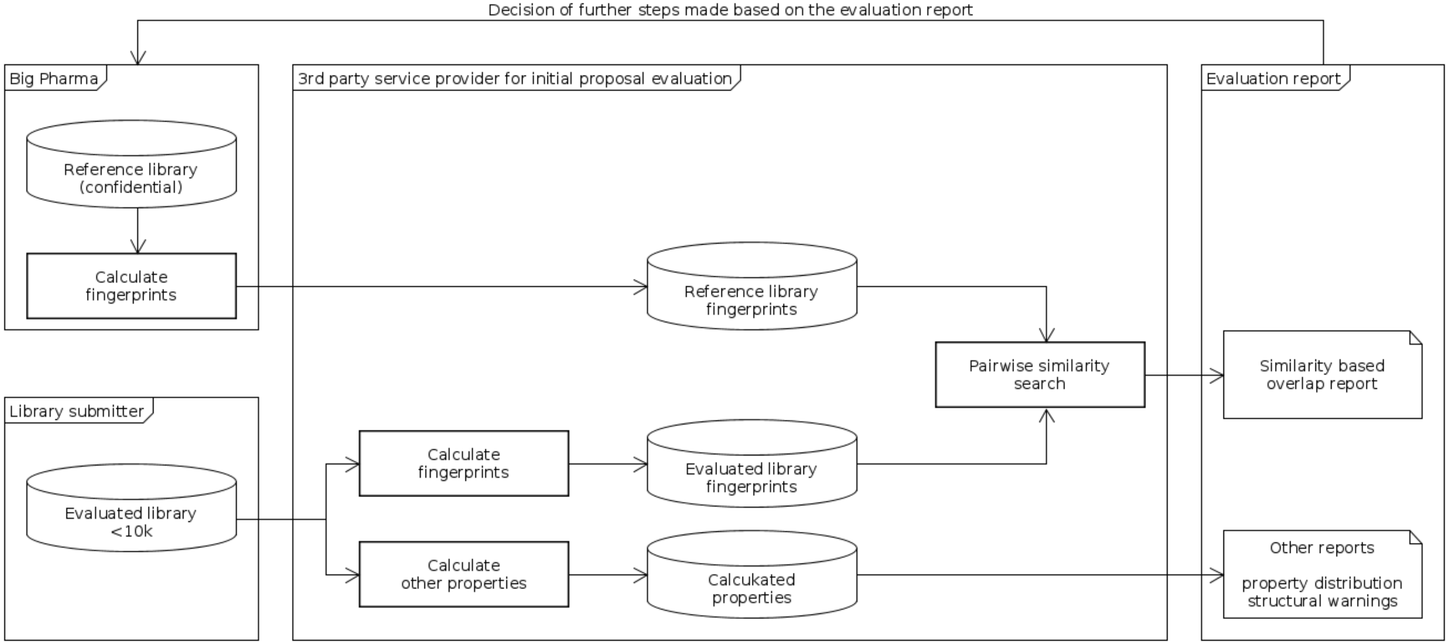
Distribution of the similarity of the most similar references

Notes

Calculation performance includes structure preprocessing and fingerprint generation.
Using 1024 bit binary path based fingerpints, small molecules from publicly available sources.
Using i7-4790 desktop, EC2 c3.8xlarge and x1.32xlarge instances.
Comparison performance is measured for most similar search using multiple (few 100) queries.
2.2G comparison/sec is equivalent with <8 min per million by million exhaustive search.

ChemAxon

Library evaluation


Decision of further steps made based on the evaluation report
Big Pharma
Reference library (confidential)
Calculate fingerprints
3rd party service provider for initial proposal evaluation
Reference library fingerprints
Pairwise similarity search
Evaluation report
Similarity based overlap report
Library submitter
Evaluated library <10k
Calculate fingerprints
Evaluated library fingerprints
Calculate other properties
Calcukated properties
Other reports
property distribution
structural warnings

ChemAxon

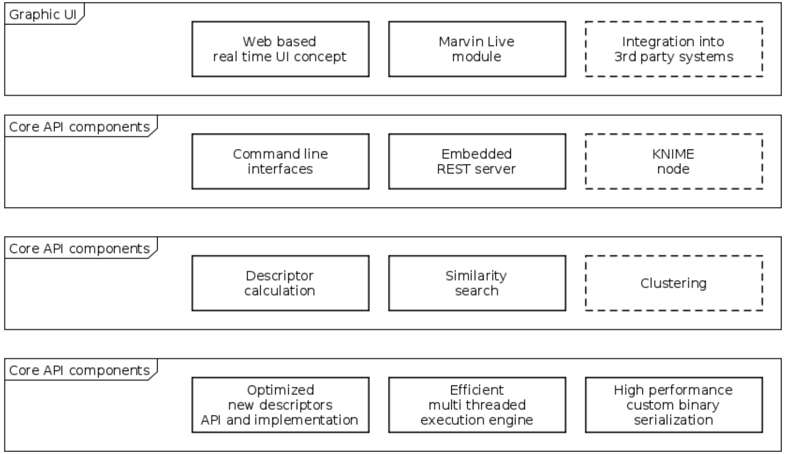
FUN, WHAT MORE DO YOU HAVE

Available components

Distribution

- Distribution for early adopters - contact us for details
- First public release is on the way

- Command line interfaces for the hardcore users
- REST server for integrators
- ML / Plexus integration
- User interface for focused chemical space analyis



Graphic UI

Web based real time UI concept

Marvin Live module

Integration into 3rd party systems

Core API components

Command line interfaces

Embedded REST server

KNIME node

Core API components

Descriptor calculation

Similarity search

Clustering

Core API components

Optimized new descriptors API and implementation

Efficient multi threaded execution engine

High performance custom binary serialization

PLANS

Roadmap

- Interactive UI for overlap analysis
- Real time clustering
- Single desktop UI release.
- Public Java API components for developers

THANK YOU

Gábor Imre

ChemAxon