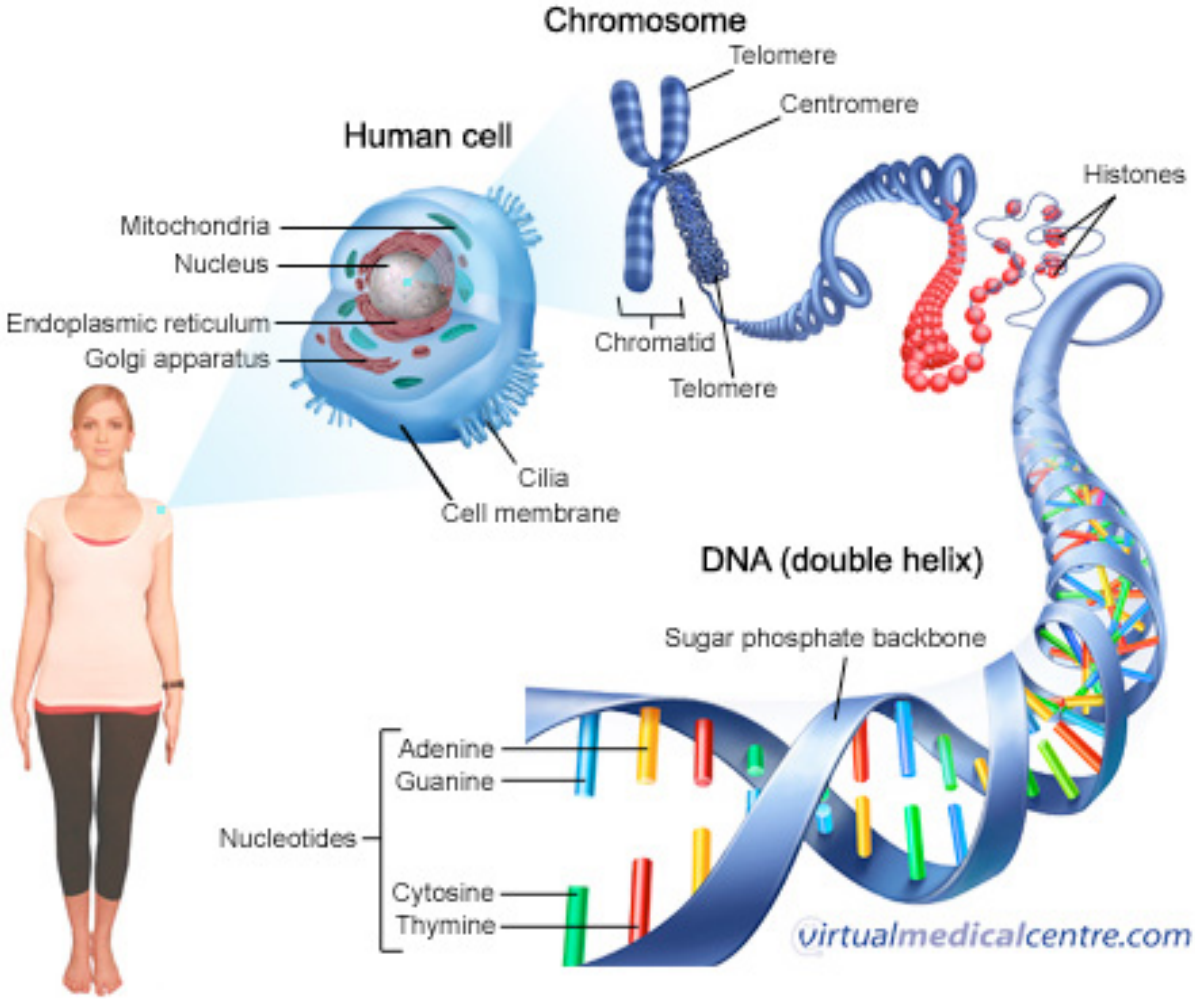
Machine Learning for Personalized Medicine

Jean-Philippe Vert

Atelier Prospectom, Grenoble, November 21, 2014

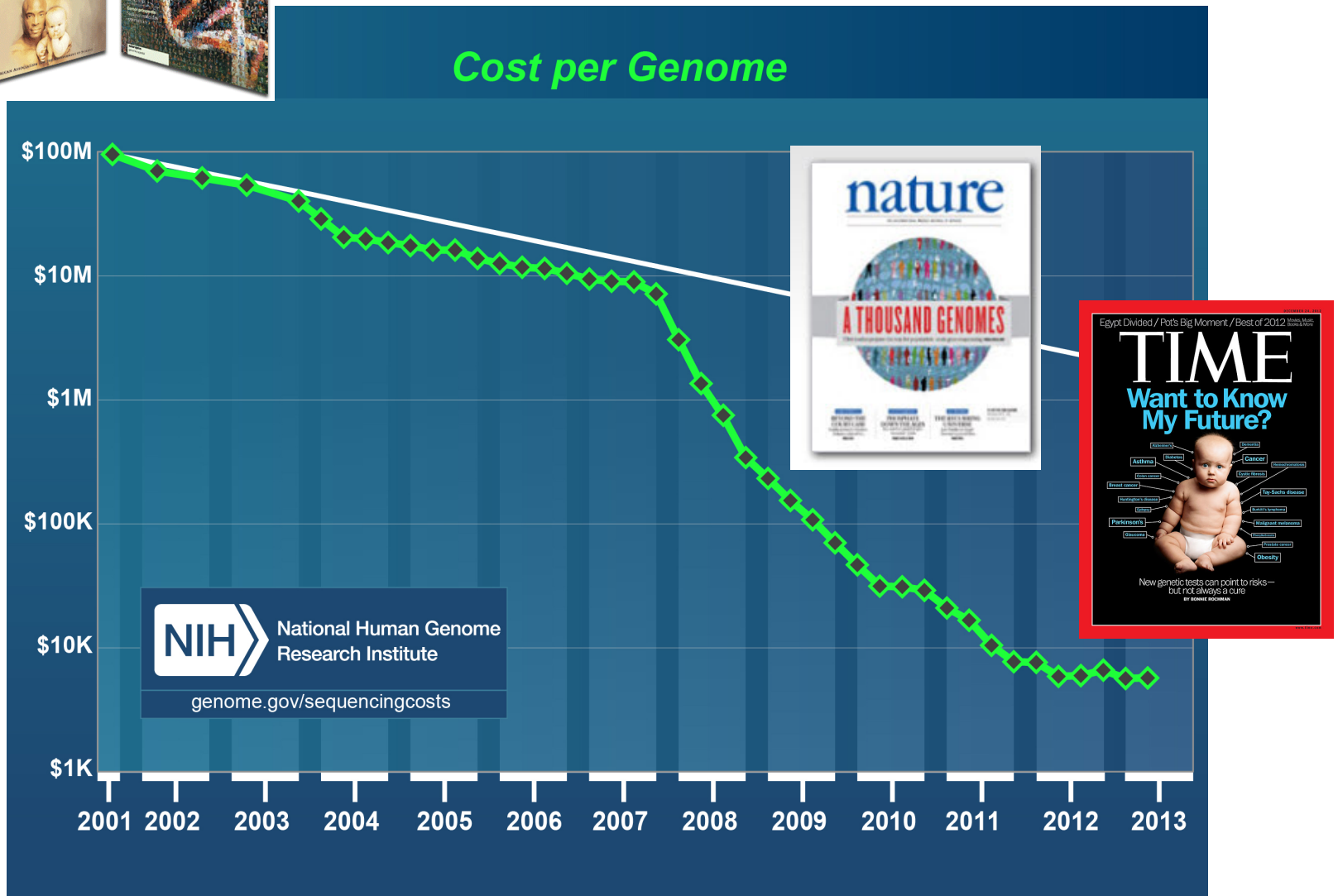
What's in your body



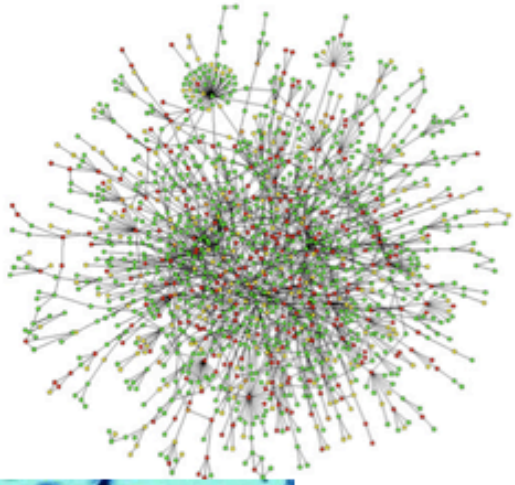
1 body = 10^{14} human cells (and 100x more non-human cells)

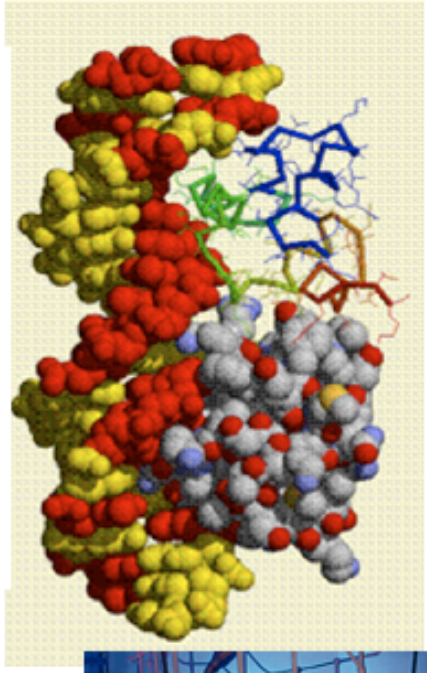
1 cell = 6×10^9 ACGT coding for $20,000$ genes

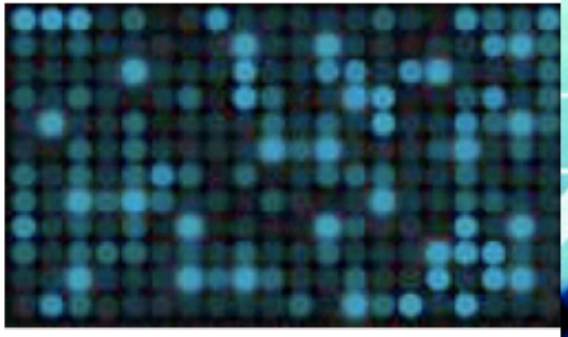
Sequencing revolution

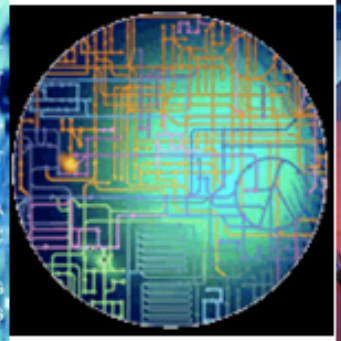


Many various data

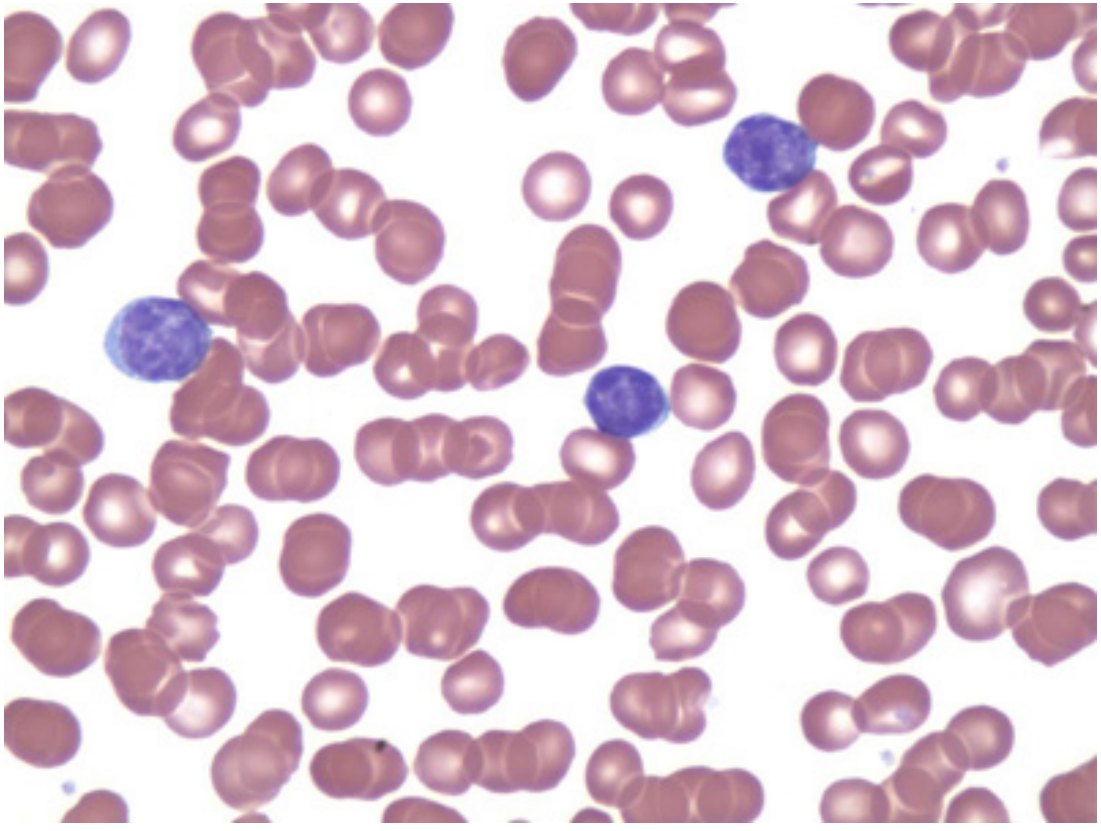




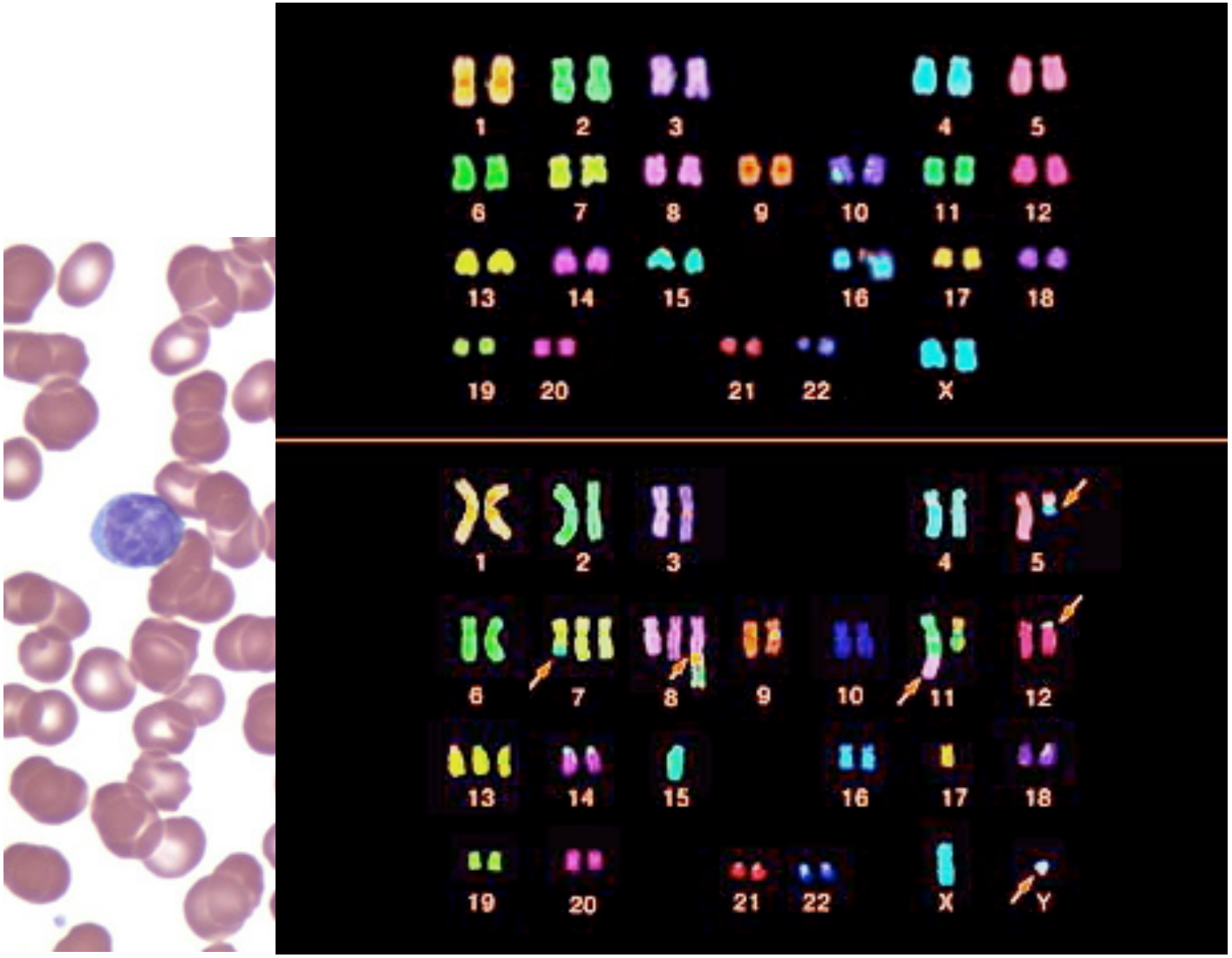




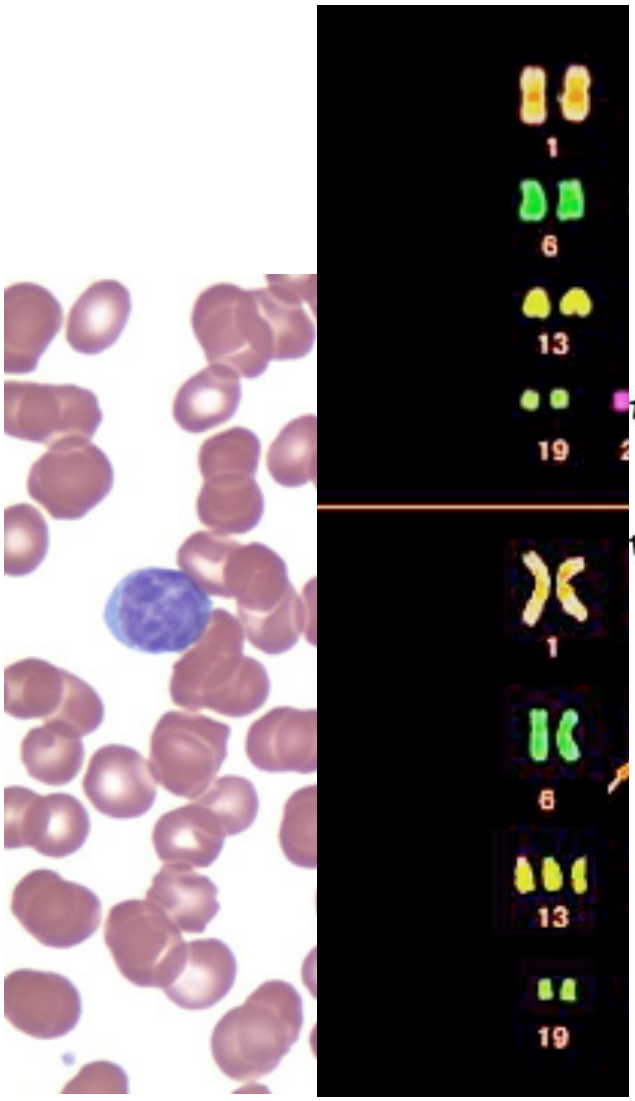
A cancer cell

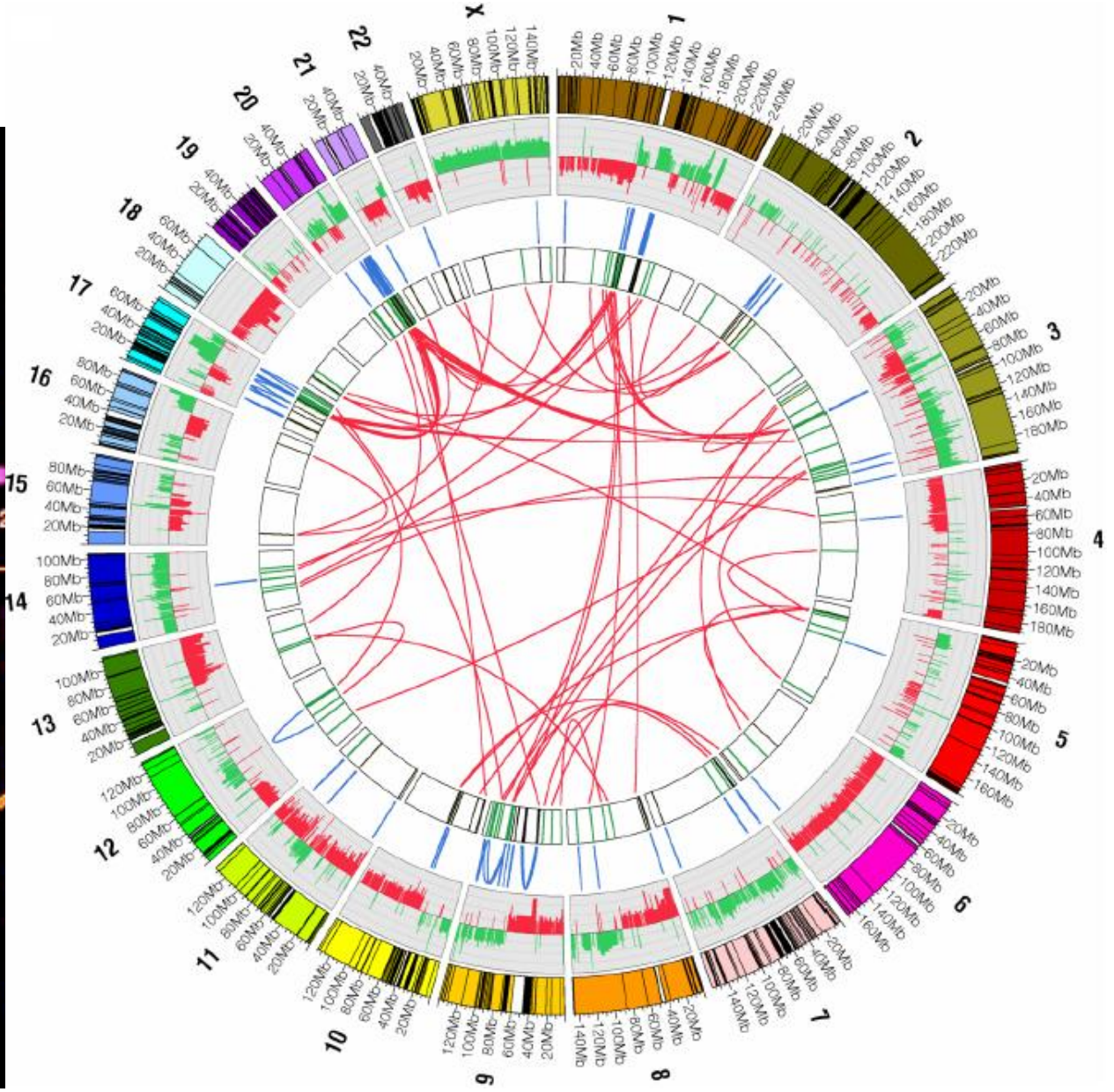


A cancer cell

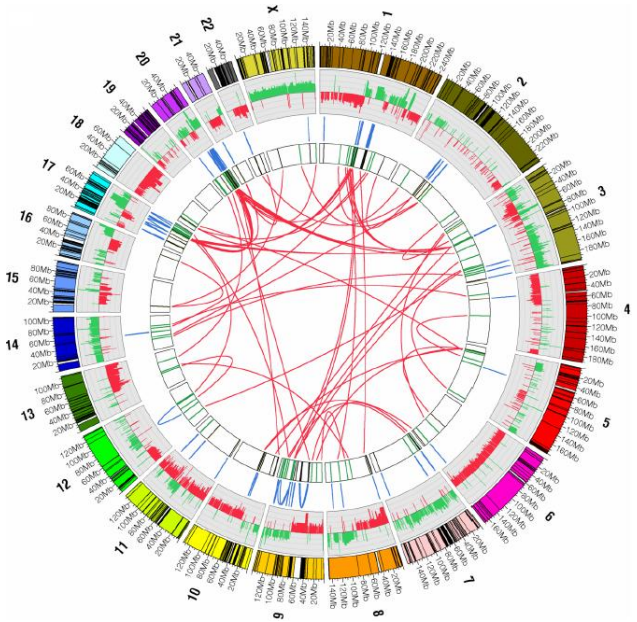


A cancer cell





Opportunities

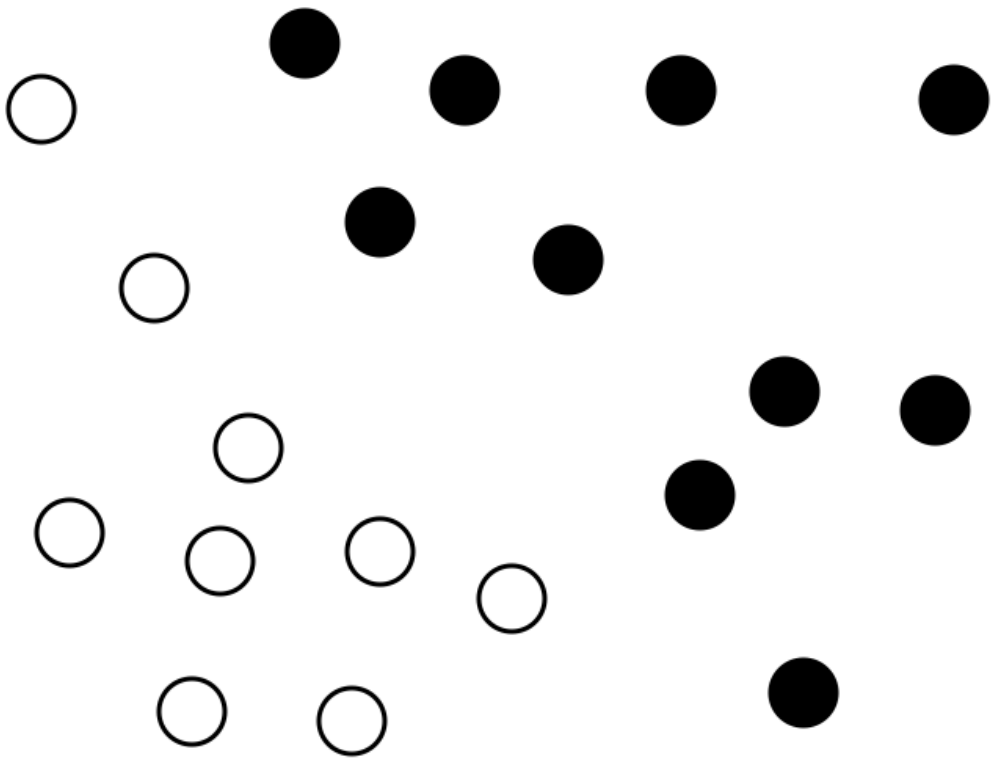


- What is your risk of developing a cancer? (*prevention*)
- After diagnosis and treatment, what is the risk of relapse? (*prognosis*)
- What specific treatment will cure your cancer? (*personalized medicine*)

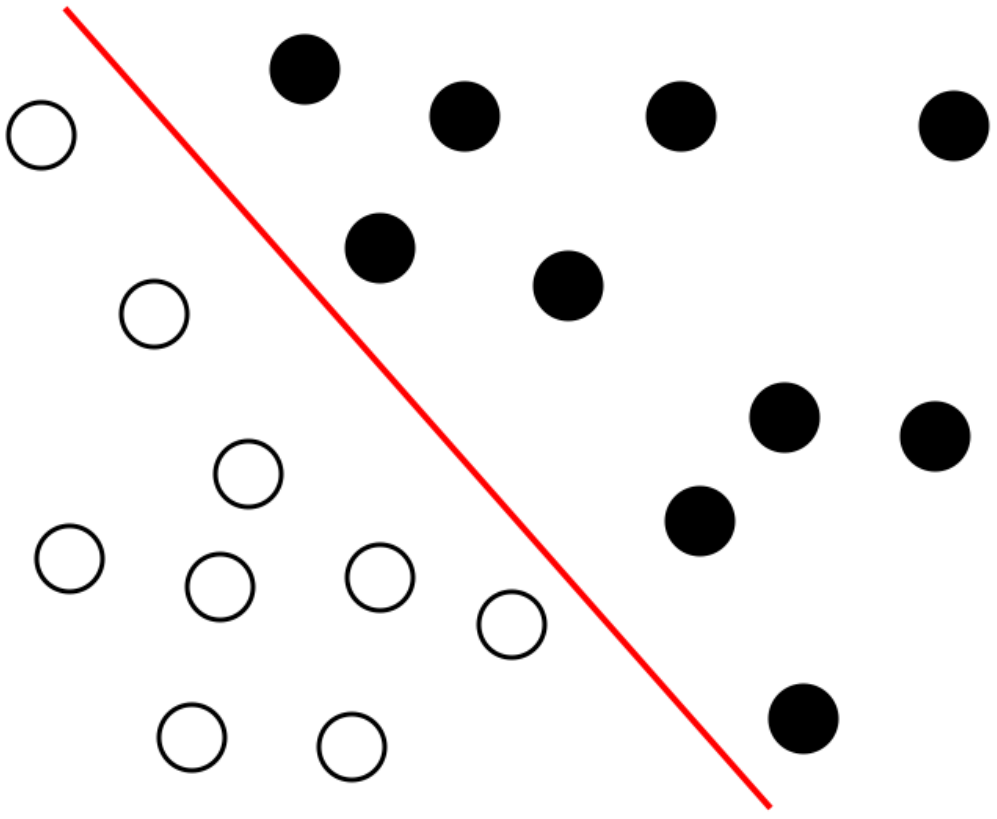
Example


Patients with same condition
DNA
Profiling
Good responders
No Responders
Bad side effects

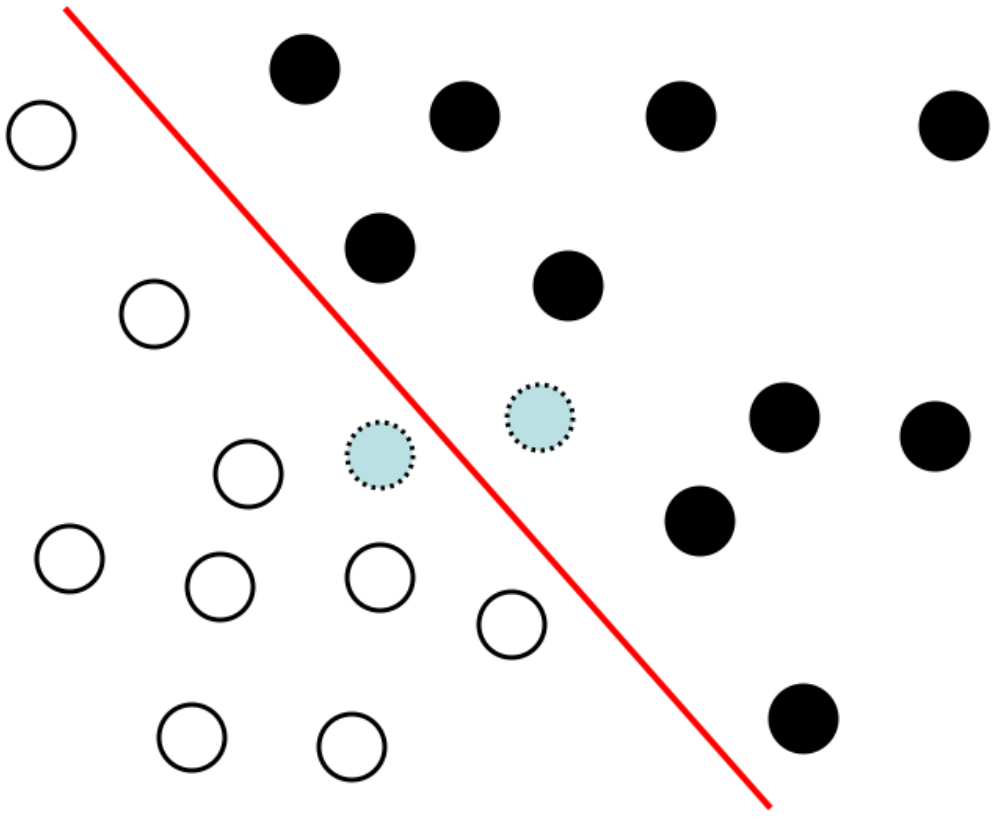
Machine learning formulation



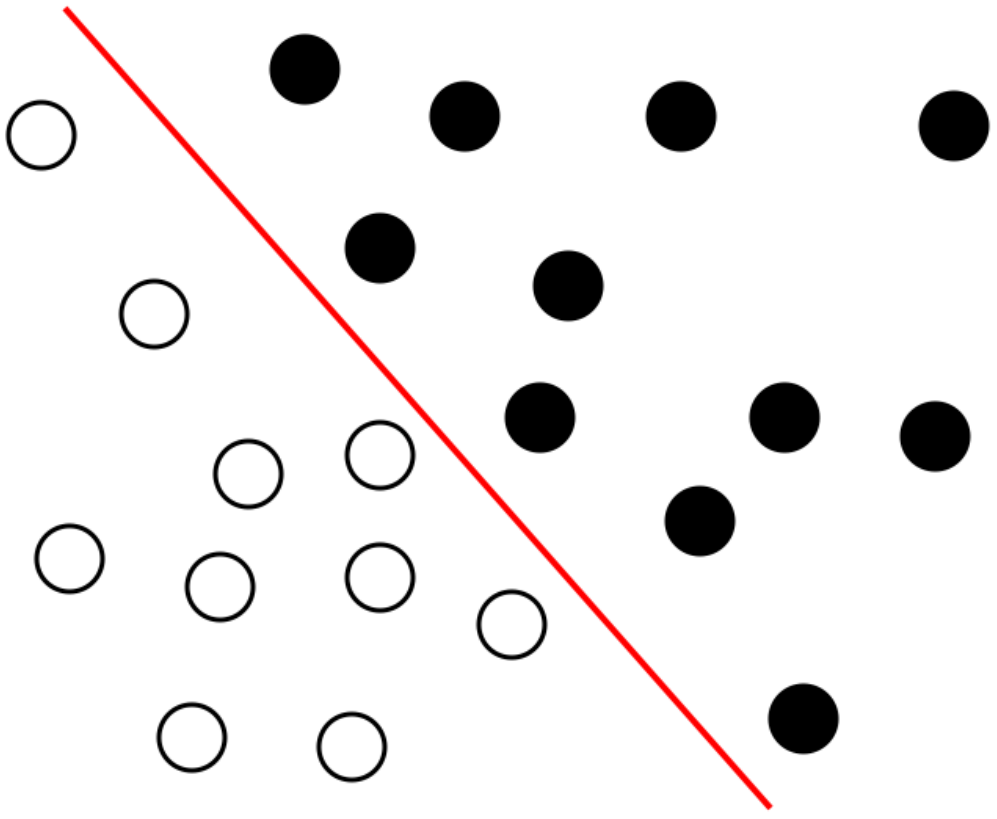
Machine learning formulation



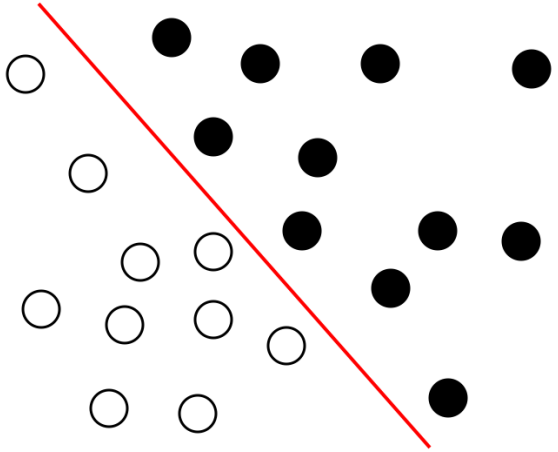
Machine learning formulation

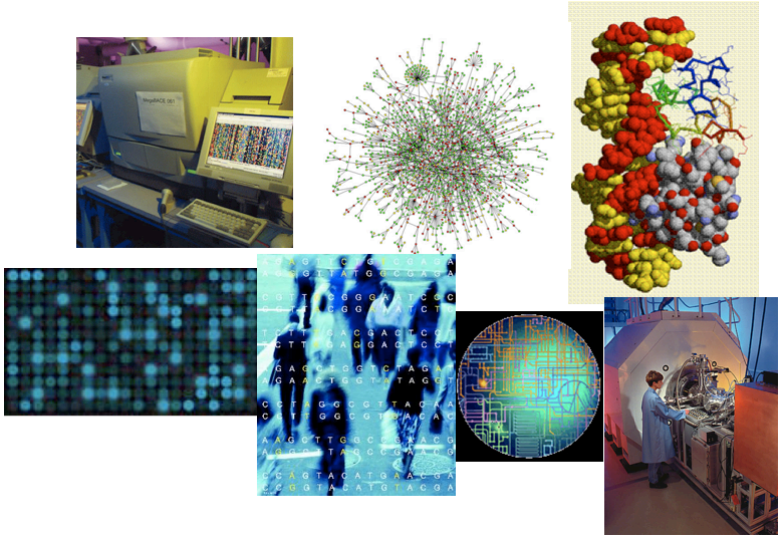


Machine learning formulation



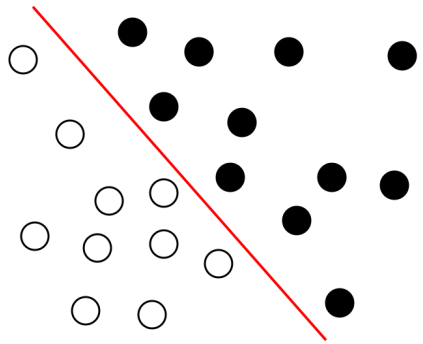
Challenges





- High dimension
- Few samples
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models

Learning with regularization



Learn

$$f_\beta(x) = \beta^\top x$$

by solving

$$\min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \Omega(\beta)$$

- $R(f_\beta)$ empirical risk
- $\Omega(\beta)$ penalty, typically:
  - $\Omega(\beta) = \sum_{i=1}^p \beta_i^2$ SVM, ridge regression, ...
  - $\Omega(\beta) = \sum_{i=1}^p |\beta_i|$ Lasso, boosting, ...

Outline

1. Learning molecular classifiers with network information
2. Kernel bilinear regression for toxicogenomics

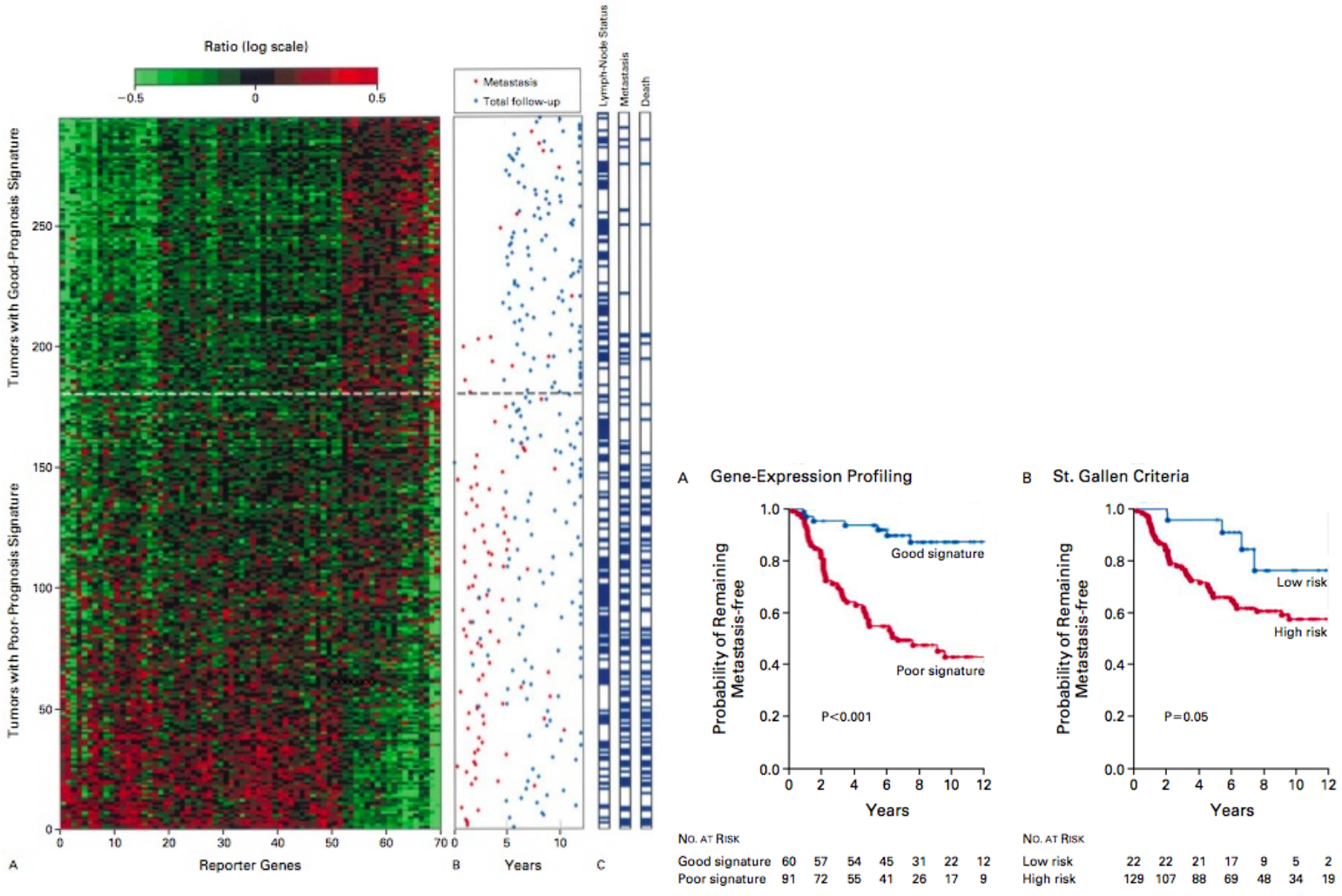
Outline

1. Learning molecular classifiers with network information
2. Kernel bilinear regression for toxicogenomics

Joint work with...

Franck Rapaport, Emmanuel Barillot, Andrei Zinovyev, Anne-Claire Haury, Laurent Jacob, Guillaume Obozinski

Breast cancer prognosis

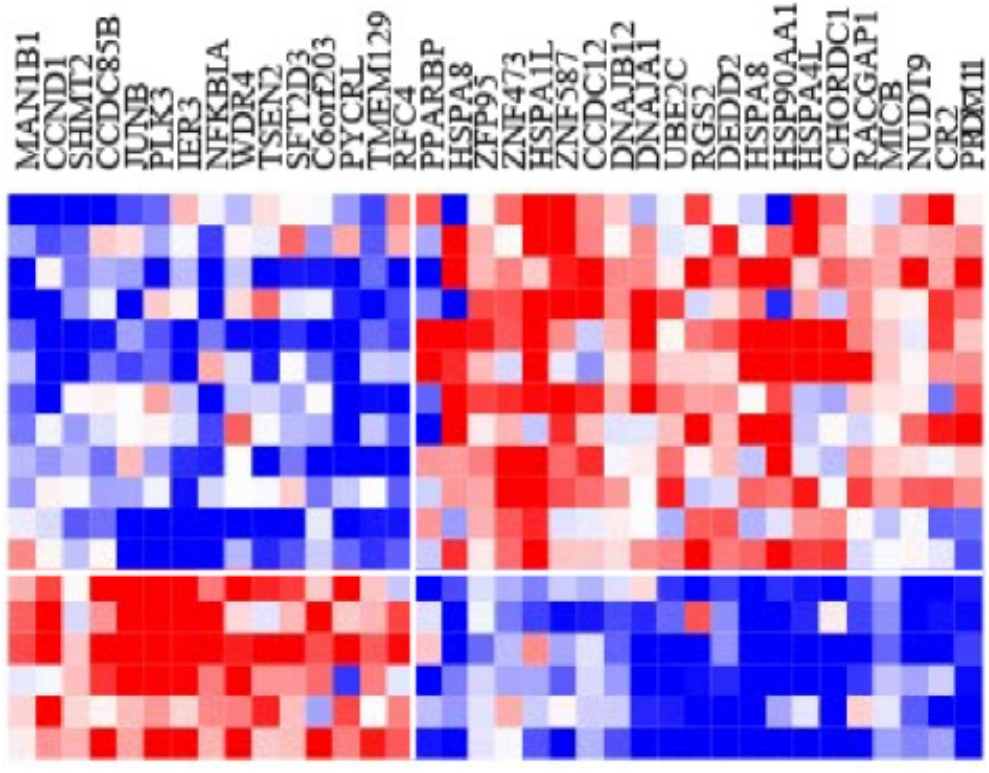


(van ’t Veer et al., 2002)

Gene selection, molecular signature

The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology



Some "surprising" results

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

70 genes (Nature, 2002)

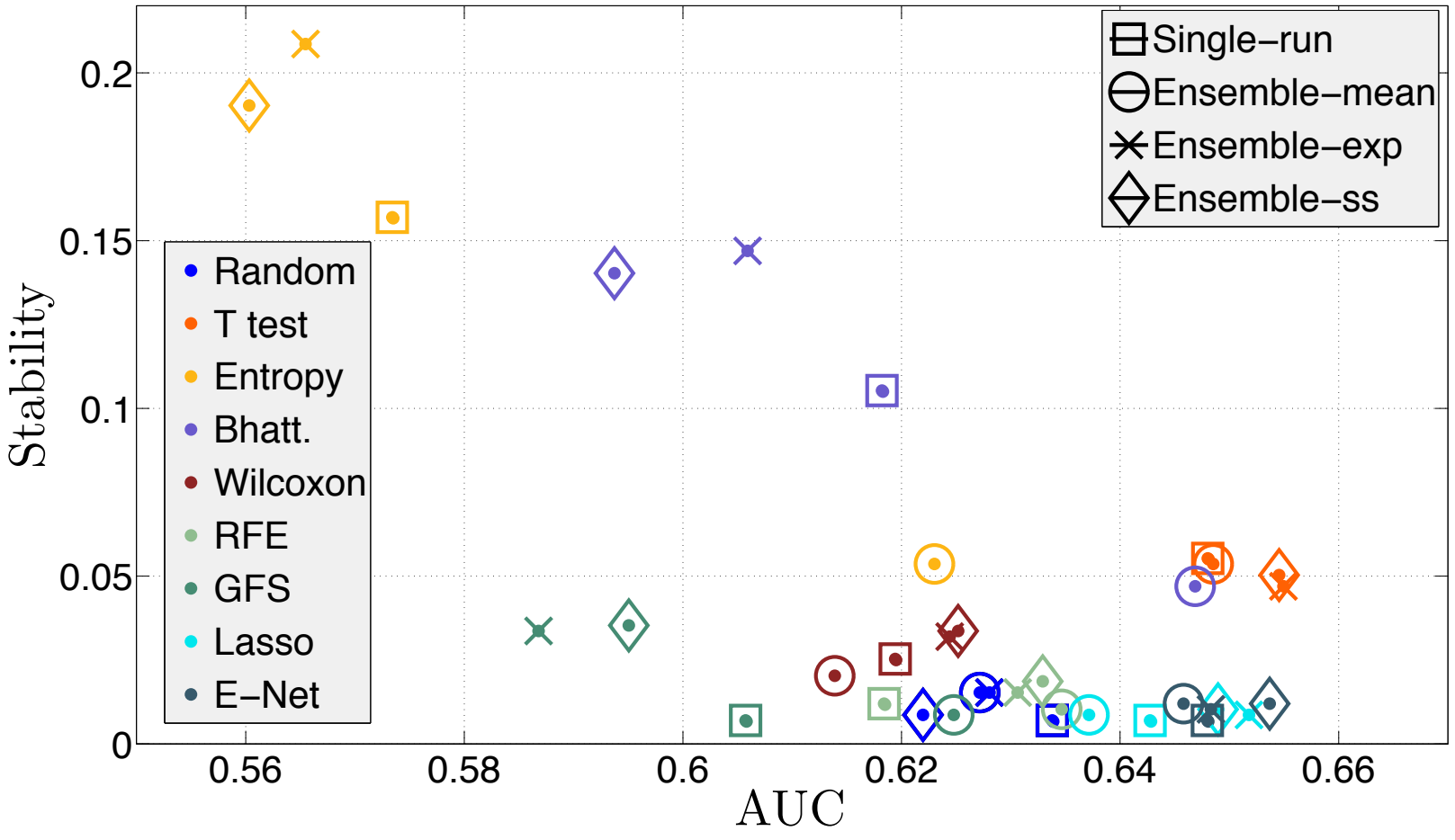
Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

76 genes (Lancet, 2005)

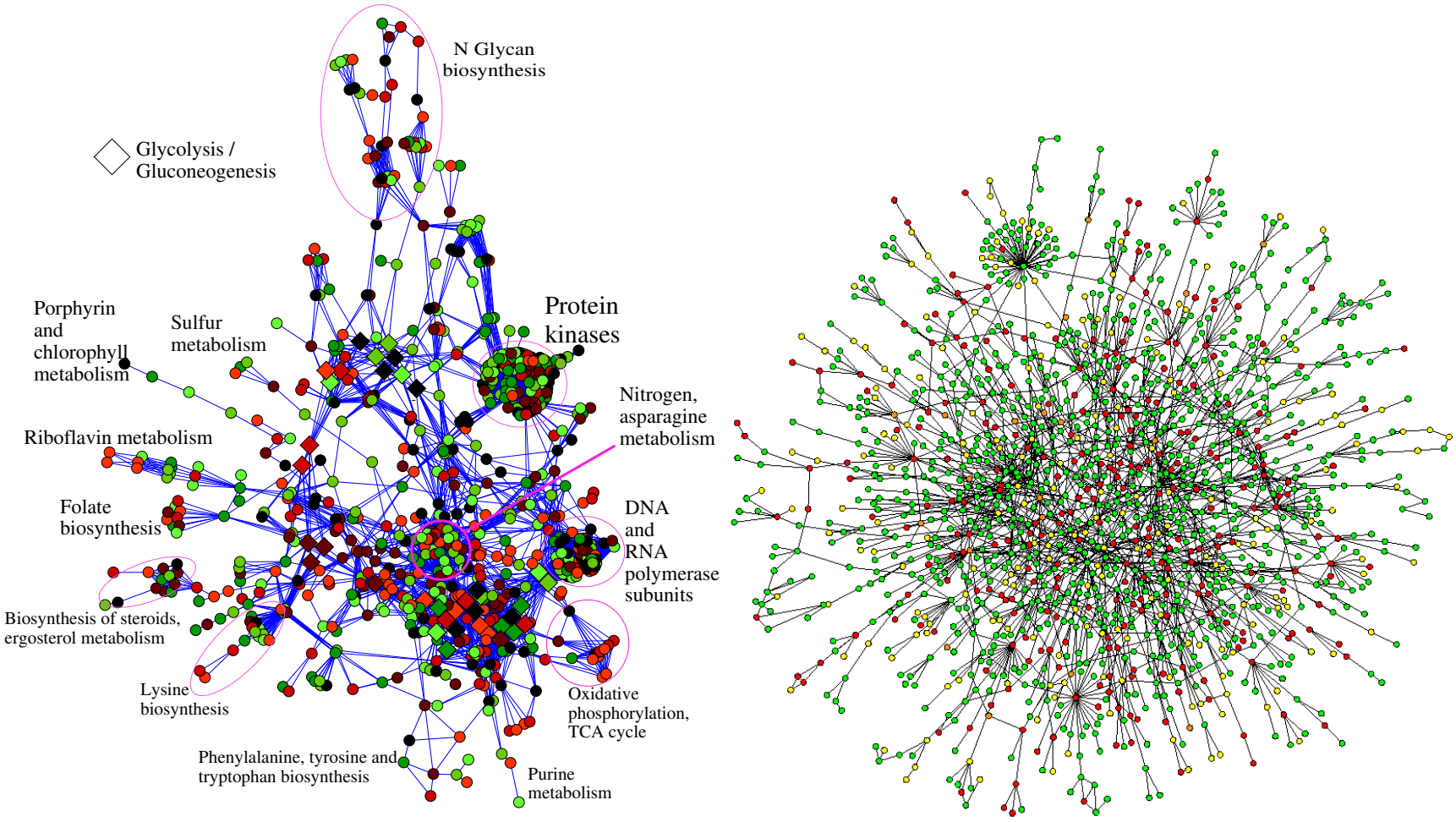
3 genes in common

Lack of stability of signatures



(Haury et al., 2011)

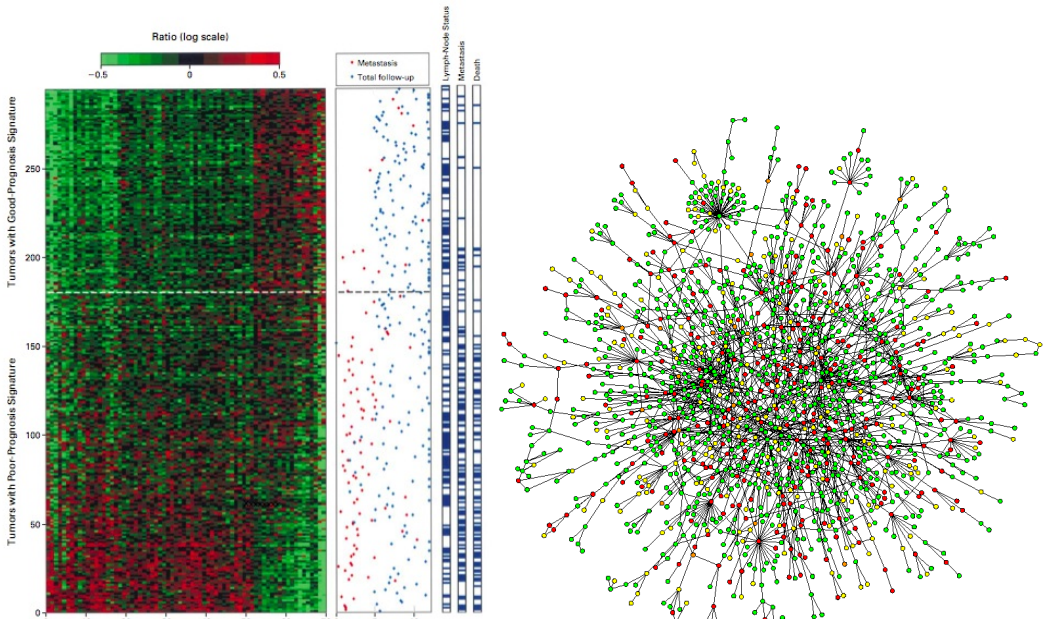
Gene networks



Gene networks and expression data

Motivation

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- Many pathways and protein-protein interactions are already known
- Hypothesis: the weights of the classifier should be “coherent” with respect to this prior knowledge



Graph based penalty

$$f_\beta(x) = \beta^\top x \qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

$$\Omega(\beta) = \sum_{i\sim j}(\beta_i - \beta_j)^2 \,,$$

$$\min_{\beta\in\mathbb{R}^p} R(f_\beta) + \lambda \sum_{i\sim j}(\beta_i - \beta_j)^2 \,.$$

Graph based penalty

$$f_\beta(x) = \beta^\top x \qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

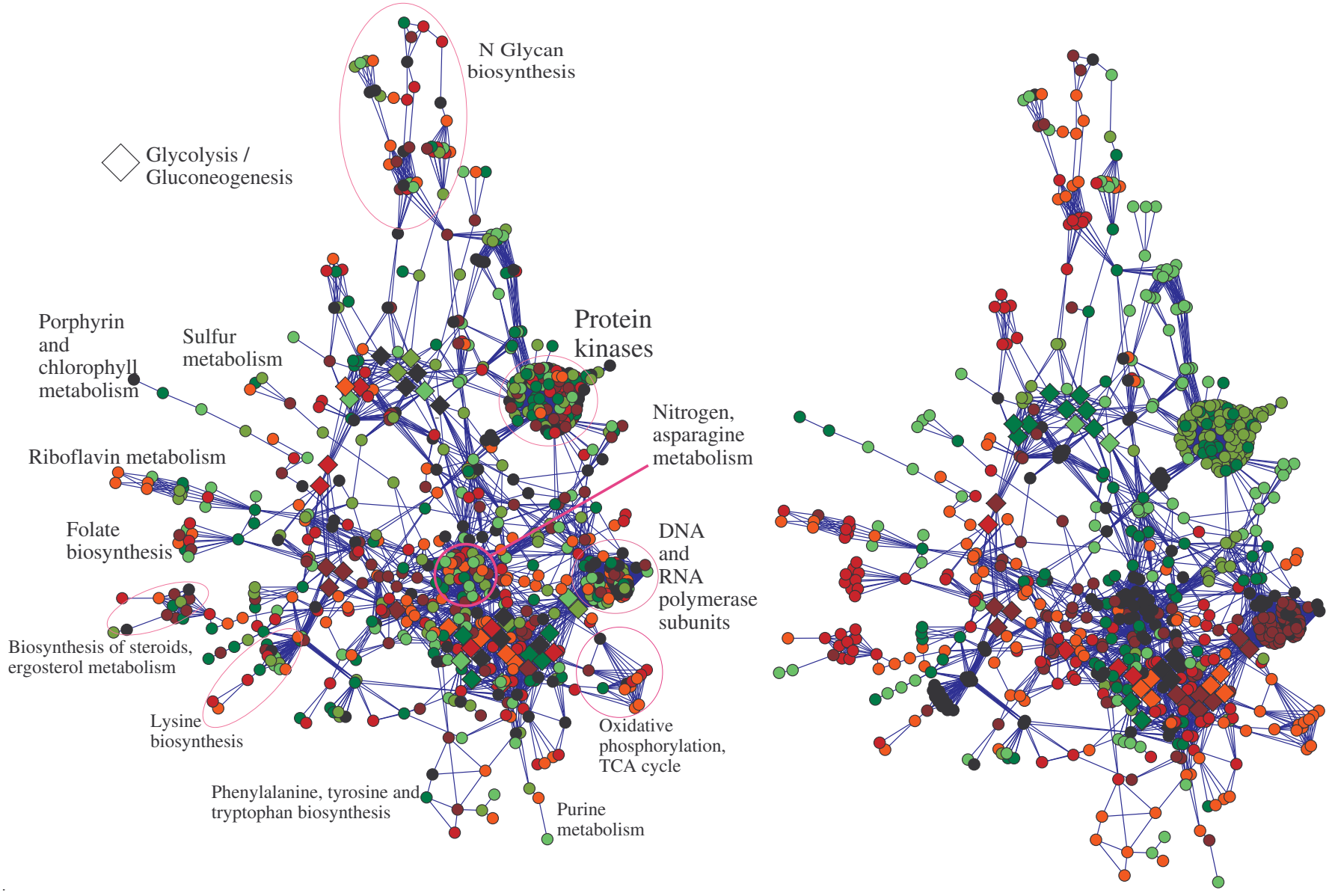
Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

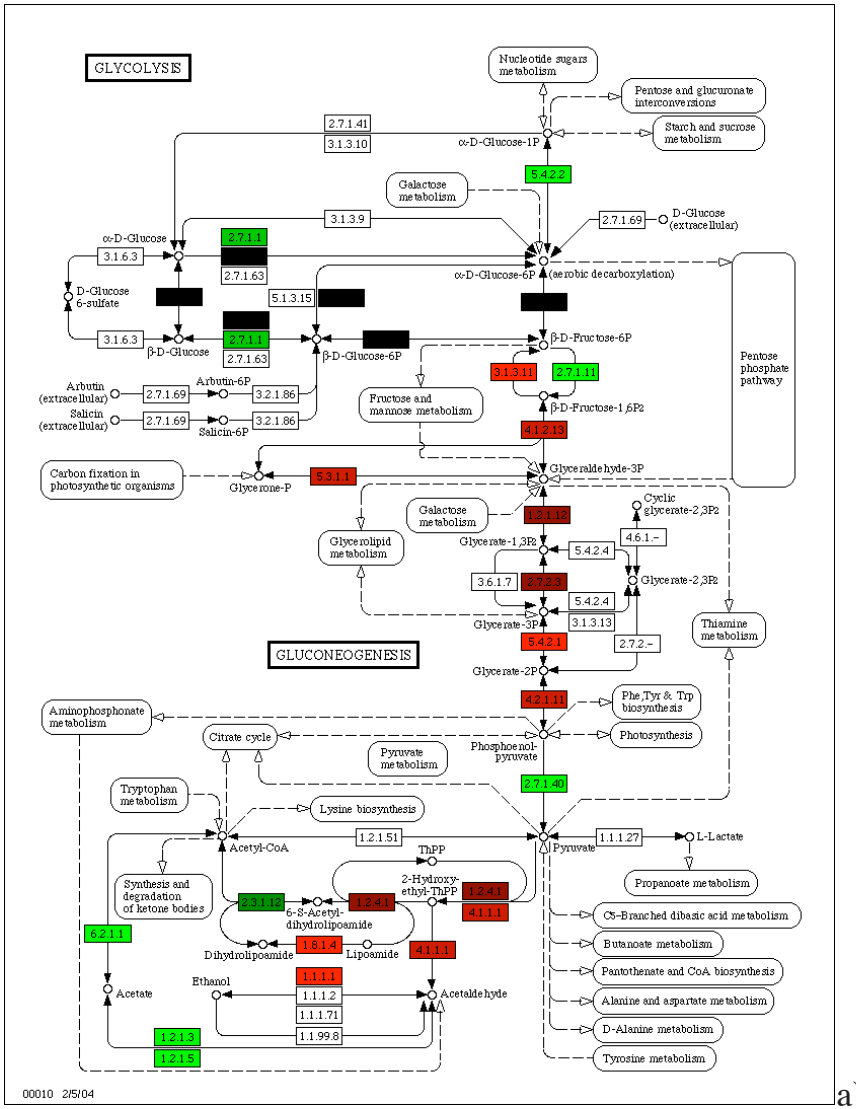
$$\Omega(\beta) = \sum_{i\sim j}(\beta_i - \beta_j)^2 \,,$$

$$\min_{\beta\in\mathbb{R}^p} R(f_\beta) + \lambda \sum_{i\sim j}(\beta_i - \beta_j)^2 \,.$$

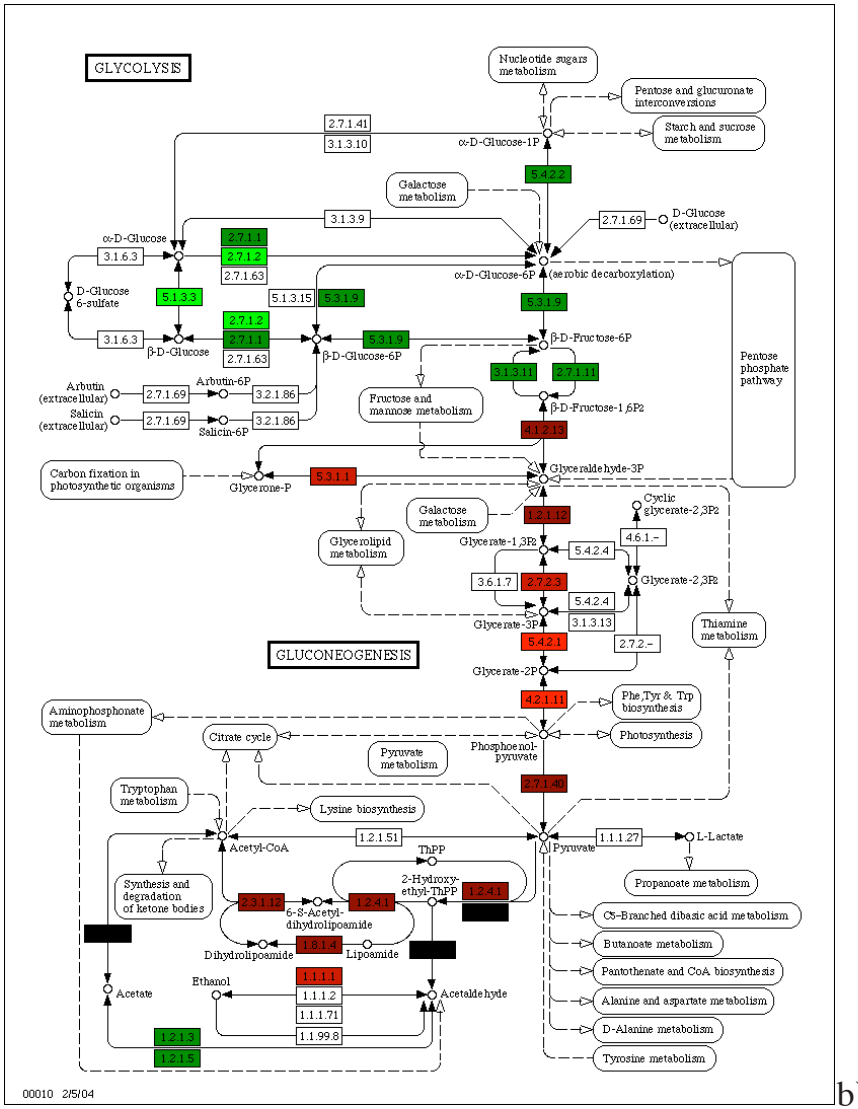
Classifiers



Classifier



a)



b)

Spectral penalty as a kernel

Theorem (Rapaport et al., 2007)

The function $f(x) = \beta^\top x$ where β is solution of

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\beta^\top x_i, y_i\right) + \lambda \sum_{i \sim j} \left(\beta_i - \beta_j\right)^2$$

is equal to $g(x) = \gamma^\top \Phi(x)$ where γ is solution of

$$\min_{\gamma \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\gamma^\top \Phi(x_i), y_i\right) + \lambda \gamma^\top \gamma \,,$$

and where

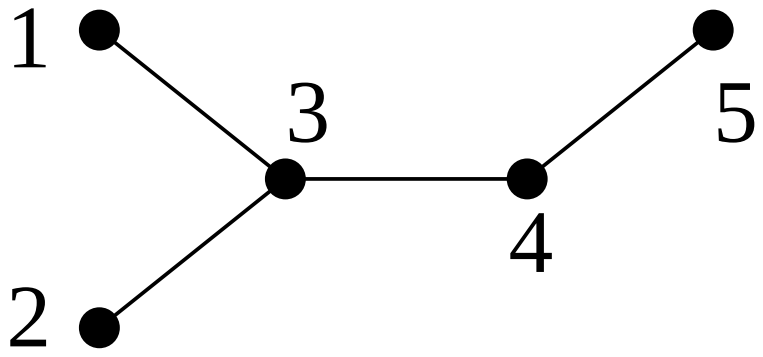
$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

for $K_G = L^*$, the pseudo-inverse of the graph Laplacian.

Graph Laplacian

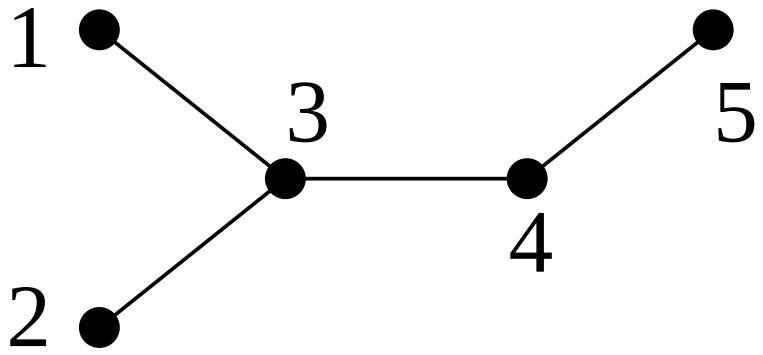
Definition

The Laplacian of the graph is the matrix $L = D - A$.



$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Pseufo-inverse of the Laplacian



$$L^* = \begin{pmatrix} 0.88 & -0.12 & 0.08 & -0.32 & -0.52 \\ -0.12 & 0.88 & 0.08 & -0.32 & -0.52 \\ 0.08 & 0.08 & 0.28 & -0.12 & -0.32 \\ -0.32 & -0.32 & -0.12 & 0.48 & 0.28 \\ -0.52 & -0.52 & -0.32 & 0.28 & 1.08 \end{pmatrix}$$

Other penalties with kernels

$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

with:

- $K_G = (c + L)^{-}1$ leads to

$$\Omega(\beta) = c \sum_{i=1}^{p} \beta_i^2 + \sum_{i \sim j} (\beta_i - \beta_j)^2 \ .$$

- The diffusion kernel:

$$K_G = \exp_M(-2tL) \, .$$

penalizes high frequencies of β in the Fourier domain.

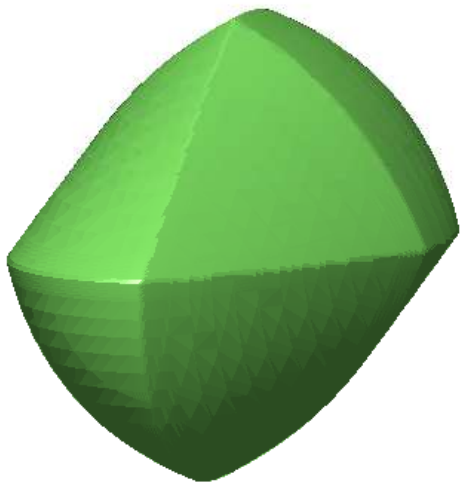
Other penalties without kernels

- Gene selection + Piecewise constant on the graph

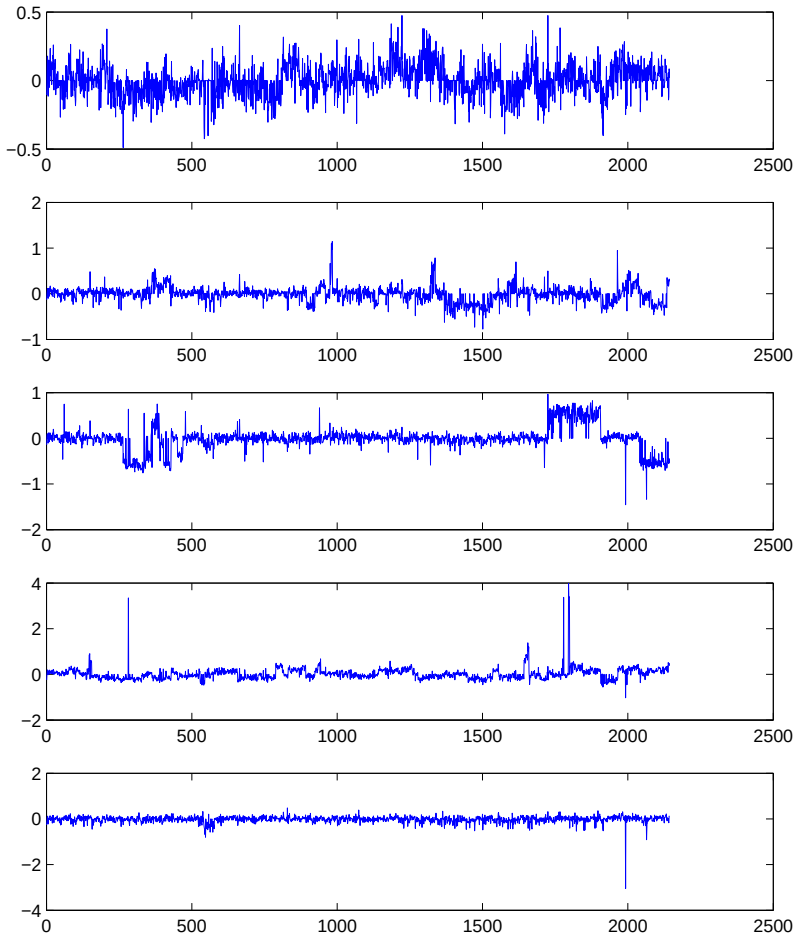
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i|$$

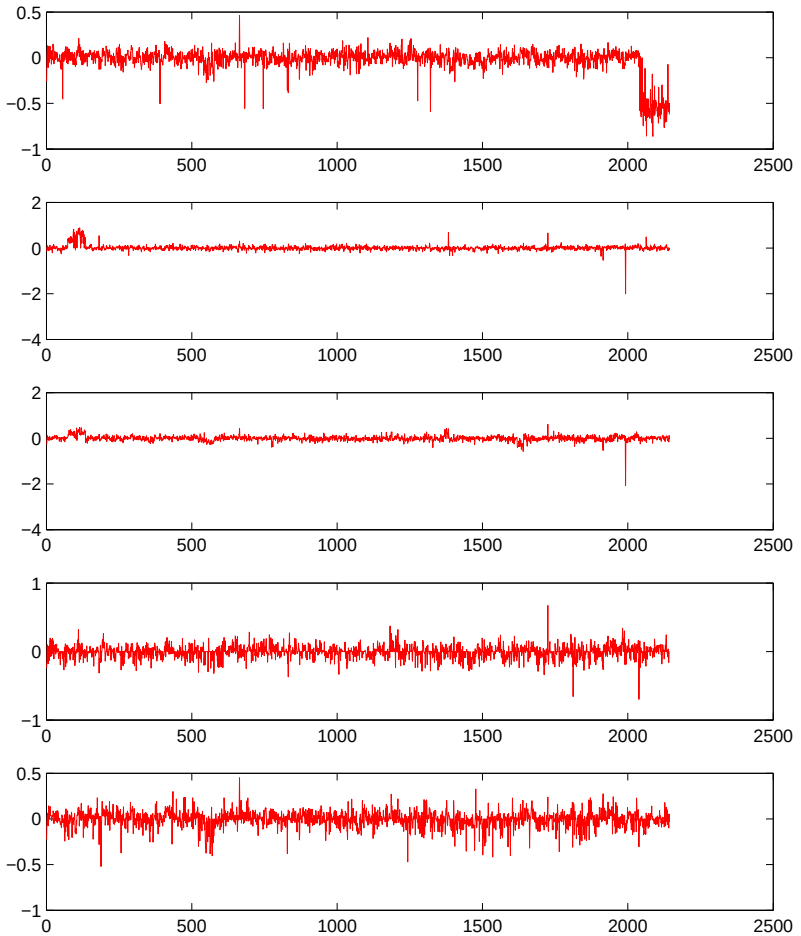
- Gene selection + smooth on the graph

$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i|$$



Example: classification of DNA copy number profiles

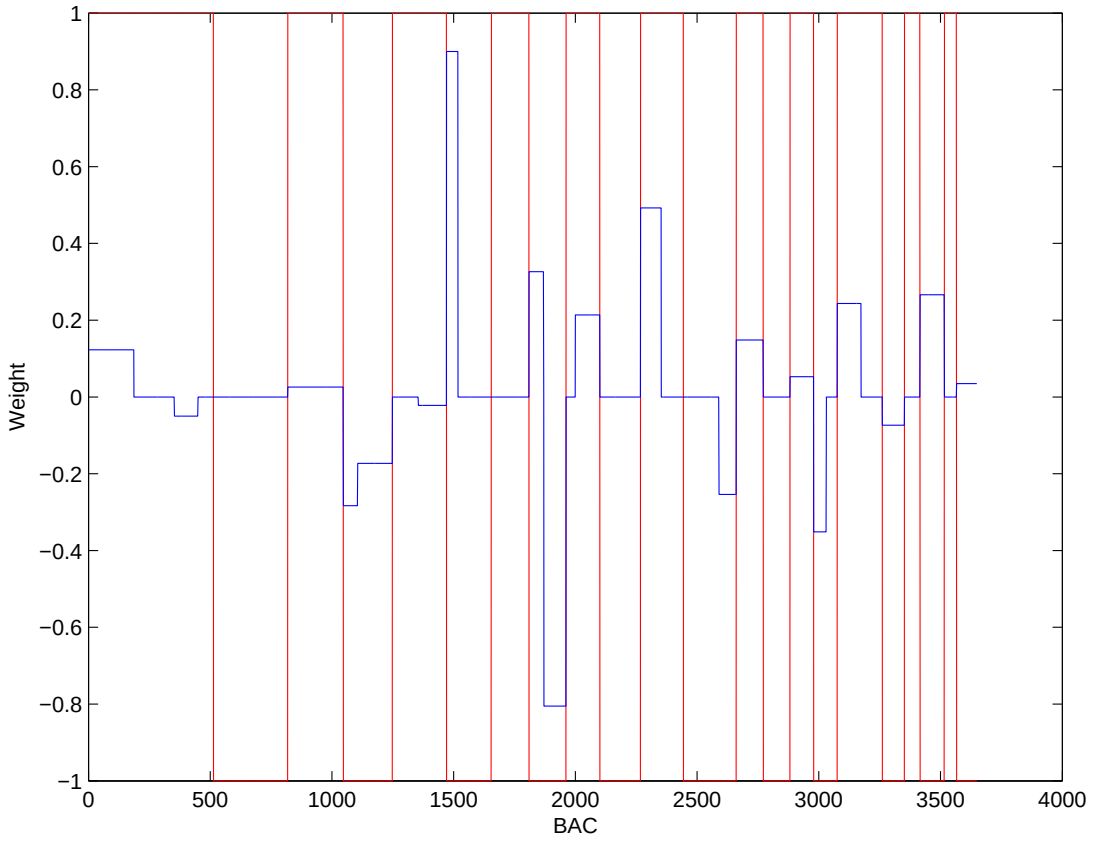




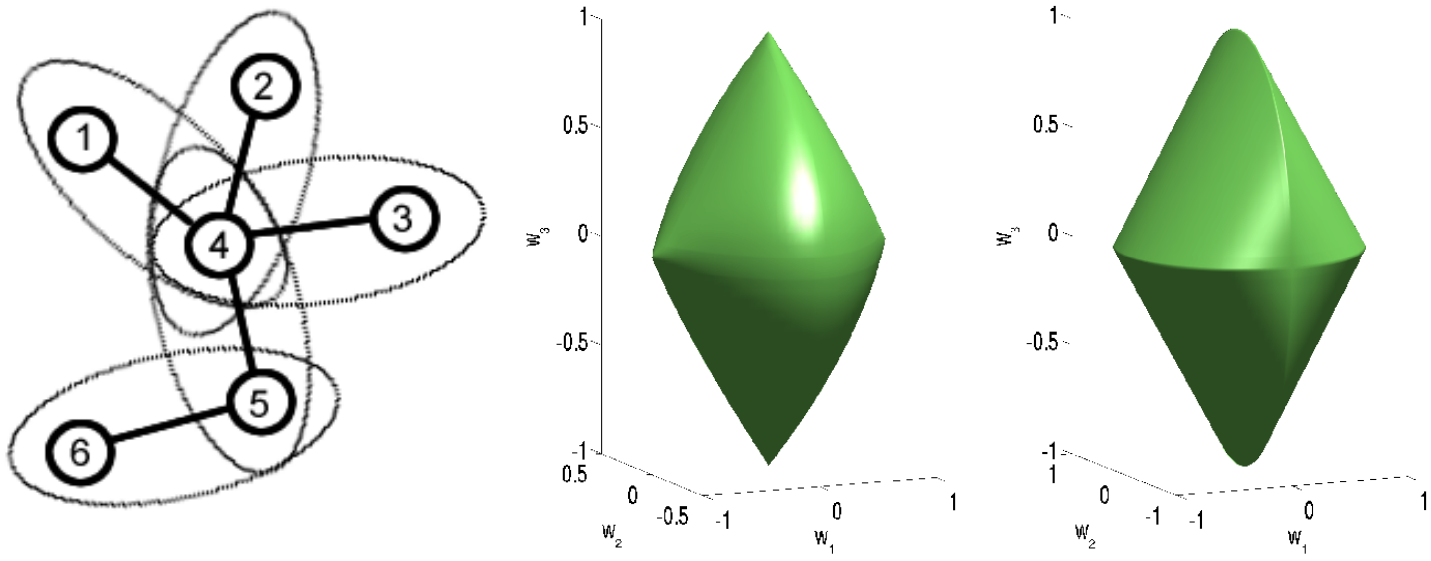
Aggressive (left) vs non-aggressive (right) melanoma

Fused lasso solution (Rapaport et al., 2008)

$$\Omega(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$



Graph-based structured feature selection

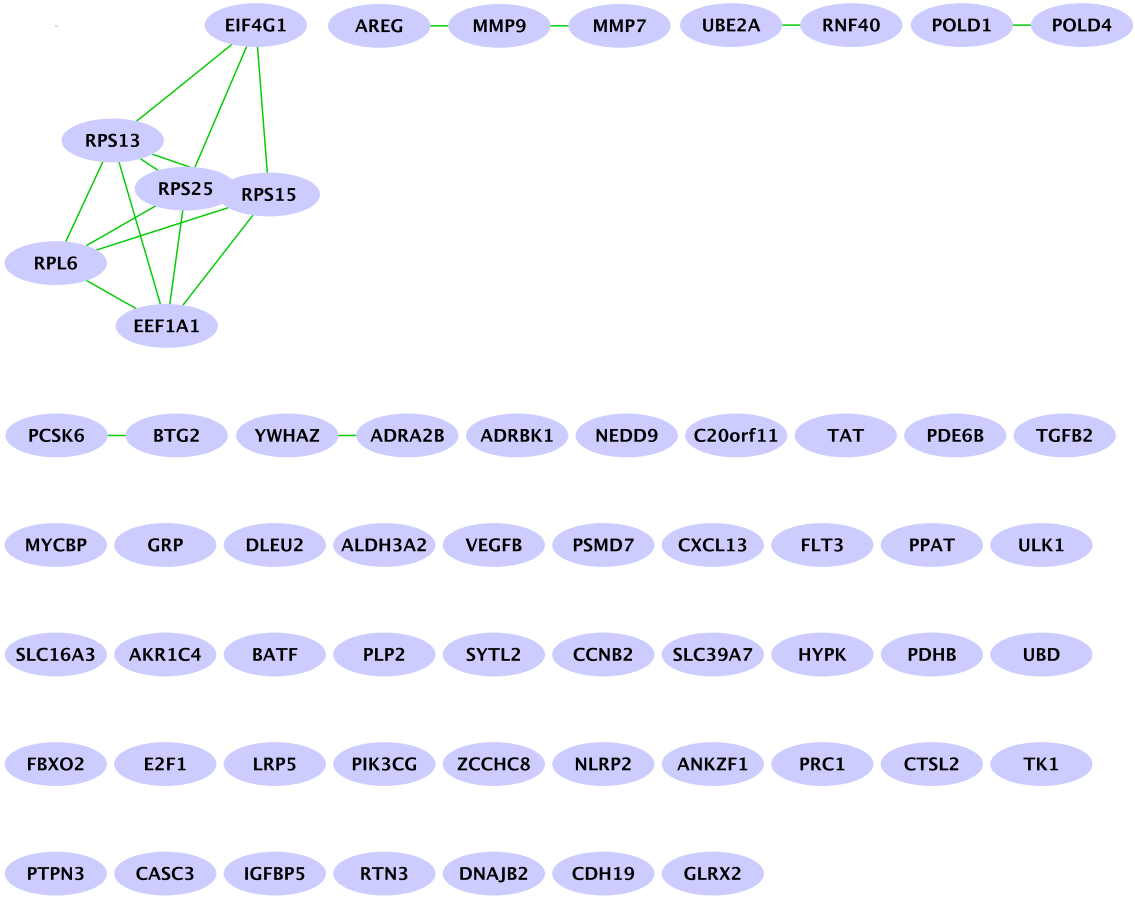


Graph lasso(s)

$$\Omega_1(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,, \quad \text{(Jenatton et al., 2011)}$$

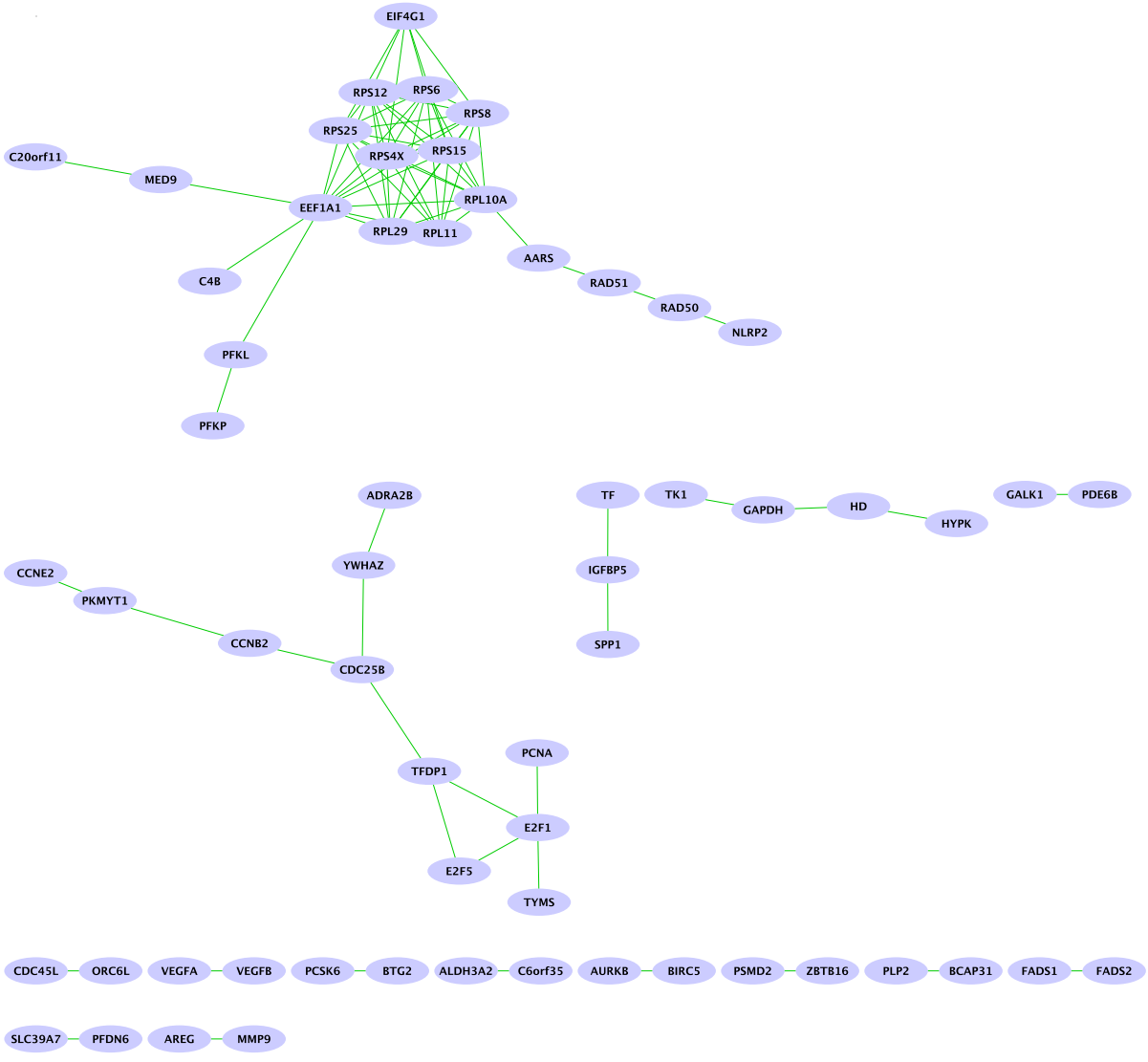
$$\Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,. \quad \text{(Jacob et al., 2009)}$$

Lasso signature (accuracy 0.61)



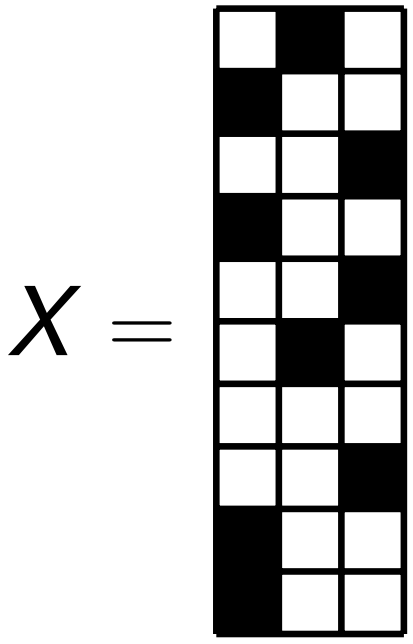
Breast cancer prognosis

Graph Lasso signature (accuracy 0.64)



Breast cancer prognosis

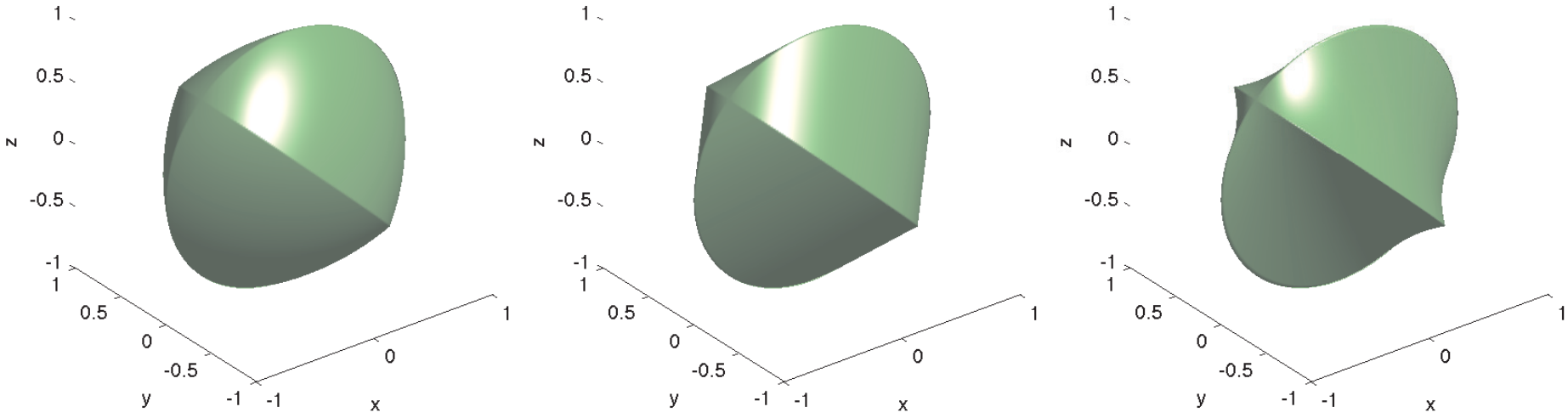
Disjoint feature selection



- Motivation: multiclass or multitask classification problems where we want to select features specific to each class or task
- Example: recognize identify and emotion of a person from an image (Romera-Paredes et al., 2012), or hierarchical coarse-to-fine classifier (Xiao et al., 2011; Hwang et al., 2011)

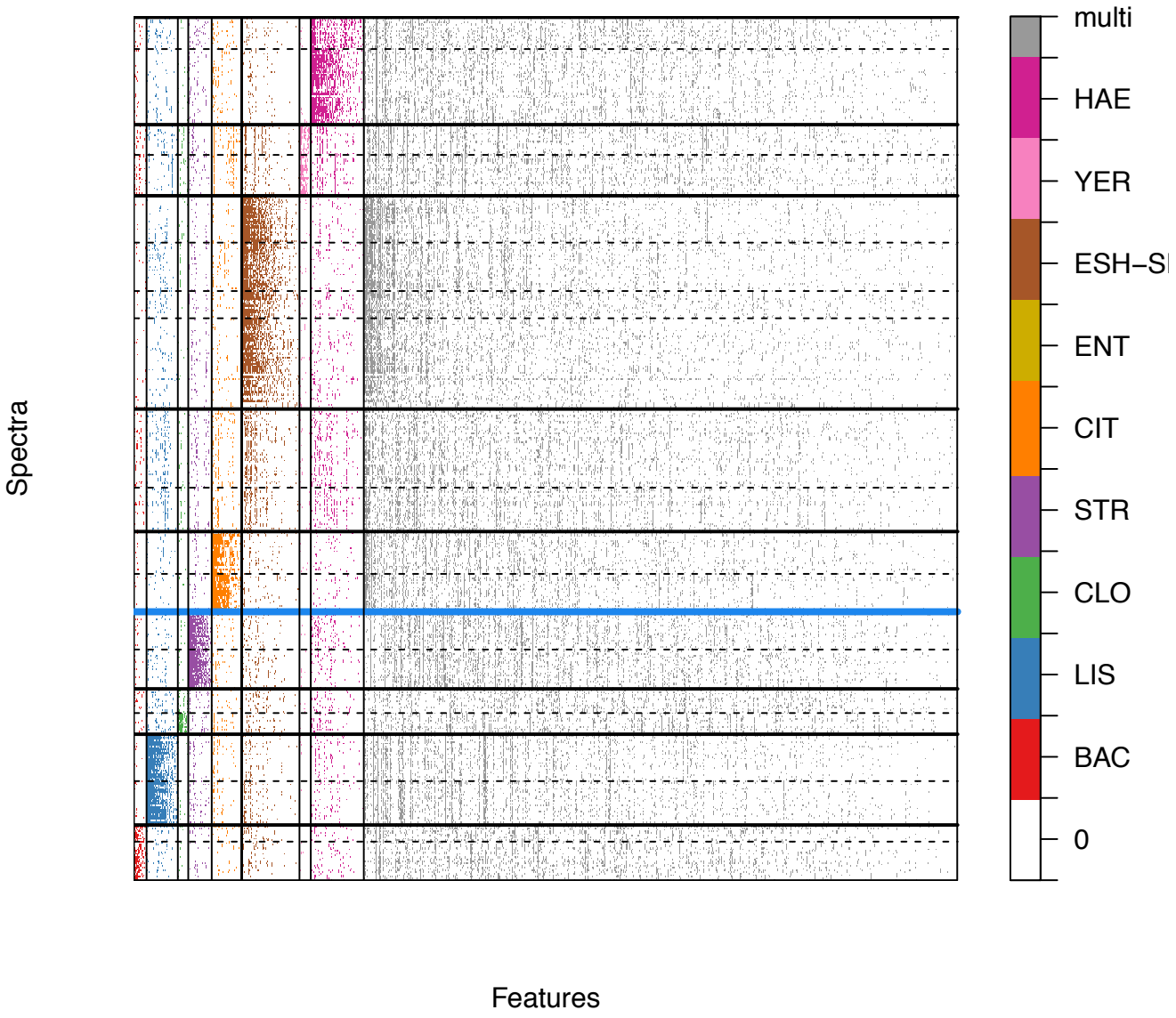
Disjoint feature selection

$$W = (w_i)_{i \in V} \in \mathbb{R}^{p \times V} \qquad \Omega(W) = \min_{-H \leq W \leq H} \sum_{i \sim j} K_{ij} \left| h_i^\top h_j \right|$$



(Vervier et al., 2014)

Example: multiclass classification of MS spectra



(Vervier et al, 2014)

Outline

1 Learning molecular classifiers with network information

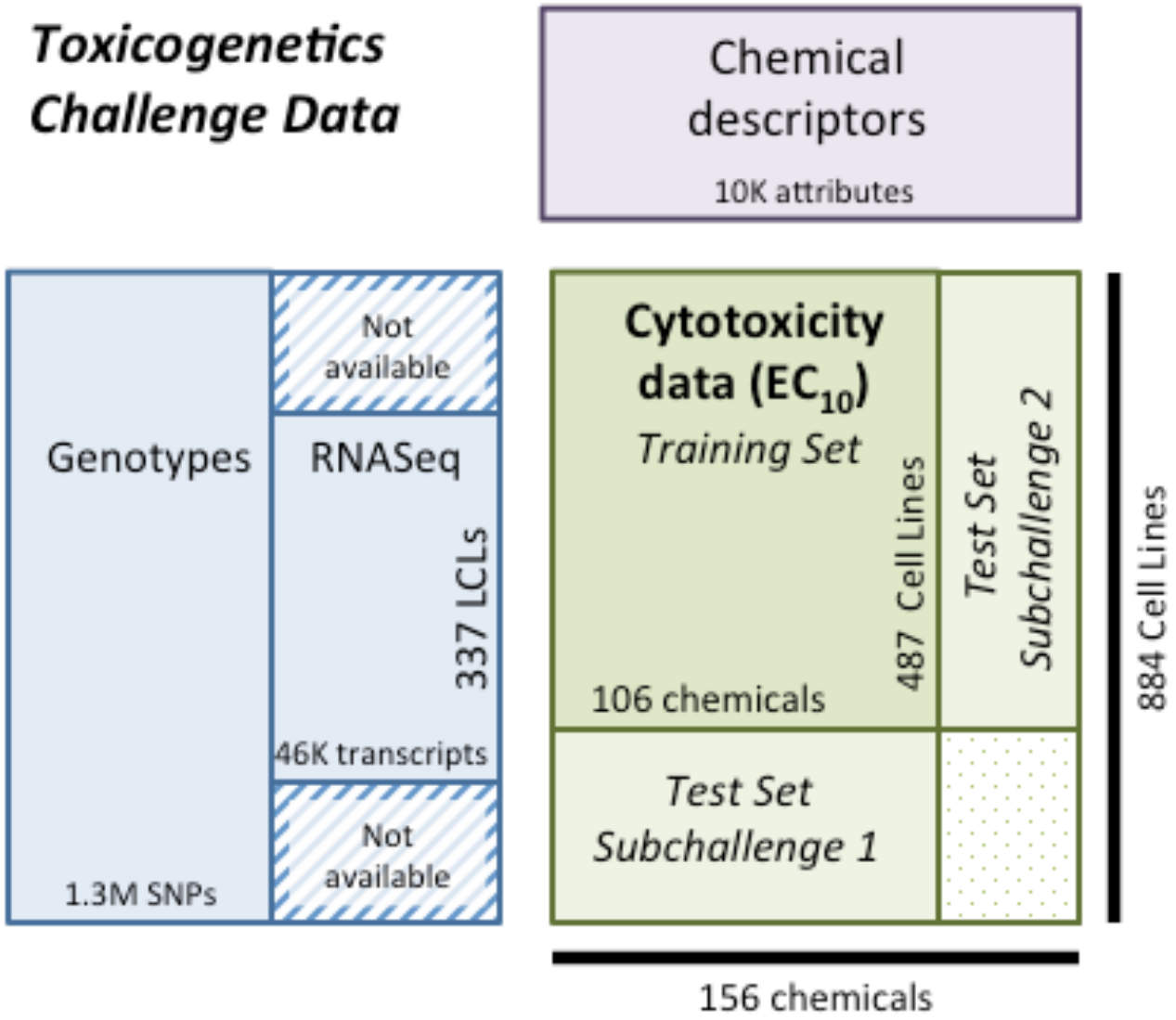
2 Kernel bilinear regression for toxicogenomics

Joint work with...

Elsa Bernard, Erwan Scornet, Yunlong Jiao, Véronique Stoven, Thomas Walter

Pharmacogenomics / Toxicogenomics

DREAM8 Toxicogenetics challenge



Genotypes from the 1000 genome project
RNASeq from the Geuvadis project

Bilinear regression

- Cell line X, chemical Y, toxicity Z.
- Bilinear regression model:

$$Z = f(X, Y) + b(Y) + \epsilon\,,$$

- Estimation by kernel ridge regression:

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}^p} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(f(x_i, y_j) + b_j - z_{ij} \right)^2 + \lambda \|f\|^2\,,$$

Solving in $O(max(n,p)^3)$

Theorem 1. *Let $Z \in \mathbb{R}^{n\times p}$ be the response matrix, and $K_X \in \mathbb{R}^{n\times n}$ and $K_Y \in \mathbb{R}^{p\times p}$ be the kernel Gram matrices of the n cell lines and p chemicals, with respective eigenvalue decompositions $K_X = U_X D_X U_X^\top$ and $K_Y = U_Y D_Y U_Y^\top$. Let $\gamma = U_X^\top \mathbf{1}_n$ and $S \in \mathbb{R}^{n\times p}$ be defined by $S_{ij} = 1/\left(\lambda + D_X^i D_Y^j\right)$, where D_X^i (resp. D_Y^i) denotes the i-th diagonal term of D_X (resp. D_Y). Then the solution (f^*, b^*) of (2) is given by*

$$b^* = U_Y Diag\left(S^\top \gamma^{\circ 2}\right)^{-1}\left(S^\top \circ \left(U_Y^\top Z^\top U_X\right)\right)\gamma \tag{3}$$

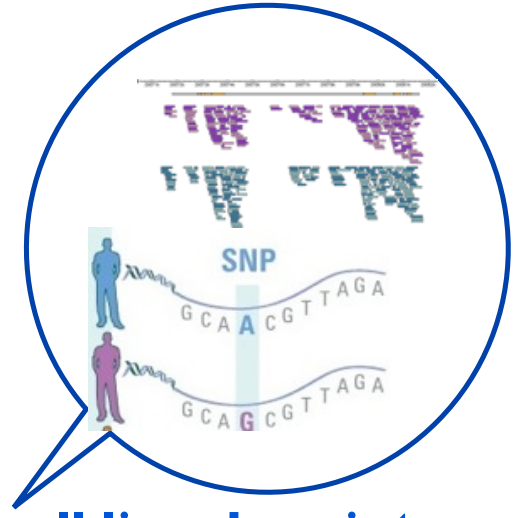
and

$$\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad f^*(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{p} \alpha_{i,j}^* K_X(x_i,x) K_Y(y_i,y)\,, \tag{4}$$

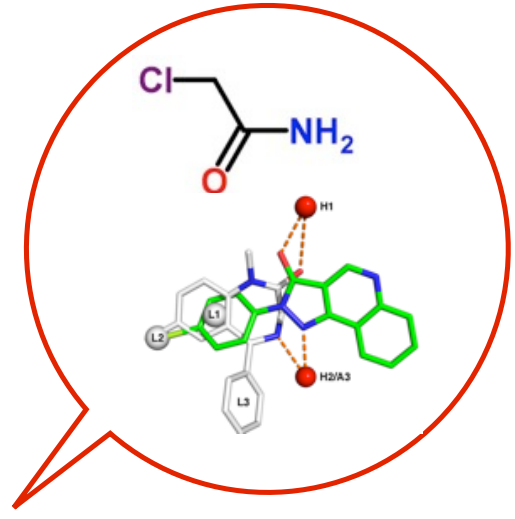
where

$$\alpha^* = U_X\left(S \circ \left(U_X^\top\left(Z - \mathbf{1}_n b^{*\top}\right)U_Y\right)\right)U_Y^\top\,. \tag{5}$$

Kernel Trick

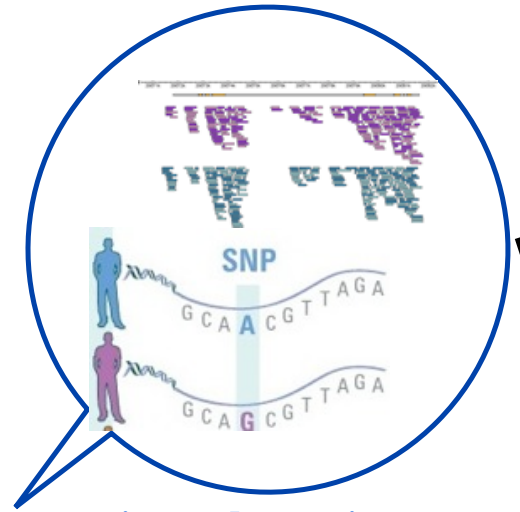


cell line descriptors

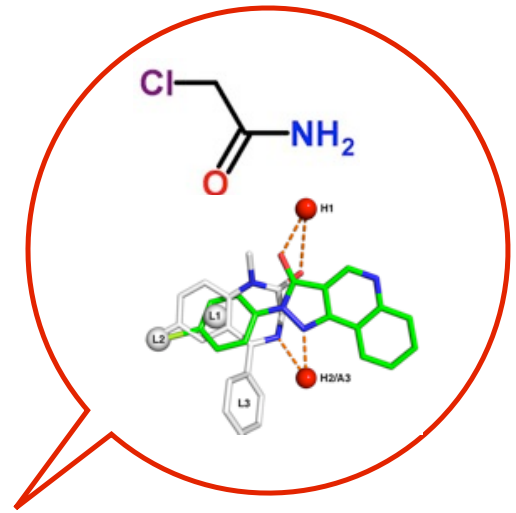


drug descriptors

Kernel Trick



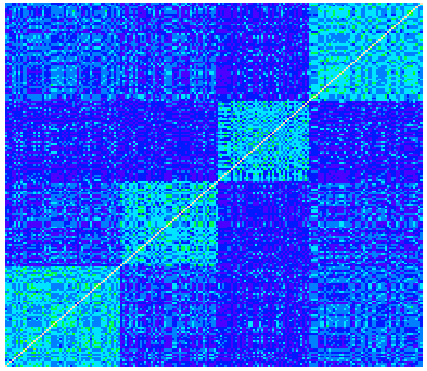
cell line descriptors



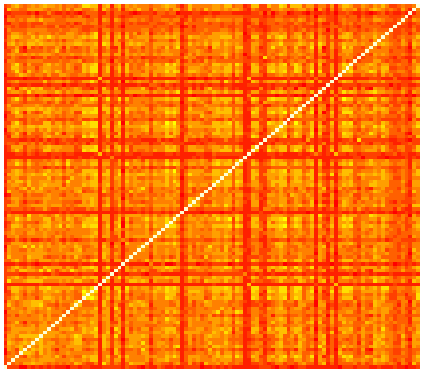
drug descriptors

kernelized →

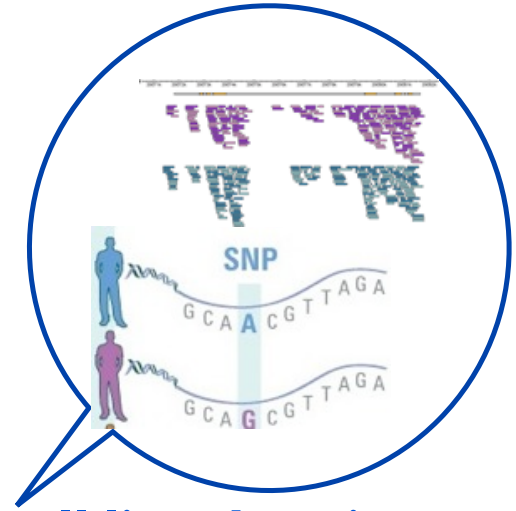
Kcell



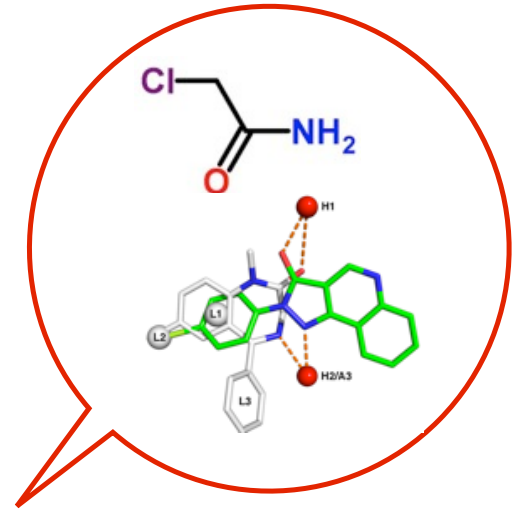
Kdrug



Kernel Trick



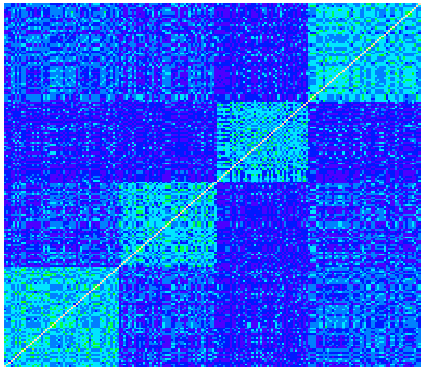
cell line descriptors



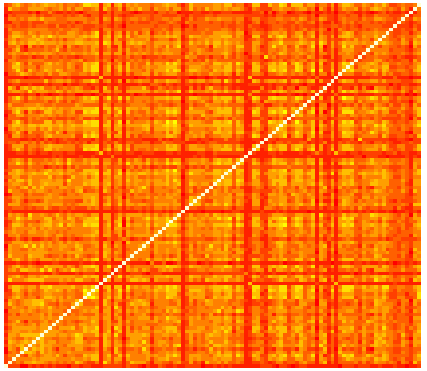
drug descriptors

kernelized →

Kcell

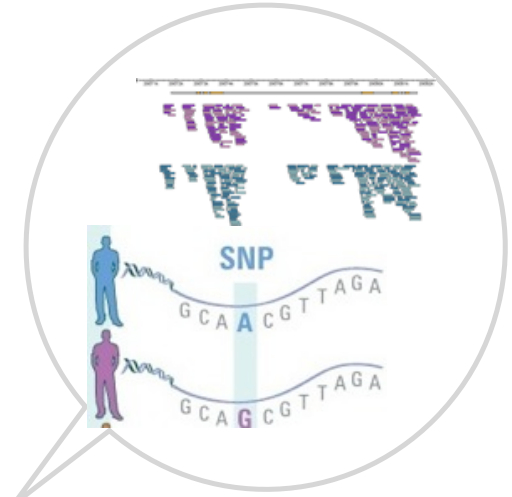


Kdrug



kernel bilinear regression → $\hat{f}$

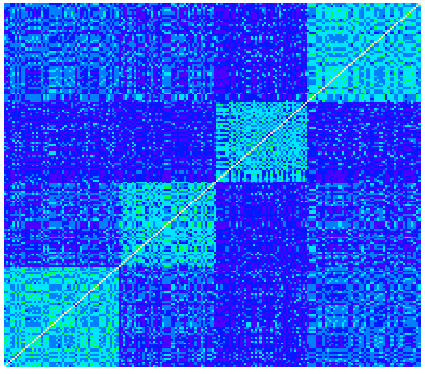
Kernel Trick



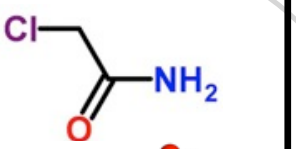
cell line descriptors

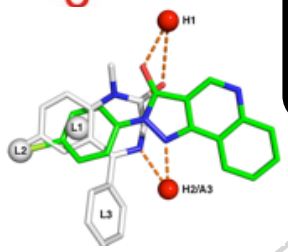
kernelized →

Kcell



kernel bilinear regression → $\hat{f}$



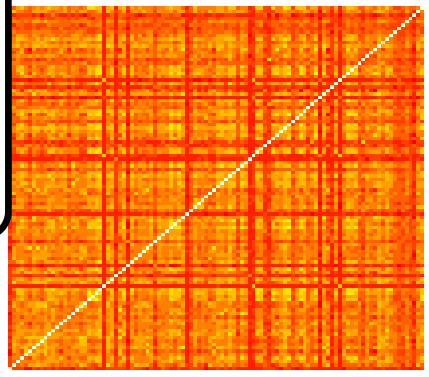


drug descriptors

Kernel choice?

. descriptors
. data integration
. missing data

Kdrug



Kernel choice

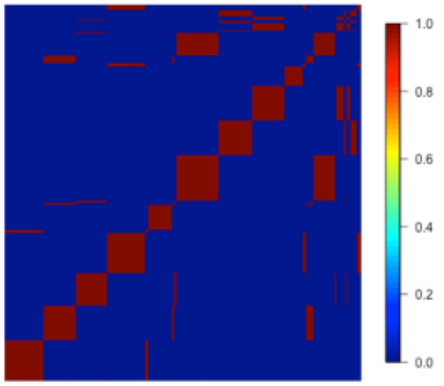
1. $\mathbf{K}_{\text{cell}}$:
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. $\mathbf{K}_{\text{drug}}$:
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Kernel choice

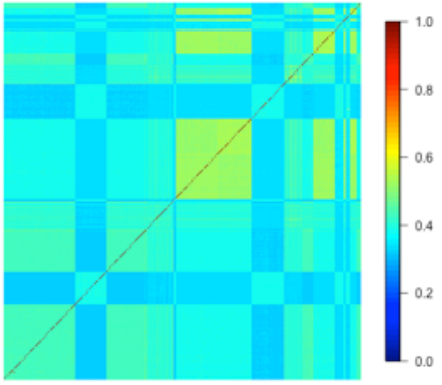
1. $\mathbf{K}_{\text{cell}}$:
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. $\mathbf{K}_{\text{drug}}$:
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Cell line data integration

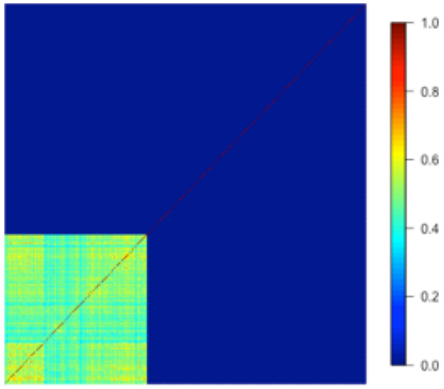
Covariates
. linear kernel



SNPs
. 10 gaussian kernels

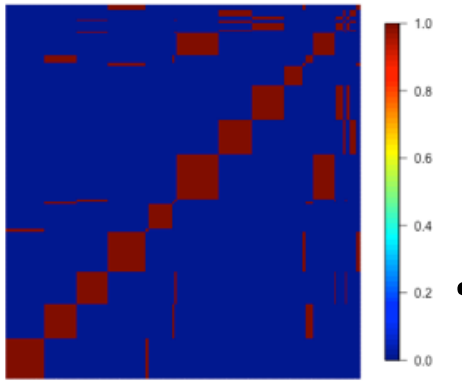


RNA-seq
. 10 gaussian kernels

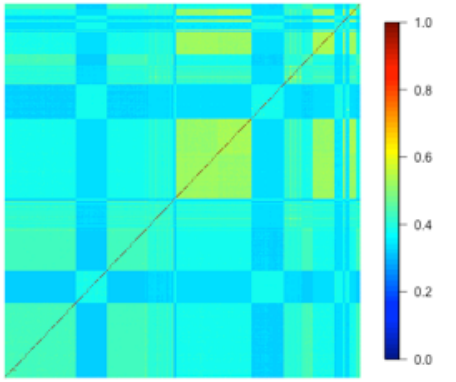


Cell line data integration

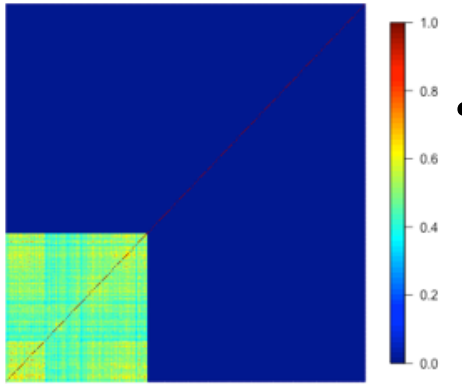
Covariates
. linear kernel



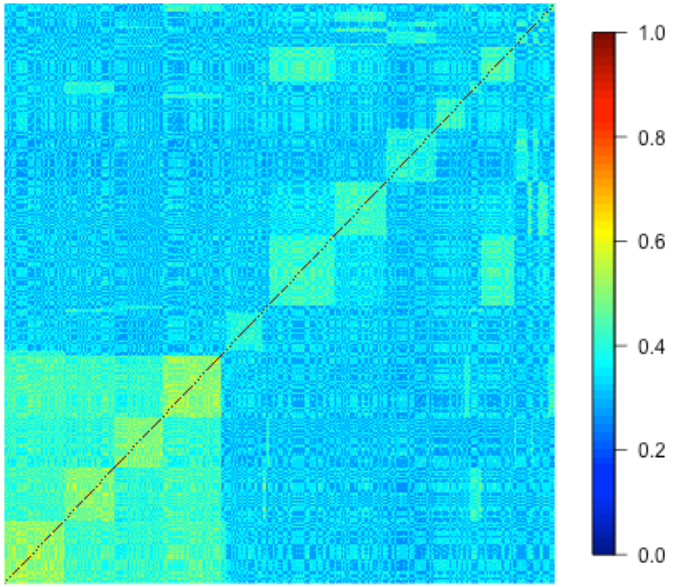
SNPs
. 10 gaussian kernels



RNA-seq
. 10 gaussian kernels

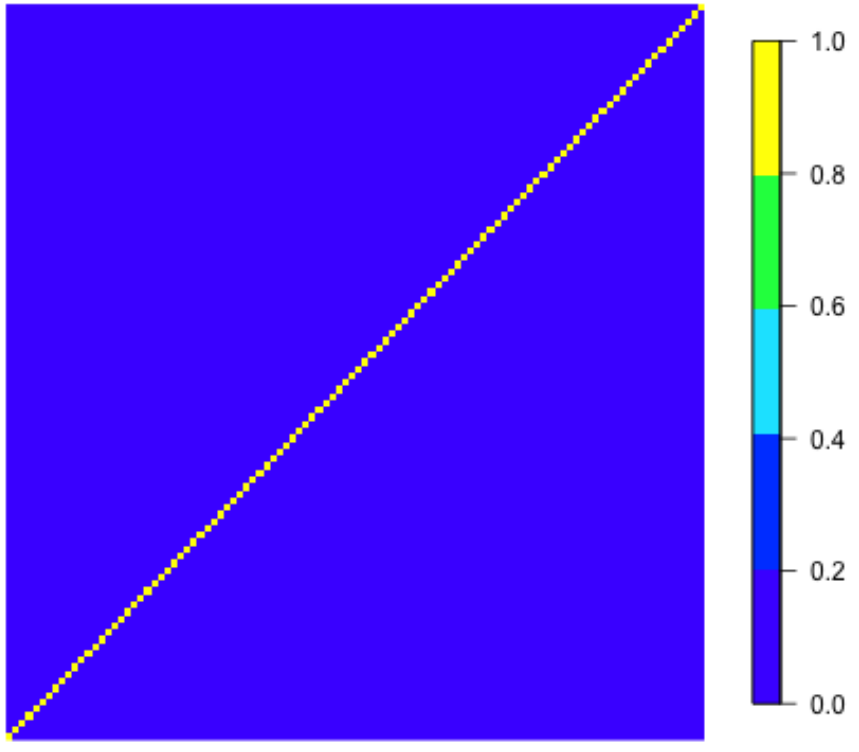


Integrated kernel



Multi-task drug kernels

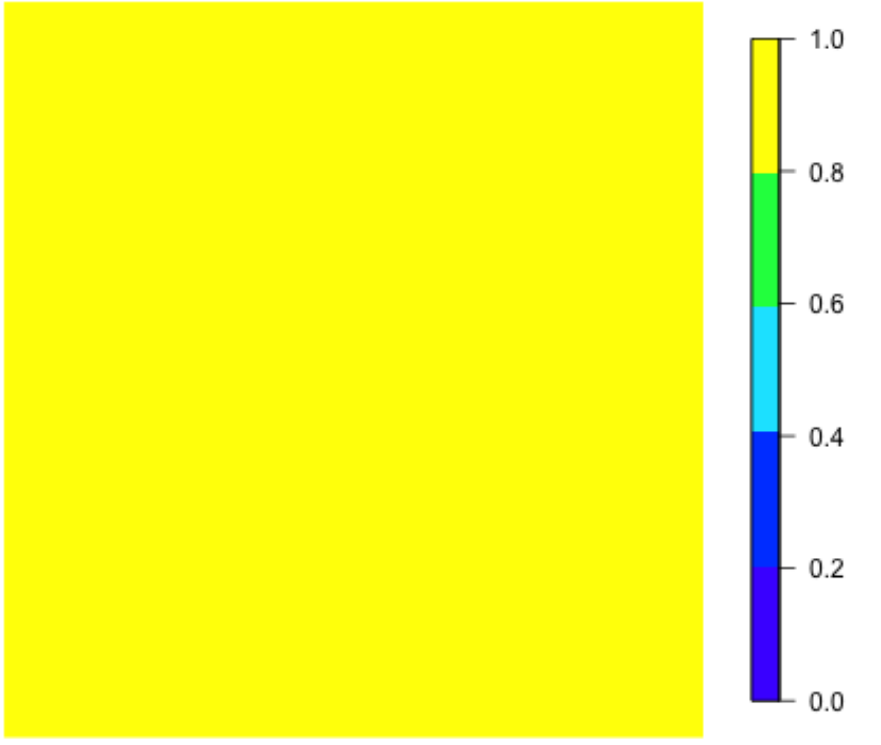
1. **Dirac**
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated



independent regression for each drug

Multi-task drug kernels

1. Dirac
2. **Multi-Task**
3. Feature-based
4. Empirical
5. Integrated



sharing information across drugs

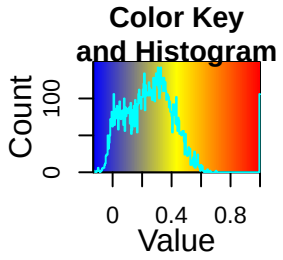
Multi-task drug kernels

1. Dirac
2. Multi-Task
3. **Feature-based**
4. Empirical
5. Integrated

Linear kernel and 10 gaussian kernels based on features:

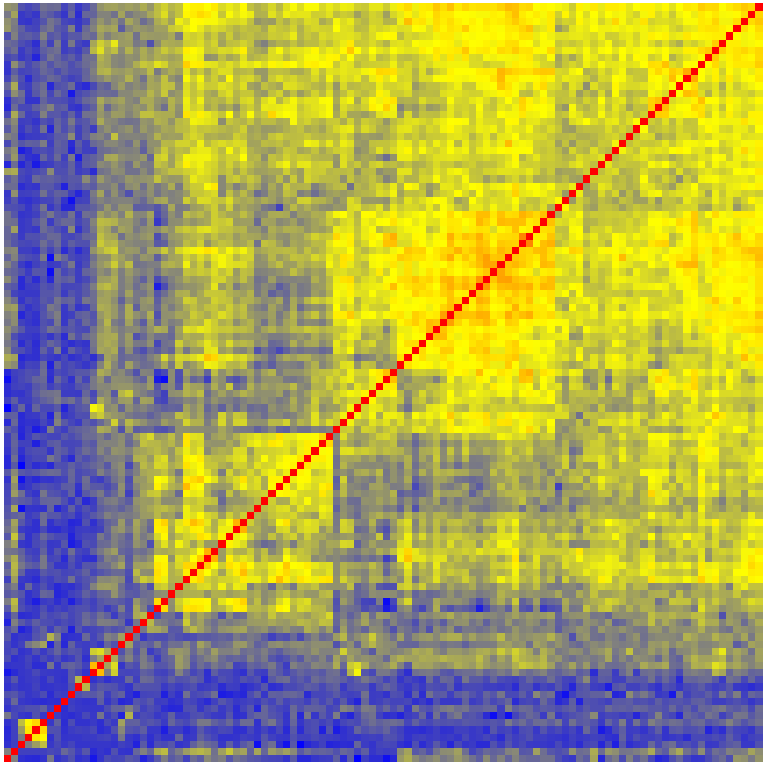
- CDK (160 descriptors) and SIRMS (9272 descriptors)
- Graph kernel for molecules (2D walk kernel)
- Fingerprint of 2D substructures (881 descriptors)
- Ability to bind human proteins (1554 descriptors)

Multi-task drug kernels



Empirical correlation

1. Dirac
2. Multi-Task
3. Feature-based
4. **Empirical**
5. Integrated



Multi-task drug kernels

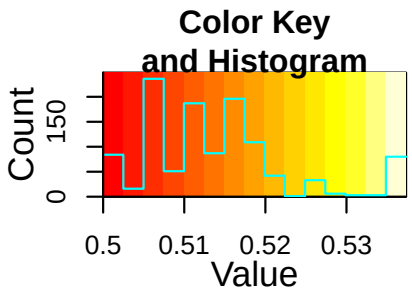
1. Dirac
2. Multi-Task
3. Feature-based
4. Empirical
5. **Integrated**

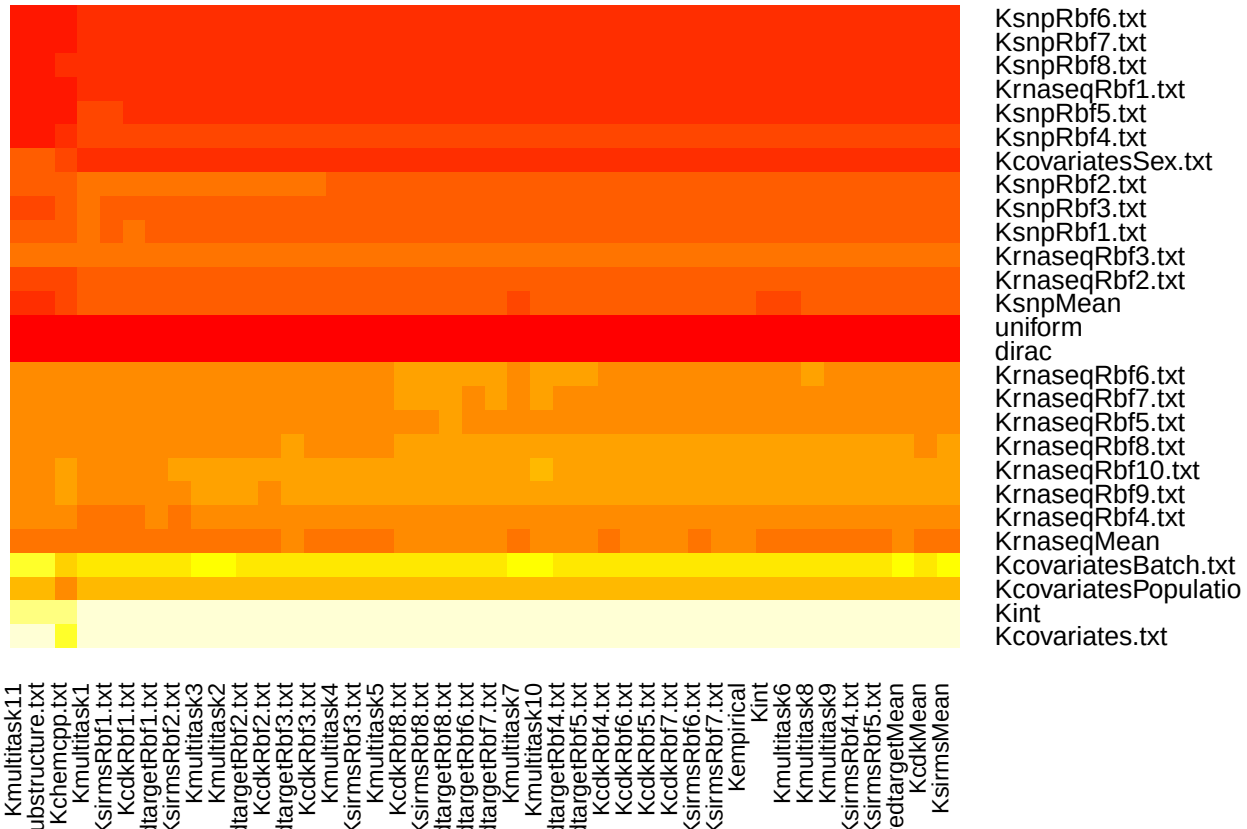
$$K_{int} = \sum_i K_i$$

Integrated kernel:

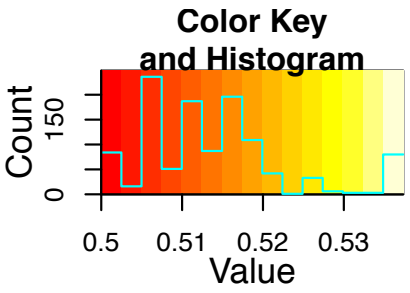
- Combine all information on drugs

29x48 kernel combinations: CV results

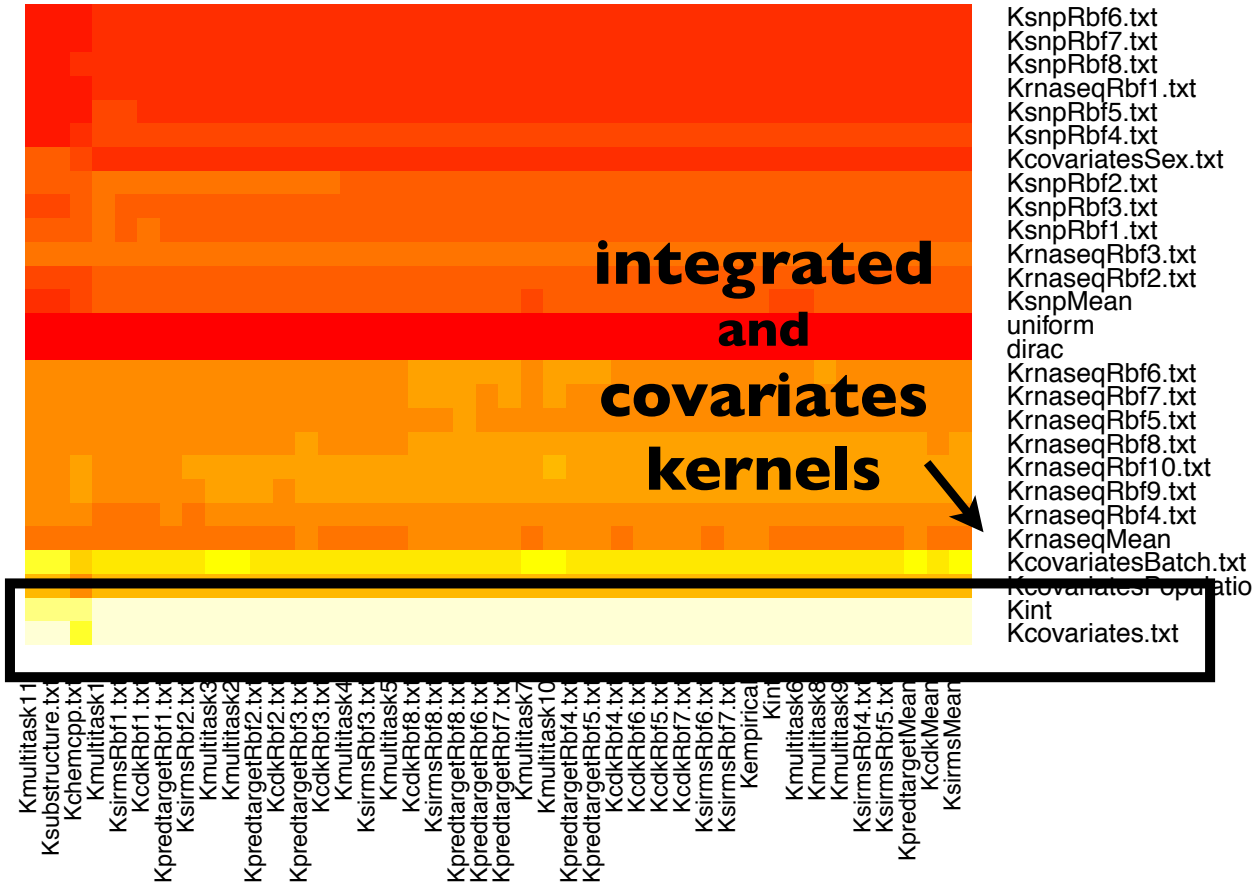




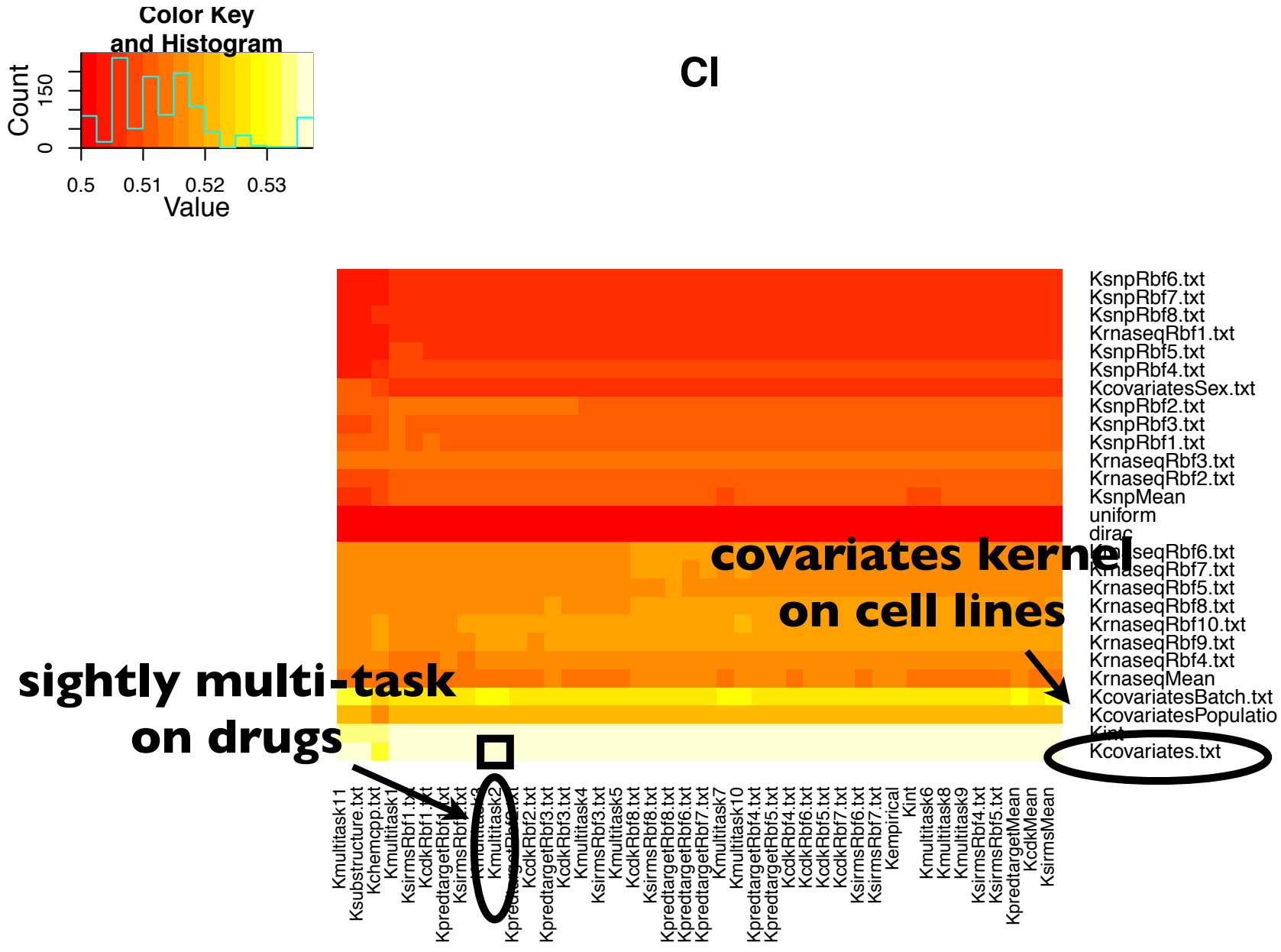
29x48 kernel combinations: CV results



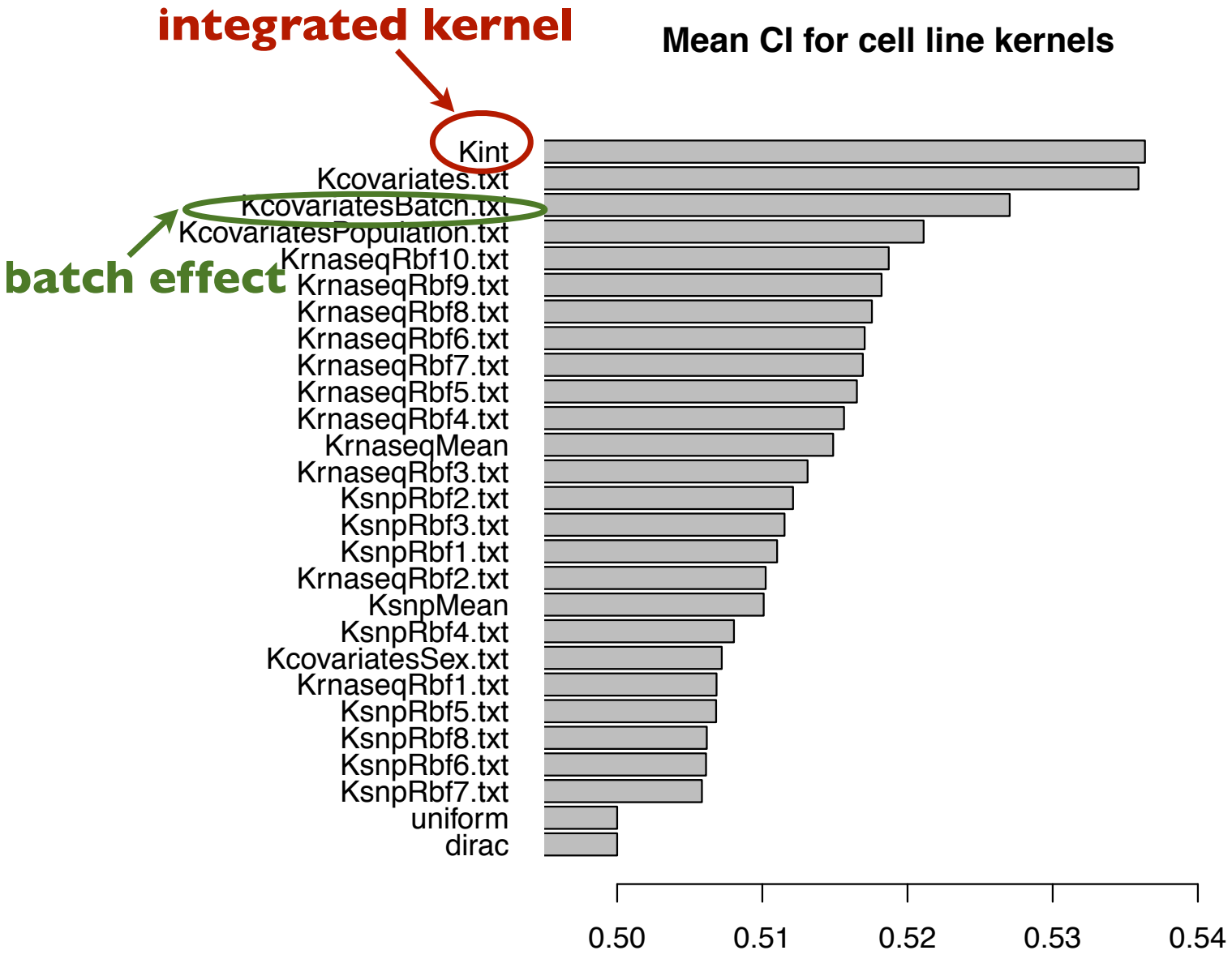
CI



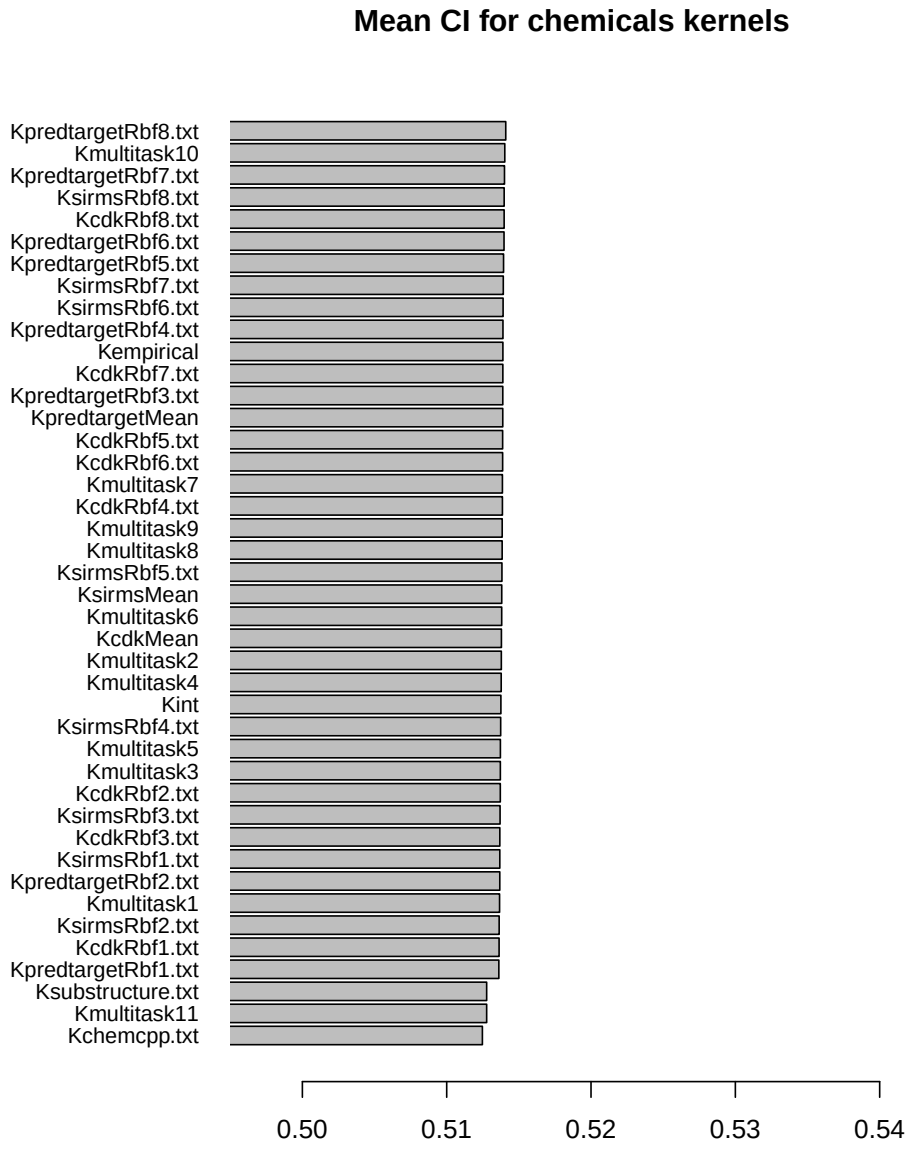
29x48 kernel combinations: CV results



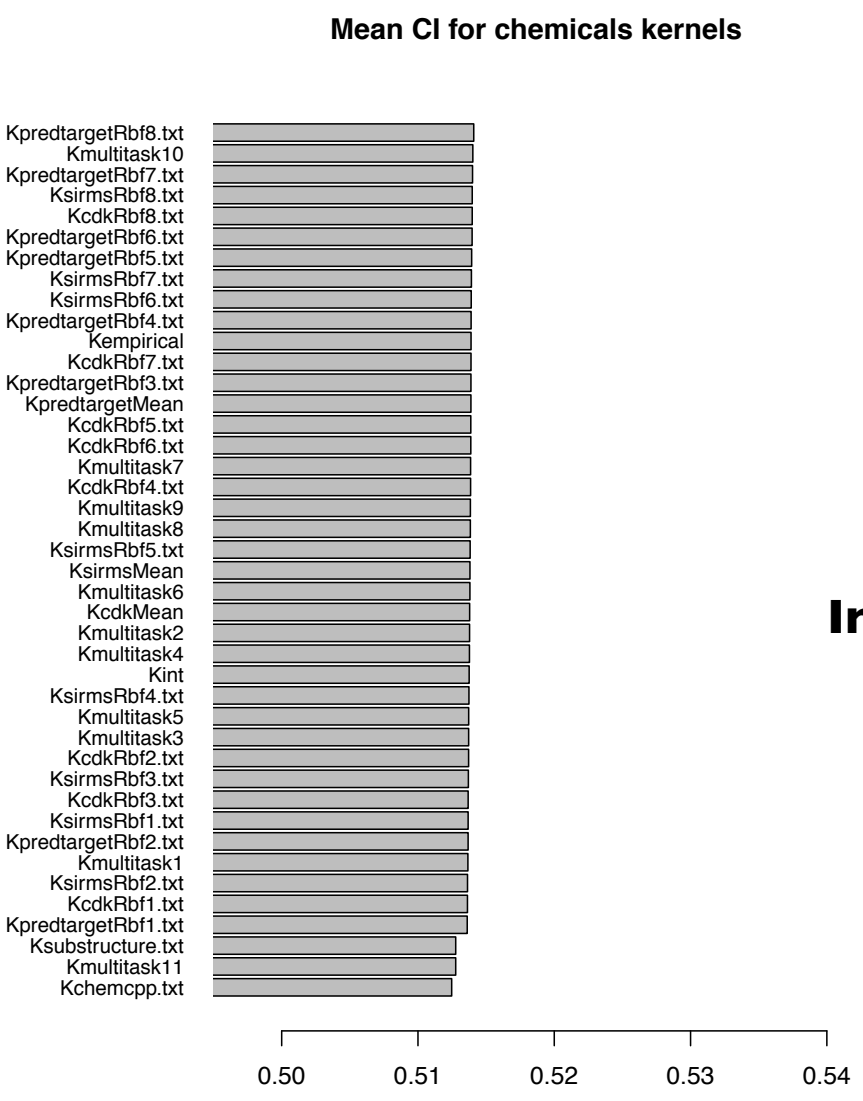
Kernel on cell lines: CV results



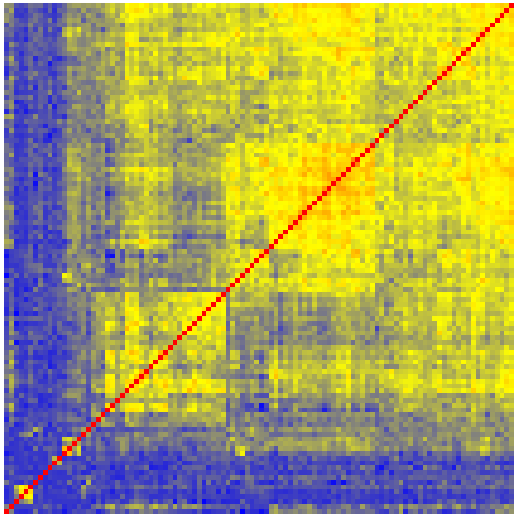
Kernel on drugs: CV results



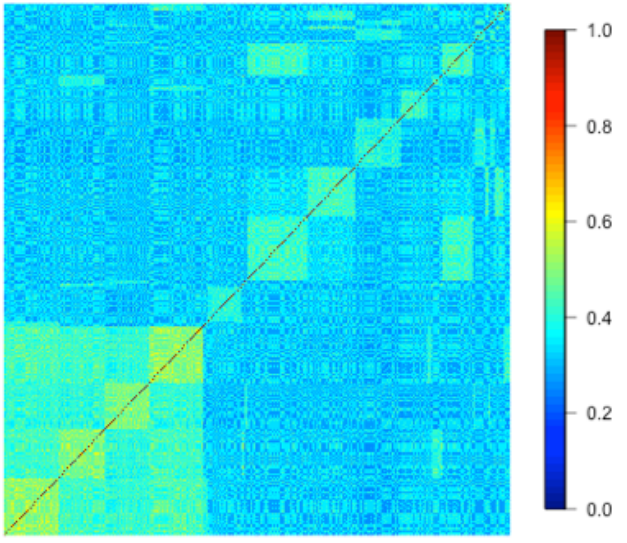
Final Submission (ranked 2nd)



Empirical kernel on drugs



Integrated kernel on cell lines



Conclusion

- Many new problems and lots of data in computational genomics
- Computational constraints $\Longrightarrow$ fast sparse models (FlipFlop)
- Small n large p $\Longrightarrow$ regularized models with prior knowledge
- Heterogeneous data integration $\Longrightarrow$ kernel methods
- Personalized medicine promising but difficult!

Thanks

institutCurie

European Research Council

erc

Inserm

Institut national
de la santé et de la recherche médicale

MINES
ParisTech

References I

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 6(12):e28210.

Hwang, S. J. J., Grauman, K., and Sha, F. (2011). Learning a tree of metrics with disjoint visual features. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 621–629.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA. ACM.

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824.

Rapaport, F., Barillot, E., and Vert, J.-P. (2008). Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382.

Rapaport, F., Zynoviev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35.

Romera-Paredes, B., Argyriou, A., Berthouze, N., and Pontil, M. (2012). Exploiting unrelated tasks in multi-task learning. *J. Mach. Learn. Res. - Proceedings Track,*, 22:951–959.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536.

References II

Vervier, K., Mahé, P., DâĂŹAspremont, A., Veyrieras, J.-B., and Vert, J.-P. (2014). On learning matrices with orthogonal columns or disjoint supports. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8726 of *Lecture Notes in Computer Science*, pages 274–289. Springer Berlin Heidelberg.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D., and Foekens, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, 365(9460):671–679.

Xiao, L., Zhou, D., and Wu, M. (2011). Hierarchical classification via orthogonal transfer. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011.*, pages 801–808. Omnipress.