ChemAxon

# MADFAST SIMILARITY SEARCH

Gábor Imre

Budapest Annual Meeting

2016

# Why MadFast?

# Bit of history

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

Two approaches were explored:

- Use clustering based heuristics to reduce set sizes.
- Optimize multi query similarity search.

# Who won?

Today’s multi core machines are fast and can have huge memory.

- Exhaustive search won.
- Can do the original 12M x 5k in a few minutes.

# So what?

In-memory storage and optimized multithreaded similarity search is faster than expected. How can we use it?

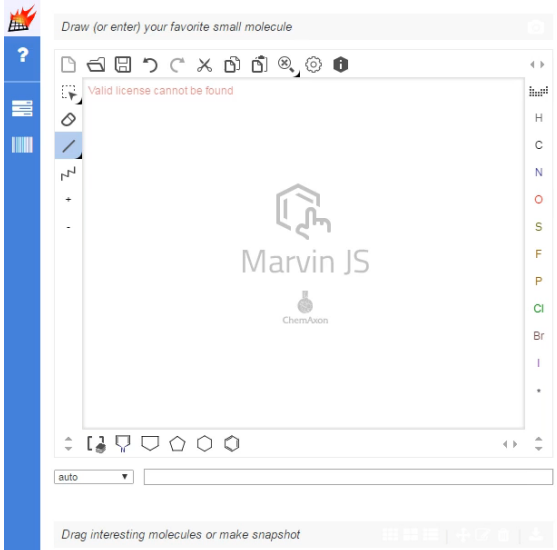
- Overlap analysis of large sets?
- Push the limits of similarity based clustering?
- Real time similarity search?

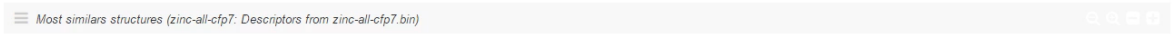
# Real time search

# Demo

Lets see a real time search on a large set:

- Searching against the Zinc database with 16M structures
- On an Amazon EC2 virtual machine (32vCPU class)


Draw (or enter) your favorite small molecule
Valid license cannot be found
Marvin JS
ChemAxon
H
C
N
O
S
F
P
Cl
Br
I
auto
Drag interesting molecules or make snapshot


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

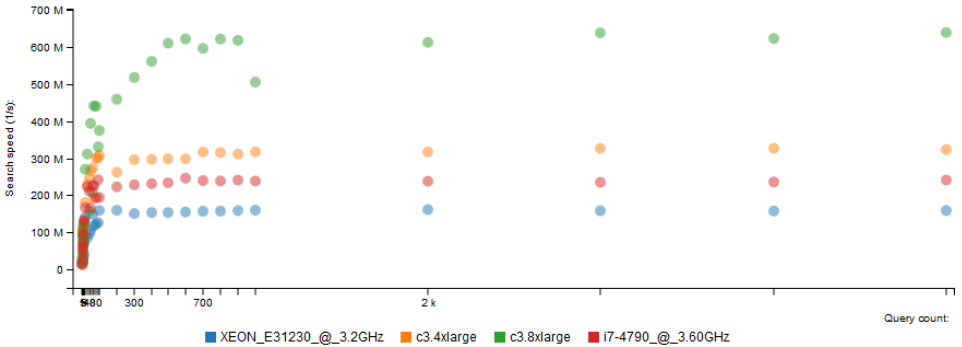


ChemAxon

# Demo

- Search time was ~0.08 sec (80ms) per query.
- Results shown as you type/draw.

- Searching multiple queries
- Most similar search speed using 1024 bit fingerprints
- Using >600 query batches >600M comparison per second (~1.5ns/comparison) sustained on a c3.8xlarge instance



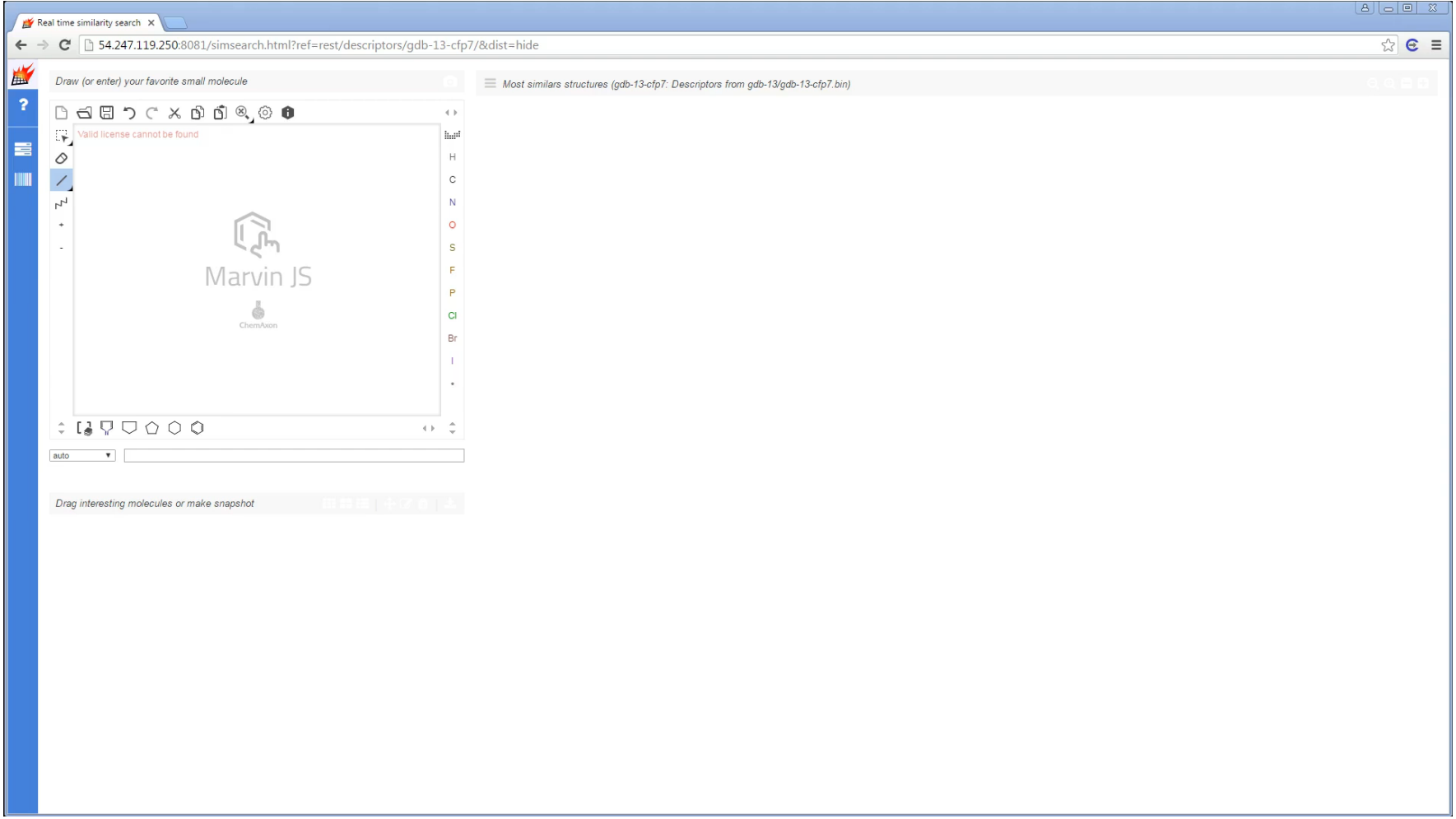



ChemAxon

# Go larger

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

- Remember the slide about fast machines with huge memory. For example Amazon EC2 provides r3.8xlarge instance with 32 vCPU and 244GB memory
- GDB-13 is the largest publicly available small organic molecule database containing 977M structures. (*Small organic molecules enumeratod 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
Marvin JS
ChemAxon
H
C
N
O
S
F
P
Cl
Br
I
auto
Drag interesting molecules or make snapshot
Most similars structures (gdb-13-cfp7: Descriptors from gdb-13/gdb-13-cfp7.bin)

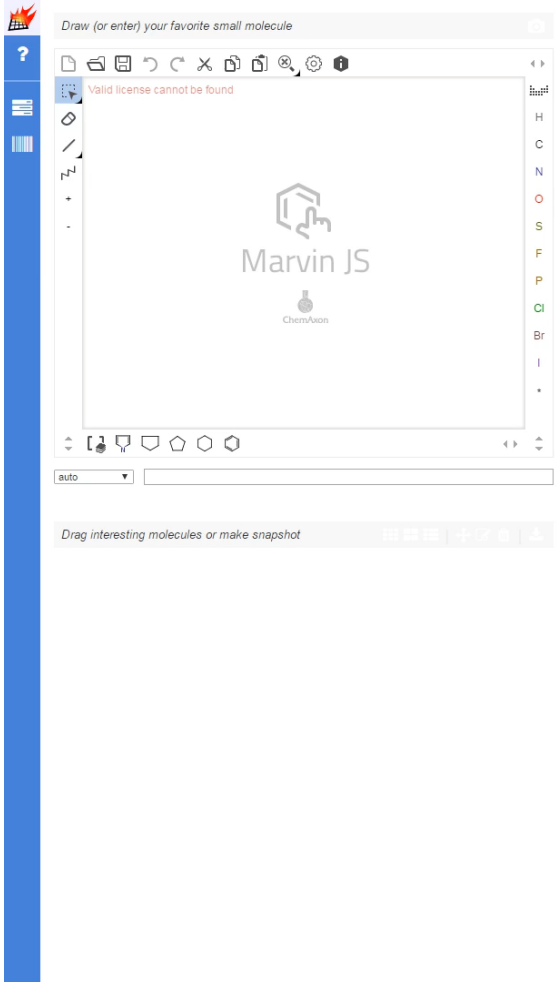


ChemAxon

# Fun, what more do you have?

# Components

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


Draw (or enter) your favorite small molecule
Valid license cannot be found
Marvin JS
ChemAxon
H
C
N
O
S
F
P
Cl
Br
I
auto
Drag interesting molecules or make snapshot
Most similars structures (zinc-all-cfp7: Descriptors from zinc-all-cfp7.bin)
Dissimilarity distribution (zinc-all-cfp7: Descriptors from zinc-all-cfp7.bin)
No query structure set.
Draw, enter or drag a structure into the sketcher component.



ChemAxon

# Plans

# Short term plans

- UI for overlap analysis.
- Real time clustering.
- Desktop UI release.
- Components for developers.

# THANK YOU

ChemAxon