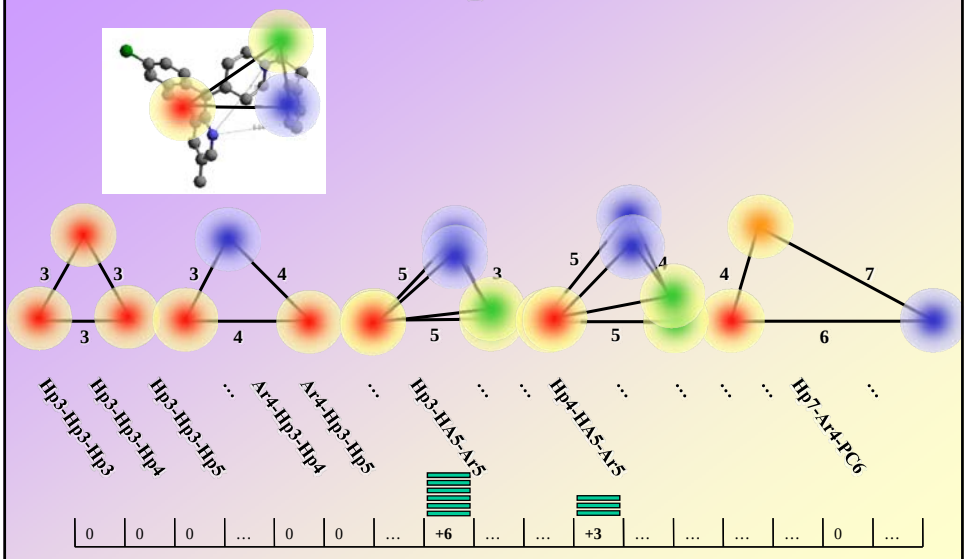
VSEngine - Similarity & Property prediction based Virtual Screening Web Services

Dragos Horvath, Gilles Marcou & Alexandre Varnek
UMR 7177 CNRS-ULP
Laboratoire d'InfoChimie, Univ. Louis Pasteur Strasbourg
horvath@chimie.u-strasbg.fr

Why yet another « Virtual Lab »?

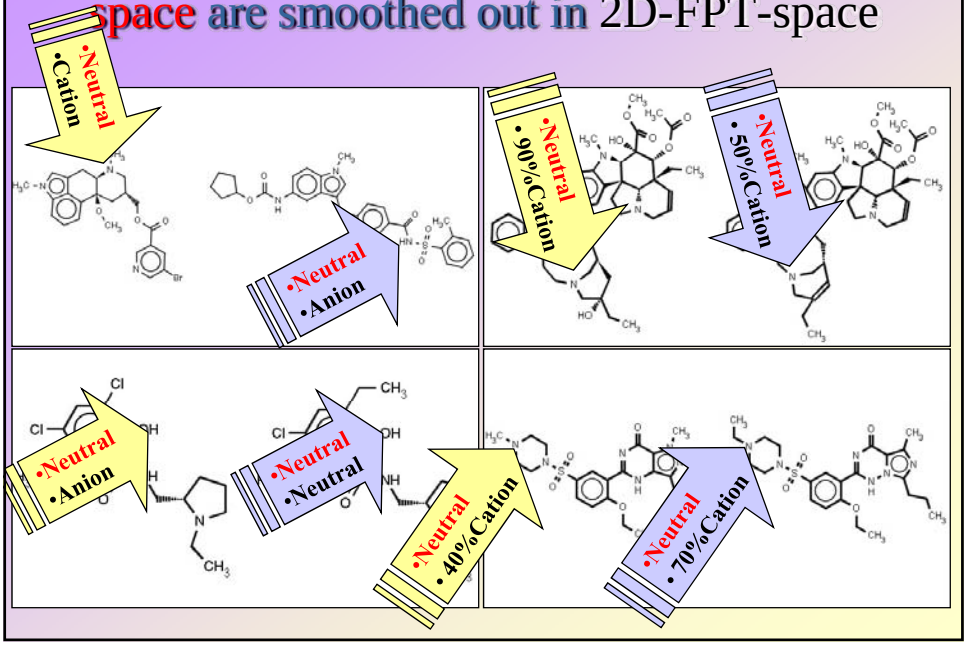
- To open up the access to the proprietary tools developed within UMR 7177:
  - Tunable Fragments (atom/bond sequence and augmented atom)
  - Fuzzy Pharmacophore Triplets & adapted similarity metrics
  - Original QSAR-building Expertise:
    - "aggressive" problem space searches, supporting non-linear model fitting
    - original applicability domain definitions…
    - potentially useful property prediction tools…
- To make all these tools, issued from different environments & backgrounds, work together
  - ChemAxon "Toolbox" welcome
- But do they work better than "Saint Tanimoto 0.85" ?
  - **Don't know – finding it out is the user's part of the deal…**

(ChemAxon-Powered) Fuzzy Pharmacophore Triplets



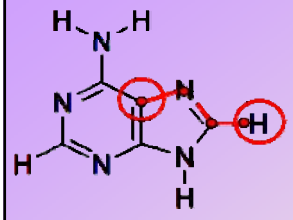
$D_i(m)$ = total occupancy of basis triplet i in molecule m.

Some 'activity cliffs' in rule-based descriptor space are smoothed out in 2D-FPT-space



Fragment Descriptors (ISIDA)

Sequences



C-N=C-H
C-N=C
N=C-N
C-N
N=C
C-H

sequence

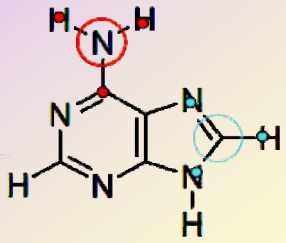
I(AB, n_{min} – n_{max})

From n_{min} to m_{max} atoms

Atoms+Bonds

n_{min} , n_{max} = 2 - ...

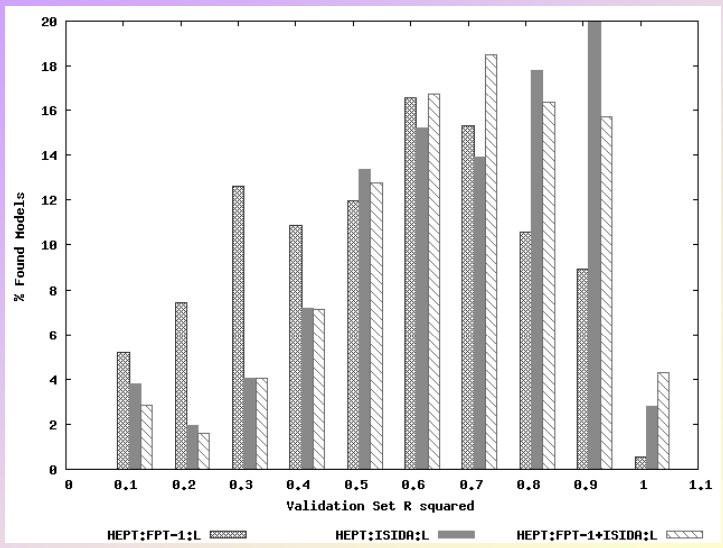
Augmented Atoms

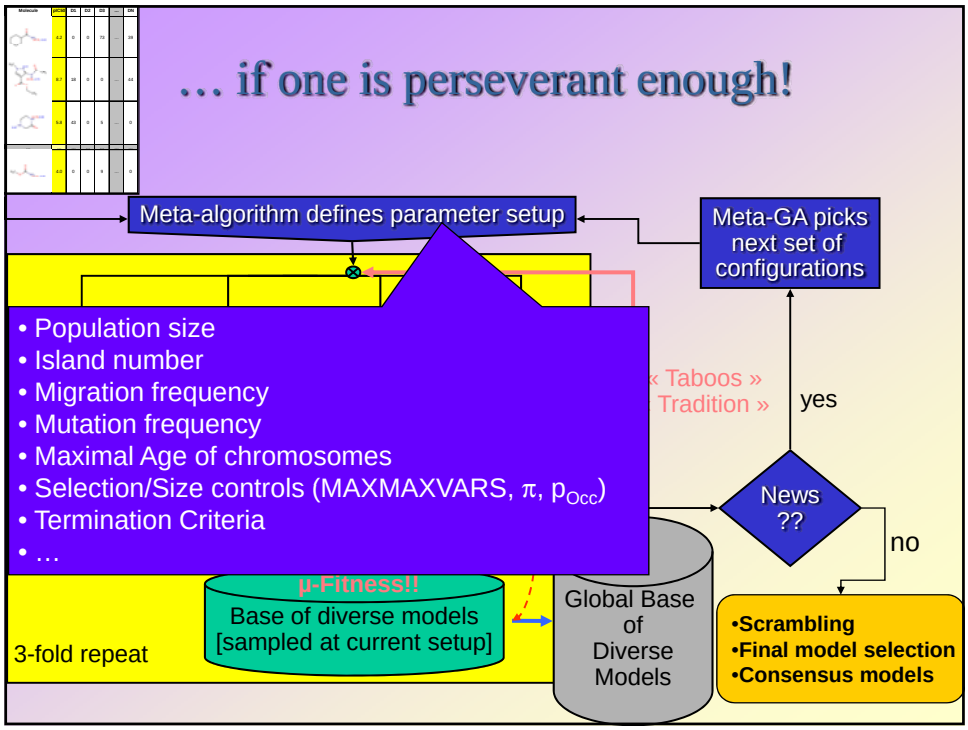


II(A) (no hybridization)

II(Hy) (hybridizarion of neighbours is taken into account)

Sharp Fragments and Fuzzy Pharmacophores, together, "crack" many QSAR problems…





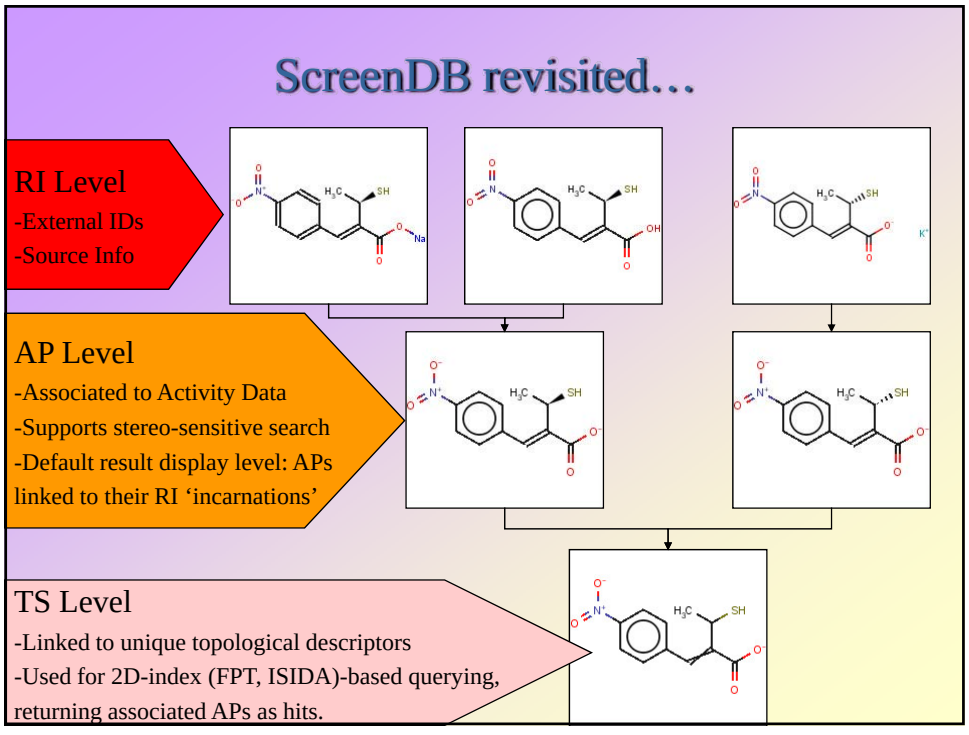
Software integration issues…

- ISIDA Fragment Descriptors & Tools originally stem from a Windows/Delphi world
  - … but Unix standalone versions have been created to facilitate integration
- The FPT generator is a standalone tool employing the ChemAxon toolkit…
  - … though not exploiting any database access facilities offered by the latter
- Don't want to risk the death of my one & only neuron while learning JSP or other weird Web freak languages…
- **Perl/cgi** and **Unix shell scripts** pilot **standalone tools** (Java, Delphi, F77, Perl, awk) having access to a central **MySQL/JChem database.**

ScreenDB, the JChem Database: Standardization, Traceability & Property Calculations

- Three levels of structure representation cohabitate:
  - **RI**: Raw Input: structures as entered by chemists
  - **AP**: Active principles: major microspecies of default tautomer – standardized, without counterions
  - **TS**: Topological Skeletons - AP stripped of stereochemical info (thanks smiles:0!)
- ChemAxon calculated properties (logP, logD, RotBonds, HBA, HBD) entered in the database upon registration
- Import & consistency check of experimental data fields
  - **QSAR set repository** & **Hit Set Property Alerts**
- ISIDA and FPT descriptors are stored in .gzipped files indexed by the JChem cd_id

ScreenDB revisited…



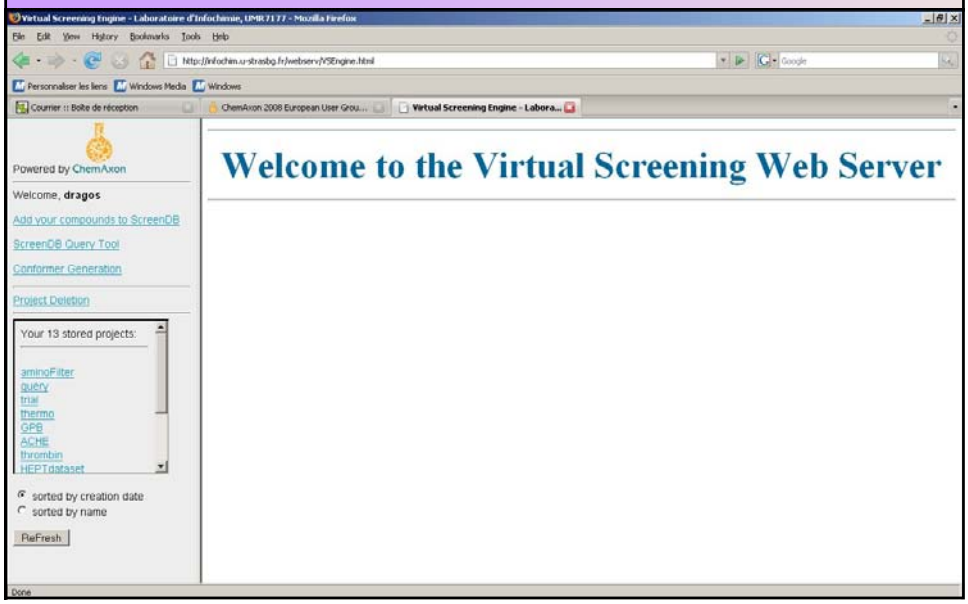
General Site Architecture…

- Apache/CGI server, with user-specific directories (pseudo-accounts)
- Users create server-side project directories, to harbor input requests (query structures, *etc.*), temporary files, flags and results.
- A project is a perennial entry, dedicated to a specific task:
  - Molecule upload into the data base
  - Database Querying
  - *Online property prediction (bypassing database storage)
  - **QSAR model building & predictions
- and associated to a stance-sensitive task-specific .cgi script
  - Interactive parameter input/instantiation of background computing jobs
  - Project job in progress…
  - Project results available or job disaster report…

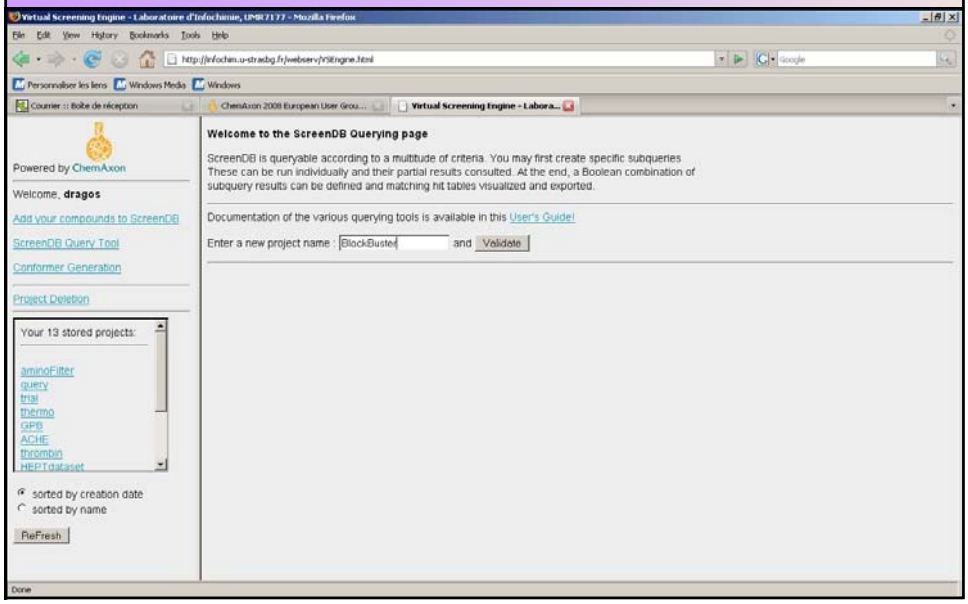
ScreenDB Querying

- Data field-based queries (in progress – including on-the-fly predicted properties from external QSARs)
- ChemAxon-driven tools (.cgi-controlled standalone java application)
  - Substructure & Unique (!!) Substructure (SMARTS)
  - Exact match or ChemAxon structural fingerprint-based default similarity search (SMILES).
- ISIDA fragment- or Fuzzy Triplet-based similarity queries
  - Support multiple molecules as a query ("AND" fingerprint), in conjunction to various descriptor instances and dissimilarity metrics (background similarity calculations using specific Unix command-line tools)
- A hit post-processing tool converts cd_id lists produced by these applications into hit display pages and hit property alerts…

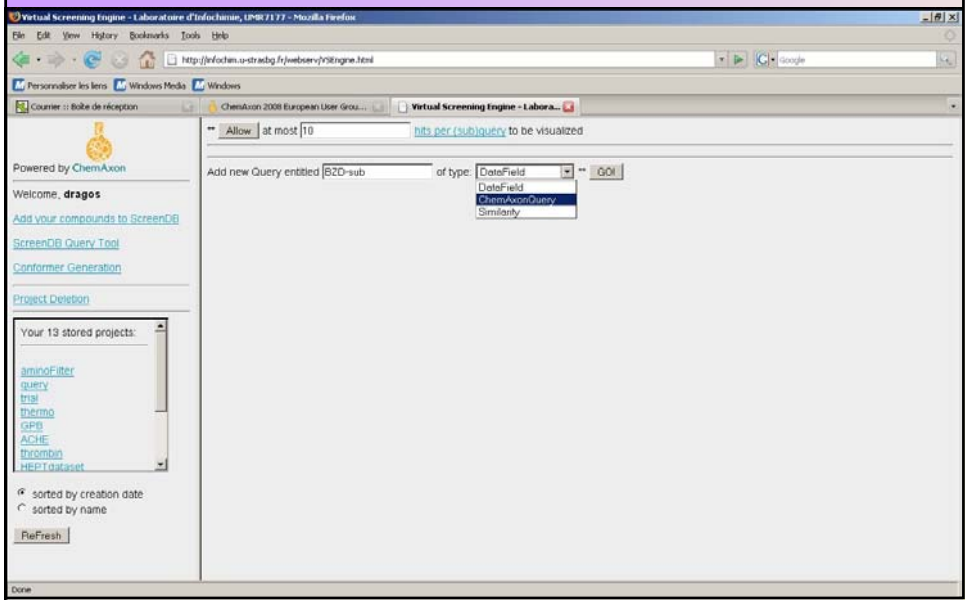
Some Screen Shots: Welcome Page…



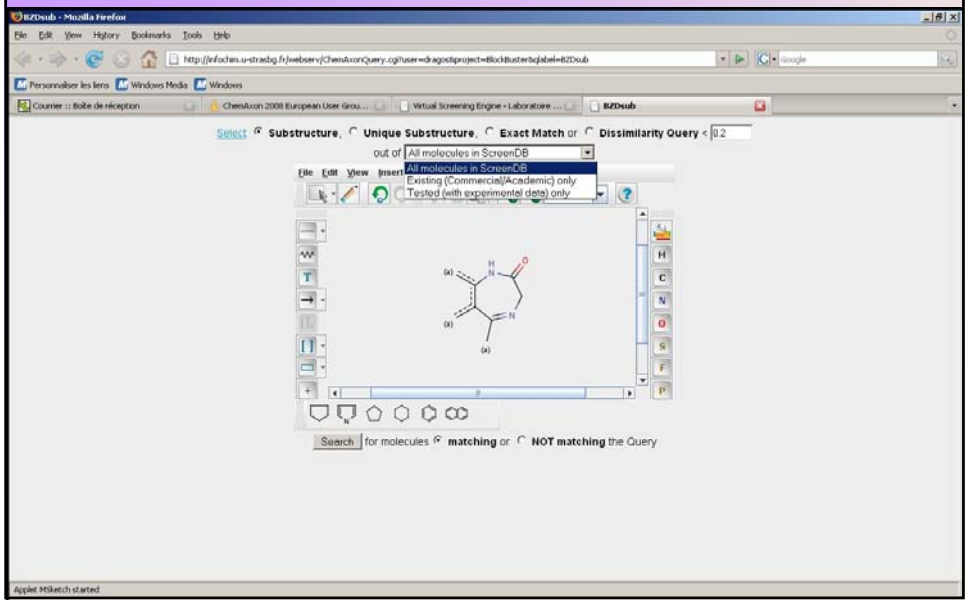
Start of a Querying Project…



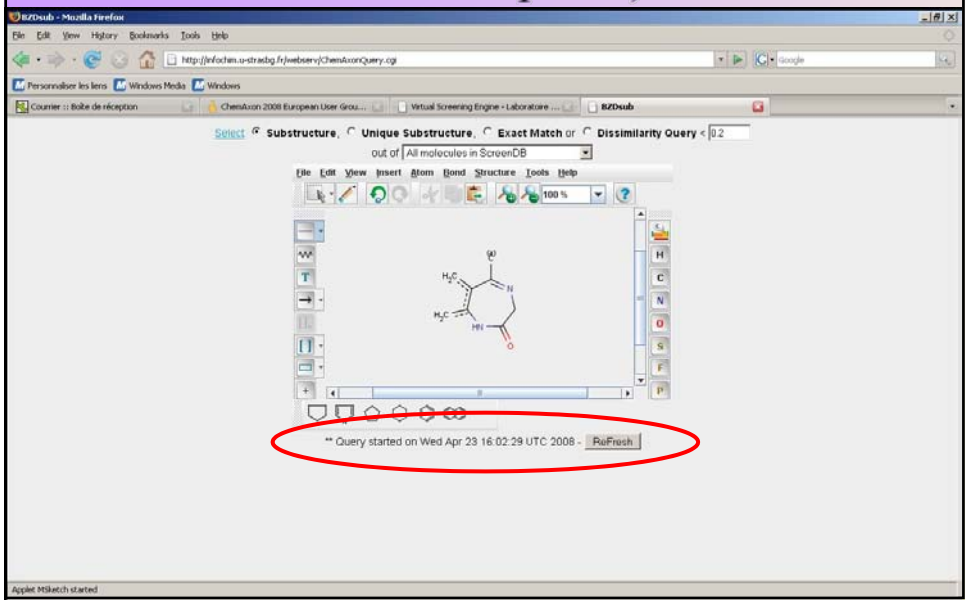
Initiate a Substructure SubQuery



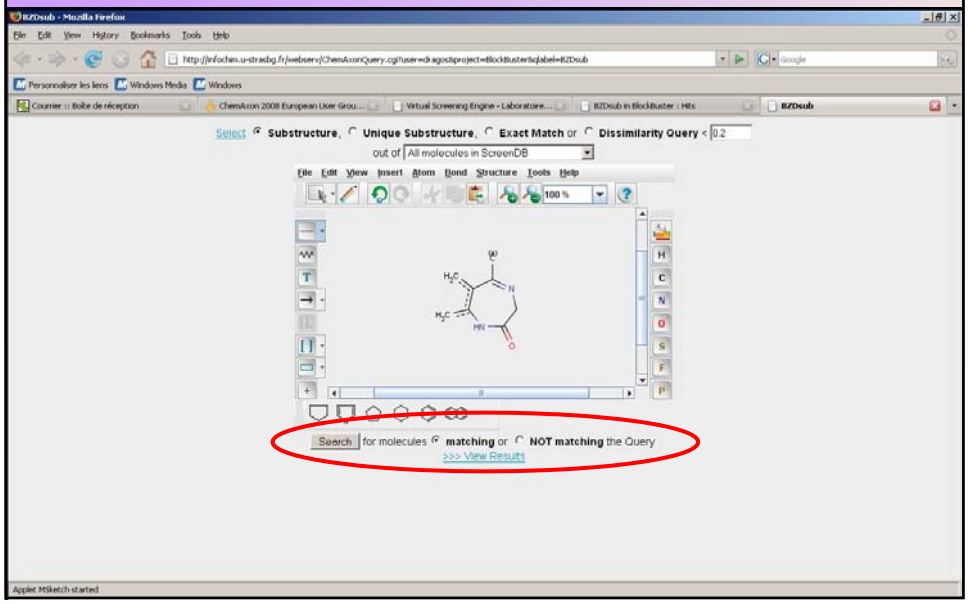
Specify the Substructure SubQuery



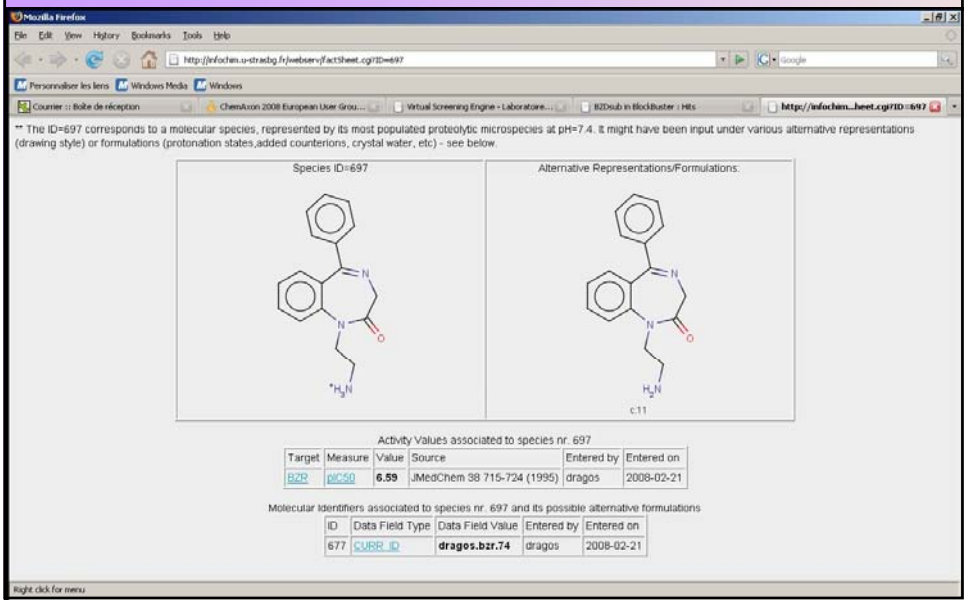
SubQuery in Progress (go start the next one while it completes)…



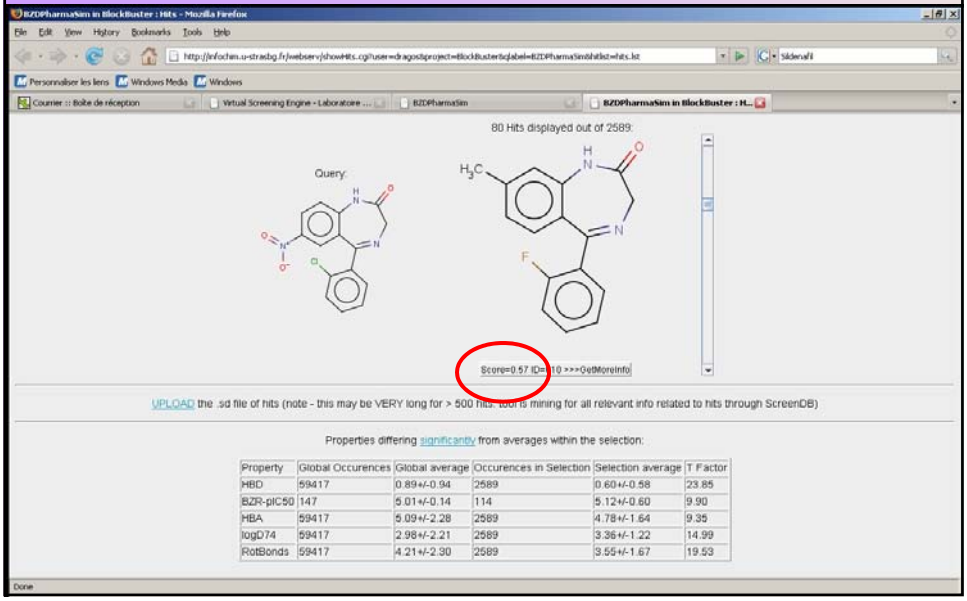
SubQuery done…



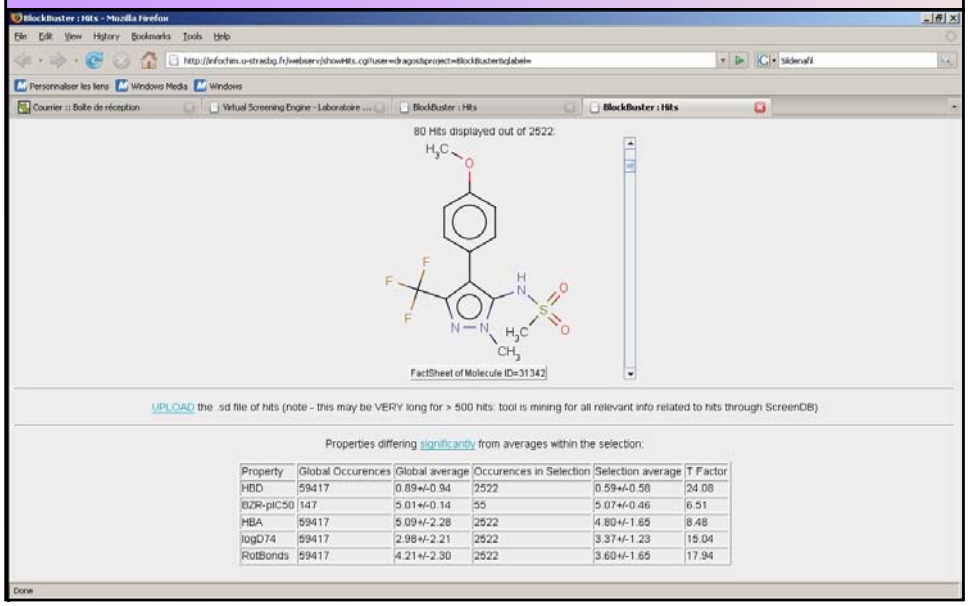
Hit display…



A Similarity SubQuery …



The Complex Query Builder…



Conclusions, Perspectives…

- Buzzword FLEXIBILITY: The ChemAxon API is a convenient way to get a heterogeneous bunch of tools talk to a MySQL chemical database – be it by means of temporary files!
  - Standardization & file format conversions are heavily used
- Buzzword CHEMICAL SENSE: structure representation, substructure querying, ionization
- Speed is not exactly a bottleneck, since we are looking at more elaborate and therefore more costly virtual screening options…
- Unix-shell centric approach has the merit to be easily deployable on GRIDS!