

Instructor: Sael Lee

CS549 Spring – Computational Biology

CLASSIFICATION OF SMALL MOLECULES BY TWO- AND THREE-DIMENSIONAL DECOMPOSITION KERNELS

Ceroni, A., Costa, F., & Frasconi, P. (2007). *Bioinformatics*, 23(16), 2038–45.

Structural bioinformatics

Classification of small molecules by two- and three-dimensional decomposition kernels

Alessio Ceroni, Fabrizio Costa* and Paolo Frasconi

Machine Learning and Neural Networks Group, Dipartimento di Sistemi e Informatica,
Università degli Studi di Firenze, Italy

Received on February 19, 2007; revised on May 14, 2007; accepted on May 28, 2007

Advance Access publication June 5, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Several kernel-based methods have been recently introduced for the classification of small molecules. Most available kernels on molecules are based on 2D representations obtained from chemical structures, but far less work has focused so far on the definition of effective kernels that can also **exploit 3D information**.

Results: We introduce new ideas for building kernels on small molecules that can effectively use **and combine 2D and 3D information**. We tested these kernels in conjunction with support vector machines for binary classification on the 60 NCI cancer screening datasets as well as on the NCI HIV data set. Our results show that 3D information leveraged by these kernels can consistently improve prediction accuracy in all datasets.

Availability: An implementation of the small molecule classifier is available from <http://www.dsi.unifi.it/neural/src/3DDK>

METHODS: BACKGROUND ON KERNEL METHODS FOR STRUCTURED DATA

A major challenge is to define an effective quantitative measure of similarity

Base Learner: **SVM**

classification function $f(x)$ is obtained from data

$$f(x) = \sum_{i=1}^m \alpha_i y_i K(x_i, x).$$

Kernels on structured data

$x \in \mathcal{X}$, suppose (x_1, \dots, x_D)

$$K(x, x') = \sum_{\substack{(x_1, \dots, x_D) \in R^{-1}(x) \\ (x'_1, \dots, x'_D) \in R^{-1}(x')}} \prod_{d=1}^D \kappa_d(x_d, x'_d)$$

where $R^{-1}(x) = \{(x_1, \dots, x_D) : R(x_1, \dots, x_D, x)\}$ denote the set of all possible decompositions of x .

A WEIGHTED DECOMPOSITION KERNEL (WDK) FOR 2D CHEMICAL STRUCTURES

The basic idea behind WDK is that each substructure in which a graph is decomposed is **enriched with its graphical context**.

characterized by a **decomposition** $R(s,z,x)$ where s is a subgraph of x called the **selector** and z is a subgraph of x called the **context** of occurrence of s in x (generally a subgraph containing s).

This setting results in the following general form of the kernel:

$$K_{2D}(x, x') = \sum_{\substack{(s,z) \in R^{-1}(x) \\ (s',z') \in R^{-1}(x')}} \delta(s, s') \kappa(z, z')$$

where, δ is the **exact matching kernel** applied to selectors and κ is a kernel on contexts.

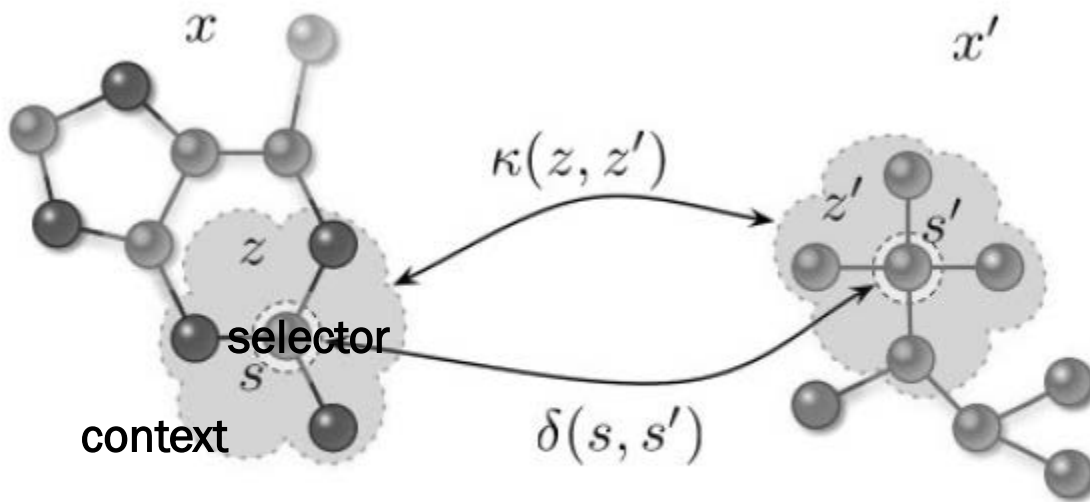


Fig. 1. Comparing substructures in a weighed decomposition kernel.

$$K_{2D}(x, x') = \sum_{\substack{(s, z) \in R^{-1}(x) \\ (s', z') \in R^{-1}(x')}} \delta(s, s') \kappa(z, z')$$

WDK PARAMETER USED IN THE ARTICLE

IDEA

- **Selectors** are always single atoms and the match
- $\delta(s, s')$ is defined by the coincidence between the type of s and s' .
- The **context kernel** κ is based on soft match between substructures, defined by the distributions of label contents after discarding topology.

SPECIFICS

1. Let L denote the total number of attributes labeling vertices and edges and for $l = 1, \dots, L$
2. Let $p_l(j)$ be the observed frequency of value j for the l – *th* attribute in a substructure z .
3. Then compare substructures by means of a **histogram intersection kernel**

$$\kappa_\ell(z, z') = \sum_{j=1}^{m_\ell} \min\{p_\ell(j), p'_\ell(j)\}$$
$$\kappa(z, z') = \sum_{\ell=1}^L \kappa_\ell(z, z').$$

Where m_l is the number of possible values of attribute l .

shall use $L = 3$: 1) **atom type**, 2) **atom charge** and 3) **bond type**, while atom coordinates are discarded for computing the WDK.

CONTEXTS ARE FORMED AS FOLLOWS

Given a vertex $v \in V$ and an integer $r \geq 0$,

- Let $x(v, r)$: substructure of x obtained by retaining all the vertices that are reachable from v by a path of length at most r , and all the edges that touch at least one of these vertices.

The **decomposition relation** R_r , dependent on the context radius r , is then defined as

$$R_r = \{(s, z, x) : x \in \mathcal{X}, s = \{v\}, z = x(v, r), v \in V\},$$

where s is the selector and z is the context for vertex v .

Weighted Decomposition Kernel (WDK)

$$K_{2D}(x, x') = \sum_{\substack{(s, z) \in R^{-1}(x) \\ (s', z') \in R^{-1}(x')}} \delta(s, s') \kappa(z, z')$$

THREE-DIMENSIONAL DECOMPOSITION KERNELS

A 3D molecular structure is interpreted here as a special kind of relational data object where atoms are related by chemical bonds but also by their spatial distances

The molecule is first decomposed into a set of overlapping 3D substructures of varied geometry, called **shapes**.

Given a molecule $x = (V, E)$, a **shape** of order n is a set of n distinct vertices

$$\sigma = \{u_1, u_2, \dots, u_i\}, \quad u_i \in V, \text{ for } i = 1, \dots, n.$$

kernel between two molecules

$$K_{3D}(x, x') = \sum_{\sigma \in \mathcal{S}_r(x)} \sum_{\sigma' \in \mathcal{S}_r(x')} k_{\text{shapes}}(\sigma, \sigma')$$

kernels between all pairs of shapes

$$k_{\text{shapes}}(\sigma, \sigma') = \prod_{i=1}^{n(n-1)/2} \delta(e_i, e'_i) e^{-\gamma(d_i - d'_i)^2}$$

KERNELS BETWEEN ALL PAIRS OF SHAPES

Given a **shape** of order n and two vertices $u, v \in \sigma$,

- let $e = (t[u], t[v], b[u, v])$ denote a **labeled edge** of the shape, formed by considering the two atom types $t[u]$ and $t[v]$ and the bond type $b[u, v]$.

Then, let $\langle e_1, \dots, e_{n(n-1)/2} \rangle$ denote the **lexicographically ordered sequence** of all labeled edges in .

For example, the shape (C1,C2,C3,O1)
for the molecule NSC_1027 yields
**lexicographically ordered sequence of all
labeled edges** (C.2,C.3,1) (C.2,C.3,1)
(C.2,O.2,2) (C.3,C.3,0) (C.3,O.2,0)
(C.3,O.2,0).

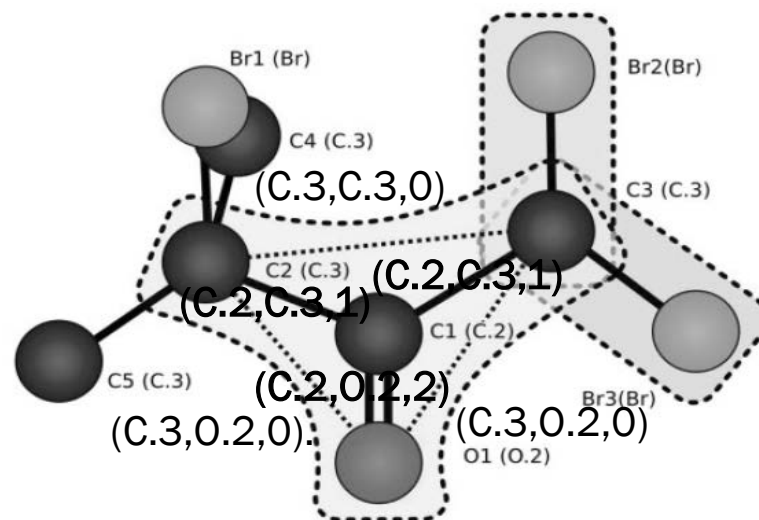


Fig. 2. Illustration of the definition of 2D-supported shapes. The three 2D-supported shapes of radius 1, anchored to atom C3 in the molecule NSC_1027 are (Br3,C3), (Br2,C3) and (C1,C2,C3,O1). Atom identifiers and types (in parentheses) are formatted according to the Tripos Sybyl MOL2 conventions.

KERNELS BETWEEN A PAIR OF SHAPES

The kernel between two shapes σ and σ' of equal order n is defined as:

$$k_{\text{shapes}}(\sigma, \sigma') = \prod_{i=1}^{n(n-1)/2} \delta(e_i, e'_i) e^{-\gamma(d_i - d'_i)^2}$$

Where γ is a kernel hyperparameter and $d_i = \|\vec{\zeta}[u_i] - \vec{\zeta}[v_i]\|$ is the length of edge e_i , i.e. the Euclidean distance between atoms u_i and v_i .

* The kernel between two shapes of different order is null.

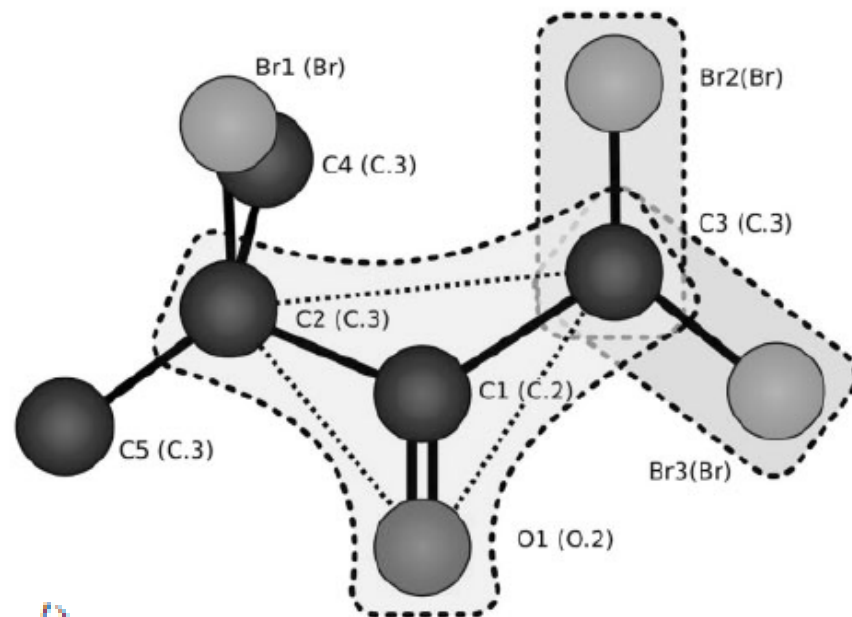
$$\sigma: \langle e_1, \dots, e_{n(n-1)/2} \rangle$$
$$\sigma': \langle e'_1, \dots, e'_{n(n-1)/2} \rangle$$

KERNELS BETWEEN ALL PAIRS OF SHAPES CONT.

3D decomposition kernels (3DDK)

Select just the adjacent list of vertices that are **within distance r** from x .

Given a vertex $v \in V$ and an integer r , a **2D-supported shape** anchored in v is a set of vertices $\sigma = \{v, w\} \cup \text{adj}[w]$ such that $w \in x(v, r)$ and $\text{adj}[w]$ is the adjacency list of w . Let $S_r(x)$ denote **shape set of radius r** of x .



$$K_{3D}(x, x') = \sum_{\sigma \in S_r(x)} \sum_{\sigma' \in S_r(x')} k_{\text{shapes}}(\sigma, \sigma')$$

DATA SET: &

NCI cancer dataset

National Cancer Institute public dataset of screening results for the ability of more than 70,000 compounds to suppress or inhibit the growth of a panel of 60 human tumor cell lines.

Subset of NCI dataset corresponding to the parameter GI50, the concentration that causes 50% growth inhibition is used.

Binary classification: cancer-inhibiting (1) or not (-1).

NCI HIV dataset

Contains 42,687 compounds evaluated for evidence of anti HIV activity from the DTP AIDS antiviral screen program of the National Cancer Institute.

Compounds are divided in **three classes:** **1)** 422 compounds are confirmed active (CA), **2)** 1081 are moderately active (CM) and **3)** 41 184 are confirmed inactive (CI).

THREE CLASS CLASSIFICATION WITH SVM

Three classification problems are formulated on this dataset:

1. (CA verses CM): positive examples are confirmed active compounds, while moderately active compounds forms the negative class;
2. (CA+CM verses CI): the positive class is formed by the combination of moderately active and confirmed active compounds and in
3. (CA verses CI): the positive examples are confirmed active compounds and the negative class is formed by confirmed inactive compounds.

COMBINING KERNELS- WDK & 3DDK

NCI cancer dataset

The WDK and 3DDK used in this experiment had both the radius $r = 3$ and no graph complement was used for the WDK. γ parameter in pair-wise shape kernel was set to 2.5.

$$K(x, x') = (1 + \kappa(x, x'))^2$$

κ is either K_{2D} or K_{3D} or $\kappa(x, x') = K_{2D}(x, x') + K_{3D}(x, x')$.

These measures were estimated by a 10-folds cross-validation

NCI HIV dataset

For the WDK, graph complement and context radius $r = 4$ is used.

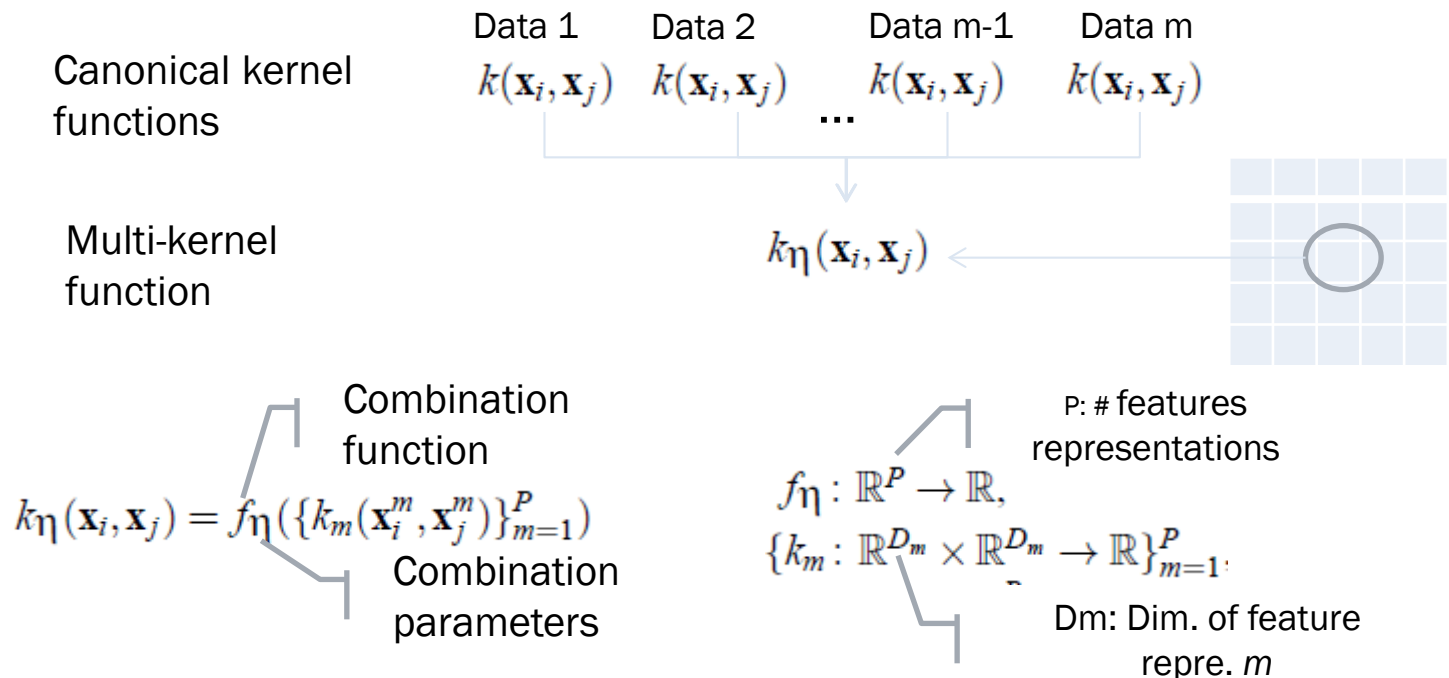
For the 3DDK, the radius to $r = 3$ is used. γ parameter in pair-wise shape kernel was set to 2.5.

$$K(x, x') = e^{-\frac{1}{2}(\kappa(x, x) + \kappa(x', x') - 2\kappa(x, x'))}$$

AUC performance was estimated by a 5-folds cross-validation

MULTIPLE KERNEL

Combine multiple canonical kernels to generate one “multiple kernel” to solve one classification problem.



Gonen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12, 2211–2268.

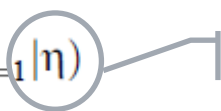
MULTIPLE KERNEL CONT.

Two usage of multiple kernels:

1. Using multiple kernels to evaluate data from one source
 - Type of kernel functions and parameters are important but non-trivial to select
2. Using multiple kernels to combine data from multiple sources and types

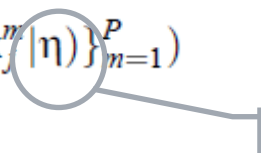
Two type of kernel combination parameter learning strategies:

1. Learn each kernel para. -> learn combination para.

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = f_{\eta}(\{k_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\}_{m=1}^P | \eta)$$


Part of combination process

2. Learn each kernel para. & combination para at all at once

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = f_{\eta}(\{k_m(\mathbf{x}_i^m, \mathbf{x}_j^m | \eta)\}_{m=1}^P)$$


Part of each canonical kernels

KEY PROPERTIES OF MULTIPLE KERNEL LEARNING

- × The Combination parameter learning method
 - + Fixed rules
 - + Heuristic
 - + Optimization
 - + Bayesian inference
 - + Iteratively adding in new kernels
- × The kernel combination functional form
 - + Linear
 - + Nonlinear
 - + Data-dependent – different weights for different data
- × The base learner
 - + Support vector machines(SVM)
 - + support vector regression (SVR)
 - + Kernel Fisher discriminant analysis (KFDA)
 - + Regularized kernel discriminant analysis (RKDA),
 - + kernel ridge regression (KRR)
 - + Multinomial probit and
 - + Gaussian process (GP)

MULTIPLE KERNEL LEARNING ALGORITHMS

Fixed Rules: functions without any parameters and do not need any training.

A valid kernel can be formed by taking the *summation* or *multiplication* of two valid kernels.

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i^1, \mathbf{x}_j^1) + k_2(\mathbf{x}_i^2, \mathbf{x}_j^2)$$

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i^1, \mathbf{x}_j^1)k_2(\mathbf{x}_i^2, \mathbf{x}_j^2).$$

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^P k_m(\mathbf{x}_i^m, \mathbf{x}_j^m)$$

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = \prod_{m=1}^P k_m(\mathbf{x}_i^m, \mathbf{x}_j^m).$$

Pairwise kernels are proposed to express the similarity between pairs in terms of similarities between individual objects.

Ex> *genomic kernel*

$$k^P(\{\mathbf{x}_i^a, \mathbf{x}_j^a\}, \{\mathbf{x}_i^b, \mathbf{x}_j^b\}) = k(\mathbf{x}_i^a, \mathbf{x}_i^b)k(\mathbf{x}_j^a, \mathbf{x}_j^b) + k(\mathbf{x}_i^a, \mathbf{x}_j^b)k(\mathbf{x}_j^a, \mathbf{x}_i^b).$$

Ex> (weighted) sum of different pairwise kernels

$$k_{\eta}^P(\{\mathbf{x}_i^a, \mathbf{x}_j^a\}, \{\mathbf{x}_i^b, \mathbf{x}_j^b\}) = \sum_{m=1}^P k_m^P(\{\mathbf{x}_i^a, \mathbf{x}_j^a\}, \{\mathbf{x}_i^b, \mathbf{x}_j^b\})$$

MULTIPLE KERNEL LEARNING ALGORITHMS

Heuristic Approaches: parameterized combination function and find the parameters of this function generally by heuristic measures on individual kernels.

Ex> weighted linear combination

$$k_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^P \eta_m k_m(\mathbf{x}_i^m, \mathbf{x}_j^m)$$

$$\eta_m = \frac{\pi_m - \delta}{\sum_{h=1}^P (\pi_h - \delta)}$$



Various **Heuristic weight functions** are possible

π_m is the accuracy obtained using only \mathbf{K}_m , and δ is the threshold that should be less than or equal to the minimum of the accuracies obtained from single-kernel learners

MULTIPLE KERNEL LEARNING ALGORITHMS

Optimization approaches: use a parameterized combination function and learn the parameters by solving an optimization problem.

EX> Lanckriet et al. (2004a) propose to optimize the kernel alignment as follows

$$\begin{aligned} & \text{maximize } A(\mathbf{K}_\eta^{\text{tra}}, \mathbf{y}\mathbf{y}^\top) \\ & \text{with respect to } \mathbf{K}_\eta \in \mathbb{S}^N \\ & \text{subject to } \text{tr}(\mathbf{K}_\eta) = 1 \\ & \quad \mathbf{K}_\eta \succeq 0 \end{aligned}$$

ideal kernel: similarity between two kernels.

$$A(\mathbf{K}, \mathbf{y}\mathbf{y}^\top) = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle_F}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle_F}} = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle_F}{N\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F}}$$

$$\text{where } \langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F = \sum_{i=1}^N \sum_{j=1}^N k_1(\mathbf{x}_i^1, \mathbf{x}_j^1) k_2(\mathbf{x}_i^2, \mathbf{x}_j^2).$$

MULTIPLE KERNEL LEARNING ALGORITHMS

Bayesian approaches: interpret the kernel combination parameters as random variables, put priors on these parameters, and perform inference for learning them and the base learner parameters.

Ex>

$$f(\mathbf{X}) = \sum_{i=0}^N \alpha_i \sum_{m=1}^P \eta_m k_m(\mathbf{X}_i^m, \mathbf{X}^m)$$

where η is modeled with a Dirichlet prior and α is modeled with a zero-mean Gaussian with an inverse gamma variance prior.

MULTIPLE KERNEL LEARNING ALGORITHMS

Boosting approaches: iteratively add a new kernel until the performance stops improving

Ex> Bennett et al. (2002)

$$f(\mathbf{x}) = \sum_{i=1}^N \sum_{m=1}^P \alpha_i^m k_m(\mathbf{x}_i^m, \mathbf{x}^m) + b.$$

parameters $\{\alpha_i^m\}_{m=1}^P$ and b of the KRR model are learned using gradient-descent

$$\text{ExpLoss}(k(\mathbf{x}_i, \mathbf{x}_j), y_i y_j) = \exp(-y_i y_j k(\mathbf{x}_i, \mathbf{x}_j))$$

$$\text{LogLoss}(k(\mathbf{x}_i, \mathbf{x}_j), y_i y_j) = \log(1 + \exp(-y_i y_j k(\mathbf{x}_i, \mathbf{x}_j))).$$

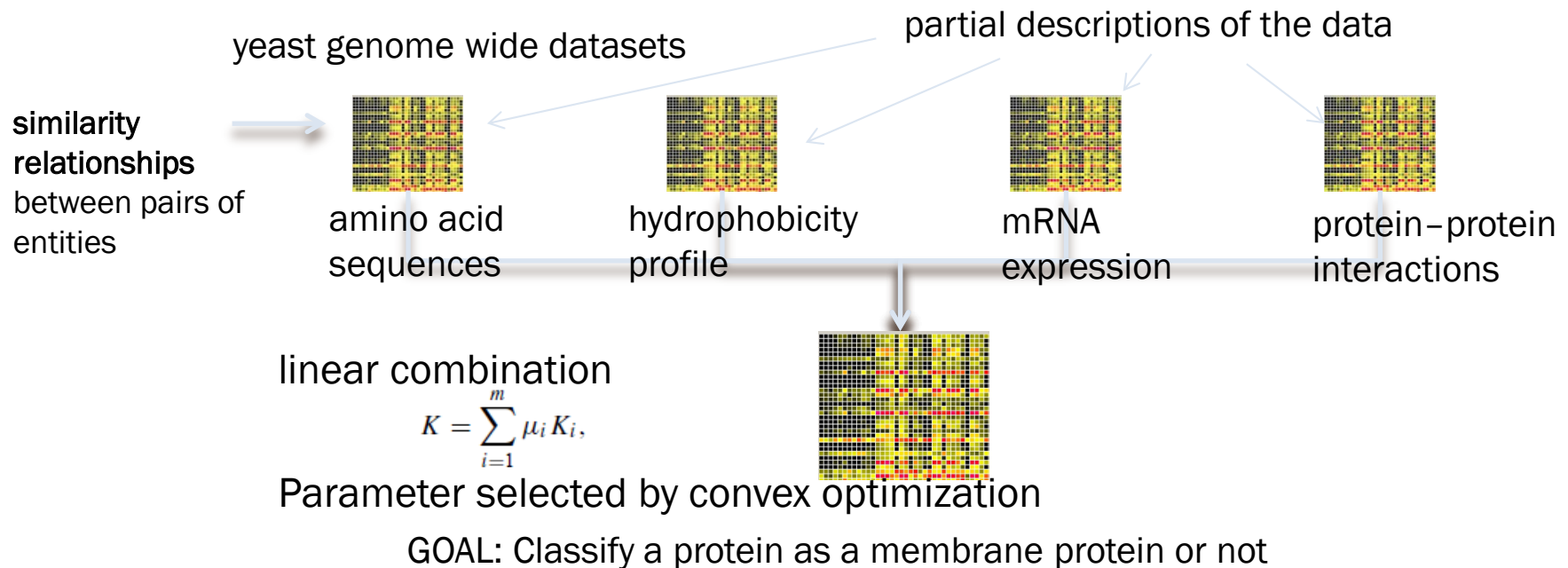
We iteratively update the combined kernel matrix using one of these two loss functions.

KRR: kernel ridge regression

MULTIPLE KERNEL: EXAMPLES

Computational framework for integrating and drawing inferences from a collection of genome-wide measurements

Each dataset is represented via a kernel function, which defines generalized similarity relationships between pairs of entities, such as genes or proteins.

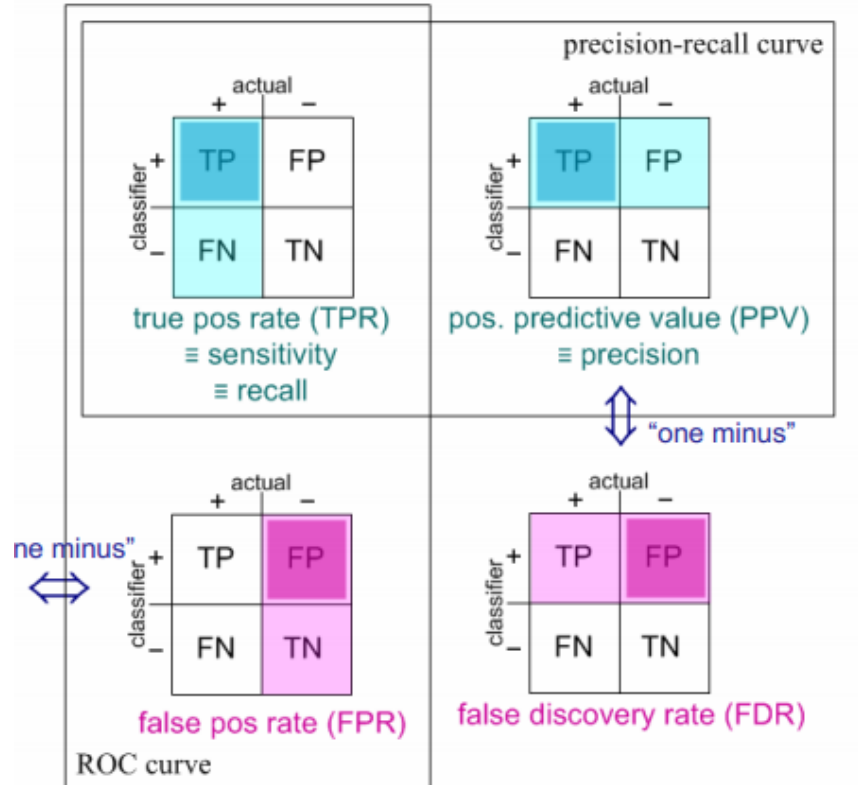


Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16), 2626–35.

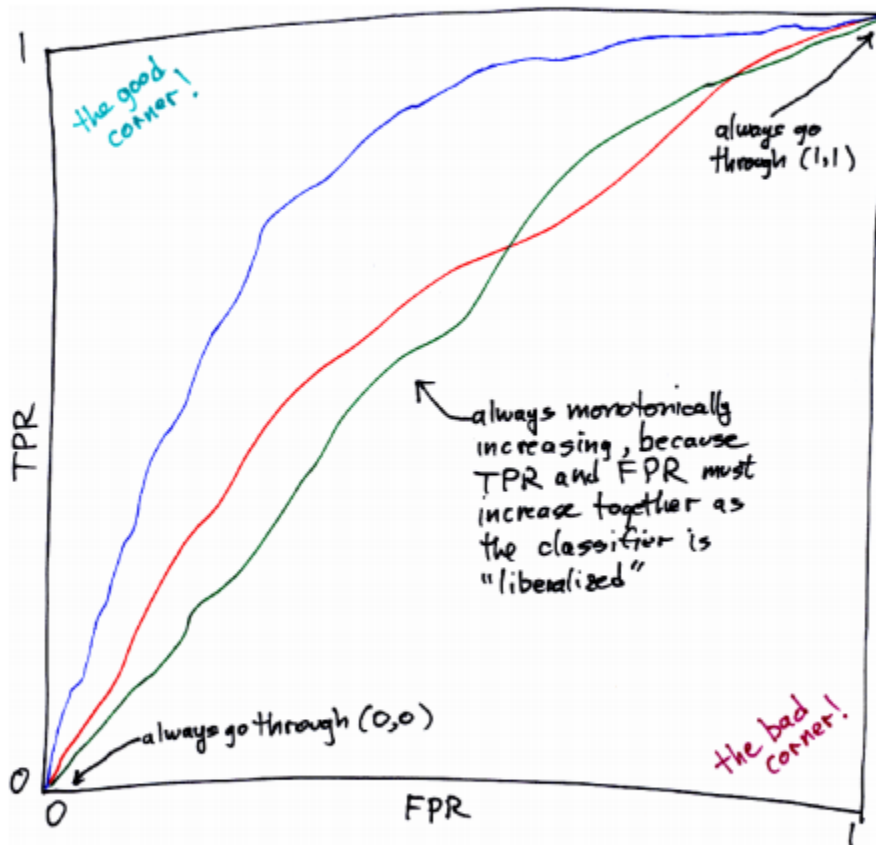
ROC VS PRECISION RECALL

		actual		
		+	-	
classifier	+	TP good!	FP bad! (Type I error)	
	-	FN bad! (Type II error)	TN good!	

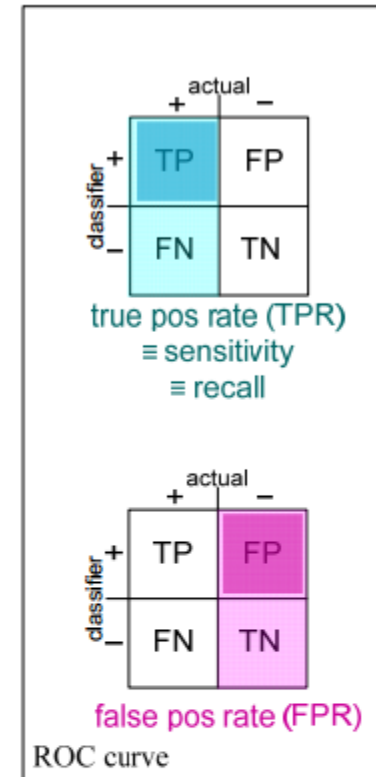
ROC (AUC)



ROC (“Receiver Operating Characteristic”) curves plot TPR vs. FPR as the classifier goes from “conservative” to “liberal”

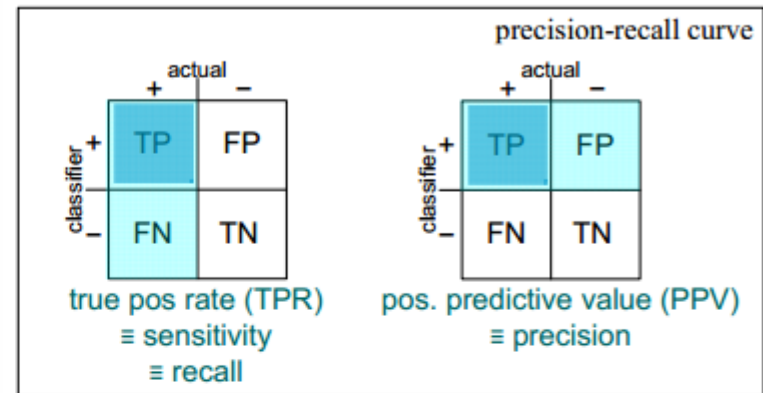
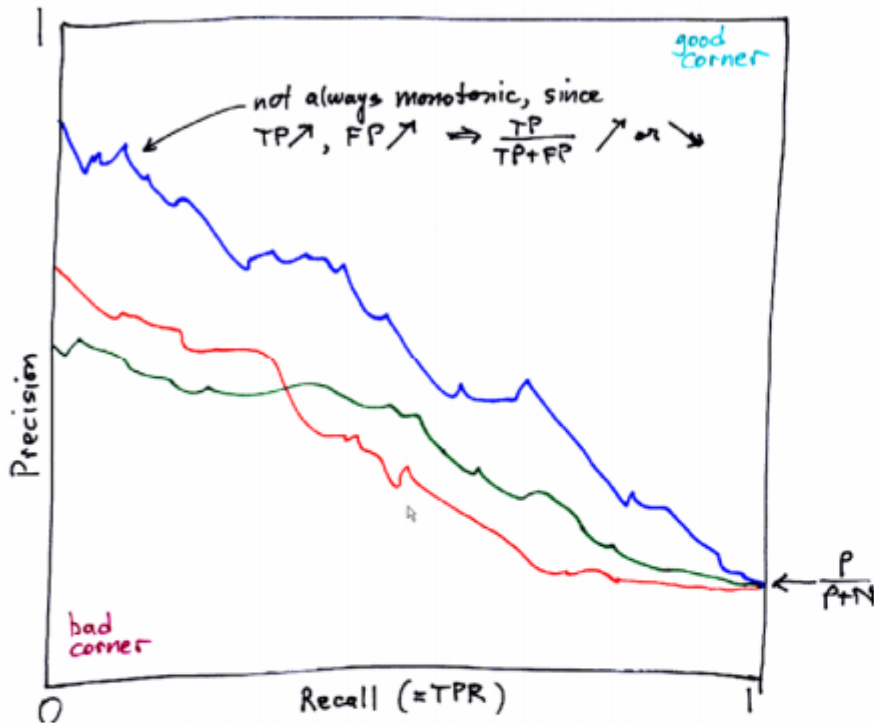


blue dominates red and green
neither red nor green dominate the other



You could get the best of the red and green curves by making a hybrid or “Frankenstein” classifier that switches between strategies at the cross-over points.

Precision-Recall curves overcome this issue by comparing TP with FN and FP

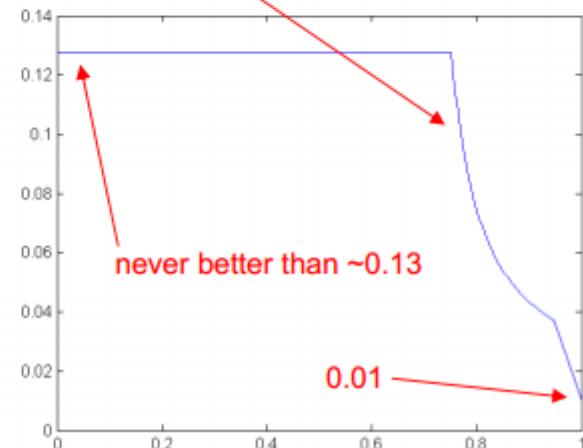


By the way, this shape "cliff" is what the ROC convexity constraint looks like in a Precision-Recall plot. It's not very intuitive.

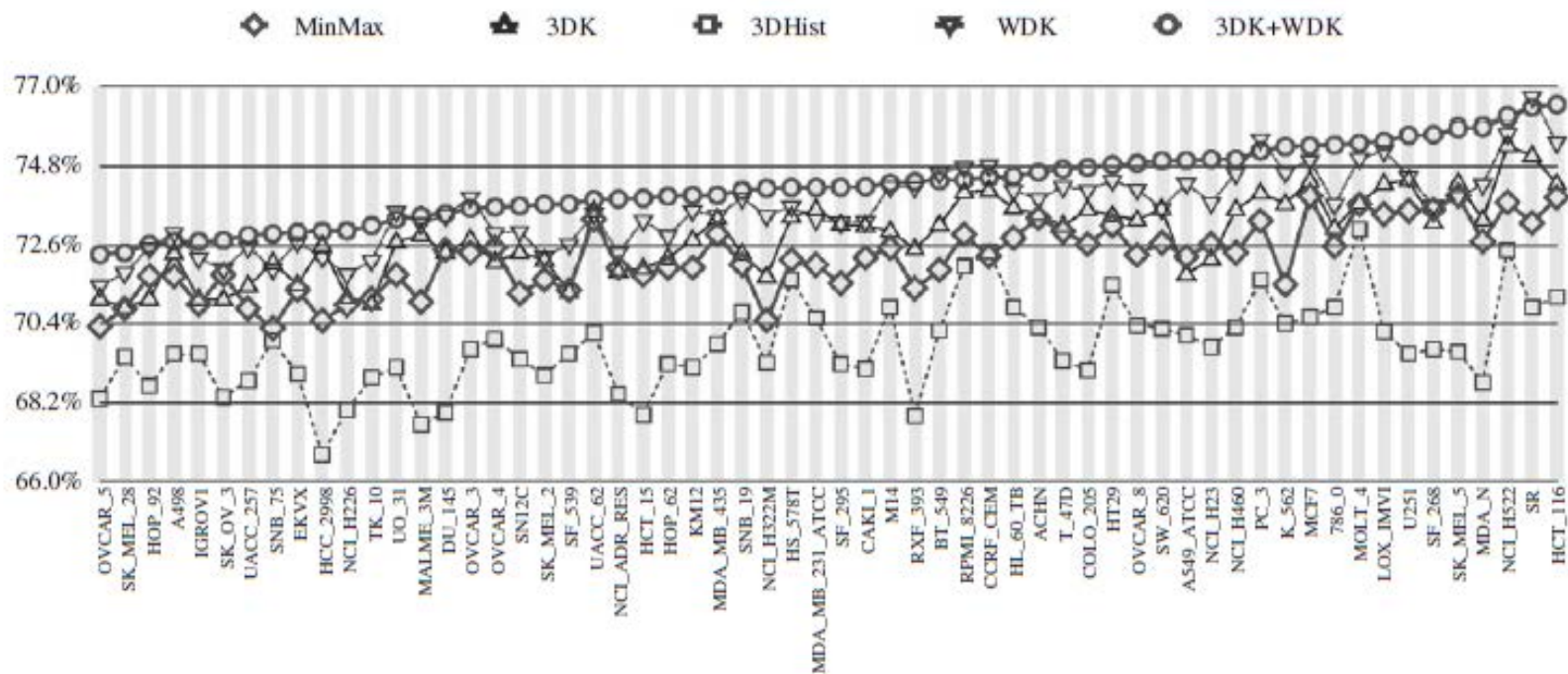
Continue our toy example:

note that P and N now enter

```
prec = tpr*100./(tpr*100+fpr*9900);
prec(1) = prec(2); % fix up 0/0
reca = tpr;
plot(reca,prec)
```

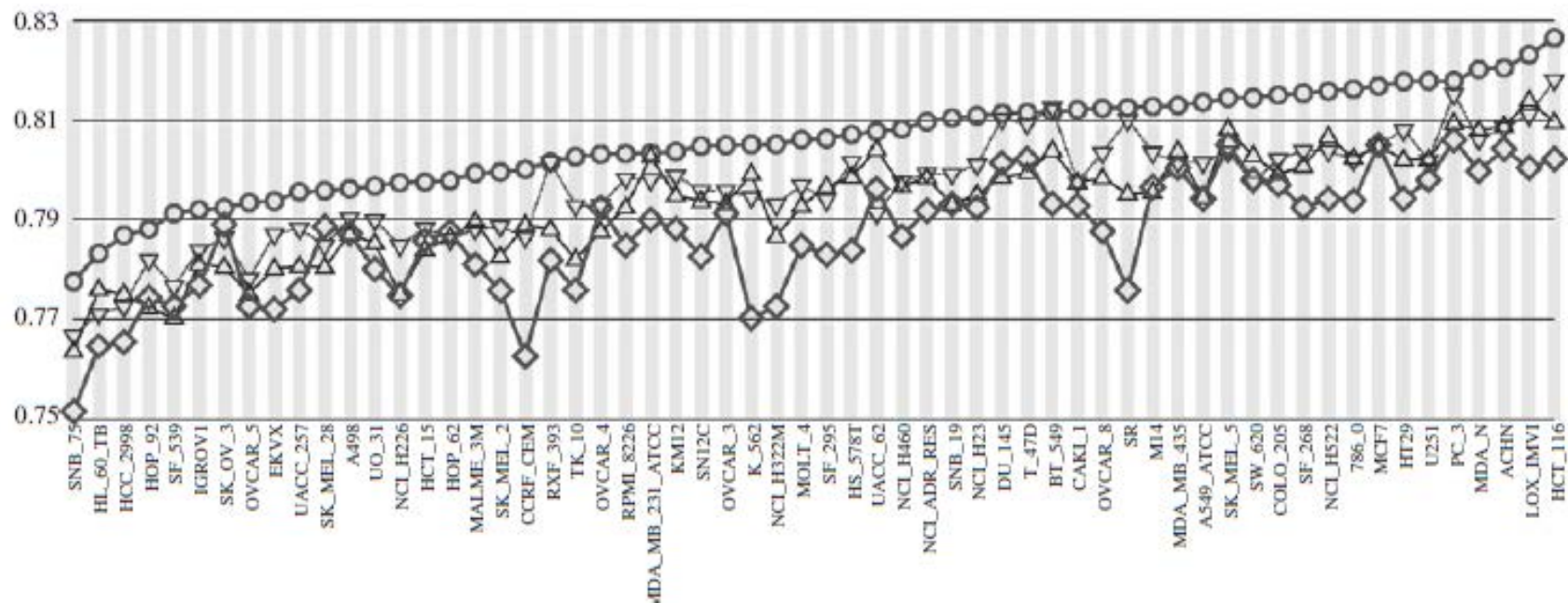


RESULTS: NCI CANCER SCREENING DATASET.



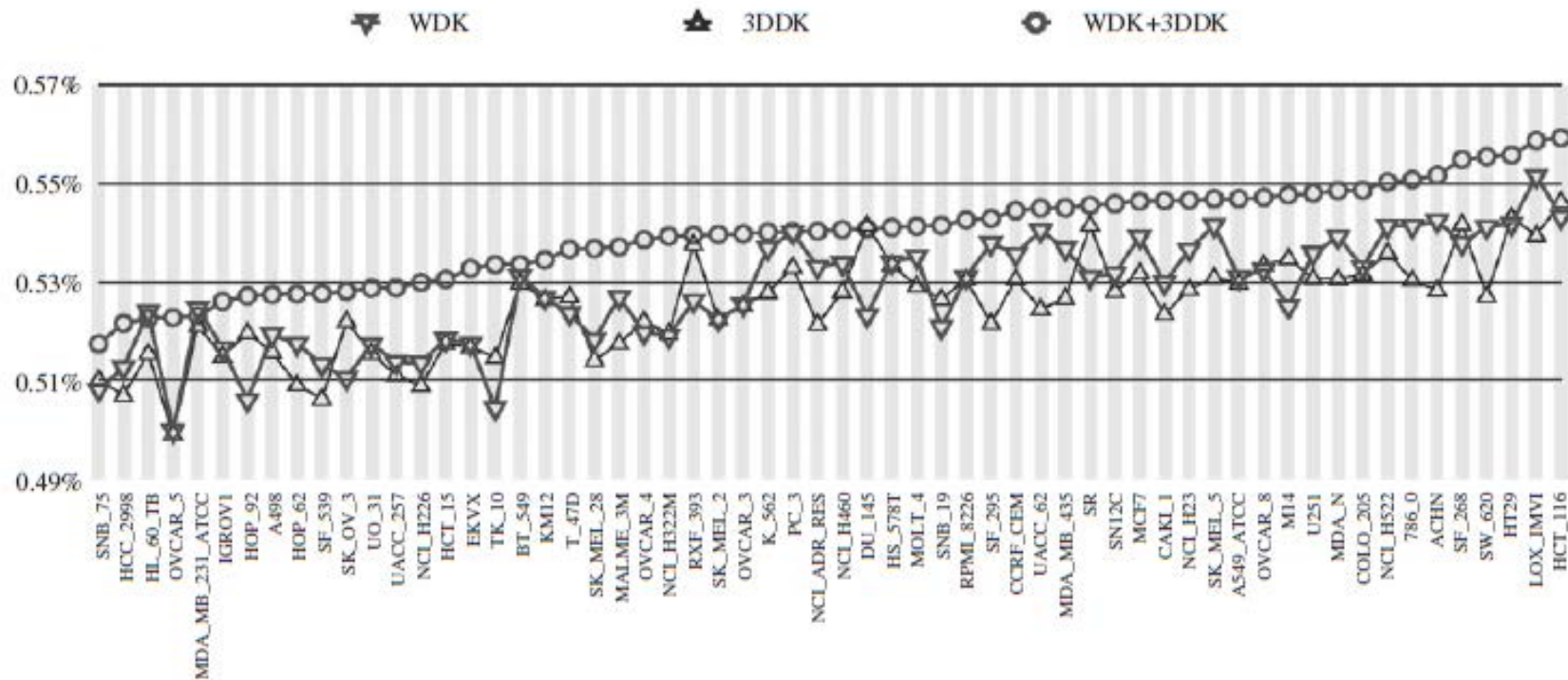
prediction accuracy

RESULTS: NCI CANCER SCREENING DATASET.



ROC AUC values

RESULTS: NCI CANCER SCREENING DATASET.



precision/recall curve values

RESULT: NCI ANTI-HIV SCREENING DATASET

Table 1. Results of the experiments on the NCI Anti-HIV screening dataset

Method	CA versus CM	CA+CM versus CI	CA versus CI
FSG	0.786	0.786	0.914
FSG+3D	0.811	0.819	0.940
γ CPK	0.840 ± 0.010	0.837 ± 0.012	0.947 ± 0.008
γ WDK	0.854 ± 0.019	0.841 ± 0.006	0.945 ± 0.009
γ 3DDK	0.853 ± 0.040	0.844 ± 0.007	0.951 ± 0.006
γ (WDK+3DDK)	0.861 ± 0.028	0.848 ± 0.009	0.951 ± 0.007

The 3DDK and WDK are compared to the frequent subgraphs approach and to the cyclic pattern kernel. The table reports the value of AUC for the various methods.