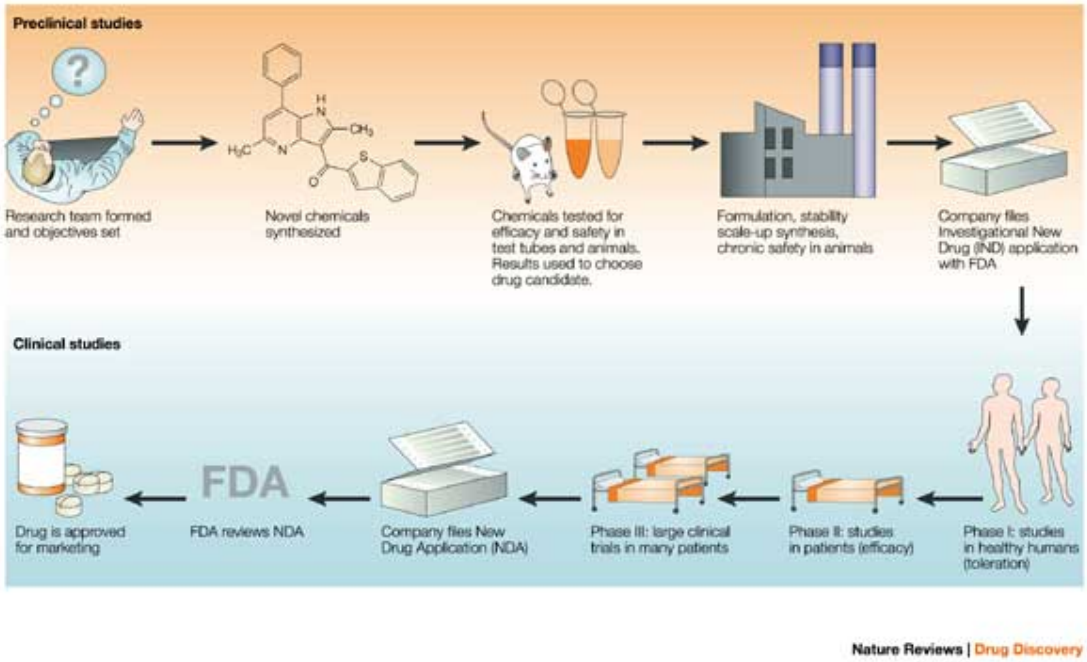
In silico virtual screening for drug discovery

Jean-Philippe Vert
Jean-Philippe.Vert@mines-paristech.fr

Mines ParisTech - INSERM - Institut Curie

"Mathematics and Industry" workshop, IHES, Bures-sur-Yvette, France, April 23-25, 2009

Drug discovery



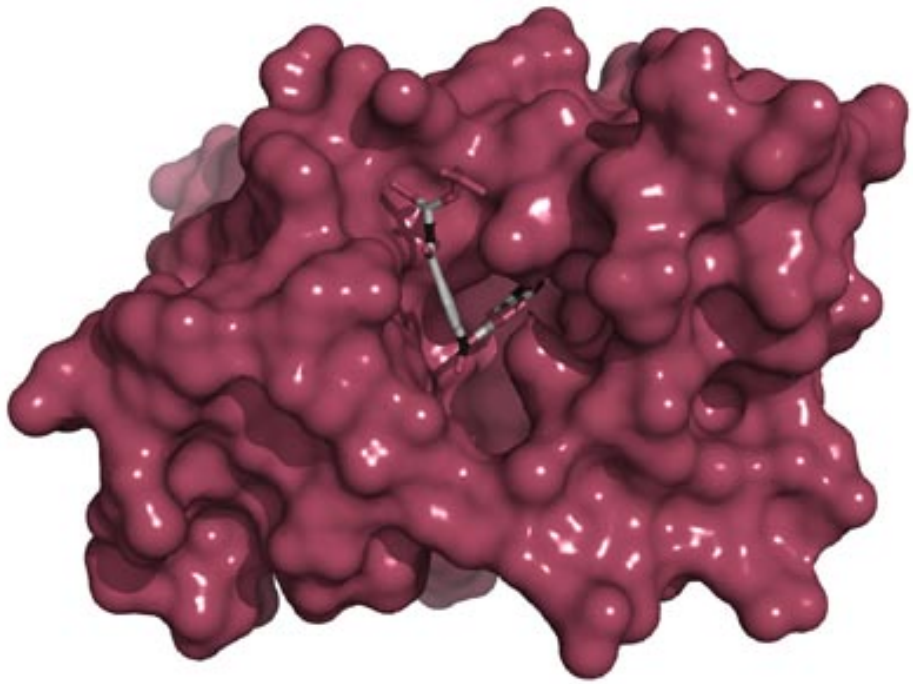
A long, expensive and risky process

- On average 15 years and $800 millions
- High attrition rate: for 10,000 molecules tested, 10 make it to clinicals, 1 to the market.
- >70% of the costs are wasted on failures

Computational approaches

The use of computers and computational methods permeates all aspects of drug discovery today, in particular for:

- Target identification
- Structure prediction, virtual screening (docking)
- Prediction of drug-likeliness of compounds



Collaboration with the industry

Practically

- Direct conctracting (leveraged through Institut Carnot M.I.N.E.S.)
- European projects
- Pôles de compétitivité

Features

- Large pharmaceutical or small biotech companies
- Shared PI or subcontracting
- Cultural shock math/info vs bio/chemistry

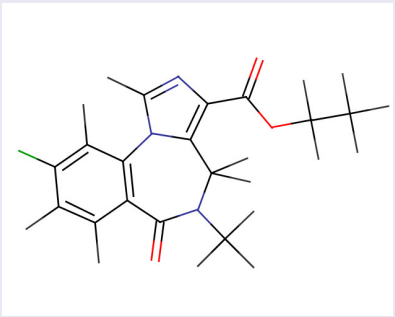
Virtual screening and predictive models

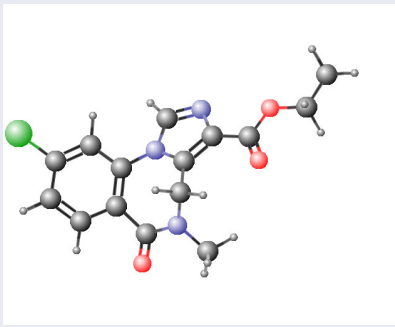
Objective

Build models to predict biochemical properties Y of small molecules from their structures X, using a training set of (X, Y) pairs.

Structures X

$C_{15}H_{14}ClN_3O_3$

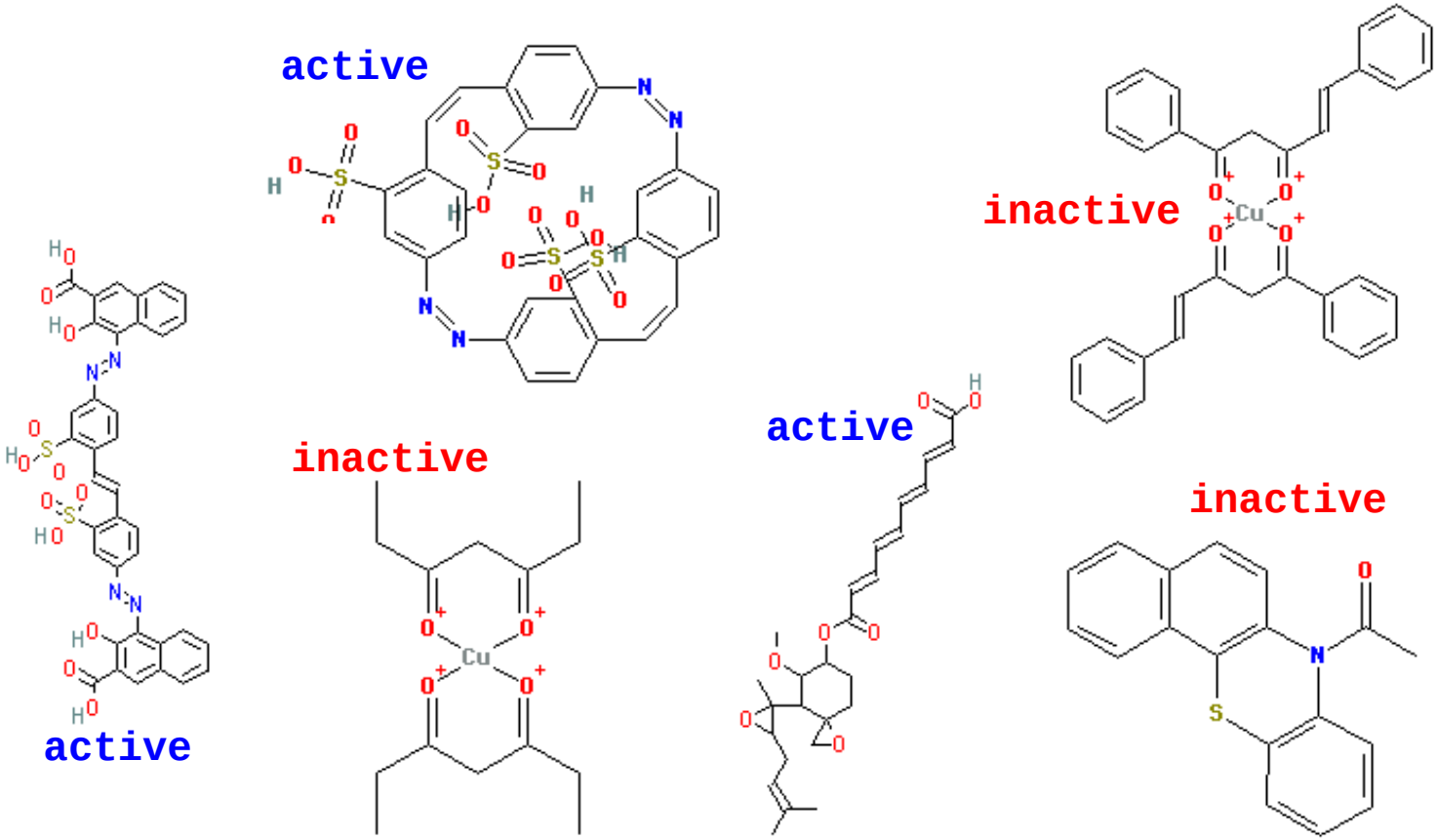




Properties Y

- binding to a therapeutic target,
- pharmacokinetics (ADME),
- toxicity...

Example : ligand-Based Virtual Screening



NCI AIDS screen results (from http://cactus.nci.nih.gov).

Formalization

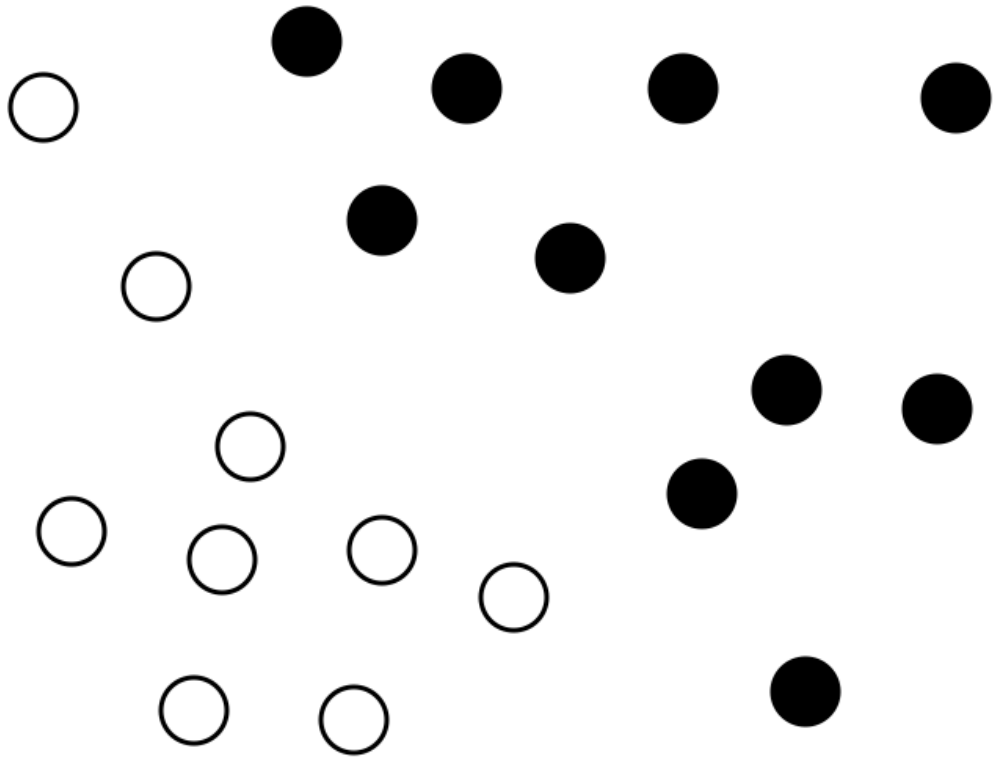
The problem

- Given a set of training instances $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i's are graphs and y_i's are continuous or discrete variables of interest,
- Estimate a function

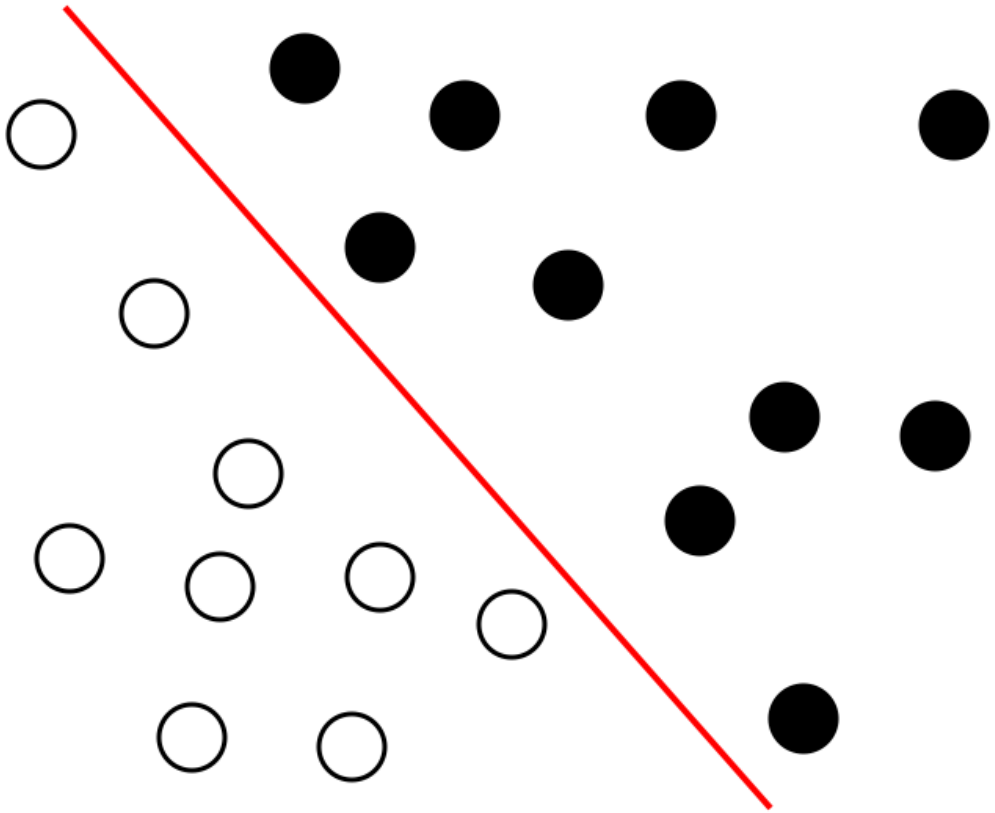
$$y = f(x)$$

  where x is any graph to be labeled.
- This is a classical regression or pattern recognition problem over the set of graphs.

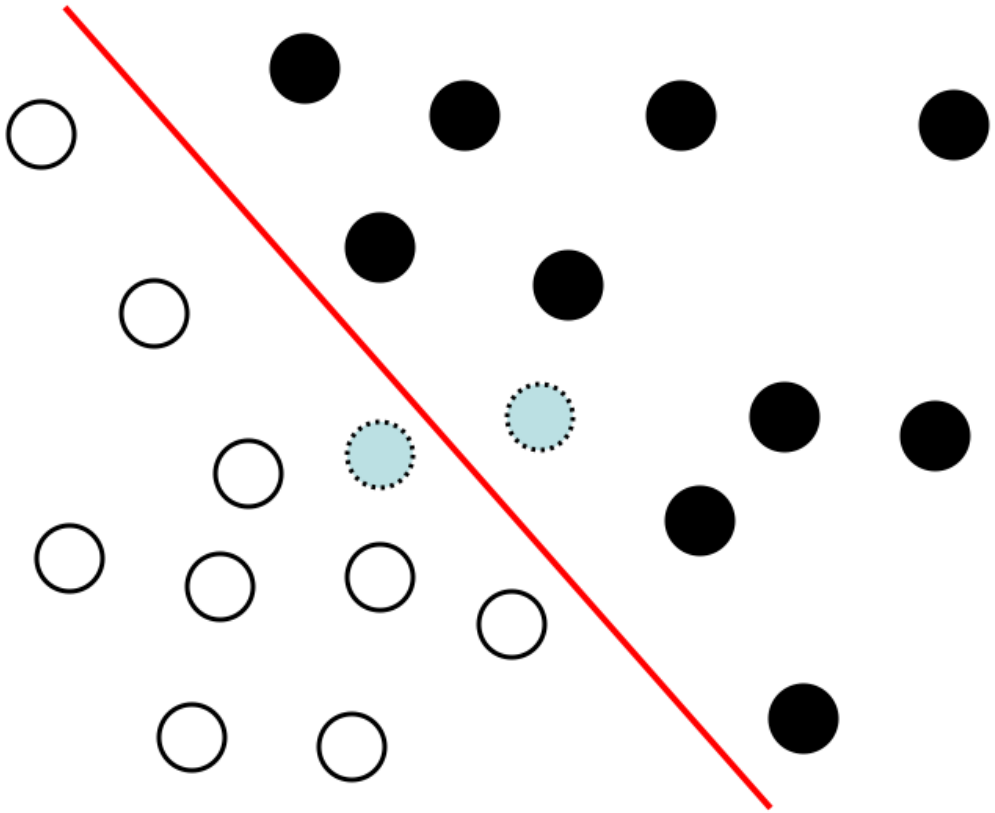
Pattern recognition



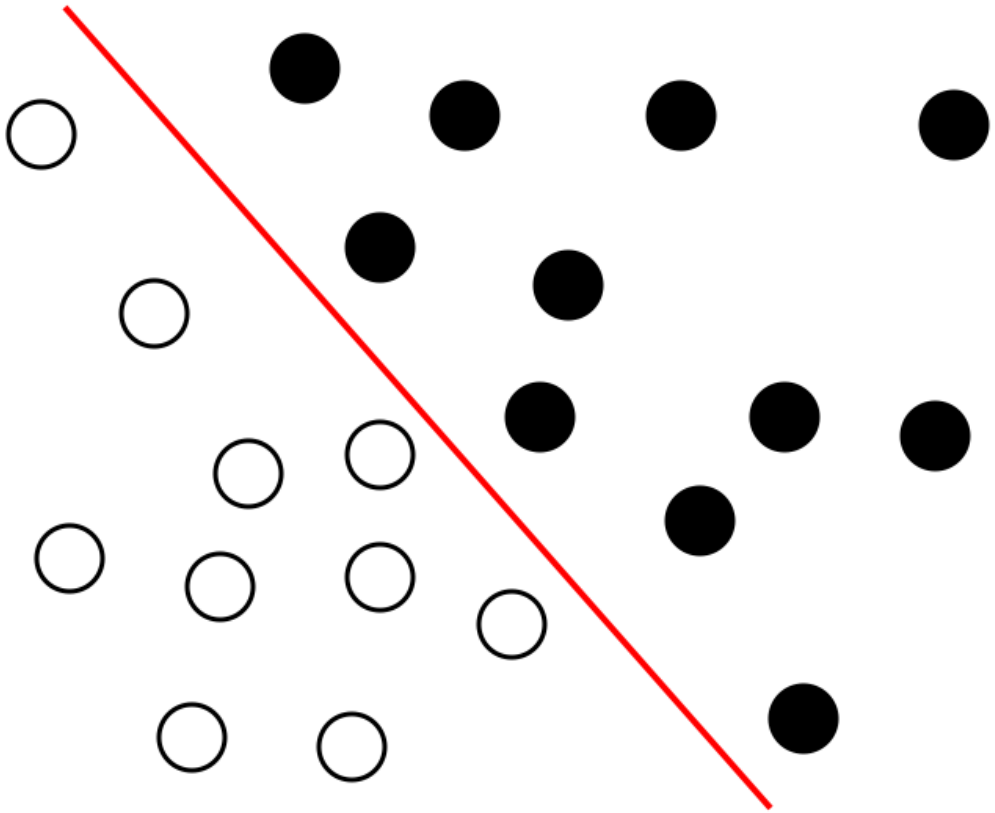
Pattern recognition



Pattern recognition



Pattern recognition

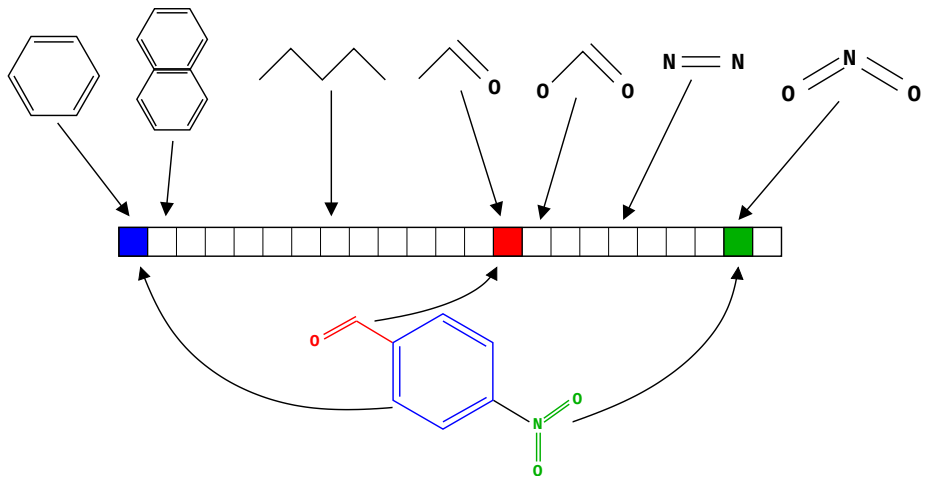


Classical approaches

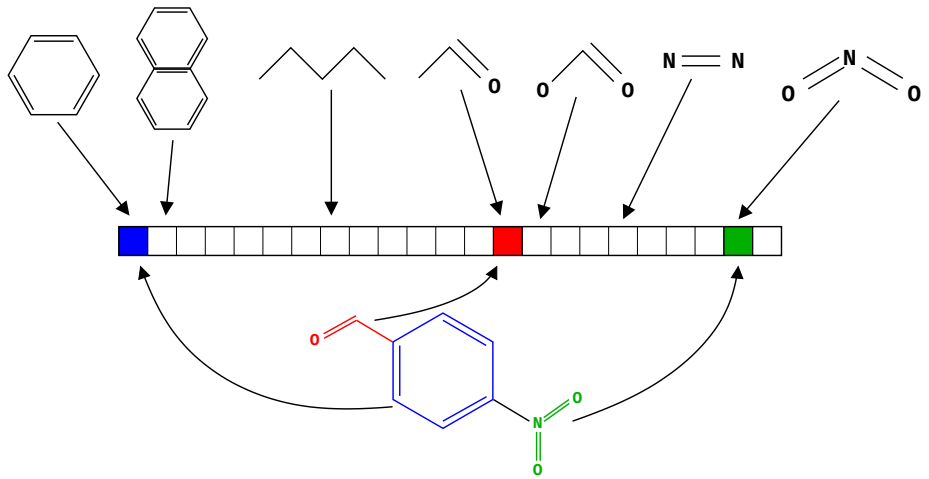
Two steps

1. Map each molecule to a vector of fixed dimension using molecular descriptors
   - Global properties of the molecules (mass, logP...)
   - 2D and 3D descriptors (substructures, fragments, ...)
2. Apply an algorithm for regression or pattern recognition.
   - PLS, ANN, ...

Example: 2D structural keys



Which descriptors?



Difficulties

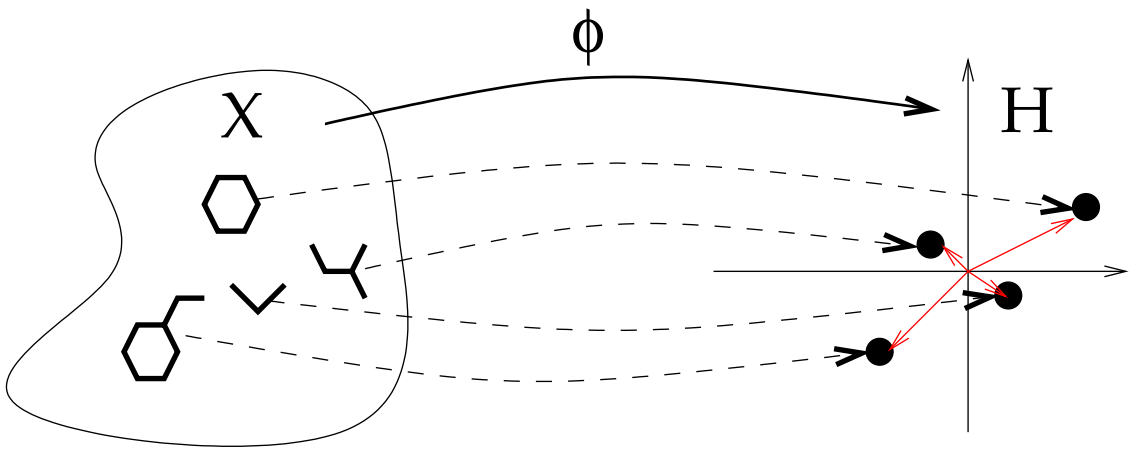
- Many descriptors are needed to characterize various features (in particular for 2D and 3D descriptors)
- But too many descriptors are harmful for memory storage, computation speed, statistical estimation

Kernels

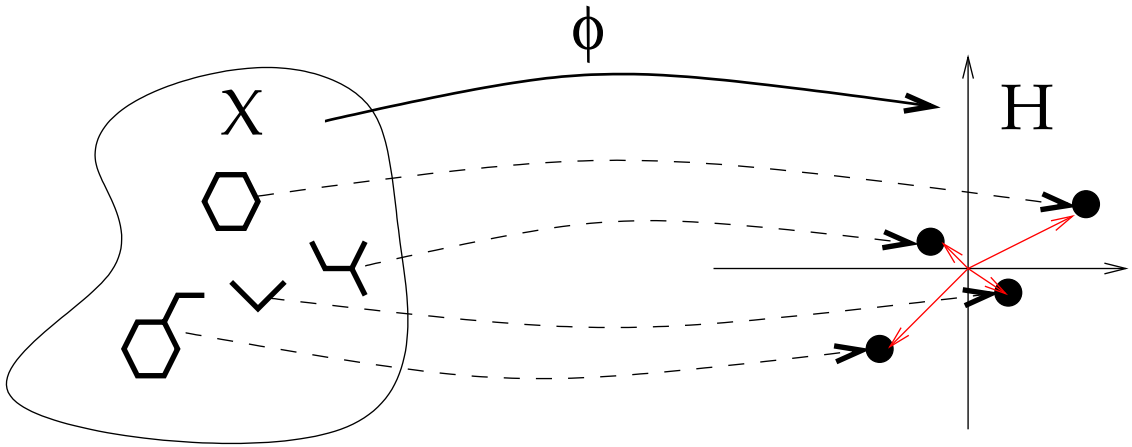
Definition

- Let $\Phi(x) = (\Phi_1(x), \ldots, \Phi_p(x))$ be a vector representation of the molecule x
- The kernel between two molecules is defined by:

$$K(x, x') = \Phi(x)^\top \Phi(x') = \sum_{i=1}^{p} \Phi_i(x) \Phi_i(x') \,.$$



The kernel trick



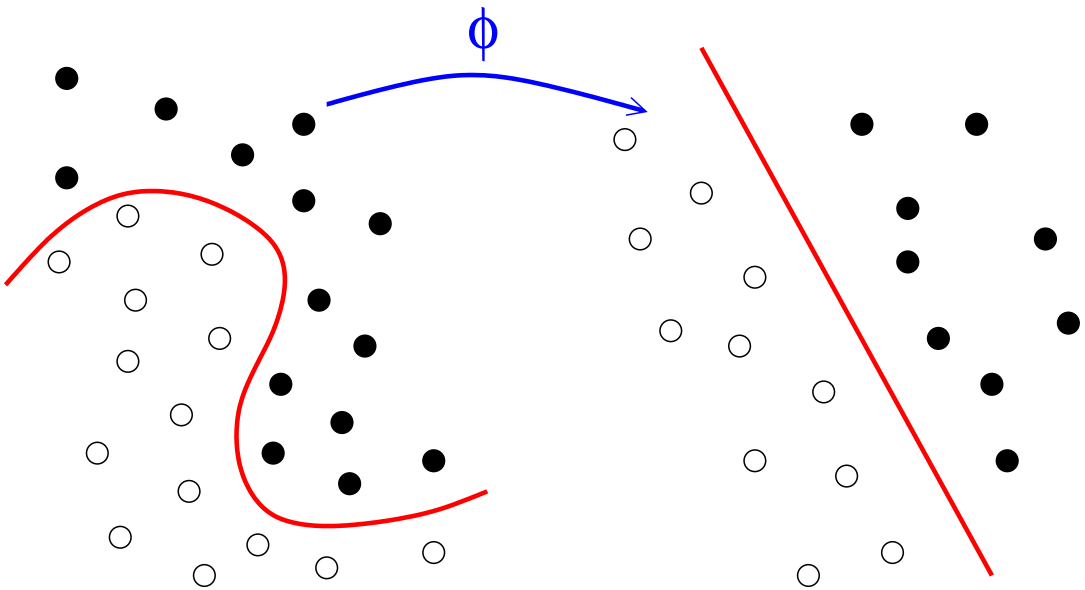
The trick

1. Any positive definite function is a valid kernel (i.e., inner product after mapping the molecules to some Hilbert space), e.g.,

$$K(x, x') = \exp\left(-\gamma \| x - x' \|^2\right) .$$

2. Many linear algorithms for regression or pattern recognition can be expressed only in terms of kernels.

Example: Support Vector Machine



$$\text{minimize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n\,,$$

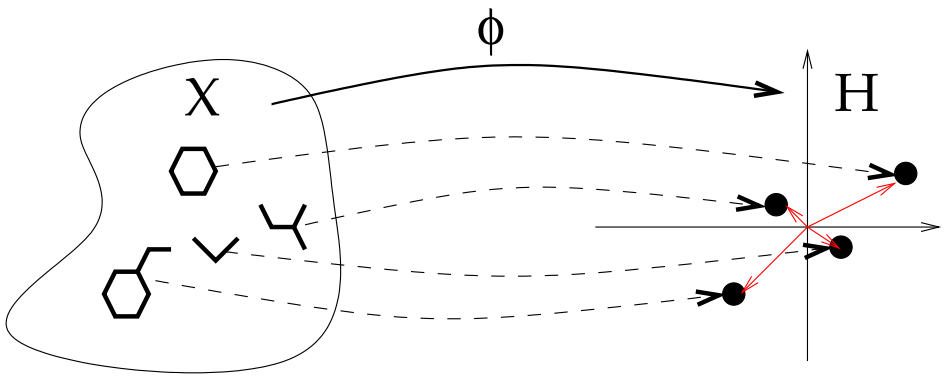
$$\sum_{i=1}^{n} \alpha_i y_i = 0\,.$$

Making kernels for molecules

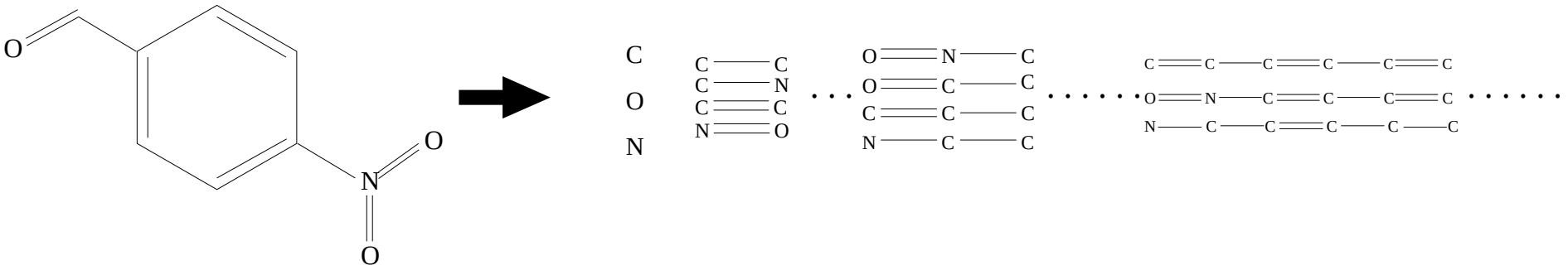
- Strategy 1: use well-known molecular descriptors to represent molecules m as vectors $\Phi(m)$, and then use kernels for vectors, e.g.:

$$K(m_1, m_2) = \Phi(m_1)^\top \Phi(m_2).$$

- Strategy 2: invent new kernels to do things you can not do with strategy 1, such as using an infinite number of descriptors. We will now see two examples of this strategy, extending 2D and 3D molecular descriptors.



Example: 2D fragment kernel



- $\phi_d(x)$ is the vector of counts of all fragments of length d:

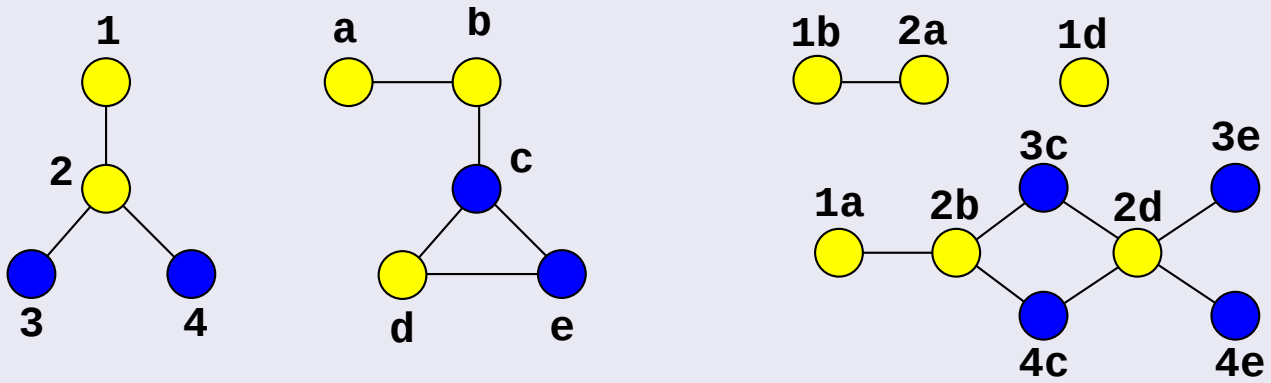
$$\phi_1(x) = (\quad \#(\mathtt{C}), \#(\mathtt{O}), \#(\mathtt{N}), \quad \ldots)^\top$$
$$\phi_2(x) = (\quad \#(\mathtt{C\text{-}C}), \#(\mathtt{C{=}O}), \#(\mathtt{C\text{-}N}), \quad \ldots)^\top \quad \text{etc...}$$

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

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

2D kernel computation trick

- $K_{fragment}$ can be computed efficiently for various weights $r(\lambda)$ although the feature space has infinite dimension.
- Rephrase the kernel computation as that as counting the number of walks on a graph (the product graph)



G1 **G2** **G1 x G2**

- The infinite counting can be factorized

$$\lambda A + \lambda^2 A^2 + \lambda^3 A^3 + \ldots = (I - \lambda A)^{-1} - I\,.$$

Experiments

MUTAG dataset

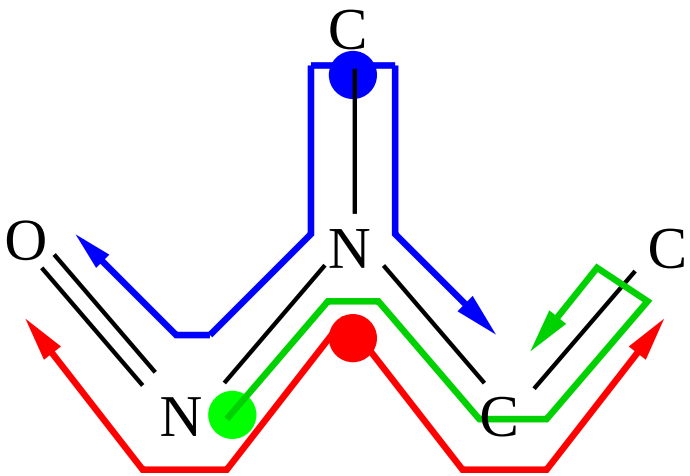
- aromatic/hetero-aromatic compounds
- high mutagenic activity /no mutagenic activity, assayed in *Salmonella typhimurium*.
- 188 compouunds: 125 + / 63 -

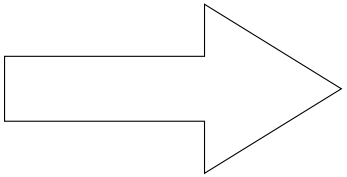
Results

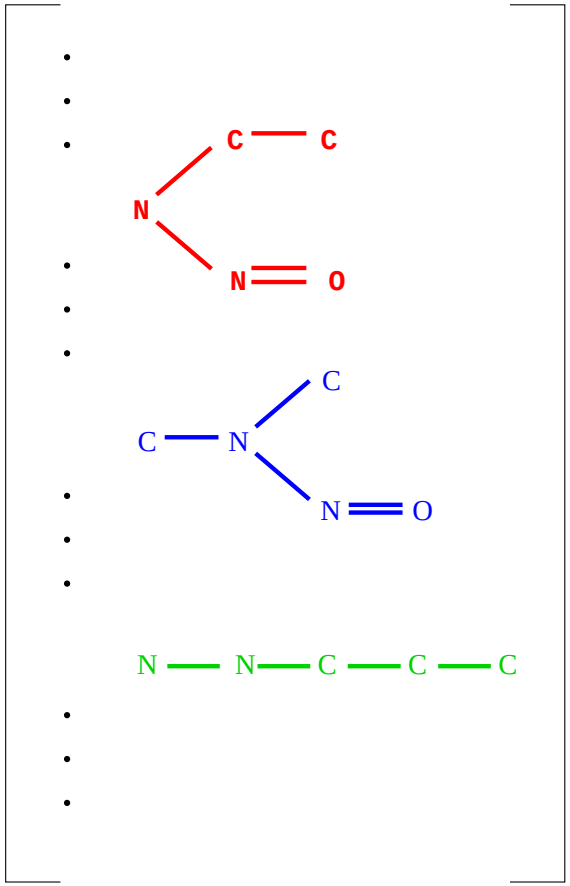
10-fold cross-validation accuracy

Method	Accuracy
Progol1	81.4%
2D kernel	91.2%

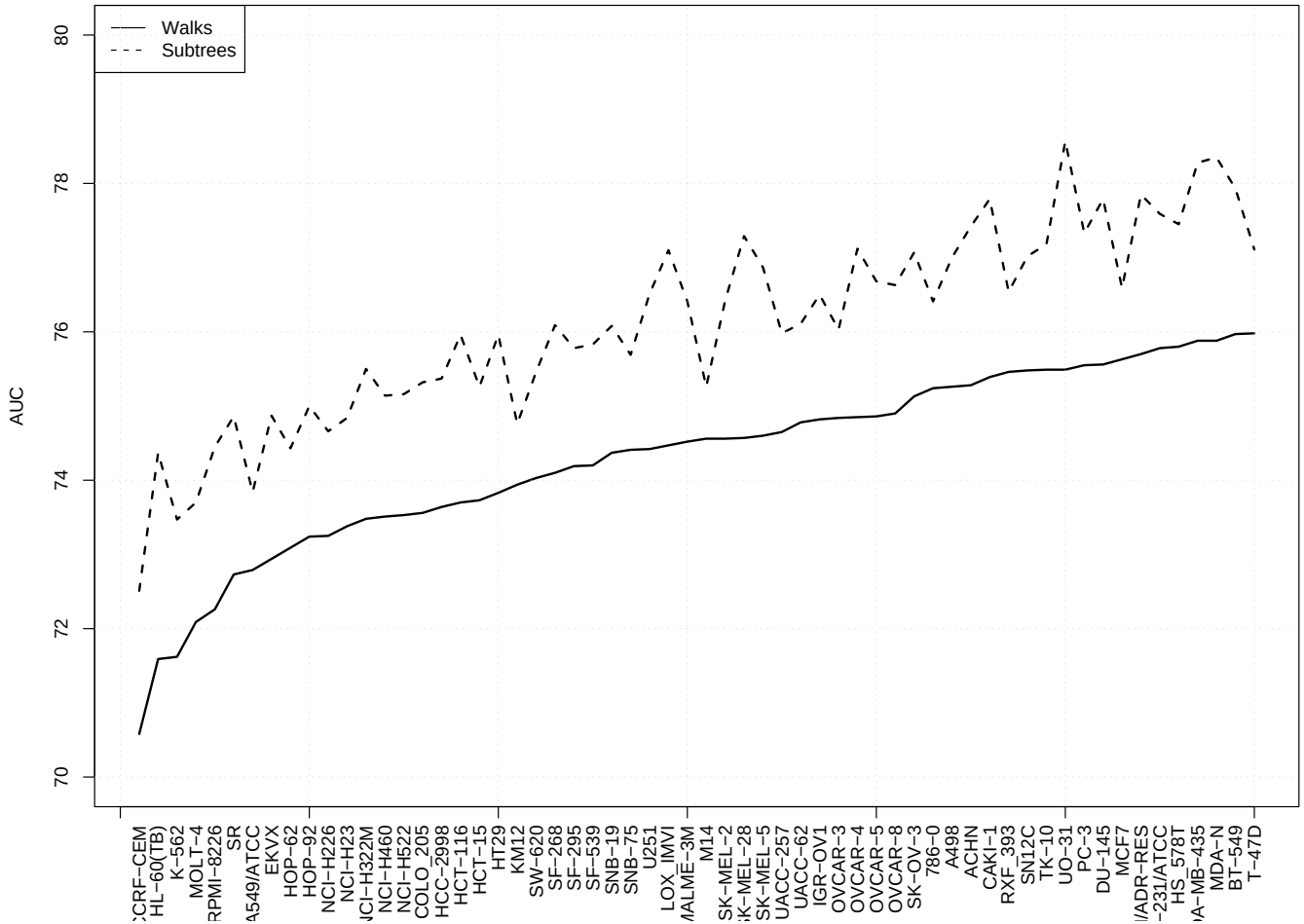
Example: 2D subtree kernel





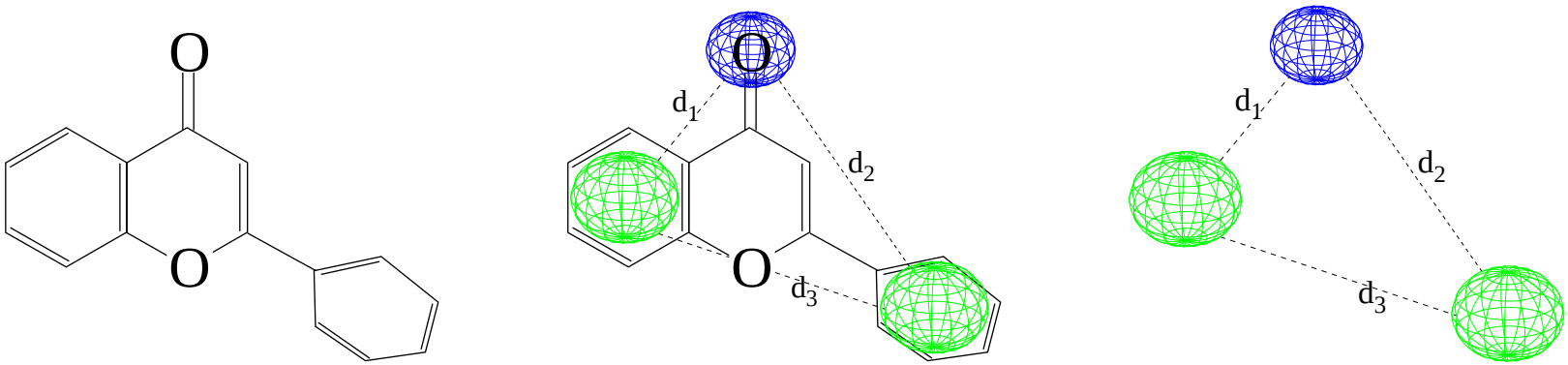


2D Subtree vs fragment kernels (Mahé and V, 2007)



Screening of inhibitors for 60 cancer cell lines (from Mahé and V., 2008)

Example: 3D pharmacophore kernel (Mahé et al., 2005)



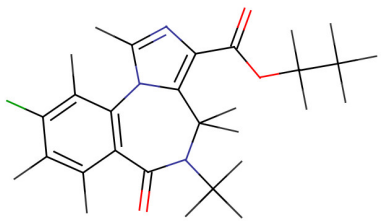
$$K(x,y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \exp\left(-\gamma d\left(p_x, p_y\right)\right) \ .$$

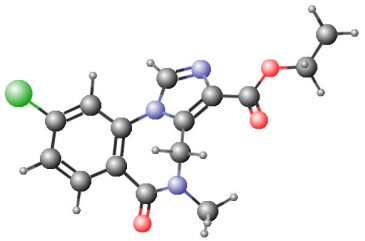
Results (accuracy)

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**

Challenges in ligand-based virtual screening

$C_{15}H_{14}ClN_3O_3$



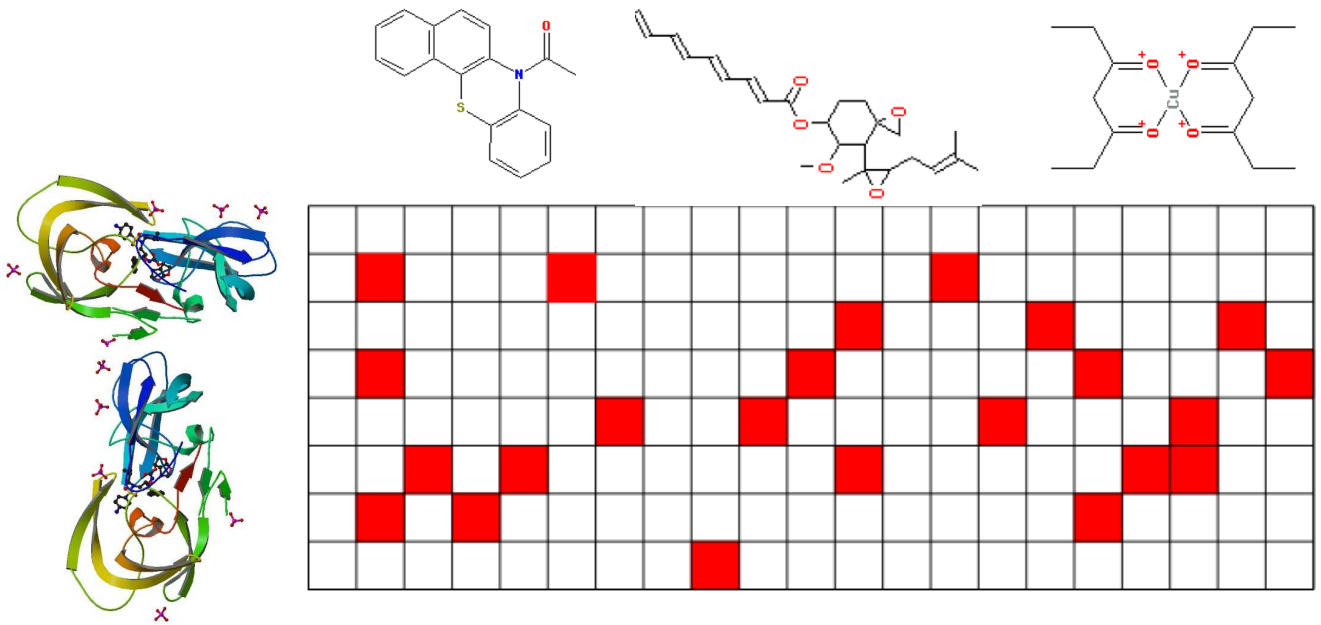


- How to embed molecules in a Hilbert space through a positive definite function $K(x, x')$ in a computationally efficient way (e.g., complete graph kernels are at least as hard as the graph isomorphism problem)?
- How to embed flexible 3D structures?
- How to extend machine learning algorithms to non-positive definite functions (e.g., edit distance between graphs)?

Towards chemogenomics

The problem

- Similar targets bind similar ligands
- Instead of focusing on each target individually, can we screen the biological space (target families) vs the chemical space (ligands)?
- Mathematically, learn $f(target, ligand) \in \{bind, notbind\}$: this is an instance of operator inference.



Conclusion

- Computational approaches have the potential to increase the global R&D productivity in pharmaceutical industry, in particular to speed up the identification of targets, hits, and decrease the attrition rate.
- Collaboration with the industry is needed to access large databases of molecules and impact drug discovery.
- Statistics and machine learning are natural approaches for ligand-based virtual screening approaches
- Molecules are complex objects, which can be represented as formula, structures, shapes. Applying state-of-the-art machine learning methods to these representations requires mathematical and computational developments.