Kernel methods for virtual screening and *in silico* chemogenomics

Jean-Philippe Vert

Curie Institute - INSERM U900 - Mines ParisTech

Computational Biology Research Center, Tokyo, Aug 7, 2009.

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national
de la santé et de la recherche médicale

MINES ParisTech

Outline

1. Kernel methods for QSAR and virtual screening
2. 2D kernels
3. 3D kernels
4. Towards *in silico* chemogenomics

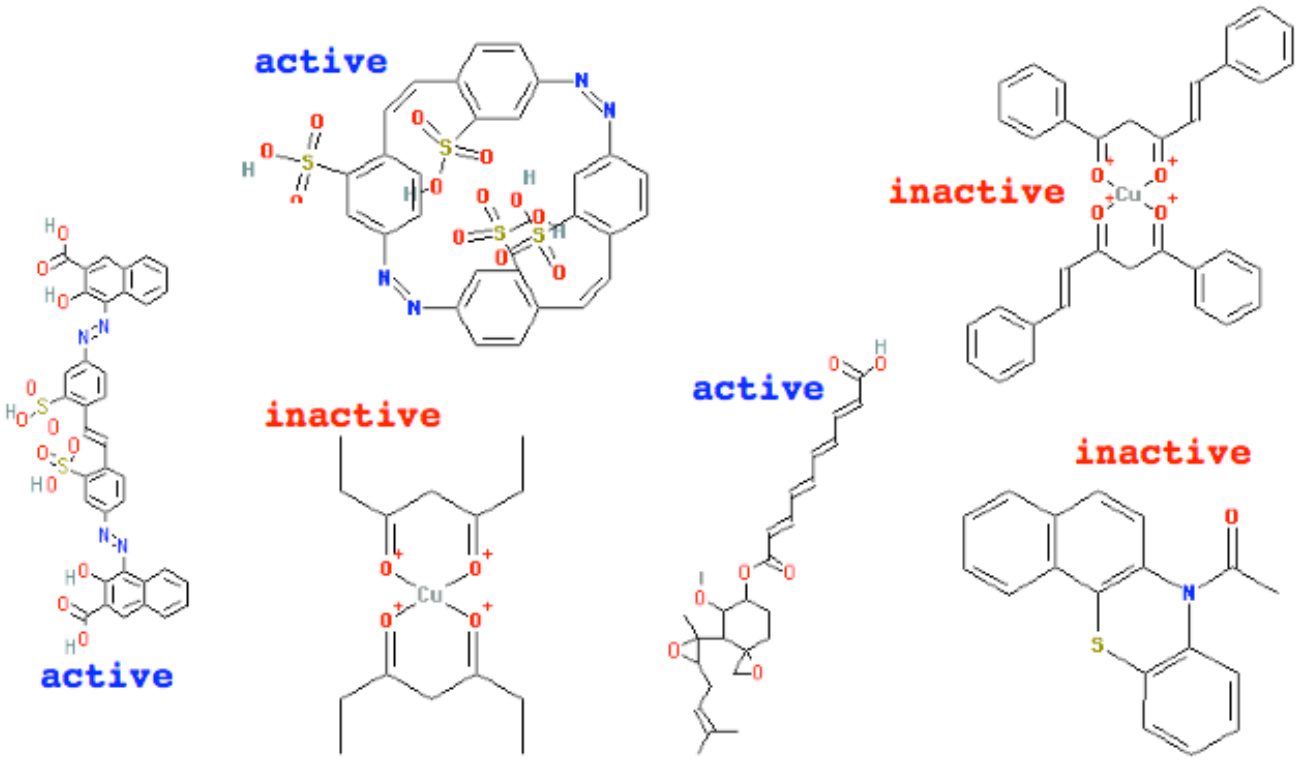
Kernel methods for QSAR and virtual screening

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national
de la santé et de la recherche médicale

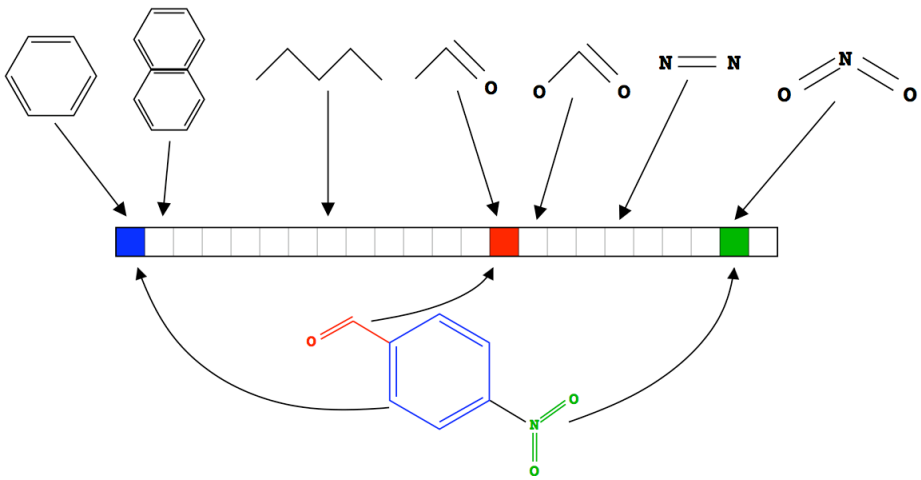
MINES ParisTech

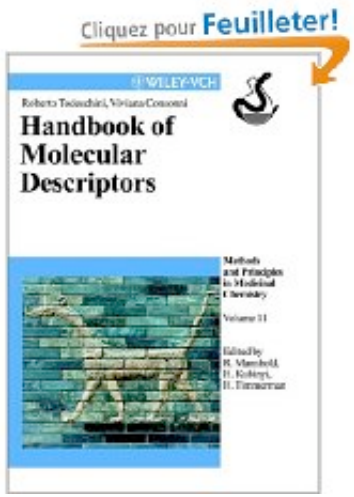
Ligand-based virtual screening / QSAR

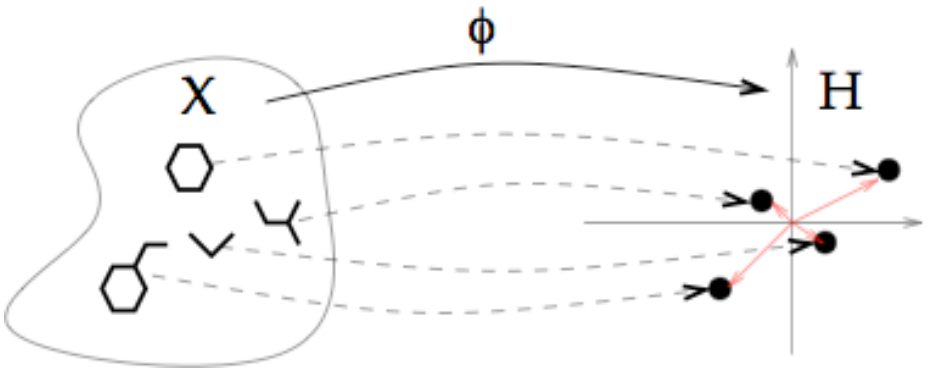


From http://cactus.nci.nih.gov

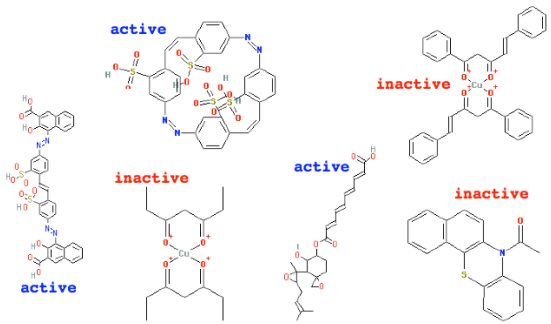
Represent each molecule as a vector…

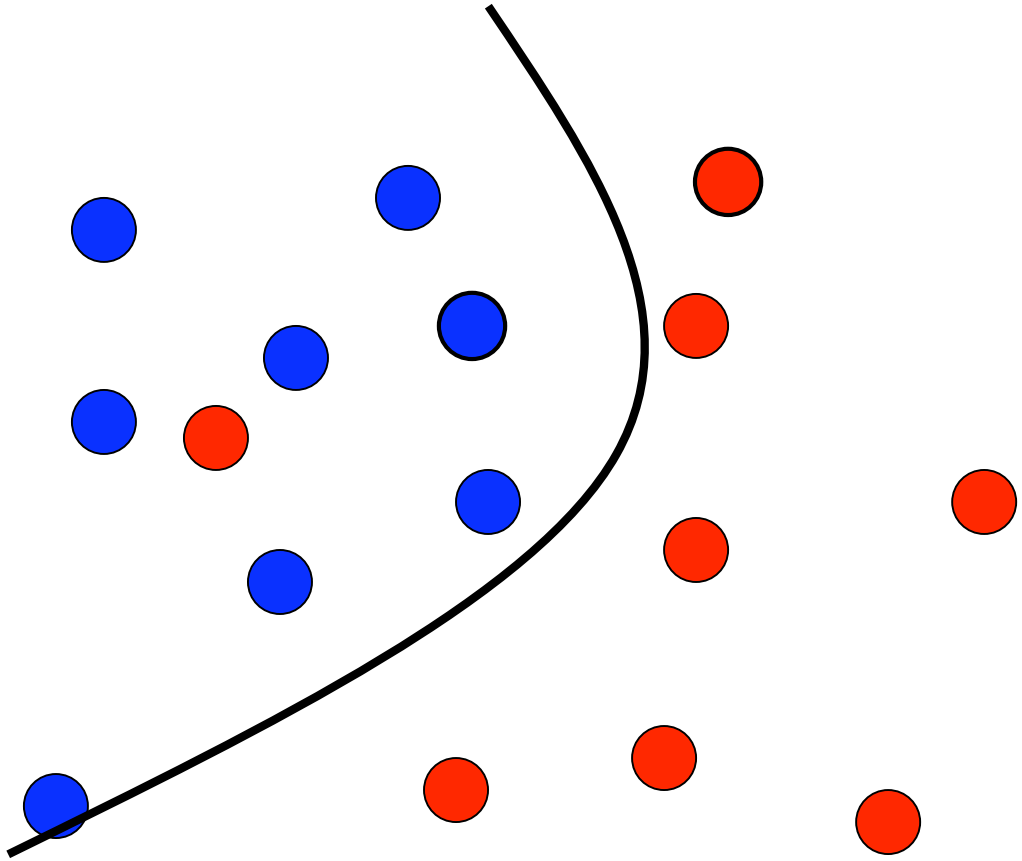






…and discriminate with machine learning





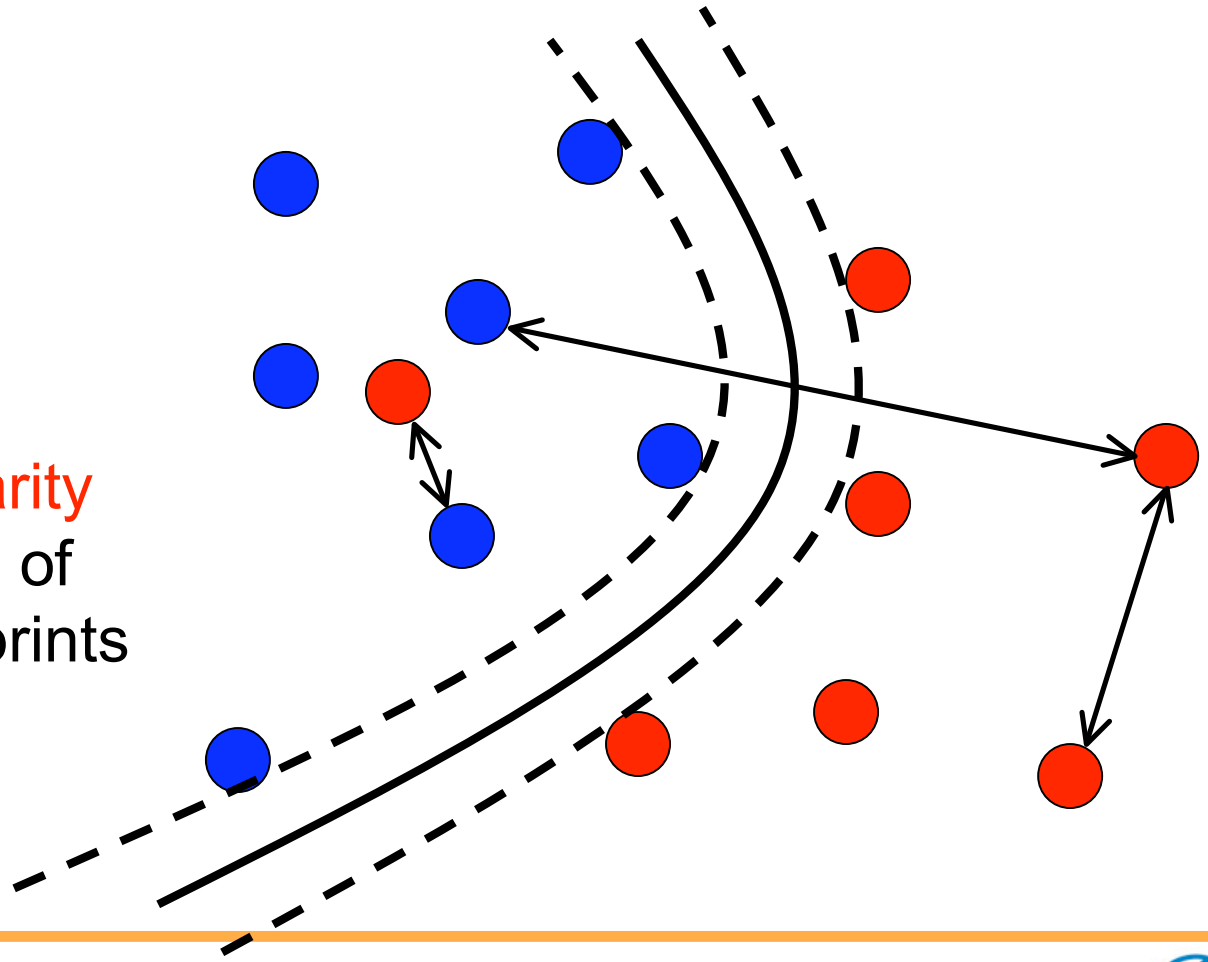
-LDA
-PLS
-Neural network
-Decision trees
-Nearest neighbour
-SVM, …

Support Vector Machine (SVM)

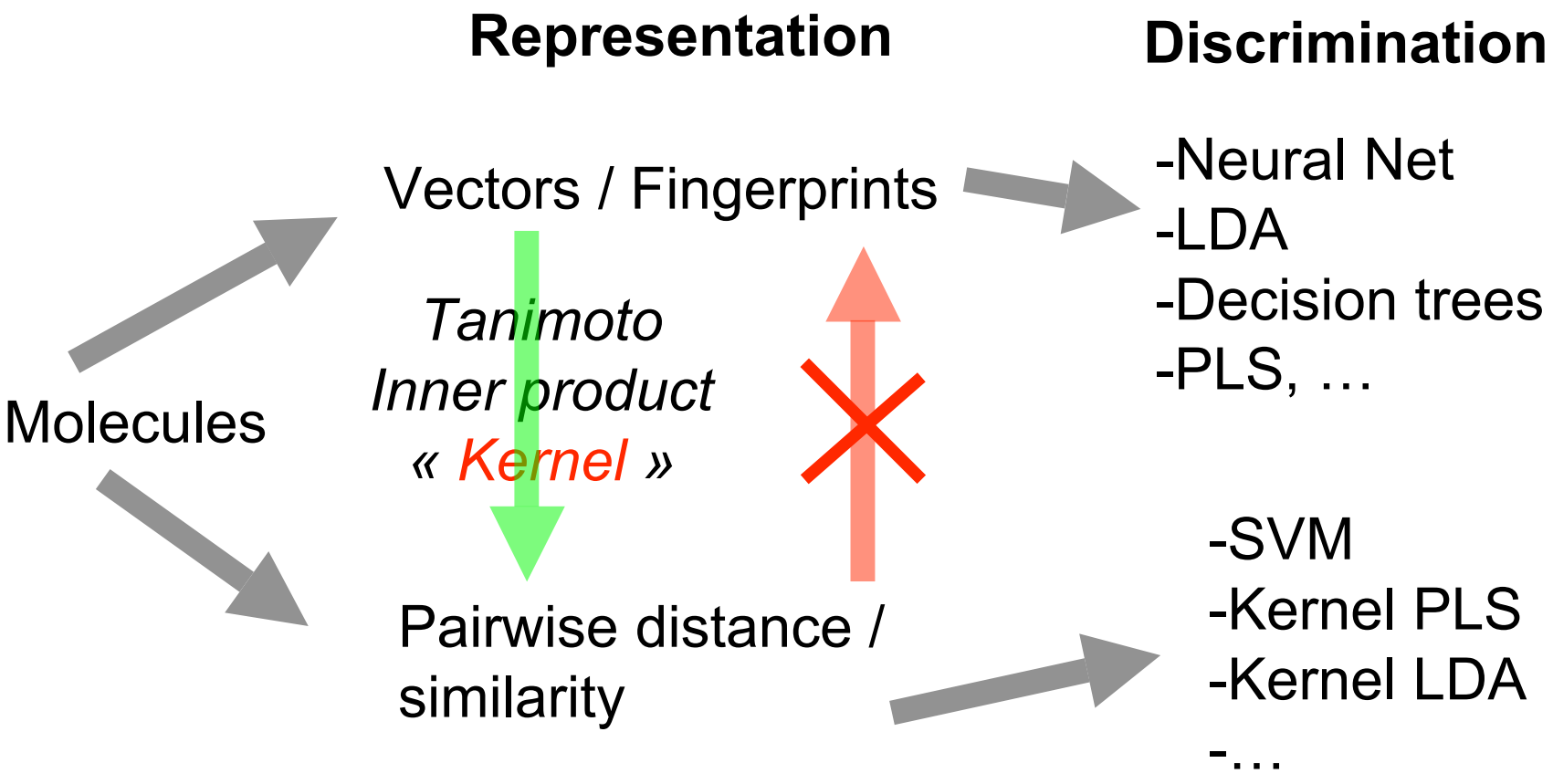
- Large margin

- Nonlinear

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

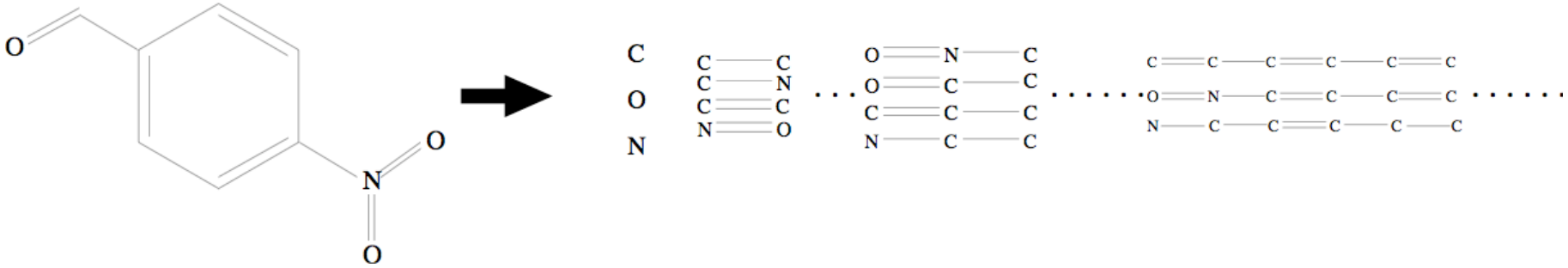


From fingerprints to similarities



2D kernels

2D fragment kernels (walks)



- For any $d > 0$ let $\phi_d(x)$ be the vector of counts of all fragments of length d:

$$\phi_1(x) = (\ \#(\mathrm{C}), \#(\mathrm{O}), \#(\mathrm{N}), \ \ldots)^\top$$
$$\phi_2(x) = (\ \#(\mathrm{C{-}C}), \#(\mathrm{C{=}O}), \#(\mathrm{C{-}N}), \ \ldots)^\top \quad \text{etc...}$$

- The 2D fingerprint kernel is defined, for $\lambda < 1$, by

$$K_{2D}(x, x') = \sum_{d=1}^{\infty} \lambda(d) \phi_d(x)^\top \phi_d(x') \ .$$

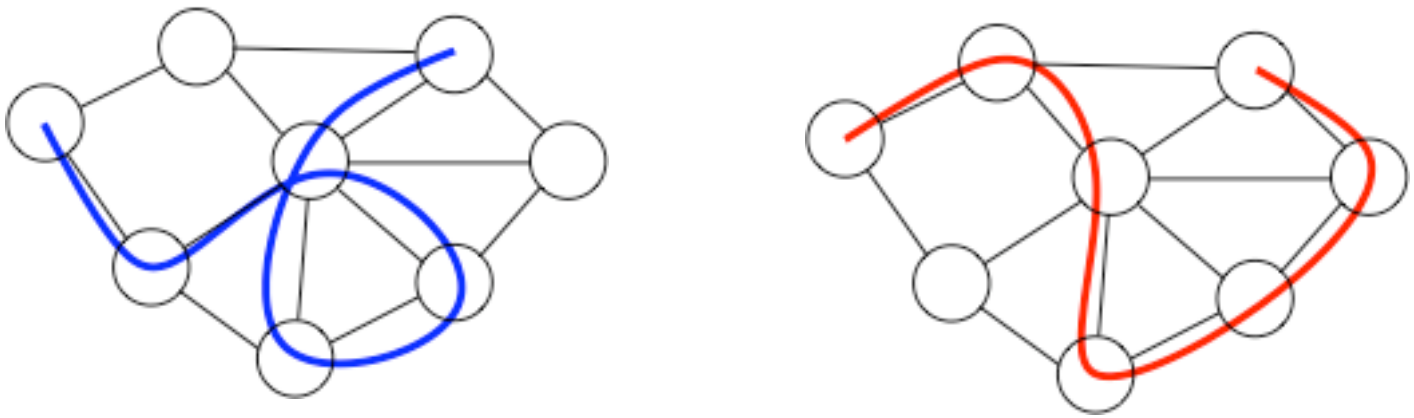
Kashima et al. (2003), Gärtner et al. (2003)

Properties of the 2D fragment kernel

- Corresponds to a fingerprint of infinite size
- Solves the problem of clashes and memory storage (fingerprints are not computed explicitly)
- Can be computed efficiently in O(|x|^3 |x'|^3) (much faster in practice)

Kashima et al. (2003), Gärtner et al. (2003)

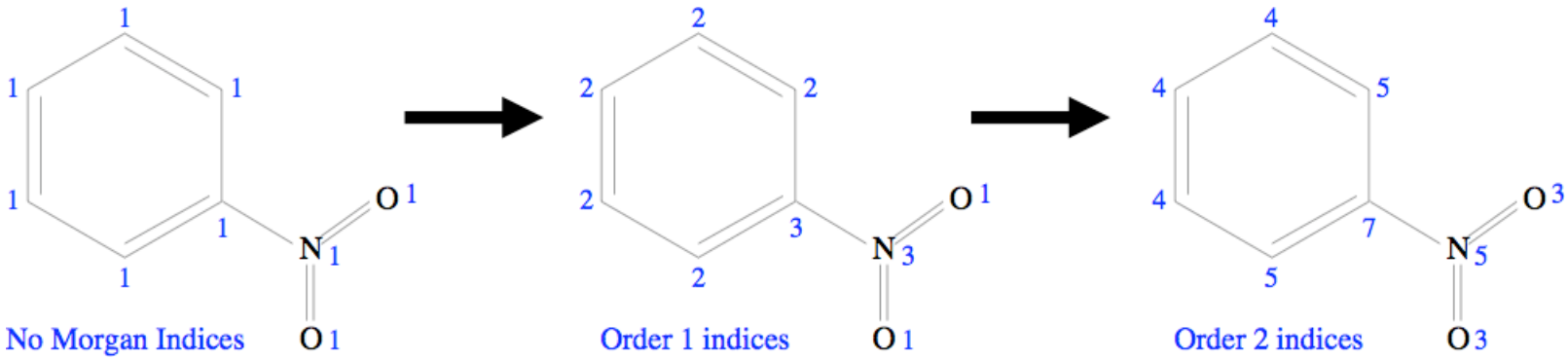
Remark: walks vs paths



Computing the path kernel is NP-hard

Gärtner et al. (2003)

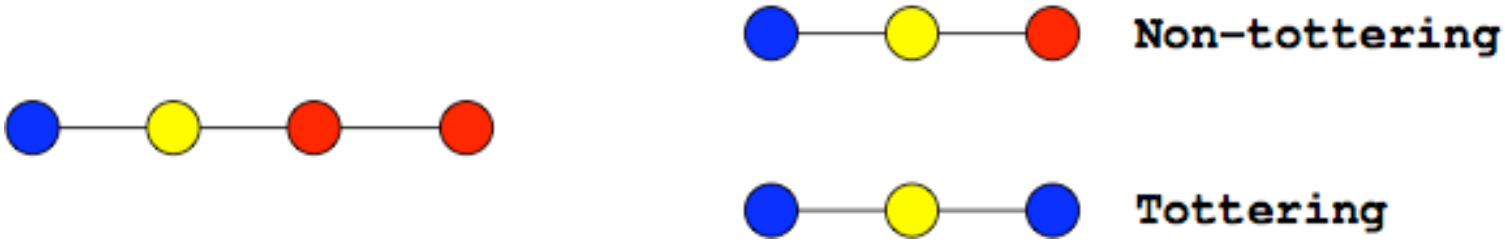
Extension 1: label enrichment



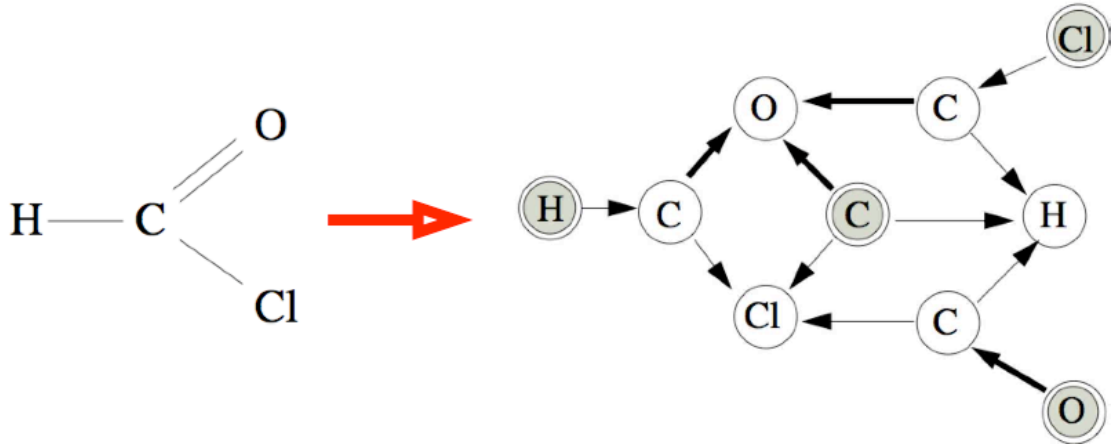
-Increases the expressiveness of the kernel
-Faster computation with more labels
-Other relabeling schemes are possible

Mahé et al. (2005)

Extension 2: removing tottering walks



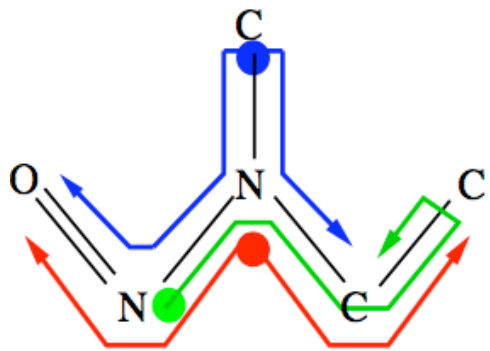
-Tottering walks are irrelevant for many applications (noise)
-Focusing on non-tottering walks only is a way to get closer to the path kernel (e.g., equivalent on trees)

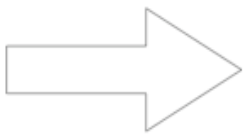


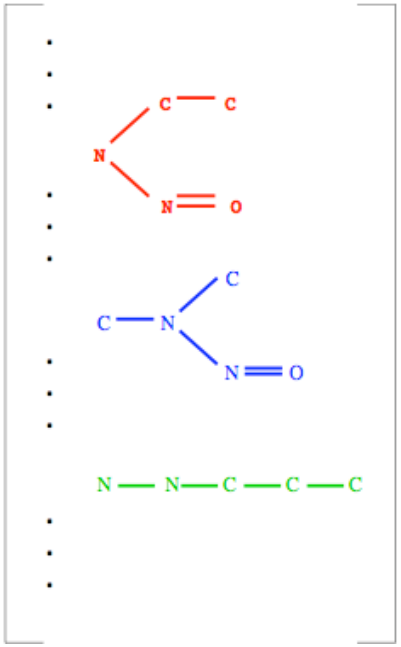
Mahé et al. (2005)

Extension 3: subtree patterns

« All subtree patterns »







Mahé and V., *Mach. Learn*, 2009.

$$\mathcal{T}(v, n+1) = \sum_{R \subset \mathcal{N}(v)} \prod_{v' \in R} \lambda_t(v, v') \mathcal{T}(v', n)$$

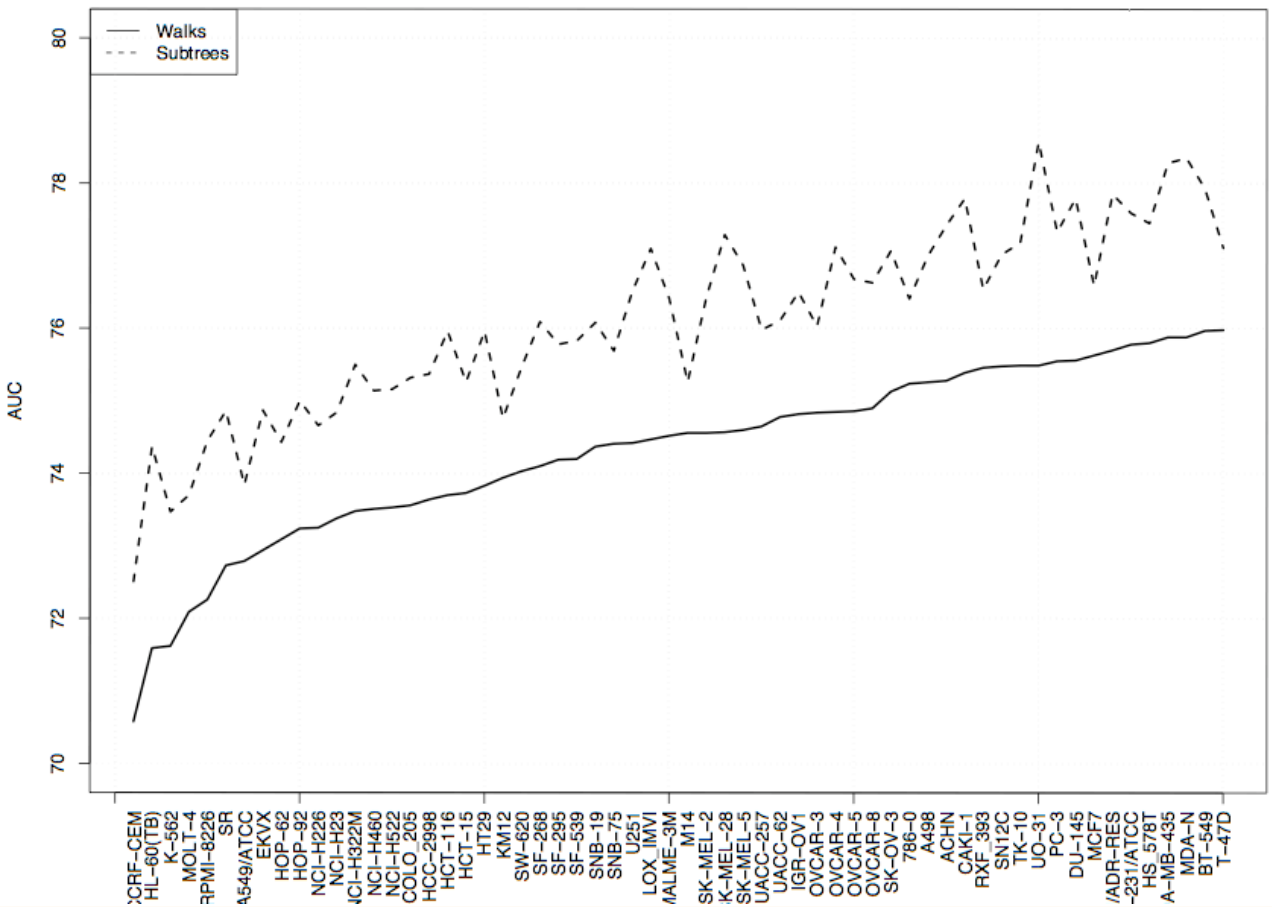
Ramon et al. (2004), Mahé & V. (2009)

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

2D subtree vs walk kernel



NCI 60 dataset

Mahé & V. (2009)

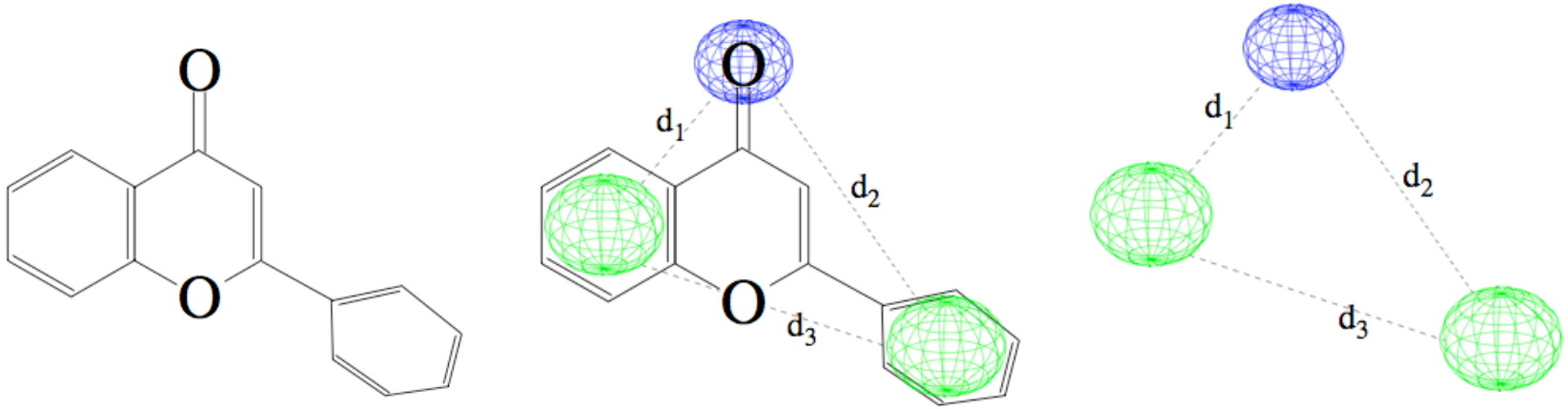
institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

3D pharmacophore kernel

3-point pharmacophores



A set of 3 atoms, and 3 inter-atom distances:

$$\mathcal{T} = \{((x_1, x_2, x_3), (d_1, d_2, d_3)), x_i \in \{\text{atom types}\}; d_i \in \mathbb{R}\}$$

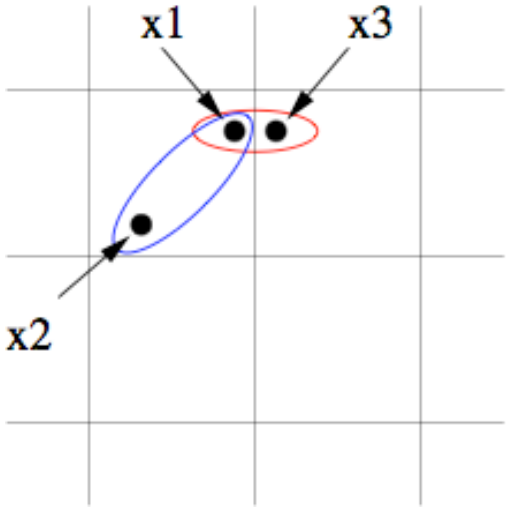
Mahé et al., *J. Chem. Inf. Model.*, 2006.

3D fingerprint kernel

1. Discretize the space of pharmacophores $\mathcal{T}$ (e.g., 6 atoms or groups of atoms, 6-7 distance bins) into a finite set $\mathcal{T}_d$
2. Count the number of occurrences $\phi_t(x)$ of each pharmacophore bin t in a given molecule x, to form a pharmacophore fingerprint.

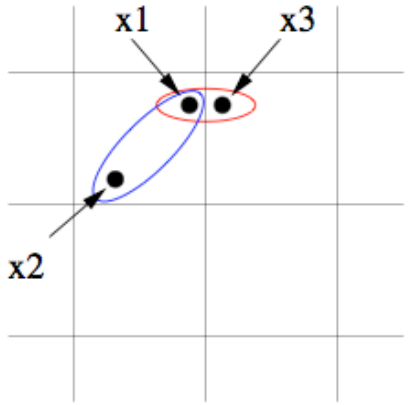
A simple 3D kernel is the inner product of pharmacophore fingerprints:

$$K(x, x') = \sum_{t \in \mathcal{T}_d} \phi_t(x)\phi_t(x') \ .$$



From the fingerprint kernel to the pharmacophore kernel

$$
\begin{aligned}
K(x,y) &= \sum_{t \in \mathcal{T}_d} \phi_t(x)\phi_t(y) \\
&= \sum_{t \in \mathcal{T}_d} \Big(\sum_{p_x \in \mathcal{P}(x)} \mathbf{1}(\mathrm{bin}(\mathbf{p_x}) = \mathbf{t}) \Big) \Big(\sum_{p_y \in \mathcal{P}(y)} \mathbf{1}(\mathrm{bin}(\mathbf{p_y}) = \mathbf{t}) \Big) \\
&= \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \mathbf{1}(\mathrm{bin}(\mathbf{p_x}) = \mathrm{bin}(\mathbf{p_y}))
\end{aligned}
$$



$$
K(x,y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \exp\left(-\gamma ||p_x - p_y||^2\right)
$$

Experiments

- BZR: ligands for the benzodiazepine receptor
- COX: cyclooxygenase-2 inhibitors
- DHFR: dihydrofolate reductase inhibitors
- ER: estrogen receptor ligands

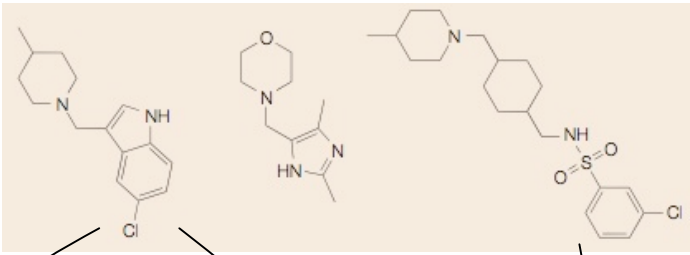
Kernel	BZR	COX	DHFR	ER
2D (Tanimoto)	71.2	63.0	76.9	77.1
3D fingerprint	75.4	67.0	76.9	78.6
3D not discretized	**76.4**	**69.8**	**81.9**	**79.8**

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

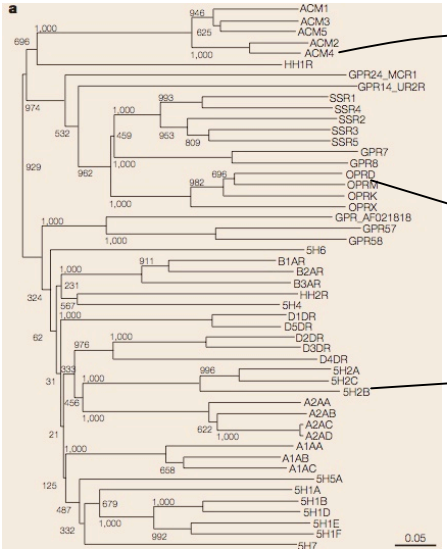
Towards *in silico* chemogenomics

Chemogenomics

Chemical space

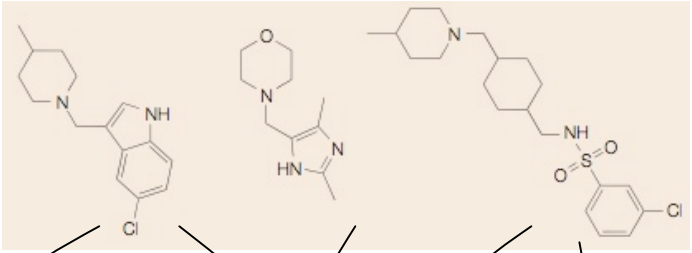


Target family

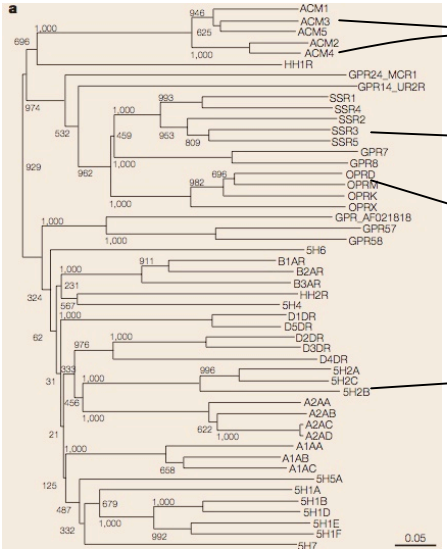


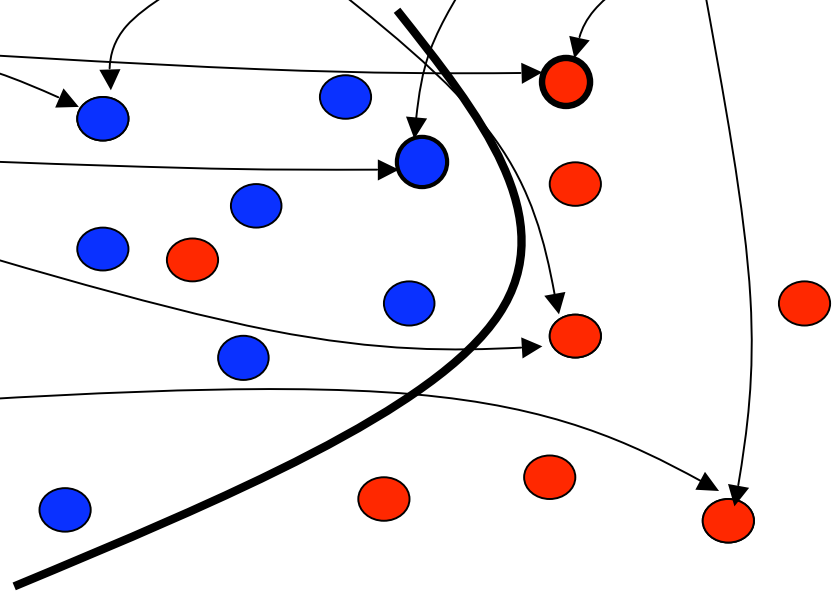
In silico Chemogenomics

Chemical space



Target family



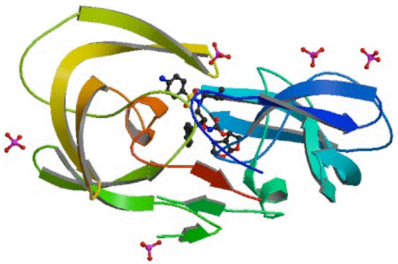


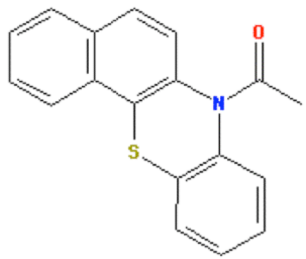
institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Fingerprint for a (target,molecule) pair?

t= 

c = 

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

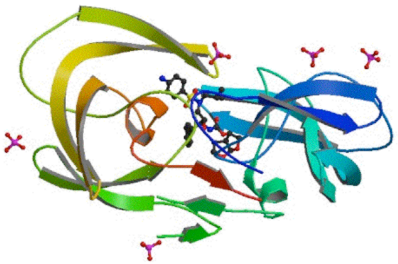
$$\Phi(c,t) = ???$$

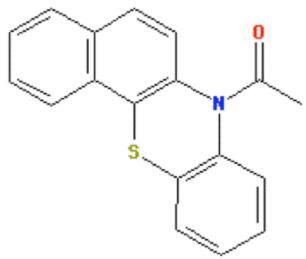
institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Fingerprint for a (target,molecule) pair?

T= 

c = 

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

$$\Phi(c,t) = \Phi_{lig}(c) \otimes \Phi_{tar}(t)$$

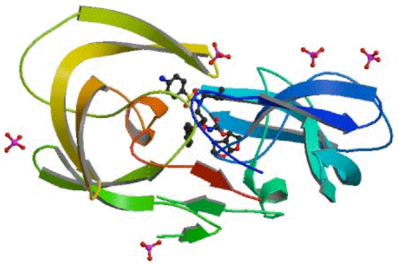
10^6 10^3 10^3

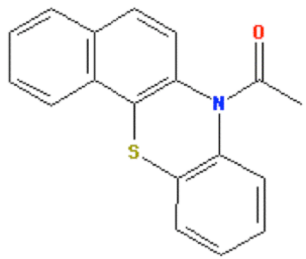
institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Similarity for (target,molecule) pairs

t= 

c = 

$$K_{target}(t,t') = \begin{cases} \text{-Sequence} \\ \text{-Structure} \\ \text{-Evolution} \\ \text{-Expression} \\ \text{-…} \end{cases} \qquad K_{ligand}(c,c') = \begin{cases} \text{-2D} \\ \text{-3D} \\ \text{-Pharmacophore} \\ \text{-logP, …} \end{cases}$$

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

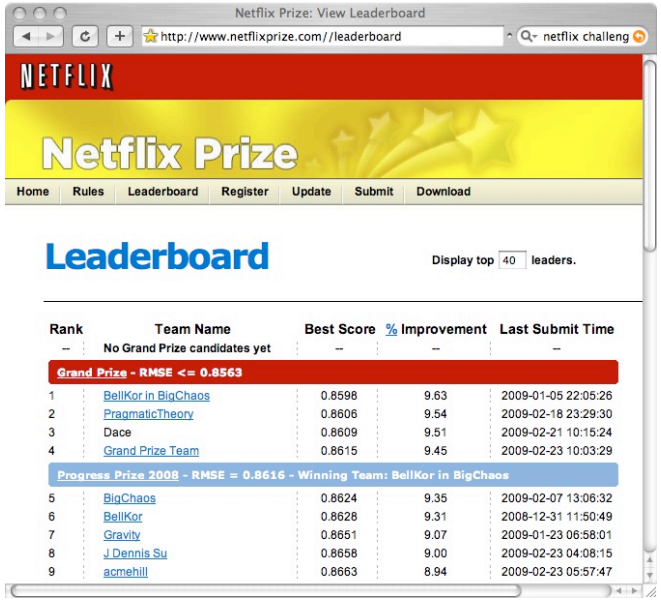
Summary: SVM for chemogenomics

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

Important remark

- New methods are being actively developed in machine learning for learning over pairs

- « Collaborative filtering », « transfer learning », « multitask learning », « MMMF », « pairwise SVM », etc…

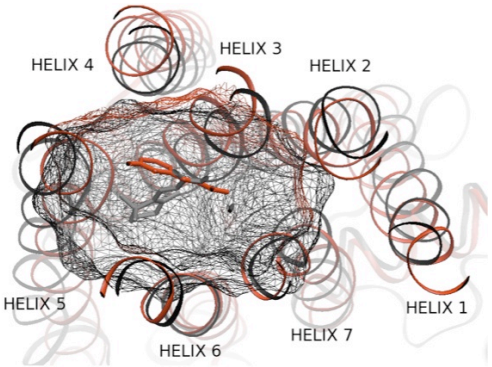


37k registered teams from 180 countries!

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)

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

Results (mean accuracy over GPCRs)

5-fold cross-validation

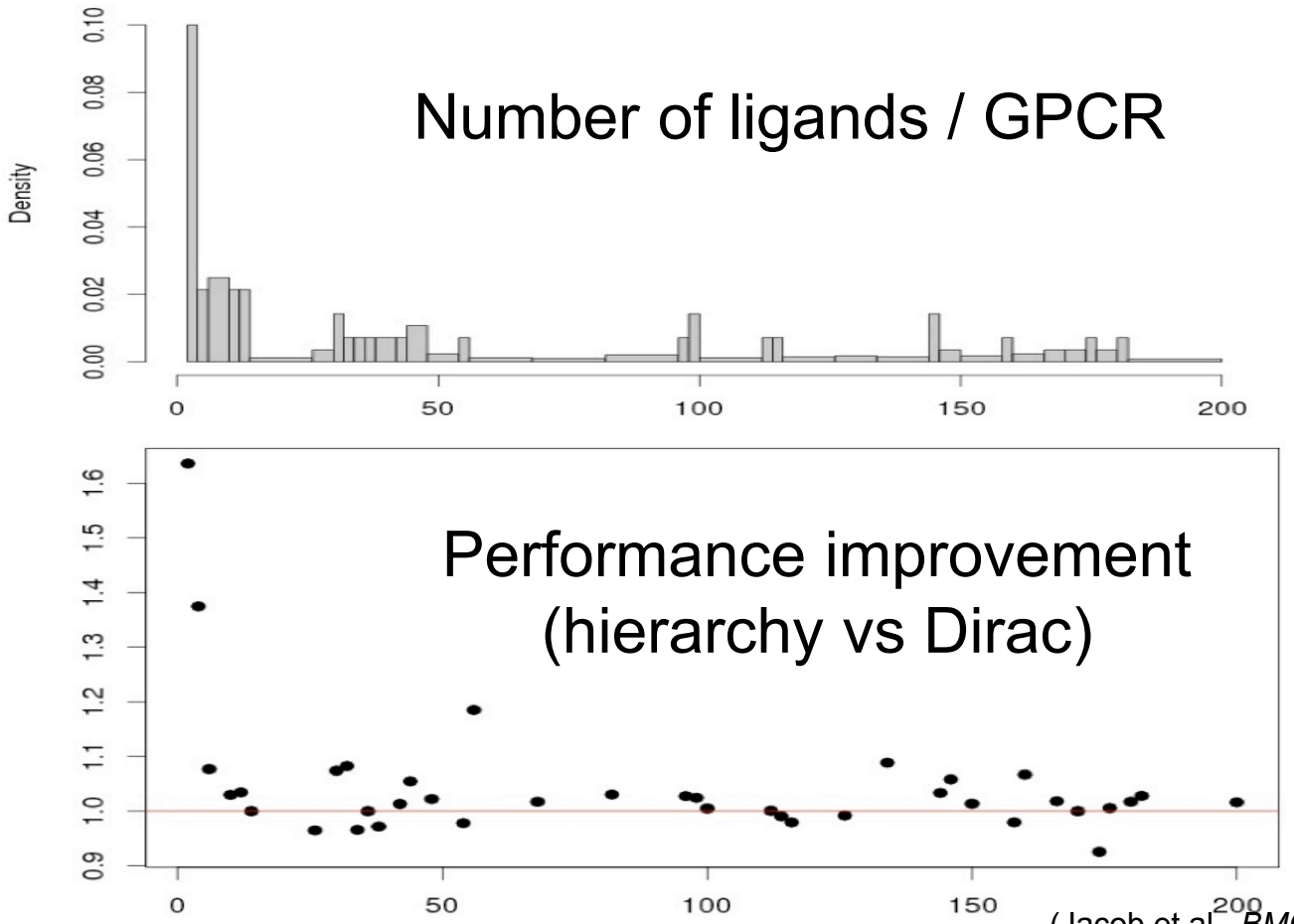
$K_{tar} \backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	86.2 ± 1.9	84.4 ± 2.0
multitask	88.8 ± 1.9	85.0 ± 2.3
hierarchy	93.1 ± 1.3	88.5 ± 2.0
binding pocket	90.3 ± 1.9	87.1 ± 2.3

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



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

institutCurie
Ensemble, prenons le cancer de vitesse.

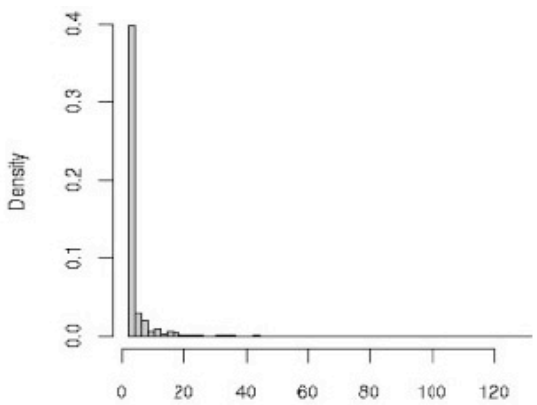
Inserm
Institut national de la santé et de la recherche médicale

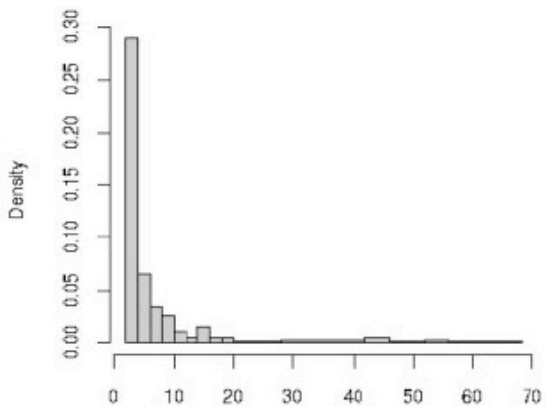
MINES ParisTech

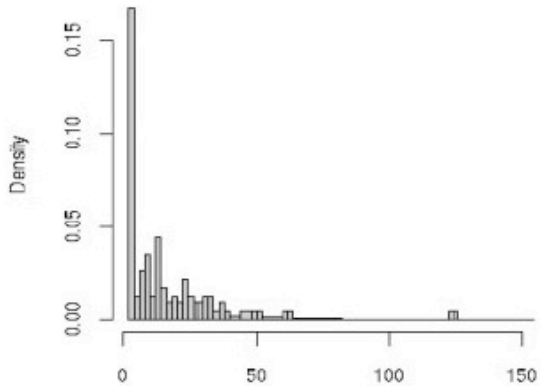
Screening of enzymes, GPCRs, ion channels

Data: KEGG BRITE database, redundancy removed

Enzymes	**GPCRs**	**Ion channels**
-675 targets	-100 targets	-114 targets
-524 molecules	-219 molecules	-462 molecules
-1218 interactions	-399 interactions	-1165 interactions
-1218 negatives	-399 negatives	-1165 negatives







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

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

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

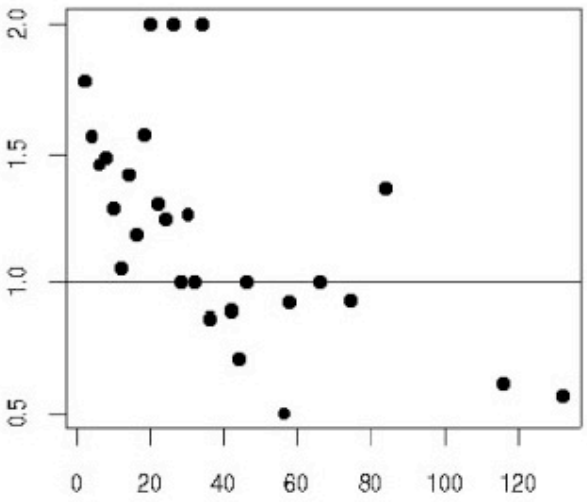
Orphan setting

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

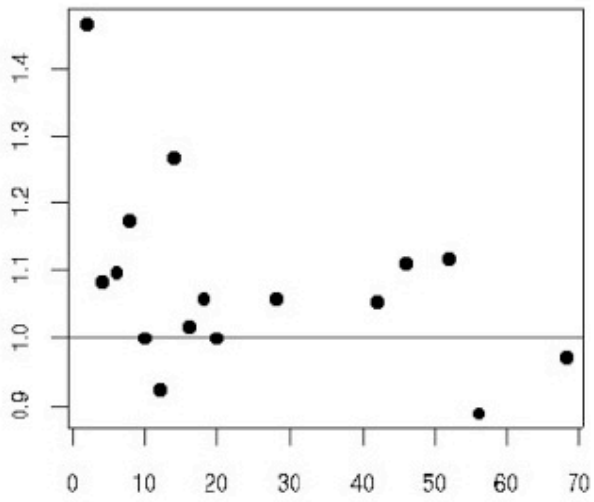
(Jacob and V., *Bioinformatics*, 2008)

Influence of the number of known ligands

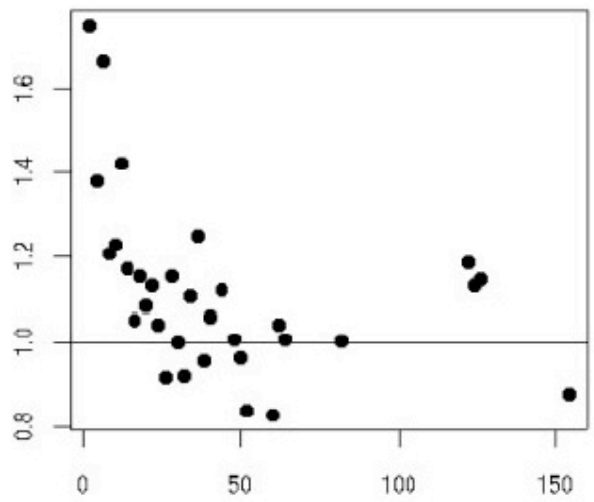
Enzymes



GPCRs



Ion channels



Relative improvement : hierarchy vs Dirac

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

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech

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
- Many more methods for « collaborative filtering » are being actively developed!

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

- P. Mahé and J.-P. Vert, "Graph kernels based on tree patterns for molecules", *Machine Learning*, 2009.
- 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", *CCHTS*, 11(8):677-685, 2008.
- P. Mahé, L. Ralaivola, V. Stoven and J.-P. Vert, "The pharmacophore kernel for virtual screening with support vector machines", *JCIM*, 46(5):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", *JCIM*, 45(4):939 -951, 2005.
- H. Kashima, K. Tsuda and A. Inokuchi, A., « Marginalized kernels between labeled graphs}. Proceedings of the 20th ICML, pp. 321-328, 2003.
- T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: hardness results and efficient alternatives. Proceedings of COLT, p.129--143, Springer, 2003.
- J. Ramon and T. Gärtner. Expressivity versus Efficiency of Graph Kernels. First International Workshop on Mining Graphs, Trees and Sequences, 2003.

Acknowledgements

Collaborators:
P. Mahé, L. Jacob, V. Stoven, B. Hoffmann

This presentation is supported by a JSPS Invitation Fellowship Program for Research in Japan, hosted by Tatsuya Akutsu (Kyoto University)

institutCurie
Ensemble, prenons le cancer de vitesse.

Inserm
Institut national de la santé et de la recherche médicale

MINES ParisTech