Using Self-Organizing maps to accelerate similarity search

Fanny Bonachera, Gilles Marcou, Natalia Kireeva, Alexandre Varnek, Dragos Horvath

Laboratoire d'Infochimie, UMR 7177. 1, rue Blaise Pascal, 67000 Strasbourg

UNIVERSITÉ DE STRASBOURG

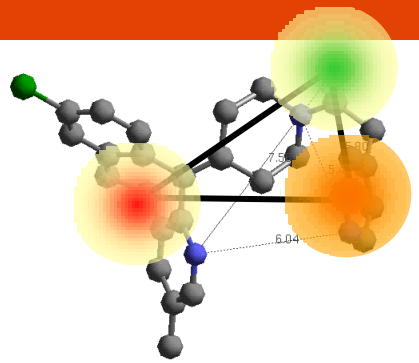
Inf Chimie STRASBOURG

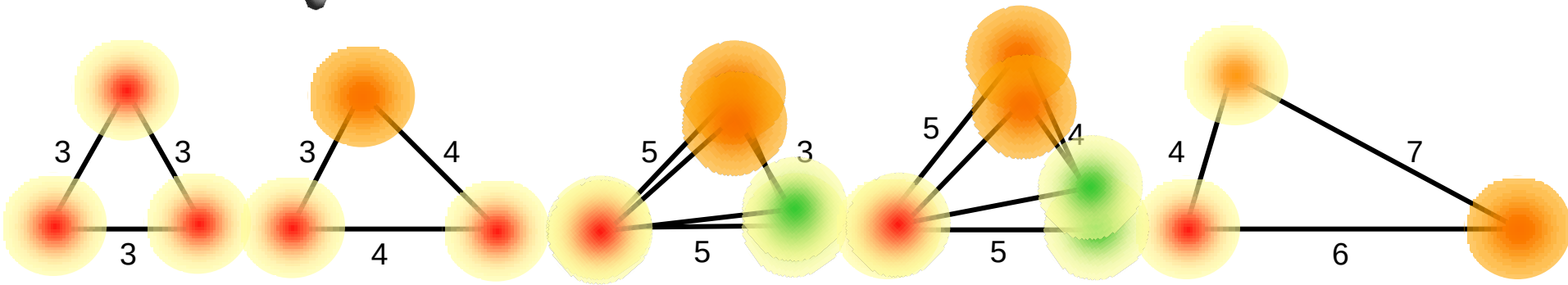
cnrs

The Problem…

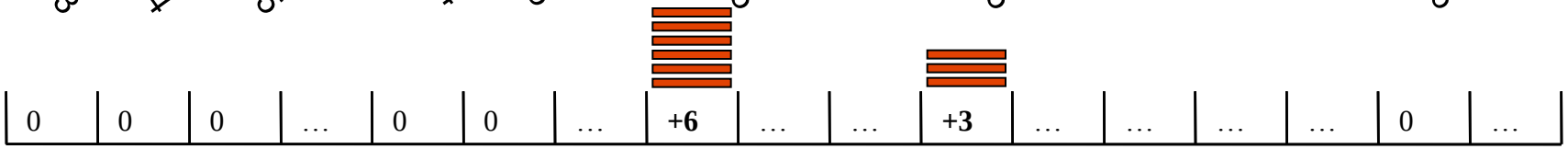
- We have used the **ChemAxon API** to develop high-quality, high information content, pH sensitive and otherwise chemically meaningful descriptors…
  - Fuzzy Pharmacophore Fingerprints (FPT), ISIDA Coloured Fragment counts…

Fuzzy Pharmacophore Triplets





Hp3-Hp3-Hp3 Hp3-Hp3-Hp4 Hp3-Hp3-Hp5 … Ar4-Hp3-Hp4 Ar4-Hp3-Hp5 … Hp3-HA5-Ar5 … … Hp4-HA5-Ar5 … … … … Hp7-Ar4-PC6 …

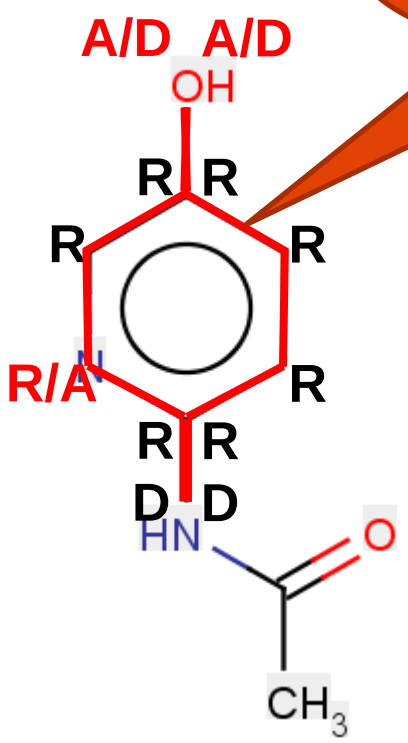


$D_i(m)$ = total occupancy of basis triplet i in molecule m.

Microspecies-Specific Labeling of Fragments…

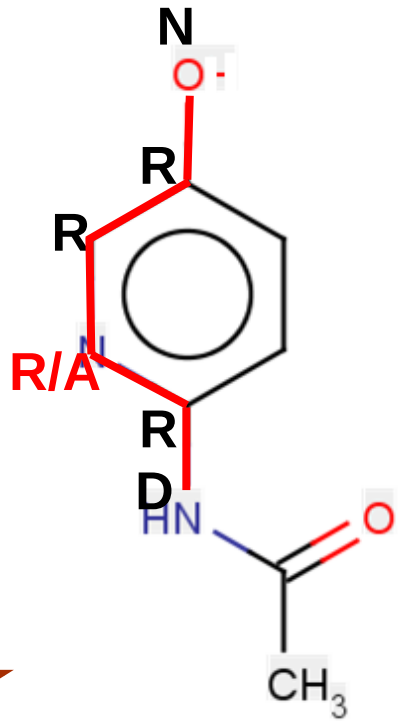
µSpecies increment … fragments by their population levels

Lower & Upper Fragment sizes are user-defined



A-R*R*R*R-D +95
D-R*R*R*R-D +95
A-R*R*R*R-D +95
D-R*R*R*R-D +95
A-R*R*A*R-D +95
D-R*R*A*R-D +95
…

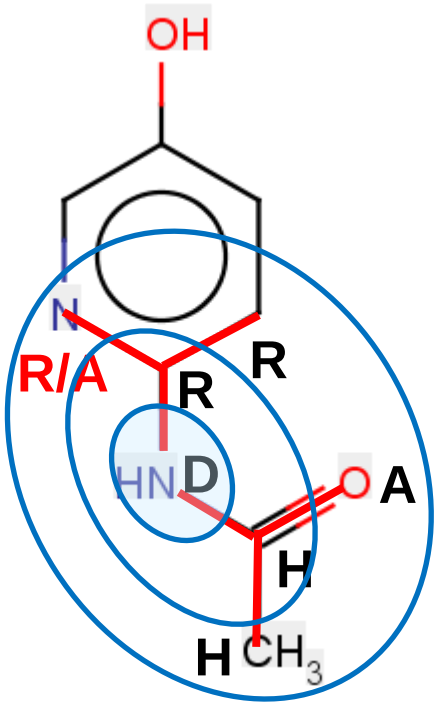
N-R*R*R*R-D +5
N-R*R*A*R-D +5
…



Population: **95%** **Molecular Fingerprint** **5%**

Augmented Atoms…

Branched fragments, representing an atom and (an user-defined number of) its successive coordination spheres



Strict Typing with Bond Info (-**b**)

D(-R(*R)*R)(-H(-H)=A)
D(-R(*R)*A)(-H(-H)=A)

Strict Typing, no Bond Info

D(R(R)R)(H(H)A)
D(R(R)A)(H(H)A)

All but Central and Terminal Atoms may be wildcards (-**b** -**w**)

D(-R(*R)*R)(-H(-H)=A)
D(-?(*R)*R)(-H(-H)=A)
D(-?(*R)*A)(-H(-H)=A)
…

"Tree" descriptors have wildcards for all but Central & Terminal:

D(-?(*R)*A)(-?(-H)=A)
…

The Problem…

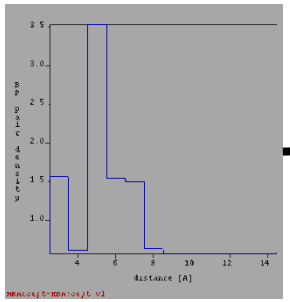
- We have used the **ChemAxon API** to develop high-quality, high information content, pH sensitive and otherwise chemically meaningful descriptors…
  - Fuzzy Pharmacophore Fingerprints (FPT), ISIDA Coloured Fragment counts…
- It is time to exploit them in similarity-driven virtual screening.
  - search for similar compounds of similar fingerprints to a given query, and therefore, hopefully, a similar activity.
- BUT, these fingerprints are neither *binary*, nor really *short* (>10^5-dimensional, with some fragmentation schemes).
  - **We are short-lived mortals who'd nevertless like to allow public web-based screening of something like the ZINC database, for mortal web site users.**

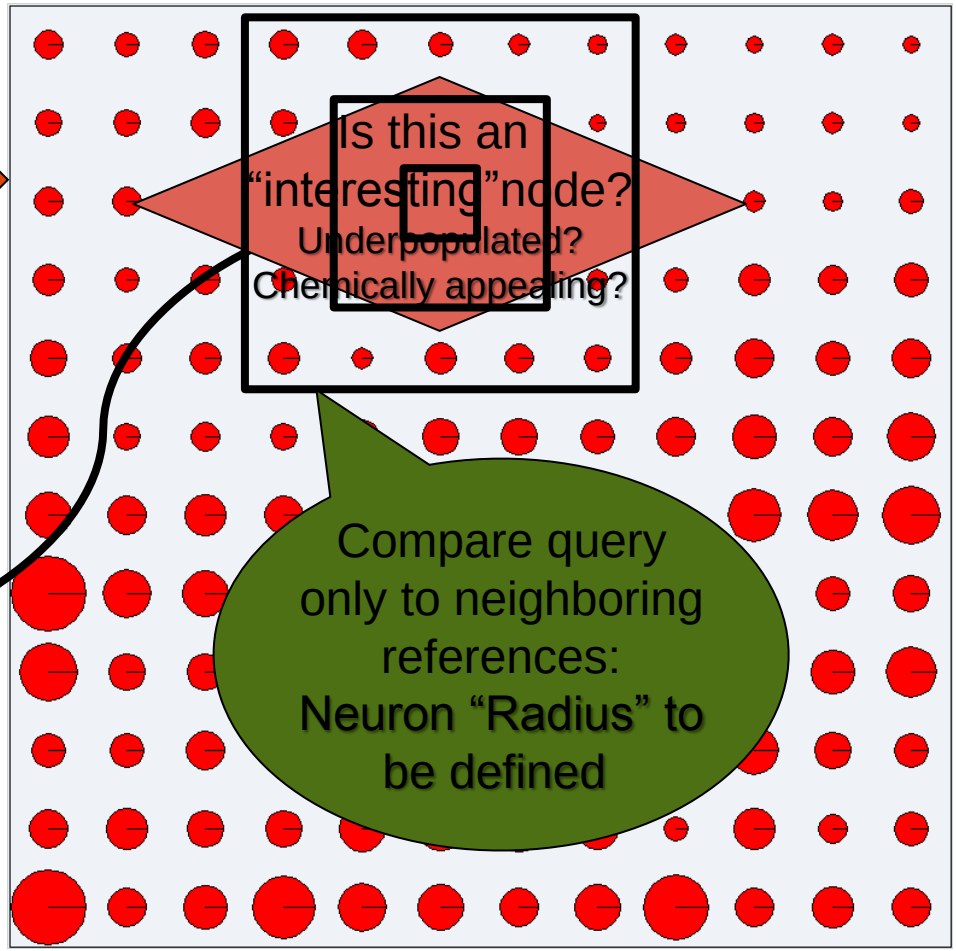
A Possible Solution…

Self Organizing Map (SOM)-enhancement: map molecules, search neighborhood

Pharmacophore patterns of database compounds (3 to 8 millions)

External compound (“query”)





Using Self-Organizing maps to accelerate similarity search

Data sets

- **Similarity Screening Sets:**

  - ***QS*** - Query Set (2000 compounds) : random subsets of 11 different analogue series used to model structure-property relationships in literature, marketed drugs and biological reference compounds and commercially available molecules (picked randomly from the ZINC database)
  - ***DB*** - Database (55613 compounds) : including the remainders of the 11 above-cited series, further marketed drugs and biological reference compounds, 1870 ligands from the Pubchem database tested on the hERG channel, and a majority of randomly picked ZINC compounds. No overlap between ***DB*** and ***QS***.
  - For each molecule in ***QS*****,** the list of its top neighbors from ***DB*** is found by classical calculation of Tanimoto and Euclidean coefficients against the entire ***DB***, then selecting top 300 hits at Tanimoto>0.75, and respectively Euclidean<9.
  - A maximum of these Tanimoto and Euclidean hits must then be found again in SOM-enhanced VS – but in a much shorter time…

Using Self-Organizing maps to accelerate similarity search

Build maps: Data sets

- **Sets for Map training:**
  - ***Extended*** - SOM training set (53206 compounds) : a subset of previous molecules (***DB+QS***), excluding the analogue series members, the Pubchem compounds and some 900 ZINC molecules.
  - ***SmallRef*** - SOM training set (11168 compounds) : features all drugs and biological reference compounds seen in *Extended*, but significantly less ZINC molecules.

- **… and external testing**
  - ***ExtDB*** - External database for real-life tests (~160000 compounds) : from the corporate collection of one of our industrial partners.
  - ***ExtQ*** - External query set for real-life tests (12491 compounds) : taken basically from *SmallRef*, and completed with randomly picked commercial compounds.

Using Self-Organizing maps to accelerate similarity search

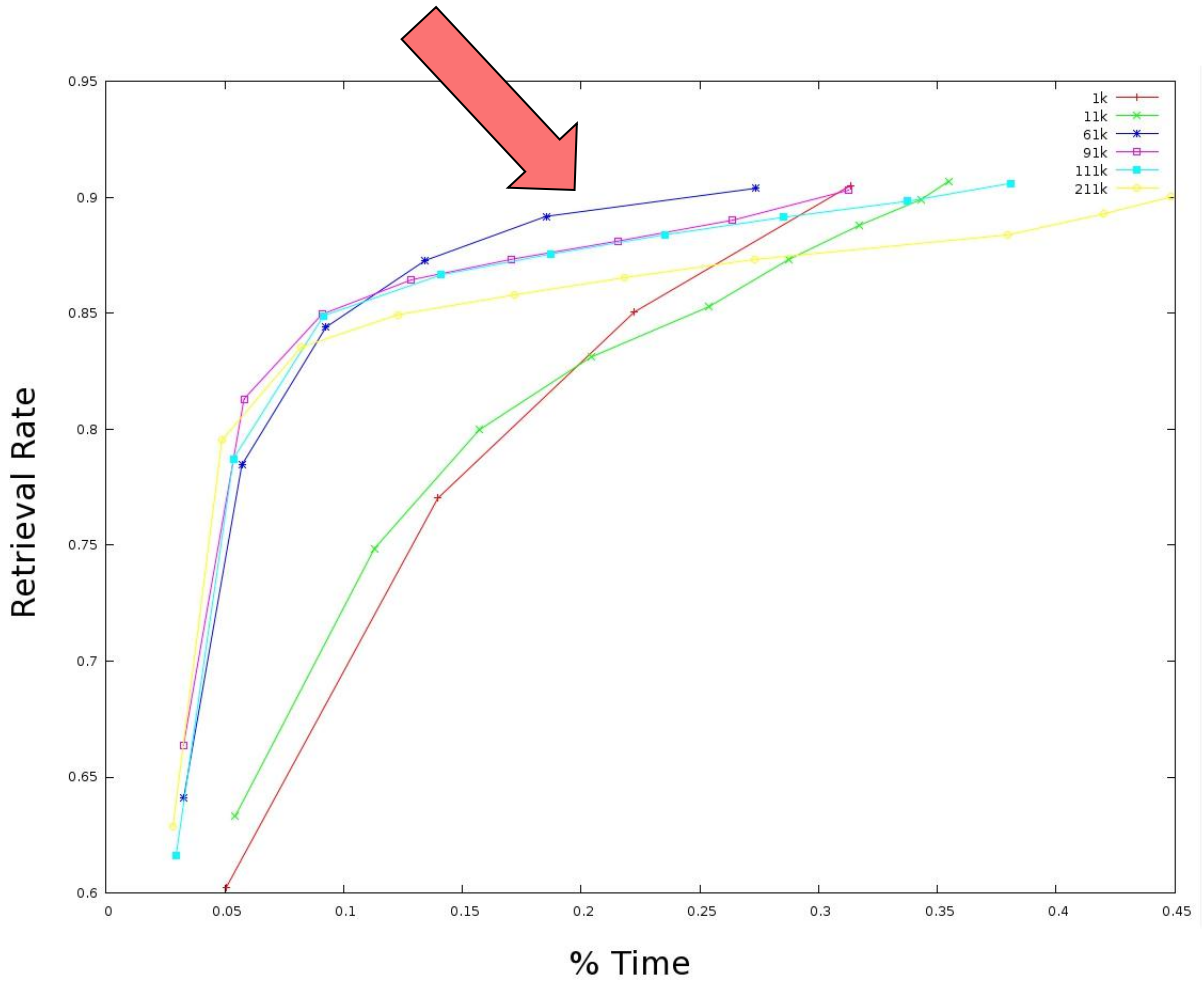
Building the SOMs

- **Generated maps :**
  - For each training set (*SmallRef* and *Extended*)
  - 36 explored geometries
  - Varying X and Y from 8x6 to 30x30

- **Map fitting steps :**
  - Explored training iterations
  - Brute training : 1000 iterations
  - Refinement : 10000 iterations
  - HyperRefinement : tests between 50000, 80000, 100000, 200000 iterations

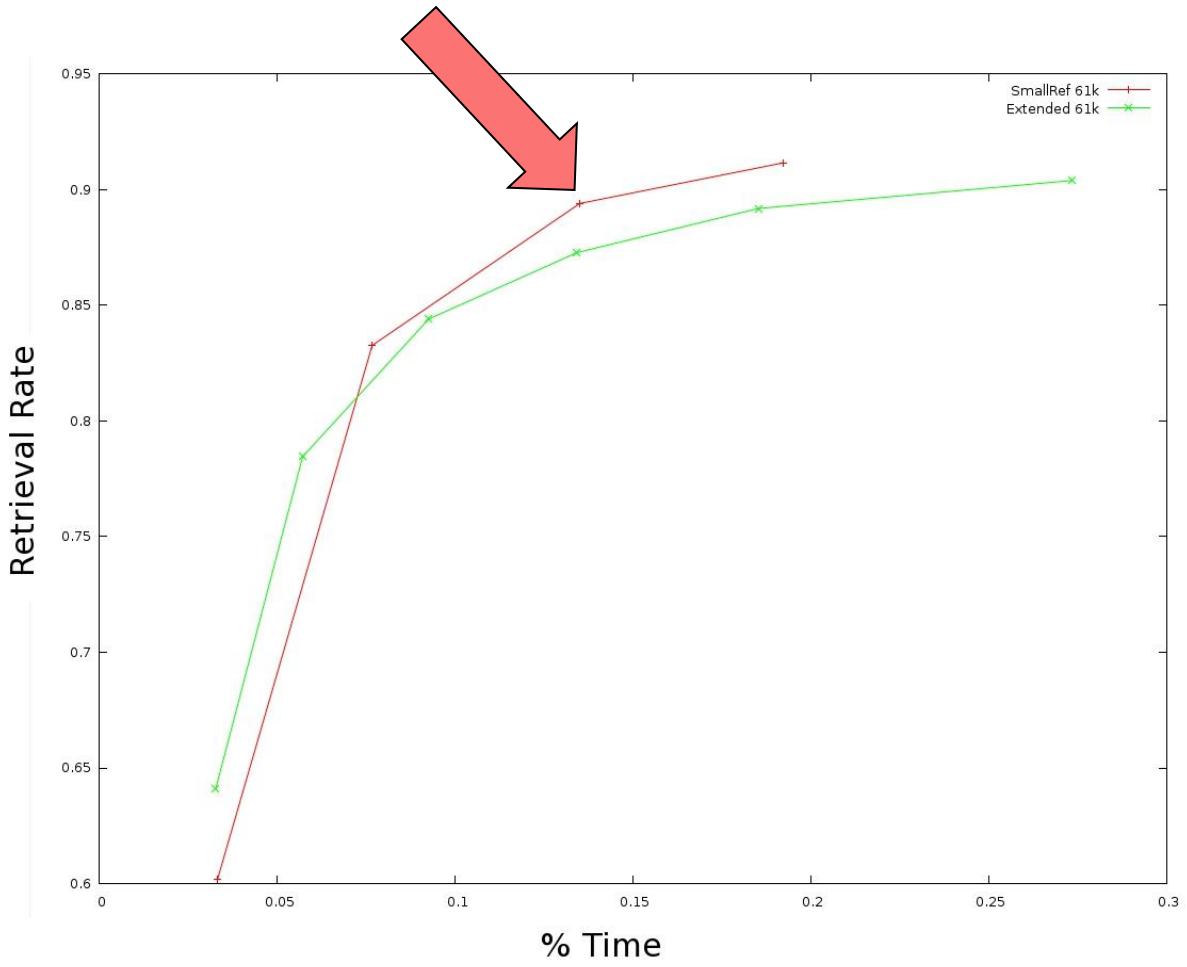
Using Self-Organizing maps to accelerate similarity search

Monitoring map convergence – 22x28 rectangle bubble trained on *Extended*



Using Self-Organizing maps to accelerate similarity search

Monitoring map convergence – 22x28 rectangle bubble : *SmallRef* vs *Ext*.



Using Self-Organizing maps to accelerate similarity search

Map-enhanced similarity searching – How do we define *Q* ?

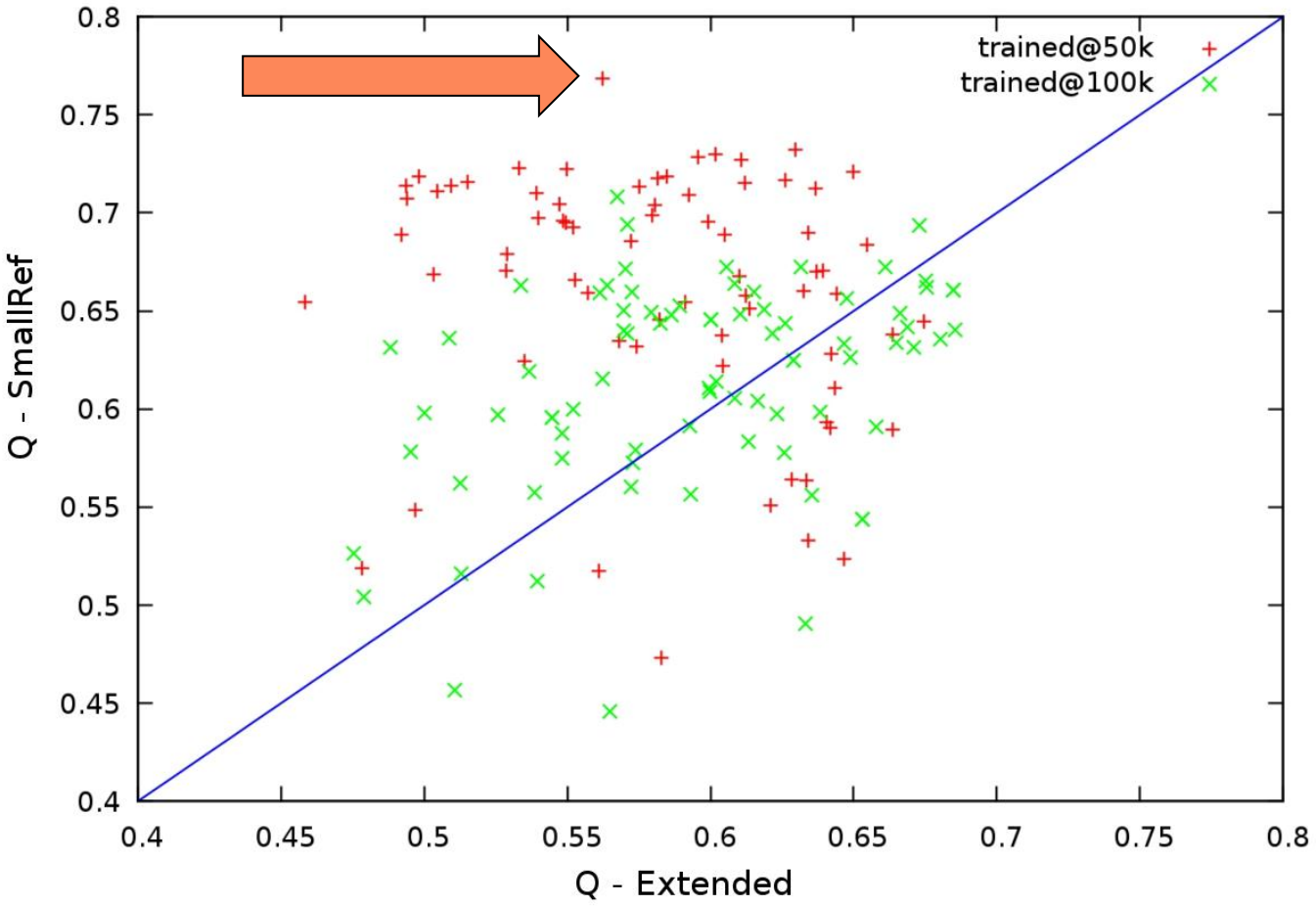
- **Map Quality criterion *Q***
  - Virtual Screening enhancement factor of a map
  - Scanning for the best time enhancement *vs.* Retrieval rate compromise over increasing Radii *R*.

$$Q_{\Sigma}^{\kappa} = \max_{R}[RR_{\Sigma}^{\kappa@R} \times (1 - f_{\Sigma}^{\kappa@R})^{2}]$$

Q, with respect to dissimilarity metric Σ, for the map κ needs to be optimized by scanning for the best time enhancement $1 - f_{\Sigma}^{\kappa@R}$ *vs.* retrieval rate compromise $RR_{\Sigma}^{\kappa@R}$ over increasing radii *R*.

Using Self-Organizing maps to accelerate similarity search

Impact of the training set size – comparison of *Q* factor



Using Self-Organizing maps to accelerate similarity search

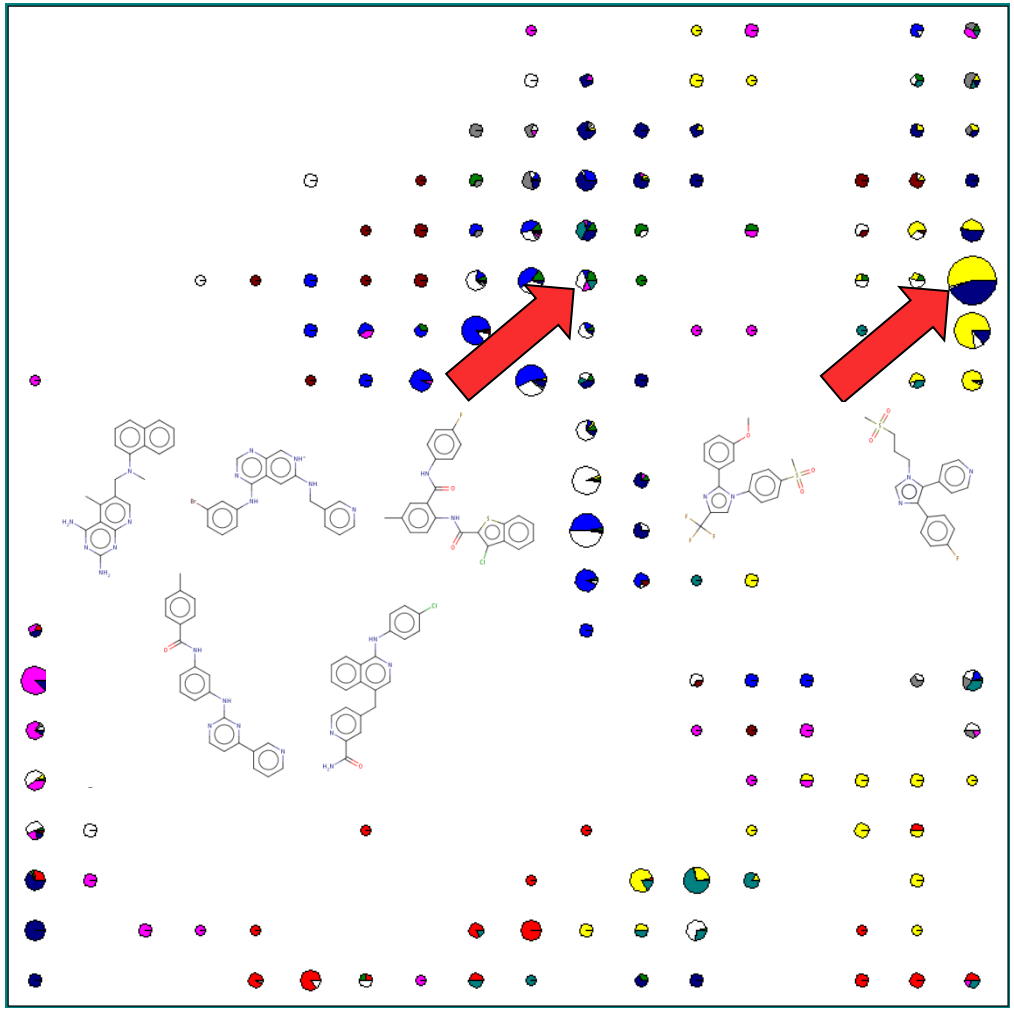
An overview of a good map

- **Best map, *Q*=0.77**

  - SOM trained on the SmallRef dataset.
  - 18x20 = 360 neurons.
  - 3 training steps (Brute + Refinement + HyperRefinement at 50k iterations)
  - Rectangular topology and Bubble neighbourhood function
  - Colored according to Lipinski-rules violations
  - (Red = 0 violations, green = 1, yellow = 2, blue = 3).
  - Mapping of DUD compounds to check consistency

Using Self-Organizing maps to accelerate similarity search

An overview of a good map



Using Self-Organizing maps to accelerate similarity search

Real-life testing

MAP	Neuron Radius	#Detected Similar	Time (mins)
top2	5*	27707	260
top2	10*	28571	597
top3	2	27837	96

The good behavior of the maps, as evidenced at their primary benchmarking stage, was confirmed.

Similar compounds not located in neighbours neurons are at risk to be dispatched Anywhere in the map – retrieving them by increasing R might be very costly.

Conclusions…

- Too much learning is harmful, even for artificial brains…
  - Maps may, and *should*, be trained on relatively small – though diverse compound sets. Too many input molecules just make convergence more difficult.
  - Good news: no need to retrain your maps when expanding your database!
  - Too much fitting of the code vectors may be detrimental – interestingly, unsupervised learning methods may suffer from overfitting too!
- SOM acceleration works in both Tanimoto and Euclidean spaces, and with different sets of descriptors: get ~90% of expected virtual hits in 10% of time.
  - … but loosing a few virtual hits was never a Greek tragedy. If this should become an issue (Greeks are unpredictable), it is enough to choose a larger Neuron Radius.