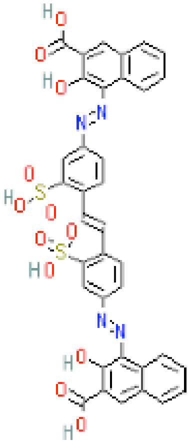
In silico chemogenomics with Support Vector Machines

Jean-Philippe Vert

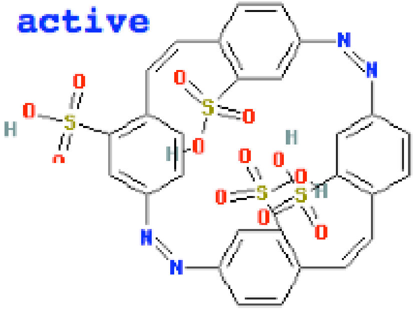
Institut Curie - U900 INSERM - Mines ParisTech

MedChem conference, Feb 22-25, 2009, Berlin, Germany.

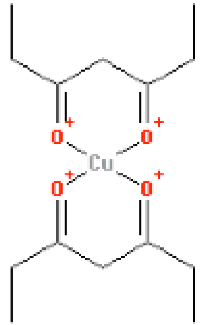
Ligand-based virtual screening / QSAR



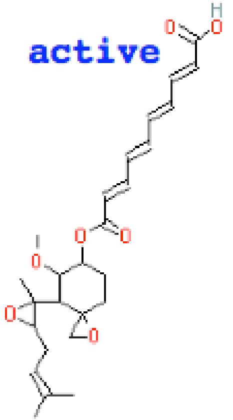
active



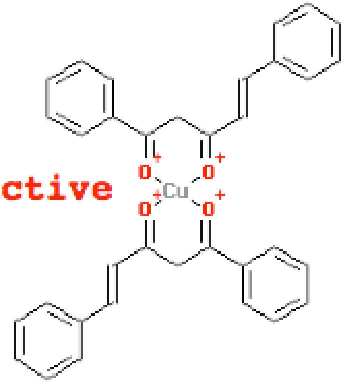
active



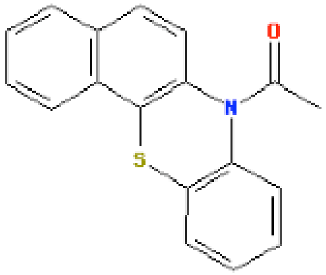
inactive



active



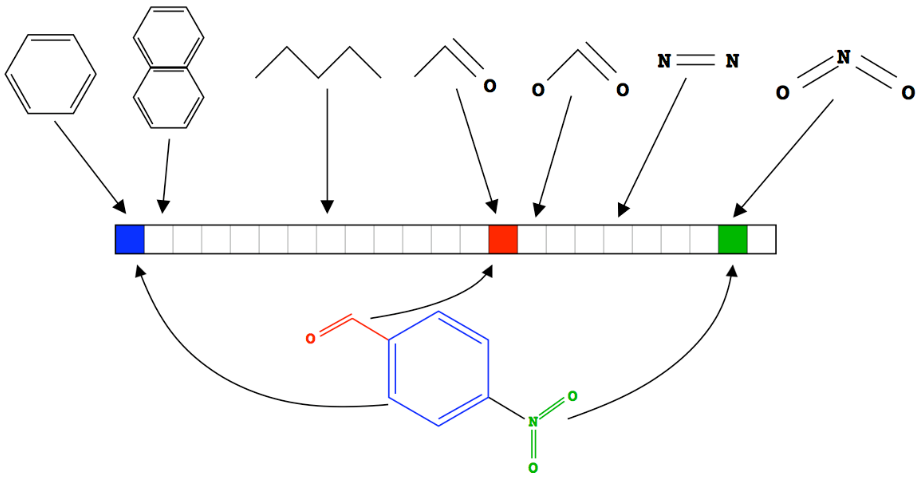
inactive

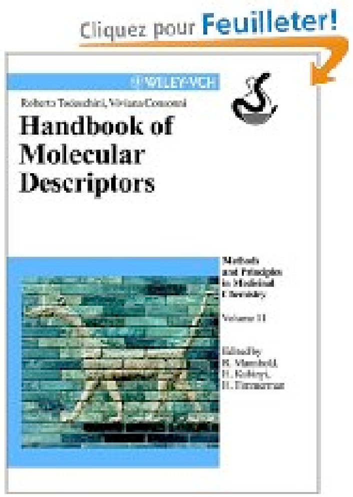


inactive

From http://cactus.nci.nih.gov

Represent each molecule as a vector…





Cliquez pour Feuilleter!

WILEY-VCH

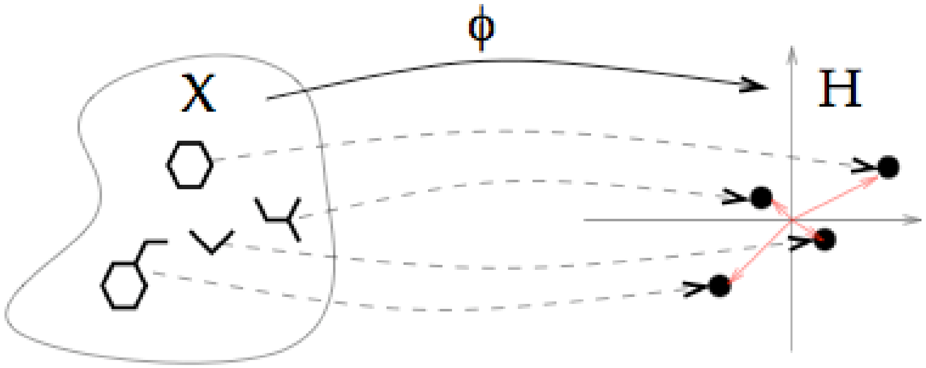
Roberto Todeschini, Viviana Consonni

Handbook of Molecular Descriptors

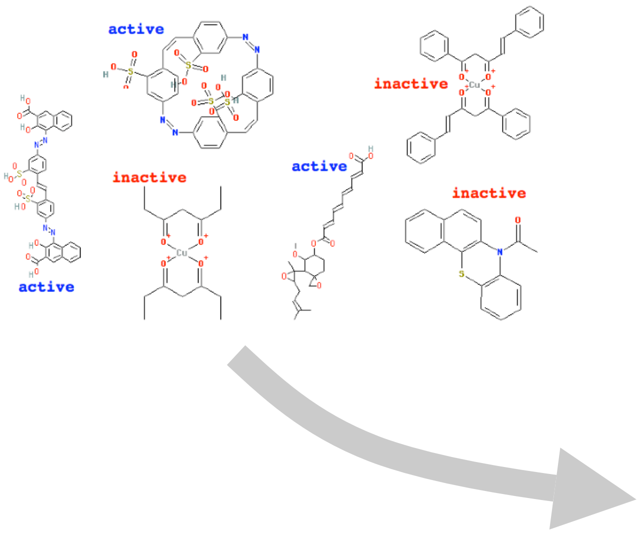
Methods and Principles in Medicinal Chemistry

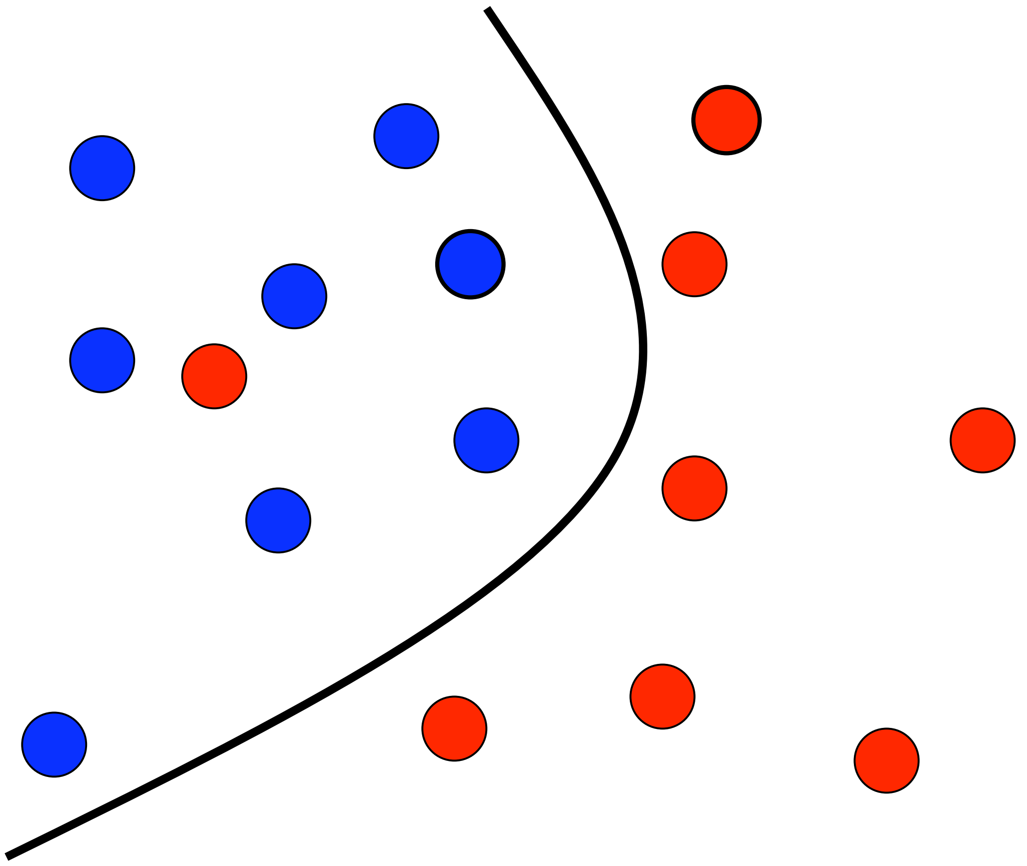
Volume 11

Edited by R. Mannhold, H. Kubinyi, H. Timmerman



…and discriminate with machine learning





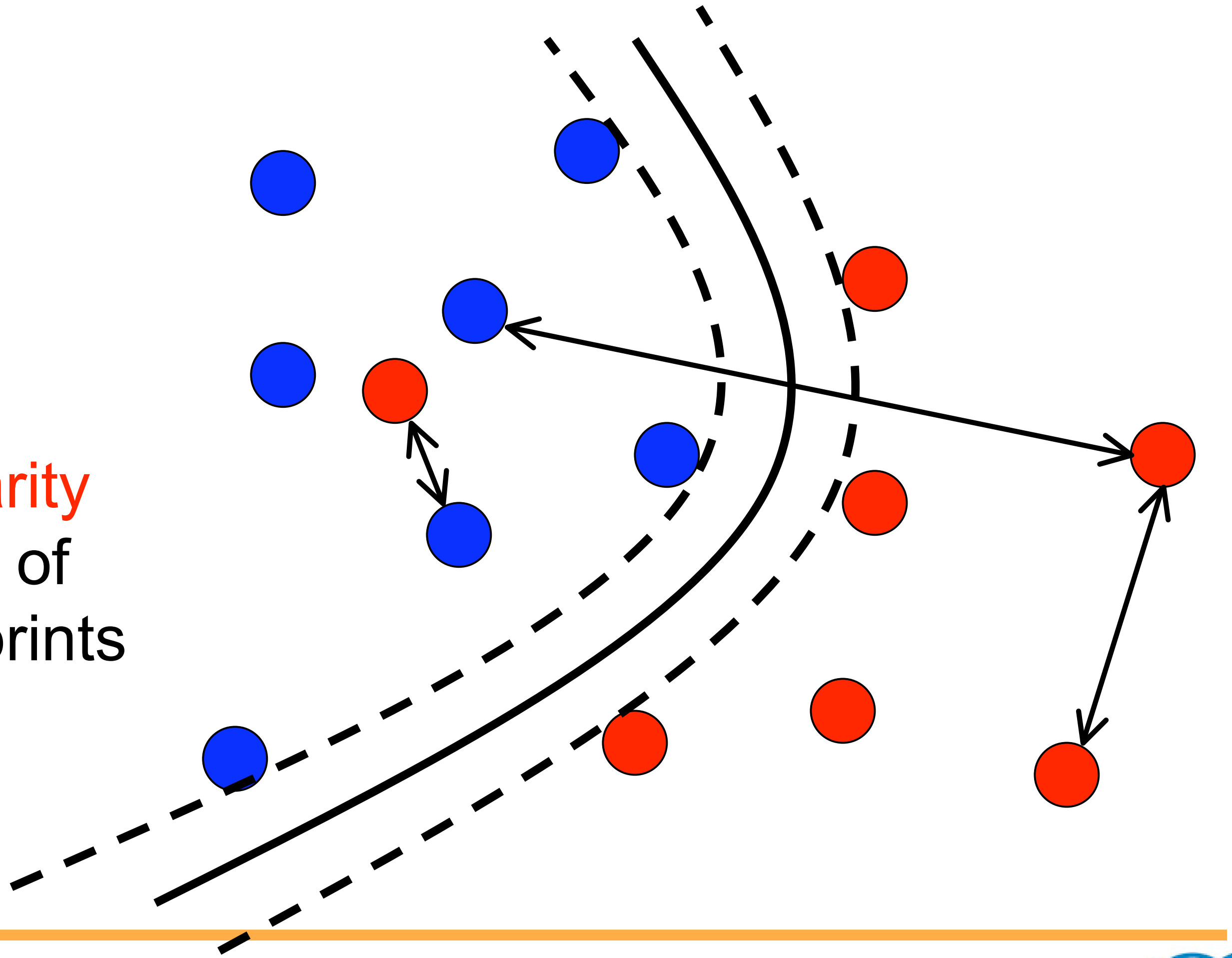
-LDA
-PLS
-Neural network
-Nearest neighbour
-SVM, …

Support Vector Machine (SVM)

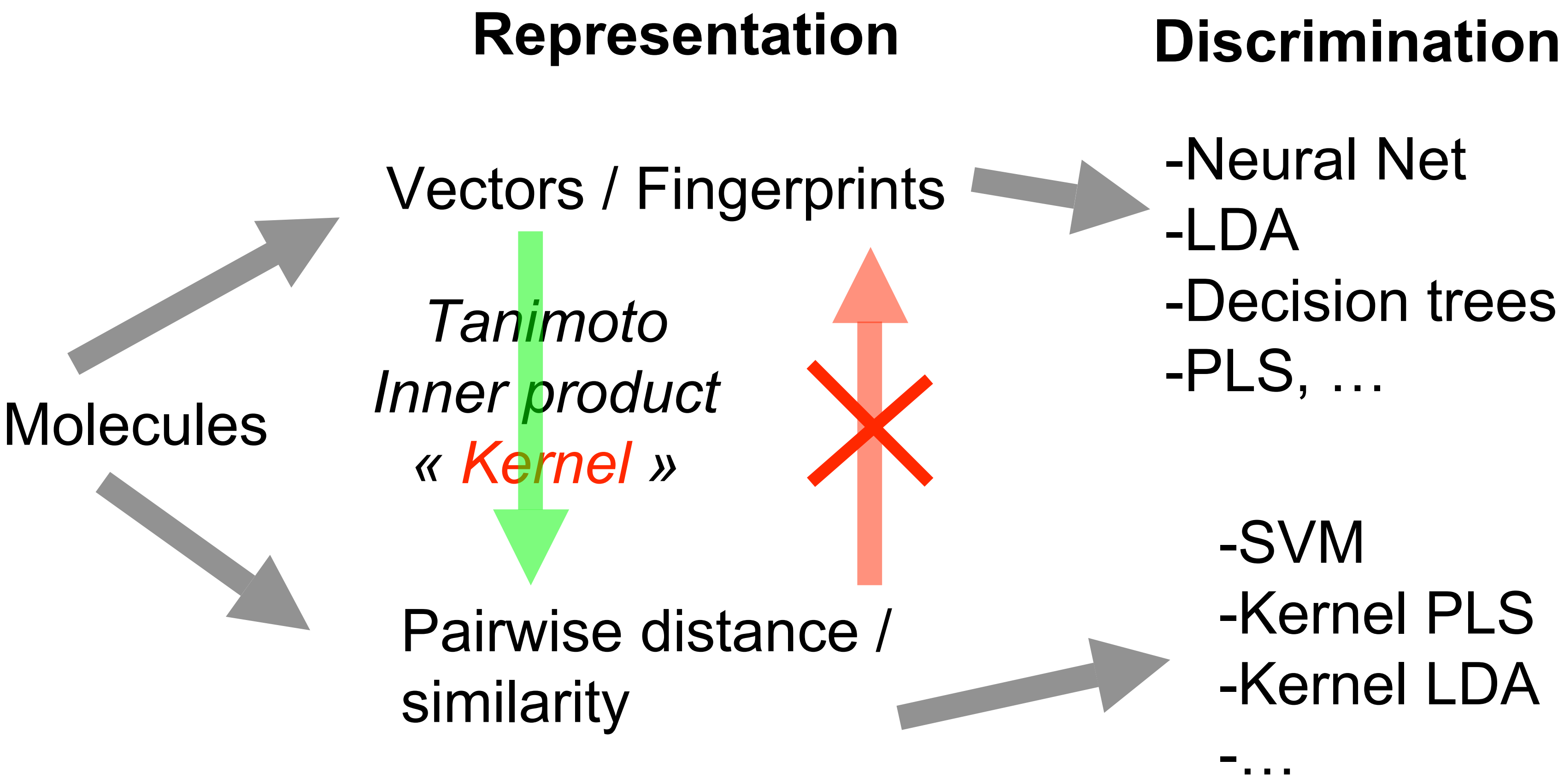
- Large margin

- Nonlinear

- Need pairwise distance / similarity as input instead of vectors / fingerprints



From fingerprints to similarities



Representation

Discrimination

Molecules

Vectors / Fingerprints

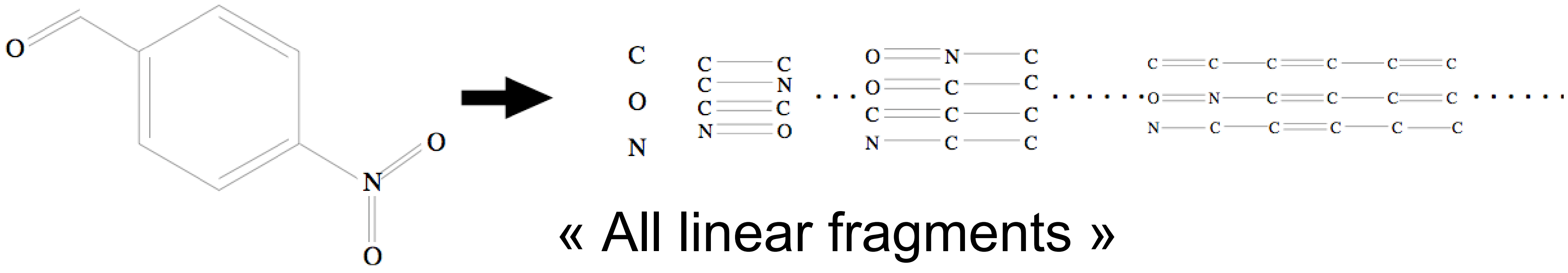
Tanimoto
Inner product
« *Kernel* »

Pairwise distance / similarity

-Neural Net
-LDA
-Decision trees
-PLS, …

-SVM
-Kernel PLS
-Kernel LDA
-…

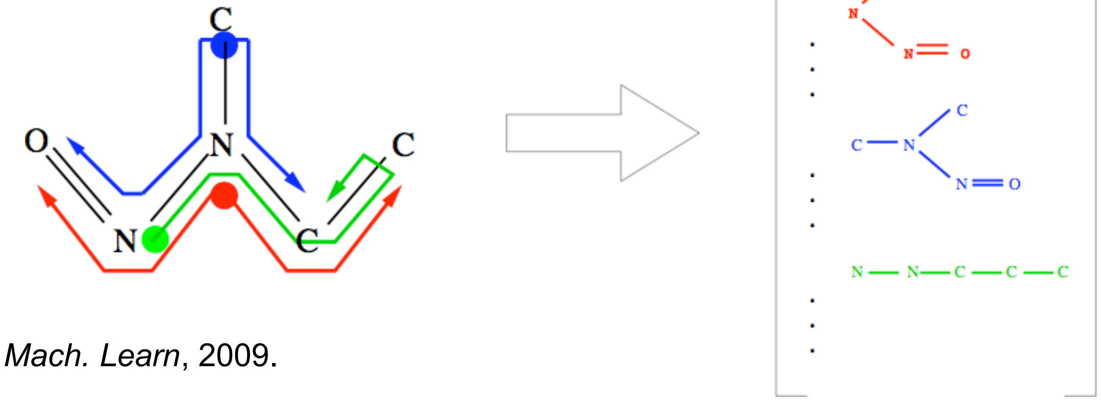
Example : 2D fragment kernel



« All linear fragments »

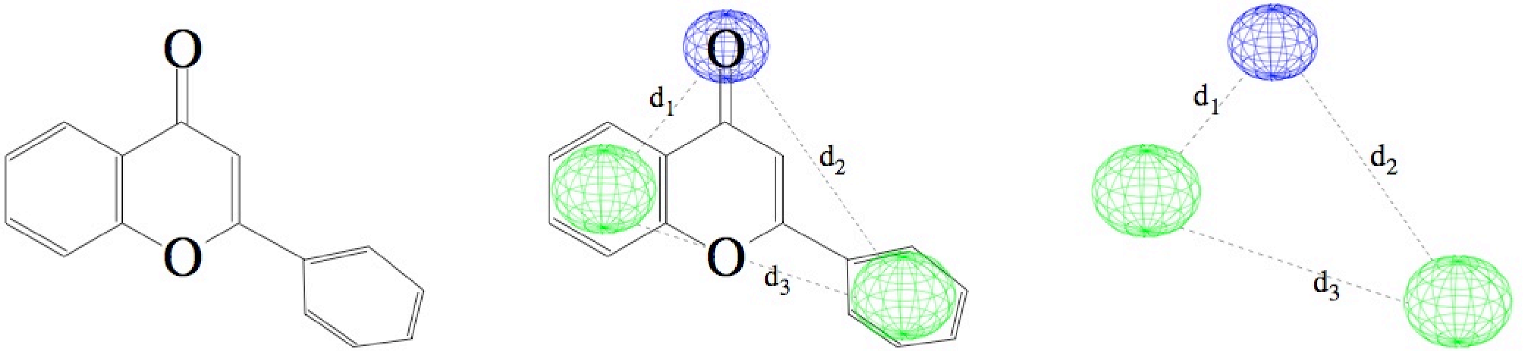
Mahé et al., *J. Chem. Inf. Model.*, 2005.

« All subtree patterns »

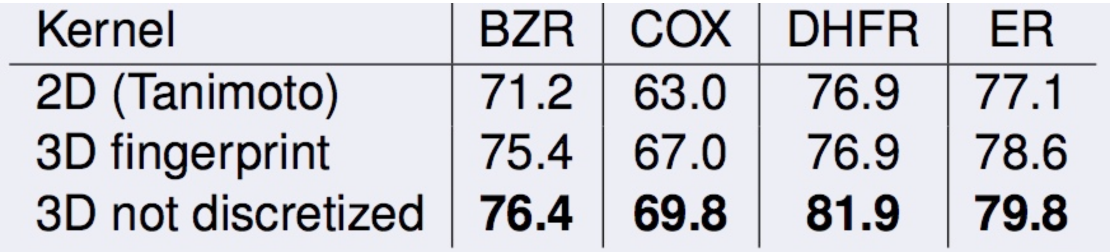


Mahé and V., *Mach. Learn*, 2009.

Example: 3D pharmacophore kernel



$$K(x,y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \exp\left(-\gamma d\left(p_x, p_y\right)\right) \; .$$



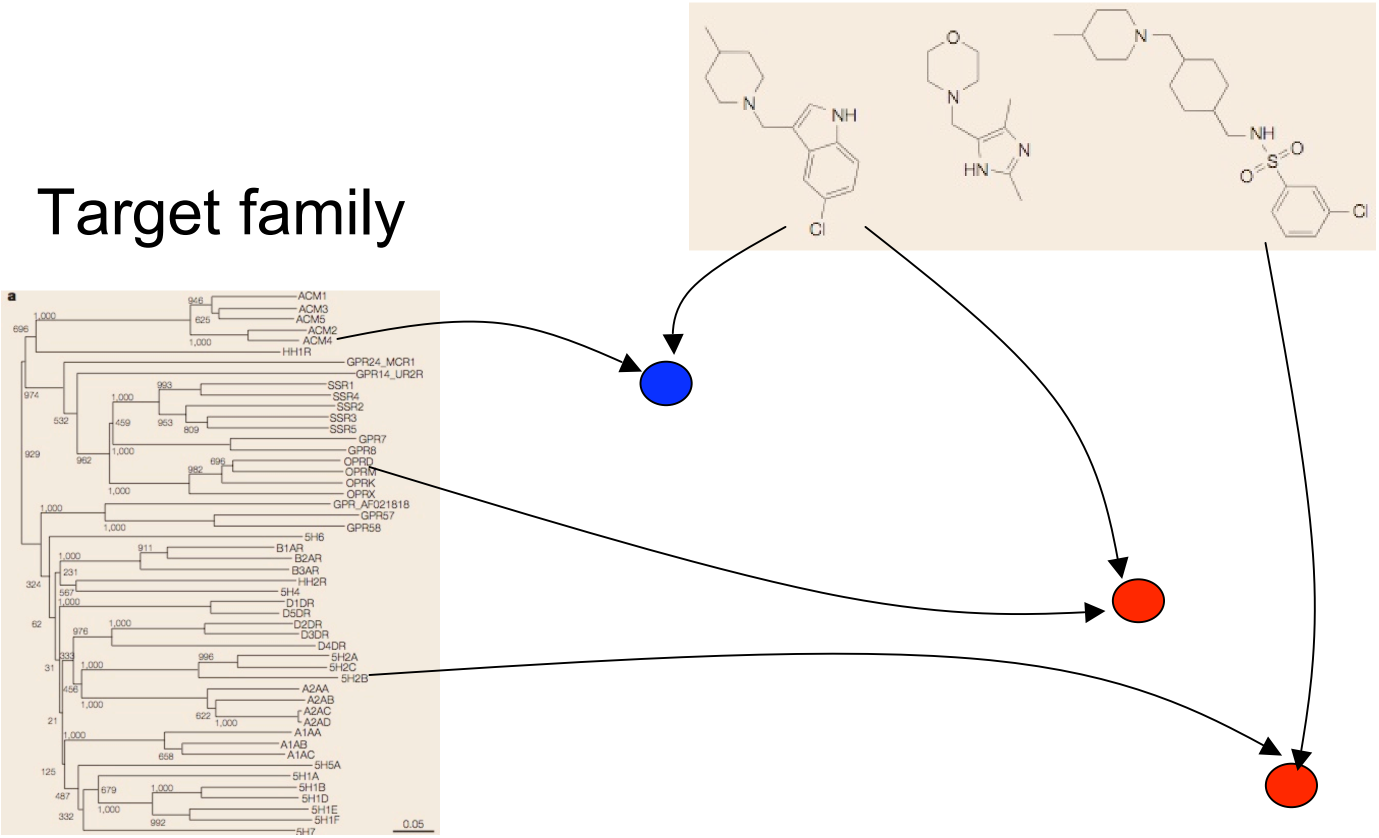
Kernel	BZR	COX	DHFR	ER
2D (Tanimoto)	71.2	63.0	76.9	77.1
3D fingerprint	75.4	67.0	76.9	78.6
3D not discretized	**76.4**	**69.8**	**81.9**	**79.8**

Mahé et al., *J. Chem. Inf. Model.*, 2006.

Summary so far…

- SVM is an algorithm for supervised classification
- SVM can be used with any « classical » vector or fingerprint description (often giving state-of-the-art performance)
- SVM can also be used with more general measures of similarity (like many related *kernel methods*)
- Much effort recently to define such kernels in bio- and chemo-informatics

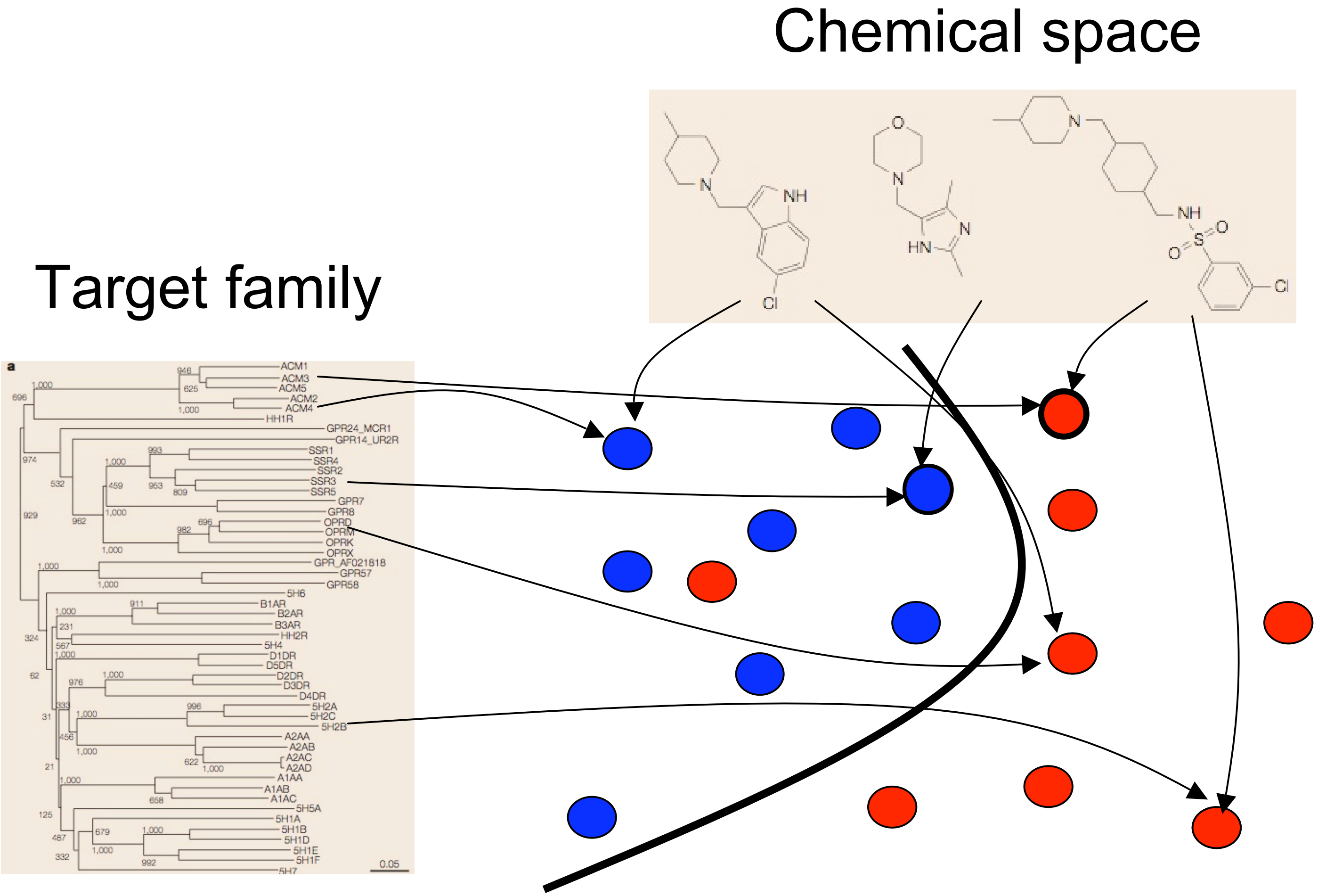
Chemogenomics



Chemical space

Target family

In silico Chemogenomics

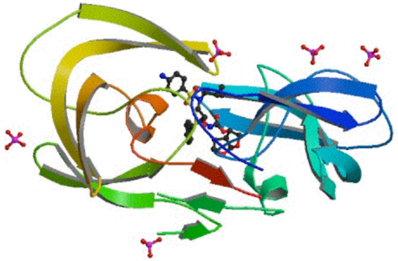


Chemical space

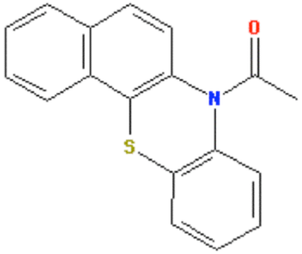
Target family

Fingerprint for a (target,molecule) pair?

t=



c =



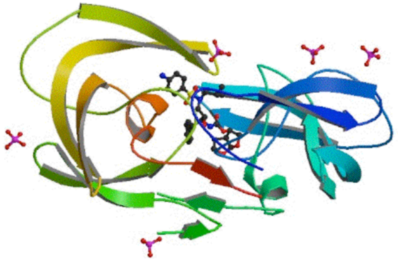
$$\Phi_{tar}(t) = \begin{cases} \text{-Sequence} \\ \text{-Structure} \\ \text{-Evolution} \\ \text{-Expression} \\ \text{-…} \end{cases}$$

$$\Phi_{lig}(c) = \begin{cases} \text{-2D} \\ \text{-3D} \\ \text{-Pharmacophore} \\ \text{-logP, …} \end{cases}$$

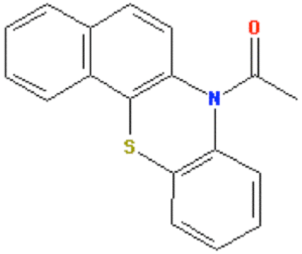
$$\Phi(c,t) = ???$$

Fingerprint for a (target,molecule) pair?

T=



c =

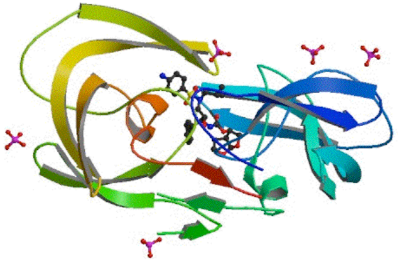


$$\Phi_{tar}(t) = \begin{cases} \text{-Sequence} \\ \text{-Structure} \\ \text{-Evolution} \\ \text{-Expression} \\ \text{-…} \end{cases} \qquad \Phi_{lig}(c) = \begin{cases} \text{-2D} \\ \text{-3D} \\ \text{-Pharmacophore} \\ \text{-logP, …} \end{cases}$$

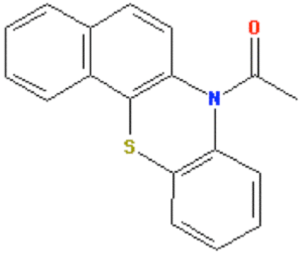
$$\underset{10^6}{\Phi(c,t)} = \underset{10^3}{\Phi_{lig}(c)} \otimes \underset{10^3}{\Phi_{tar}(t)}$$

Similarity for (target,molecule) pairs

t=



c =



$K_{target}(t,t') =$
- -Sequence
- -Structure
- -Evolution
- -Expression
- -…

$K_{ligand}(c,c') =$
- -2D
- -3D
- -Pharmacophore
- -logP, …

$$K\left((c,t),(c',t')\right) = K_{target}(t,t') \times K_{ligand}(c,c')$$

Summary: SVM for chemogenomics

1. Choose a kernel (similarity) for targets
2. Choose a kernel (similarity) for ligands
3. Train a SVM model with the product kernel for (target/ligand) pairs

Application: virtual screening of GPCR

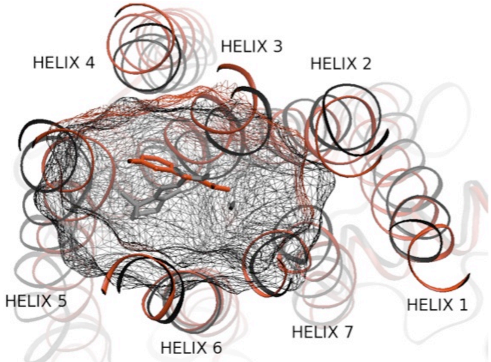
Data: GLIDA database filtered for drug-like compounds

- 2446 ligands
- 80 GPCR
- 4051 interactions
- *4051 negative interactions generated randomly*

Ligand similarity

-2D Tanimoto

-3D pharmacophore



Target similarities

-0/1 Dirac (no similarity)

-Multitask (uniform similarity)

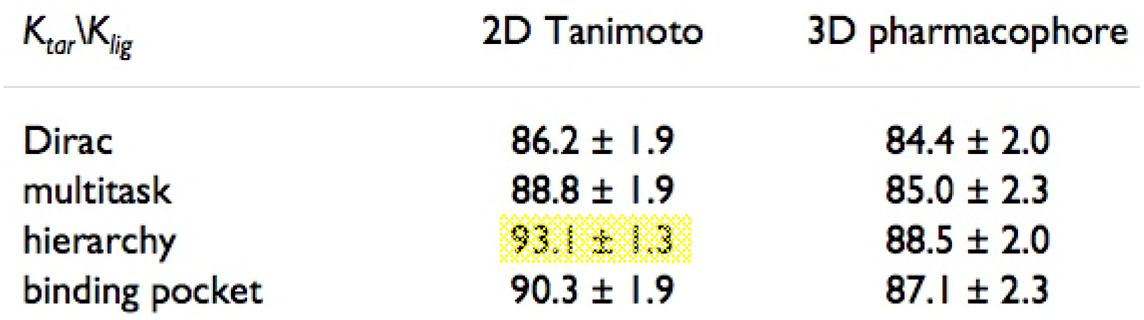
-GLIDA's hierarchy similarity

-Binding pocket similarity (31 AA)

(Jacob et al., *BMC Bioinformatics*, 2008)

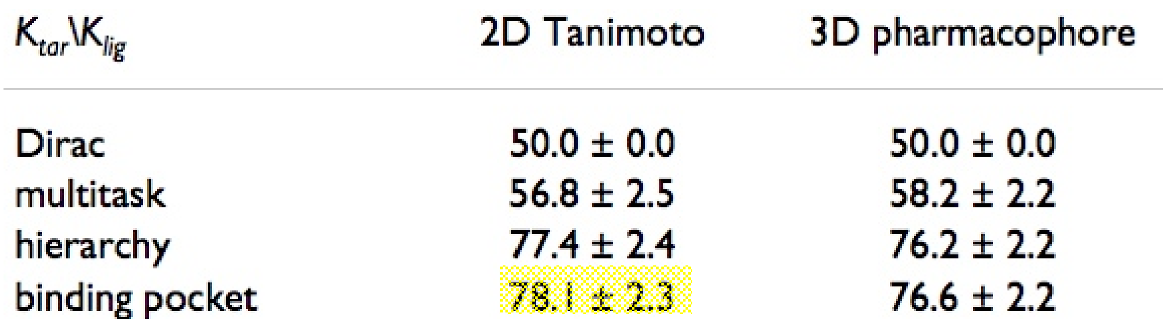
Results (mean accuracy over GPCRs)

5-fold cross-validation



$K_{tar}\backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	86.2 ± 1.9	84.4 ± 2.0
multitask	88.8 ± 1.9	85.0 ± 2.3
hierarchy	93.1 ± 1.3	88.5 ± 2.0
binding pocket	90.3 ± 1.9	87.1 ± 2.3

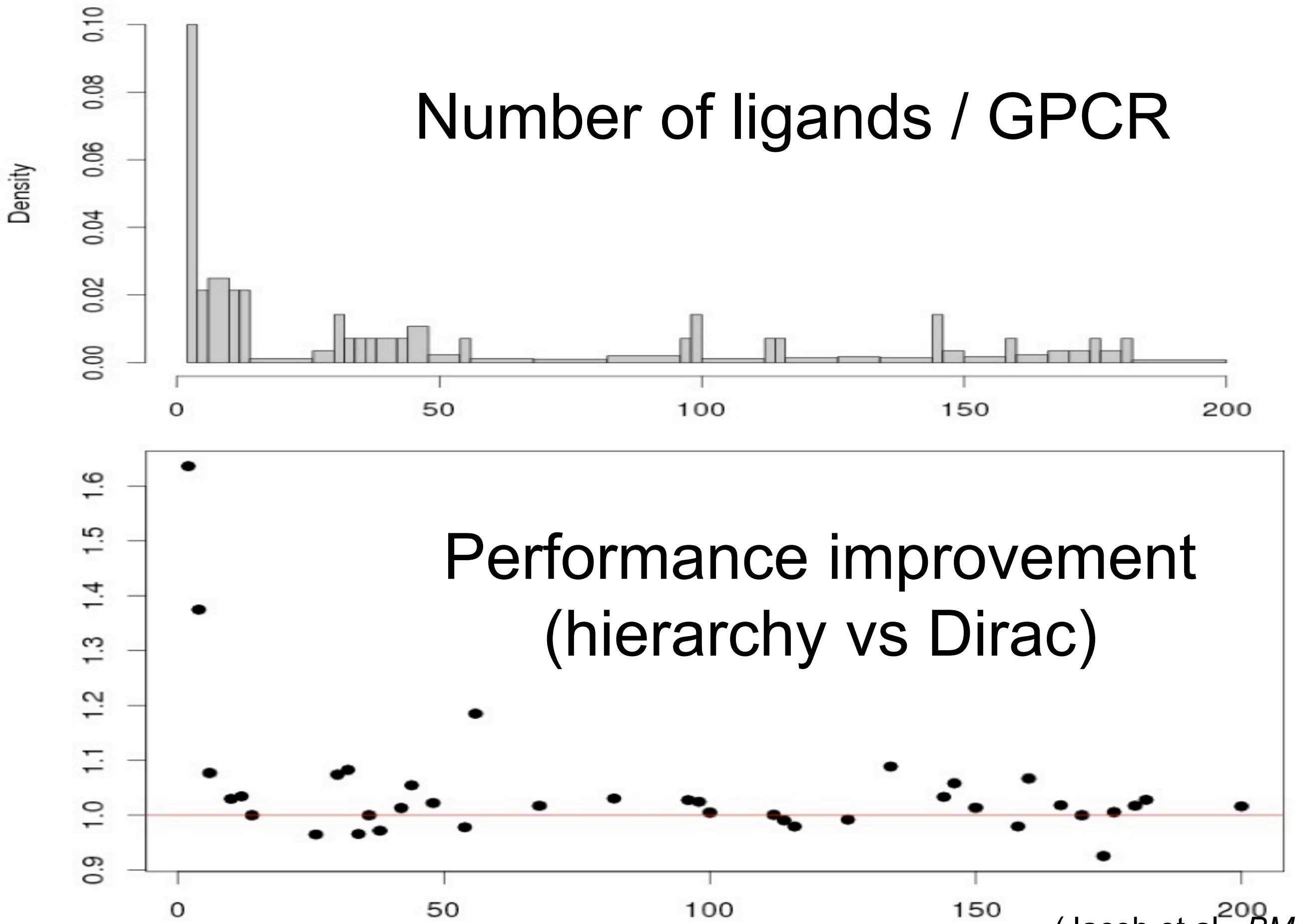
Orphan GPCRs setup



$K_{tar}\backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	50.0 ± 0.0	50.0 ± 0.0
multitask	56.8 ± 2.5	58.2 ± 2.2
hierarchy	77.4 ± 2.4	76.2 ± 2.2
binding pocket	78.1 ± 2.3	76.6 ± 2.2

(Jacob et al., *BMC Bioinformatics*, 2008)

Influence of the number of known ligands



Number of ligands / GPCR

Performance improvement (hierarchy vs Dirac)

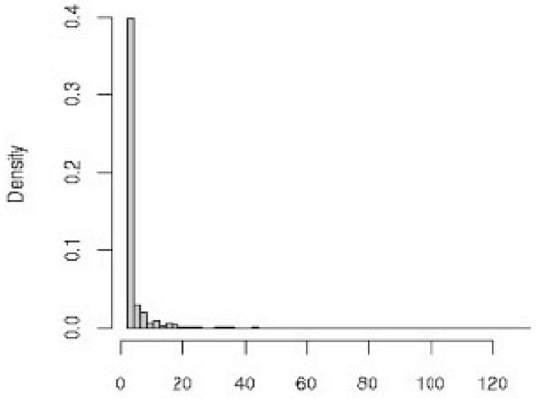
(Jacob et al., *BMC Bioinformatics*, 2008)

Screening of enzymes, GPCRs, ion channels

Data: KEGG BRITE database, redundancy removed

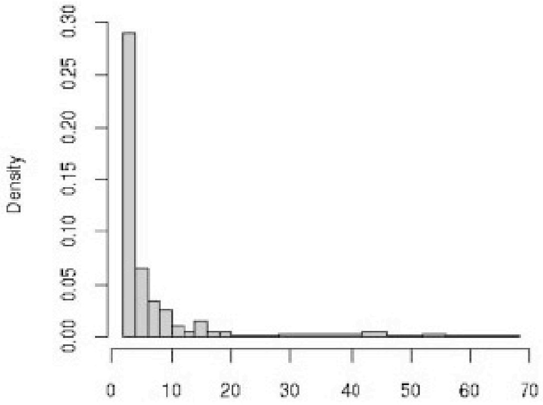
Enzymes

-675 targets
-524 molecules
-1218 interactions
-1218 negatives



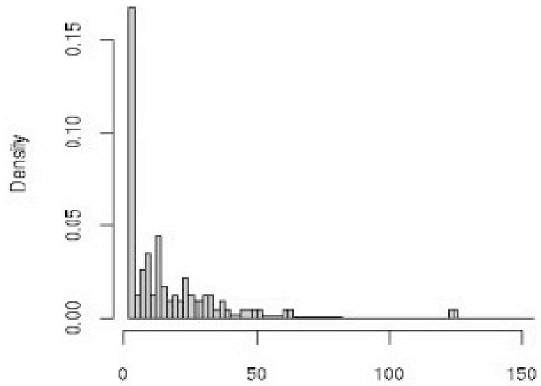
GPCRs

-100 targets
-219 molecules
-399 interactions
-399 negatives



Ion channels

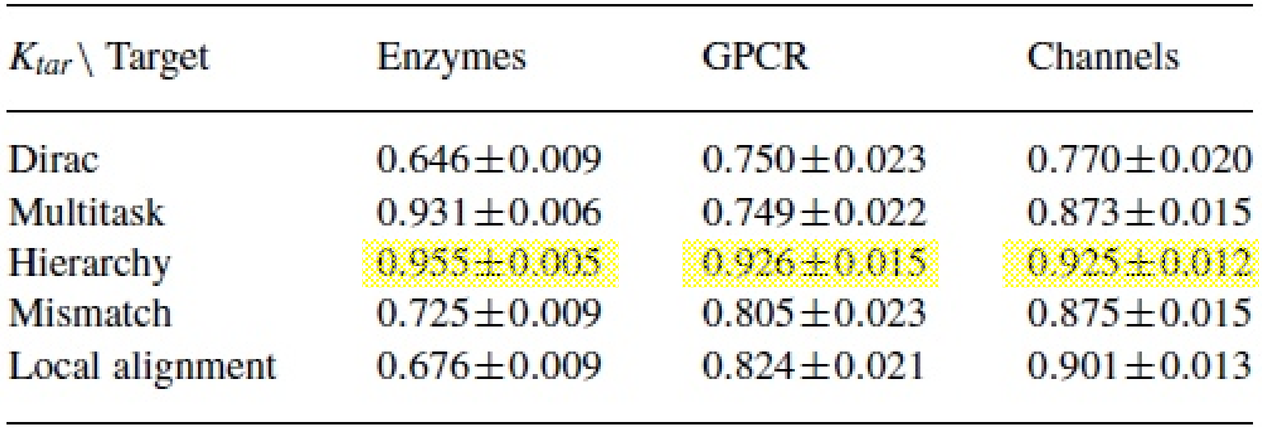
-114 targets
-462 molecules
-1165 interactions
-1165 negatives



(Jacob and V., *Bioinformatics*, 2008)

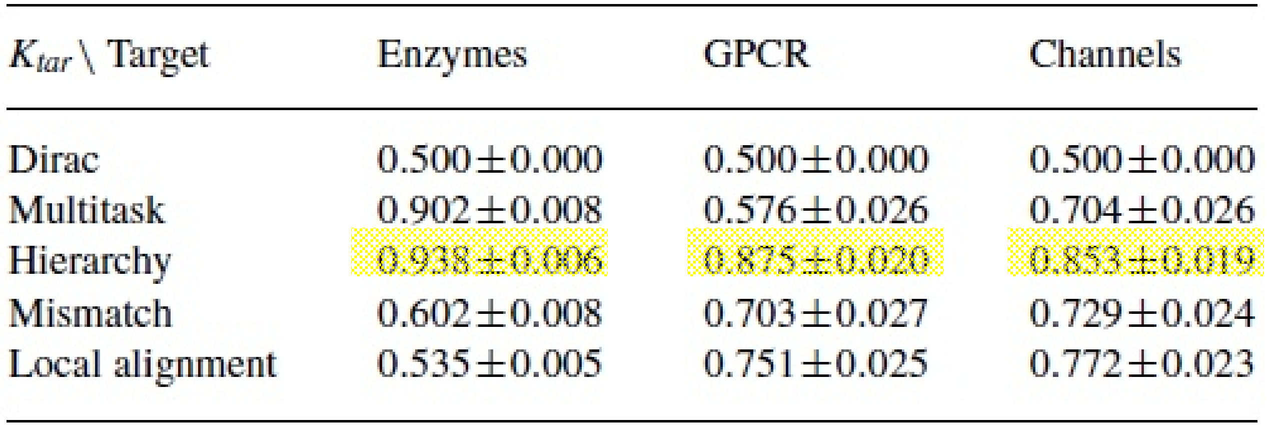
Results (mean AUC)

10-fold CV



K_{tar} \ Target	Enzymes	GPCR	Channels
Dirac	0.646±0.009	0.750±0.023	0.770±0.020
Multitask	0.931±0.006	0.749±0.022	0.873±0.015
Hierarchy	0.955±0.005	0.926±0.015	0.925±0.012
Mismatch	0.725±0.009	0.805±0.023	0.875±0.015
Local alignment	0.676±0.009	0.824±0.021	0.901±0.013

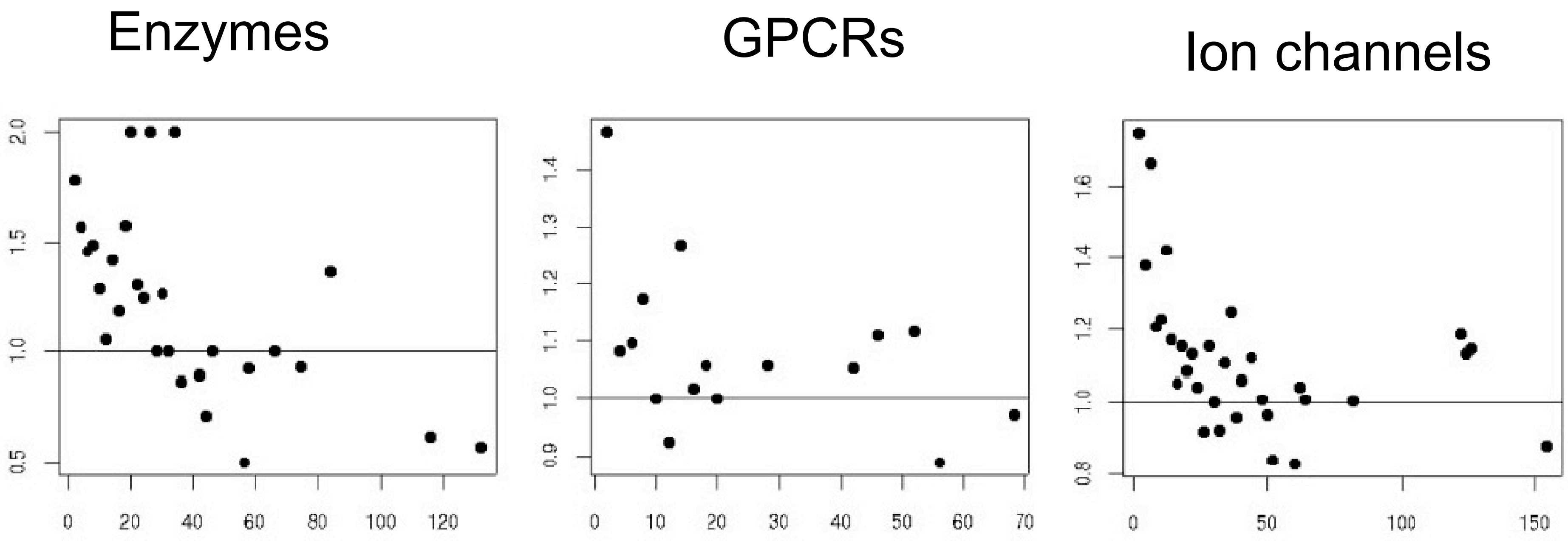
Orphan setting



K_{tar} \ Target	Enzymes	GPCR	Channels
Dirac	0.500±0.000	0.500±0.000	0.500±0.000
Multitask	0.902±0.008	0.576±0.026	0.704±0.026
Hierarchy	0.938±0.006	0.875±0.020	0.853±0.019
Mismatch	0.602±0.008	0.703±0.027	0.729±0.024
Local alignment	0.535±0.005	0.751±0.025	0.772±0.023

(Jacob and V., *Bioinformatics*, 2008)

Influence of the number of known ligands



Relative improvement : hierarchy vs Dirac

(Jacob and V., *Bioinformatics*, 2008)

Conclusion

- SVM offer state-of-the-art performance in chemo- and bio-informatics
- Much work recently to define « kernels » for small molecules and proteins
- Combining them provides a theoretically sound and computationnally efficient framework for *in silico* chemogenomics
- Promising results on several benchmarks for important target families

References : http://cbio.ensmp.fr/~jvert/

- L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach", *Bioinformatics*, 24(19):2149-2156, 2008
- L. Jacob, B. Hoffmann, V. Stoven and J.-P. Vert, "Virtual screening of GPCRs: an *in silico* chemogenomics approach", *BMC Bioinformatics*, 9:363, 2008.
- J.-P. Vert and L. Jacob, "Machine learning for *in silico* virtual screening and chemical genomics: new strategies", *Combinatorial Chemistry & High Throughput Screening*, 11(8):677-685, 2008.
- P. Mahé and J.-P. Vert, "Graph kernels based on tree patterns for molecules", to appear in *Machine Learning*, 2009.
- P. Mahé, L. Ralaivola, V. Stoven and J.-P. Vert, "The pharmacophore kernel for virtual screening with support vector machines", *Journal of Chemical Information and Modeling*, vol. 46, n.5, p.2003-2014, 2006.
- P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret and J.-P. Vert, "Graph kernels for molecular structure-activity relationship analysis with support vector machines", *Journal of Chemical Information and Modeling*, vol. 45, n. 4, 939 -951, 2005.