*Data Science Summer School 2018*

# Machine learning for genetic data

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines ParisTech – Institut Curie – INSERM U900
PSL Research University, Paris, France

June 28, 2018

http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

# Precision Medicine

- **Adapt** treatment to the **genetic specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

# Precision Medicine

- **Adapt** treatment to the **genetic specificities** of the patient.

  E.g. Trastuzumab for HER2+ breast cancer.

- **Data-driven** biology/medicine

  Identify **similarities** between patients that exhibit similar phenotypes.

# Precision Medicine

- **Adapt** treatment to the **genetic specificities** of the patient.

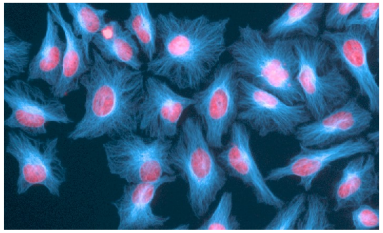
  E.g. Trastuzumab for HER2+ breast cancer.

- **Data-driven** biology/medicine

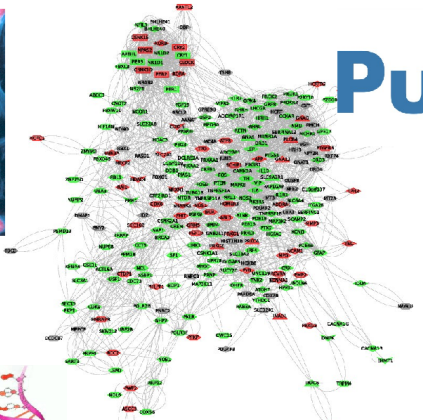
  Identify **similarities** between patients that exhibit similar phenotypes.

**Data** + **Feature Selection**

# Accumulating data



phenome

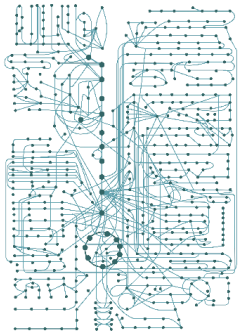


interactome

PubMed

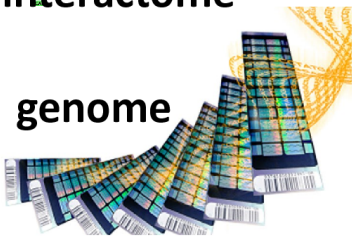
publications


transcriptome

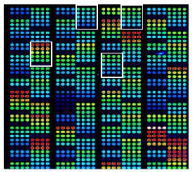


metabolome

methylome



genome



proteome

Image sources: ajc1@ flickr; Zlir'a@wikimedia

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

biobank uk
Improving the health of future generations

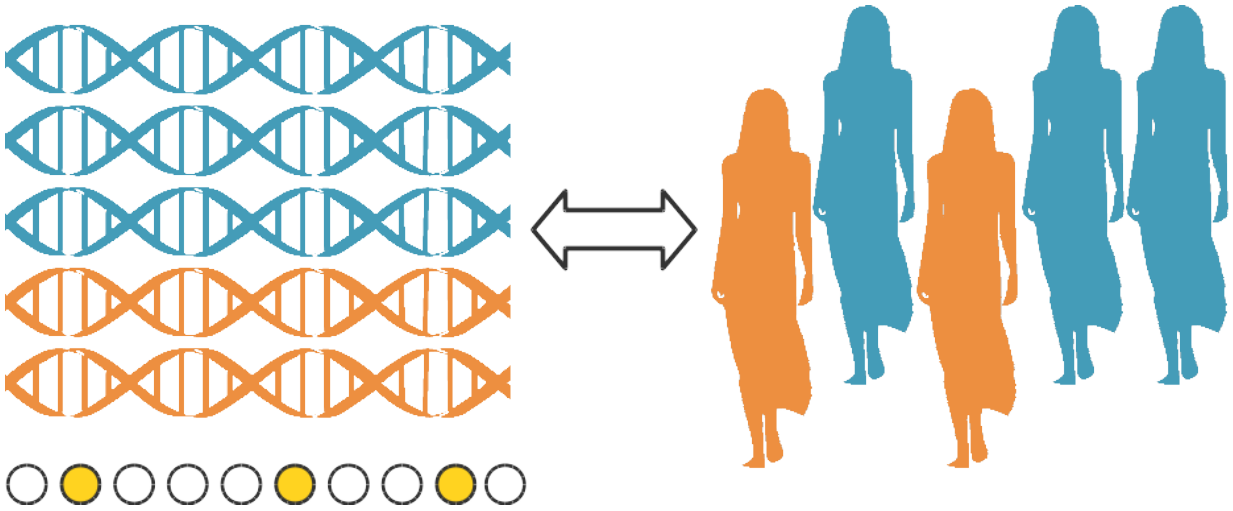
db GaP
GENOTYPES and PHENOTYPES

DREAM Challenges

*Ceci n'est pas* **du Big Data**

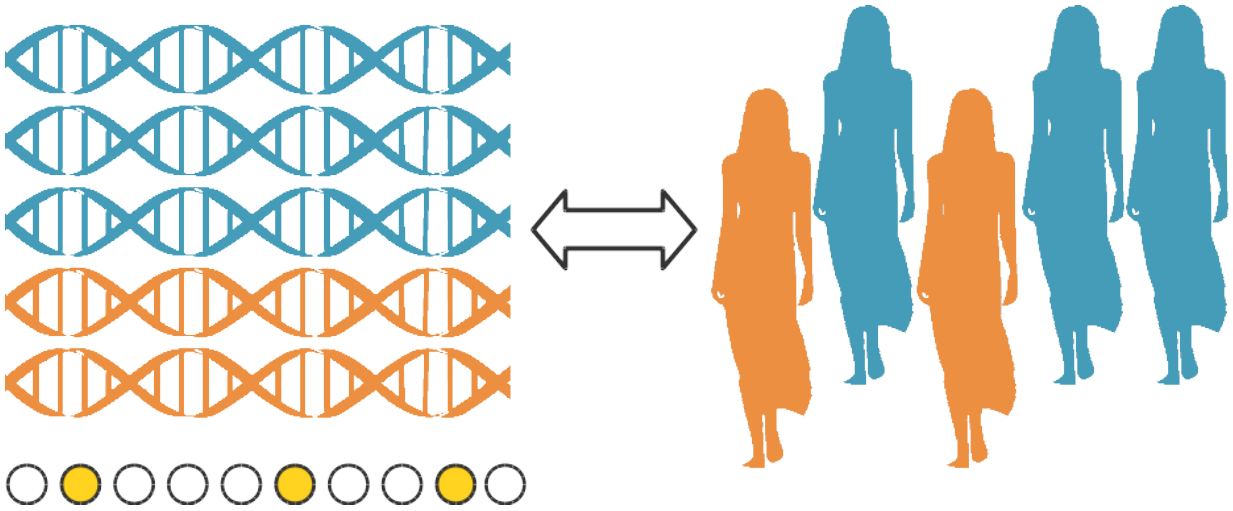
Magritte

# From genotype to phenotype



**Which genomic features explain the phenotype?**

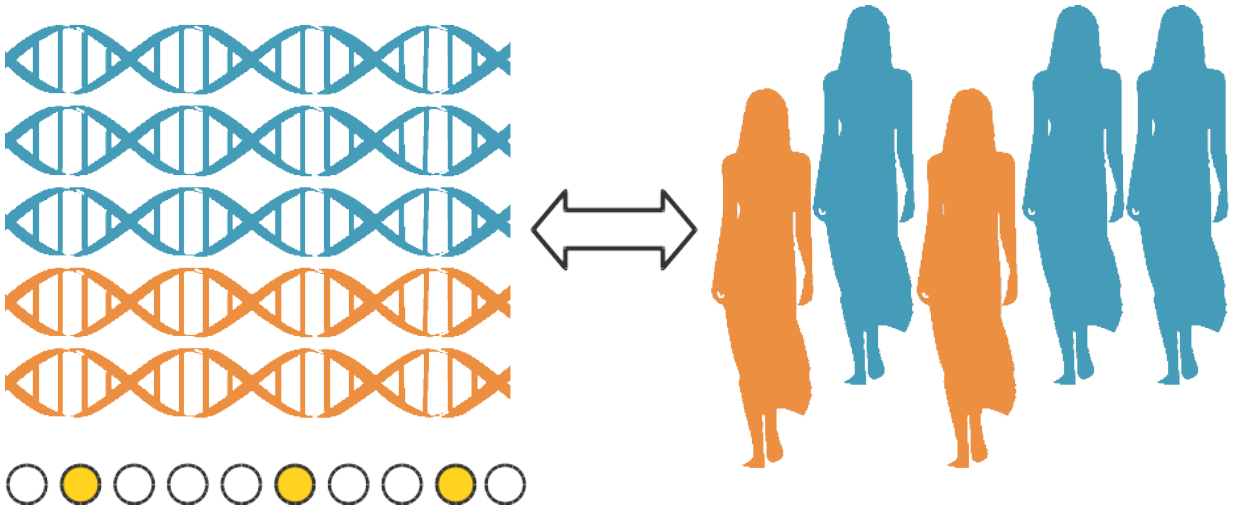
# From genotype to phenotype



**Which genomic features explain the phenotype?**

- 80 000 proteins
- 200 000 mRNA
- 10 million SNPs
- 28 million CpG islands

# From genotype to phenotype



**Which genomic features explain the phenotype?**

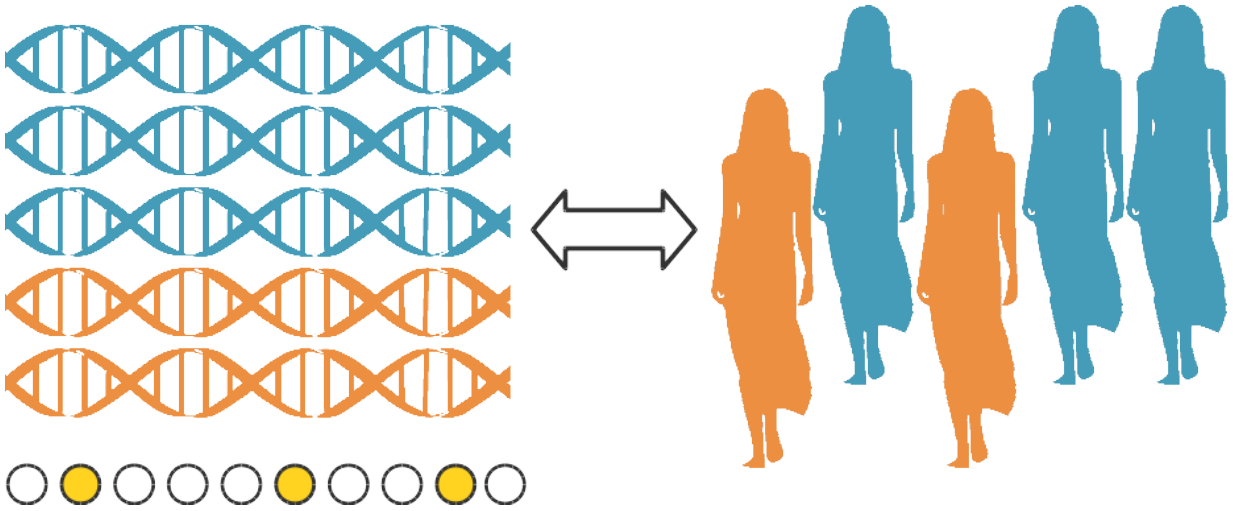
$p = 10^5 - 10^7$ **genomic features**

- 80 000 proteins
- 200 000 mRNA

$n = 10^3 - 10^5$ **samples**

- 10 million SNPs
- 28 million CpG islands

# From genotype to phenotype



## Which genomic features explain the phenotype?

p = $10^5$ – $10^7$ **genomic features**

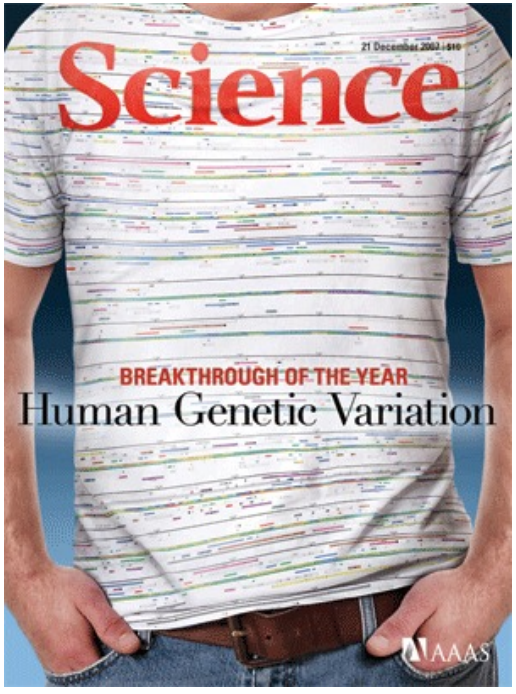
- 80 000 proteins
- 200 000 mRNA

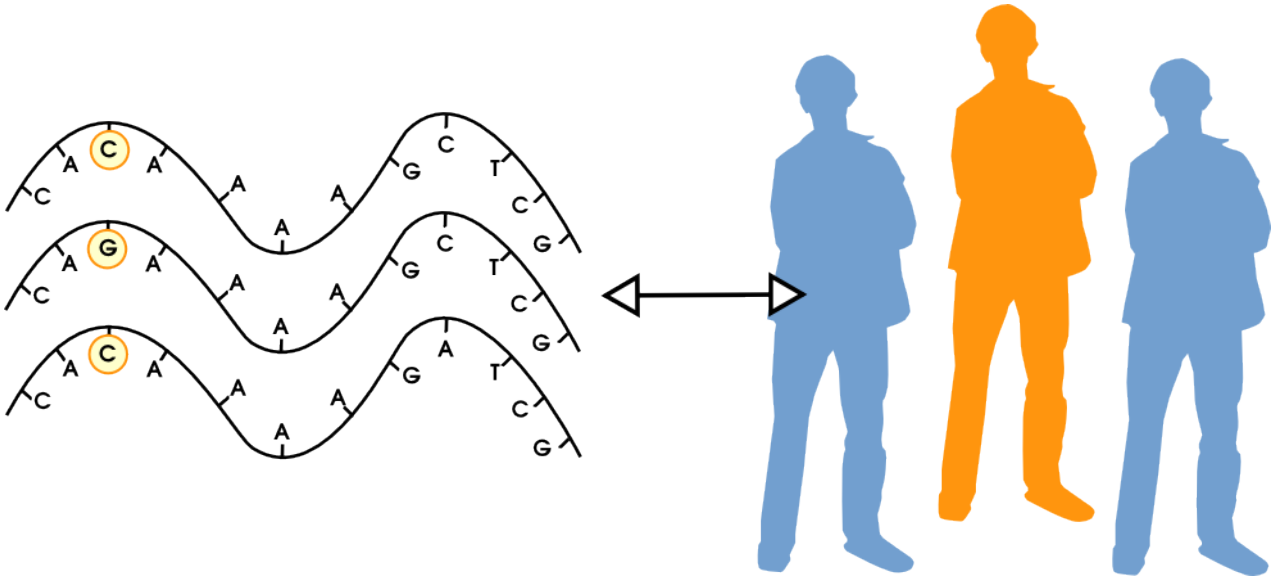
n = $10^3$ – $10^5$ **samples**

- 10 million SNPs
- 28 million CpG islands

**High-dimensional** (large p), **low sample size** (small n) data.

# GWAS: Genome-Wide Association Studies



**Which genomic features explain the phenotype?**

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)

$n = 10^2 - 10^4$ samples

[Pen07; Man13; Vis+17]

# 1. Challenges of high-dimensional data

# Challenges of high-dimensional data

- **Computational challenges** for some algorithms

  **Linear regression:** inverting $X^\top X$ takes $\mathcal{O}(p^3)$ computations.

- The **curse of dimensionality** makes it hard to learn

- **Overfitting** is more likely

- **Ill-posed problems.**

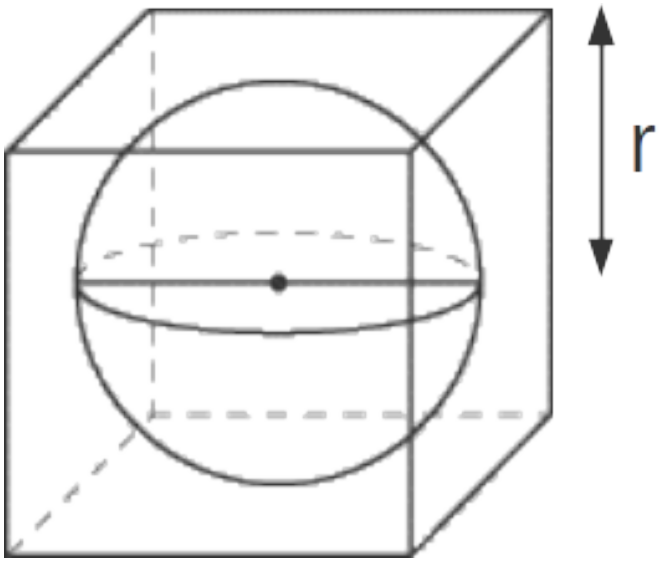
# The curse of dimensionality

- Methods and intuitions that work in low dimension **might not apply to higher dimensions.**

- **Hyperspace is very big and everything is far apart**

  Fraction of the points within a cube that fall outside the inscribed circle:

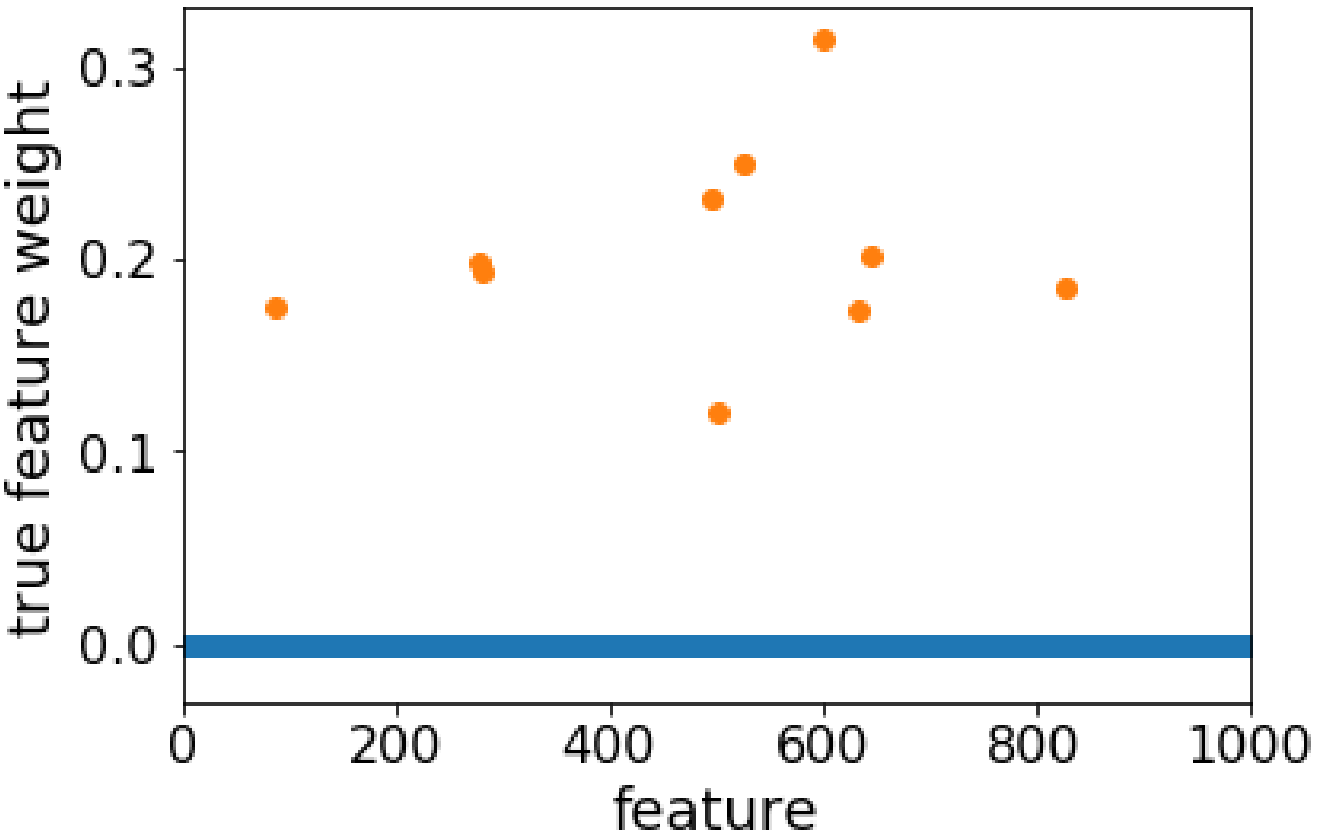
  - In two dimensions: $1 - \frac{\pi r^2}{4r^2} = 1 - \frac{\pi}{4}$
  - In three dimensions: $1 - \frac{4/3\pi r^3}{8r^3} = 1 - \frac{\pi}{6}$
  - In higher dimension: tends towards $1$.



# Large p, small n data

**Simulation:** n=150, p=1000, 10 causal features.

$$y = \sum_{j=1}^{p} w_j x_j + \epsilon$$



# Qualitative GWAS

**Binary** phenotype, i.e. case/controls encoded as 1/0.

- **Contingency table**

| | AA | Aa | aa |
|---|---|---|---|
| Cases | | | |
| Ctrls | | | |

| | 0 | 1 |
|---|---|---|
| Cases | a | b |
| Ctrls | c | d |

**Statistical tests:** $\chi^2$, Cochran-Armitage trend test, etc.

- **Logistic regression**

$$\text{logit}(p(y|X)) = \beta_0 + \beta_1 X$$

- **Odds-ratio**

$$\underbrace{\frac{\text{P(0|case)}}{\text{P(0|ctrl)}}}_{\text{odds of 0 in cases}} \Big/ \underbrace{\frac{\text{P(1|case)}}{\text{P(1|ctrl)}}}_{\text{odds of 1 in cases}} = \frac{ad}{bc}$$

# Quantitative GWAS

- **Linear regression**

$$y = \beta_0 + \beta_1 X$$

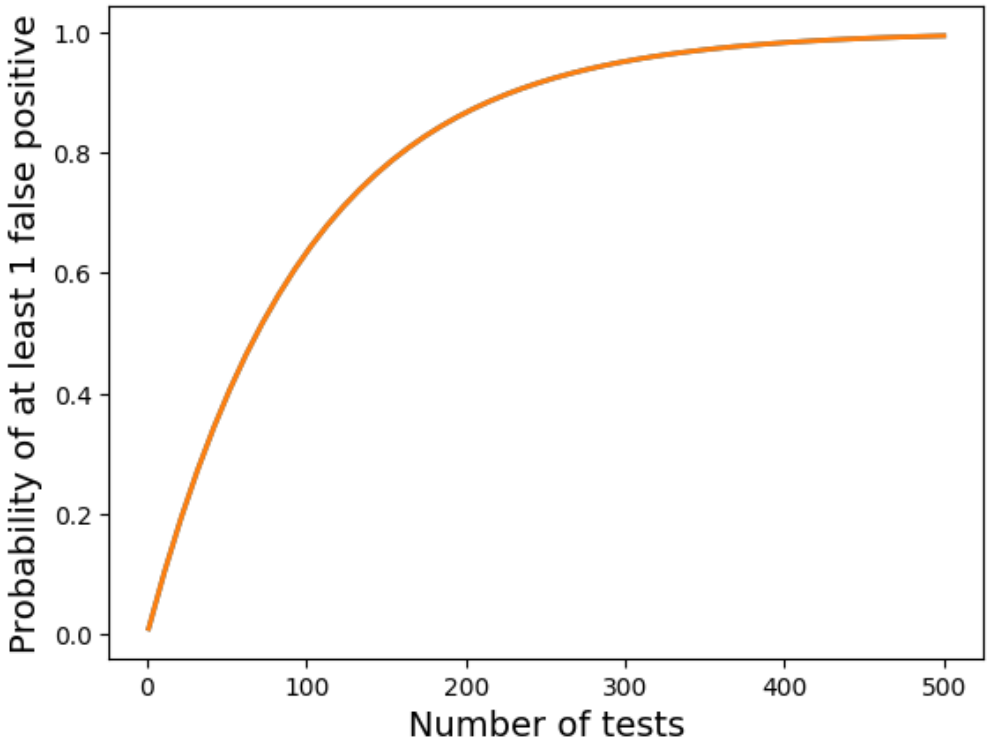
- p-value: Is $\hat{\beta}_1$ significantly different from 0?

  **Wald test:** compare $\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)}$ to a $\chi^2$ distribution.

- **Effect size:** $\beta_1$.

# Multiple Hypothesis Testing

- Probability of having **at least one false positive:**
  - For **one** test: $\alpha$
  - For **p** tests: $1 - (1 - \alpha)^p$




- Controlling **Family-Wise Error Rate** (FWER)

  $\text{FWER} = P(|\text{FP}| \geq 1)$

  FP = number of false positives (Type I errors)

- **Bonferroni** correction: $\alpha \rightarrow \frac{\alpha}{p}$

# Simulation

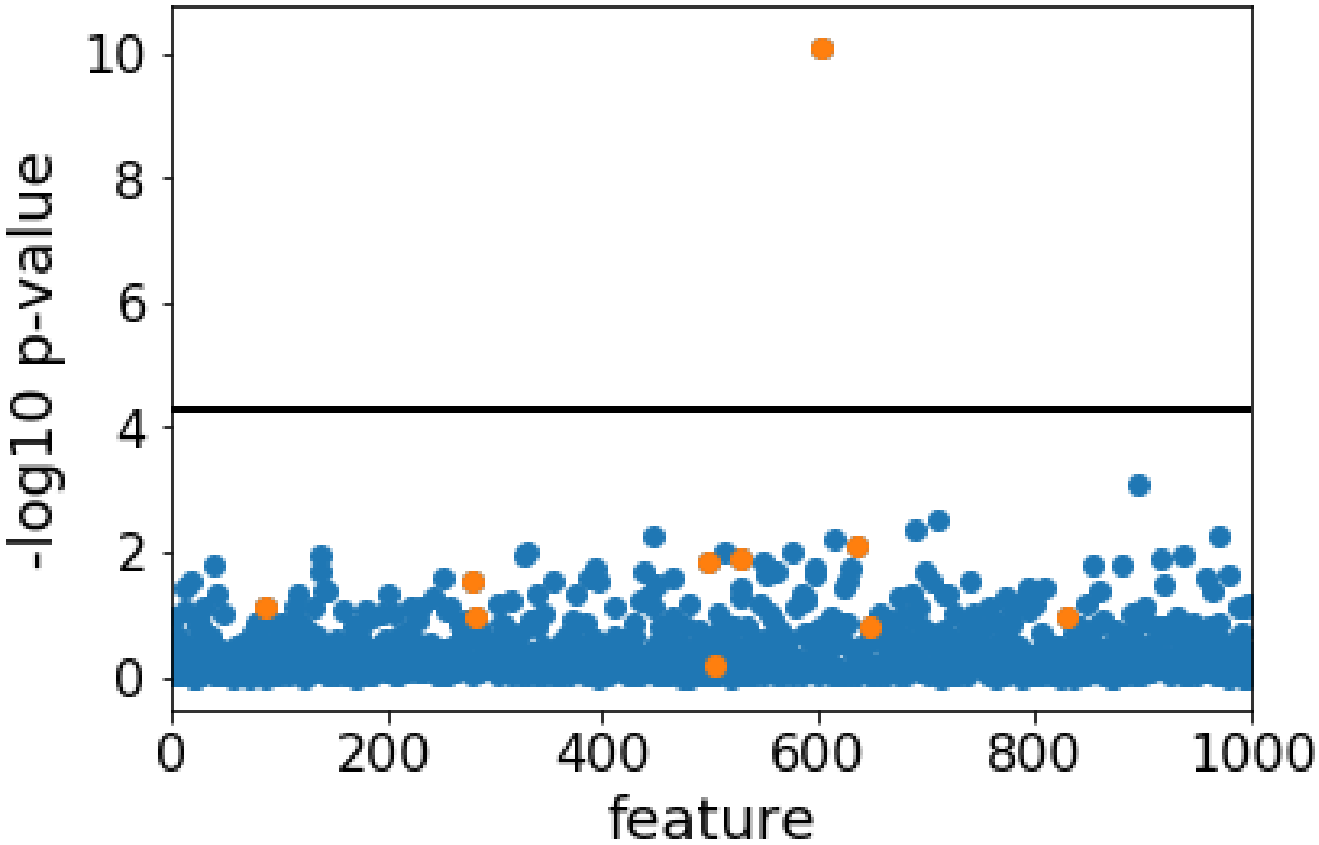
**t-test:** For each feature,

- fit $y \sim w_j x_j + b_j$
- test whether $w_j \neq 0$.

# Simulation

**t-test:** For each feature,

- fit $y \sim w_j x_j + b_j$
- test whether $w_j \neq 0$.



# Missing heritability

GWAS **fail to explain** most of the **inheritable variability** of complex traits.

Many possible reasons:

- non-genetic / non-SNP factors
- heterogeneity of the phenotype
- rare SNPs
- weak effect sizes
- **few samples in high dimension ($p \gg n$)**
- **joint effets of multiple SNPs.**

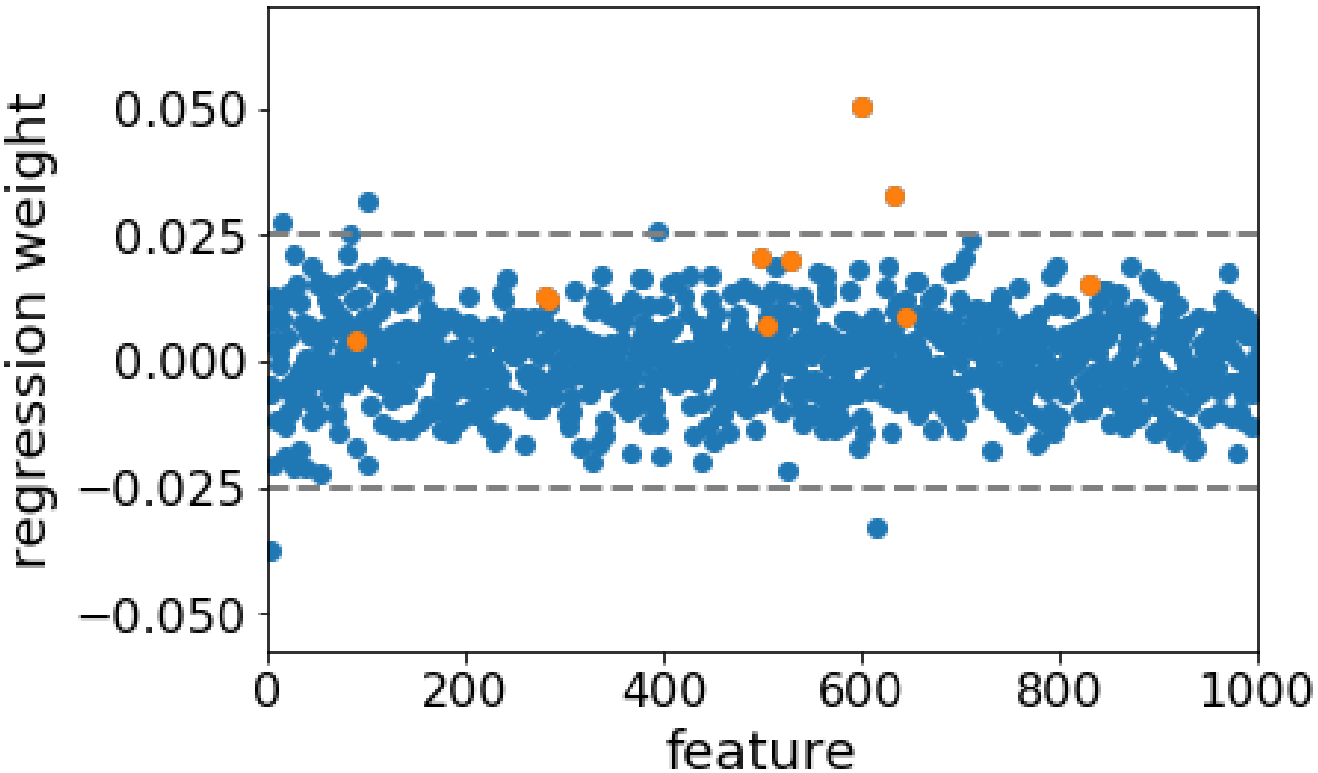
Ref: [Man+09]

# Simulation

**Linear regression:** Fit $y \sim \sum_{j=1}^{p} w_j x_j + b$.

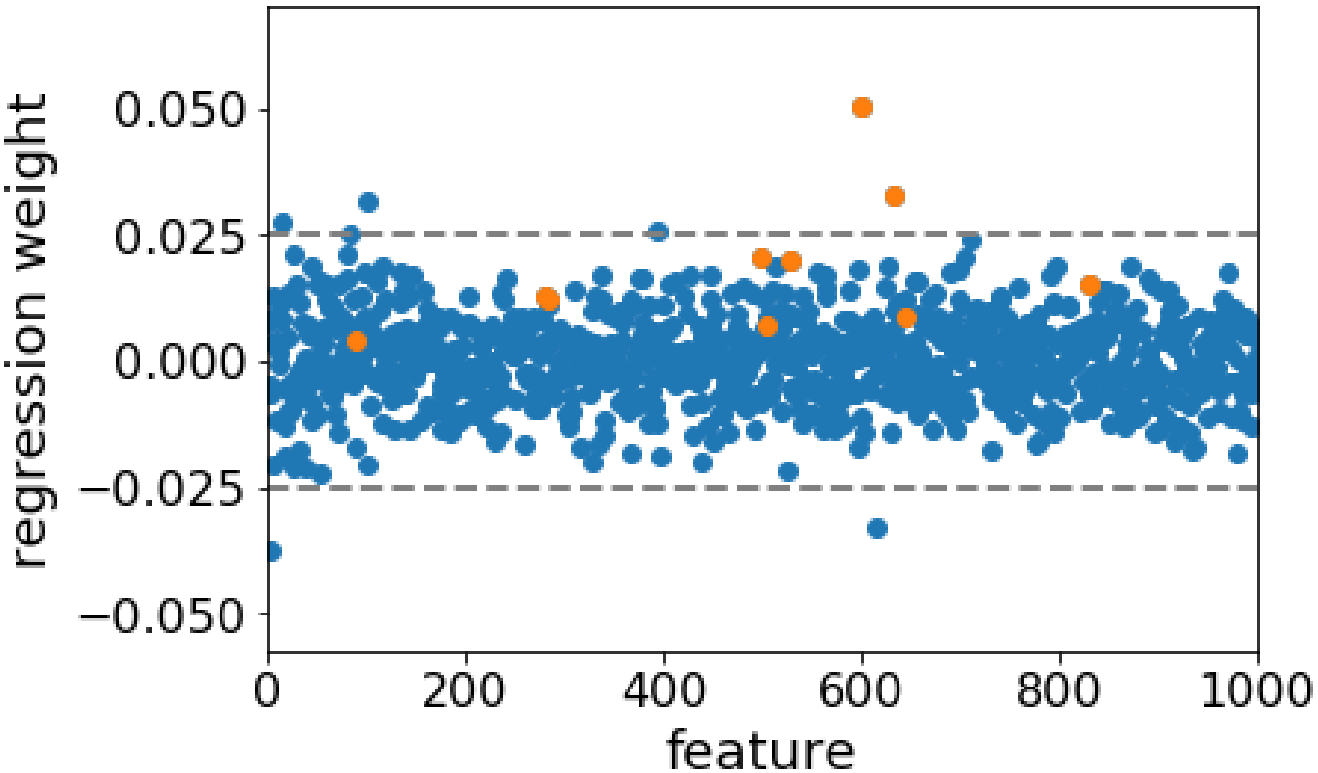
# Simulation

**Linear regression:** Fit $y \sim \sum_{j=1}^{p} w_j x_j + b$.



# Simulation

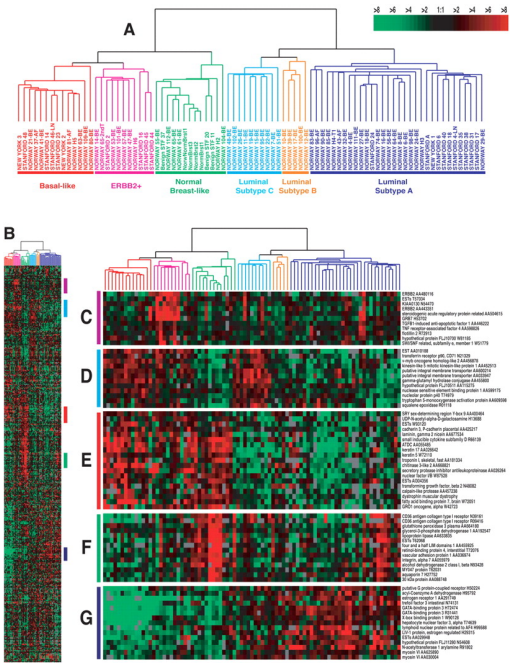
**Linear regression:** Fit $y \sim \sum_{j=1}^{p} w_j x_j + b$.

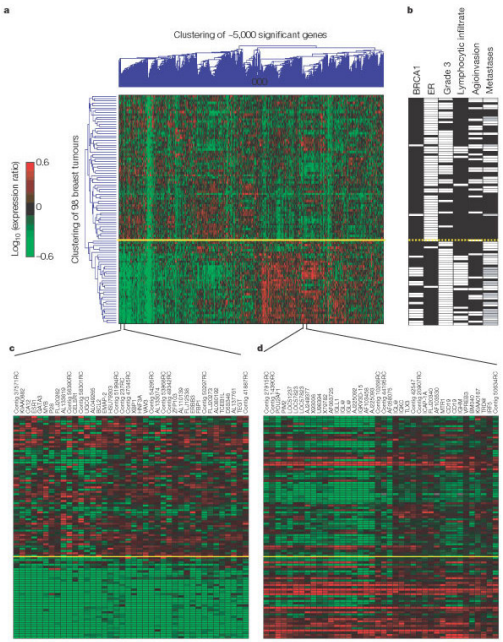



- RMSE on test set: 0.21.

# Molecular signatures stability

**Stability (robustness):** find similar answers on different data sets linked to the same biological question.




OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

## Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

# Molecular signatures stability

- **Correction** to Snyder, A. et al. "Genetic basis for clinical response to CTLA-4 blockade in melanoma", NEJM (2014)

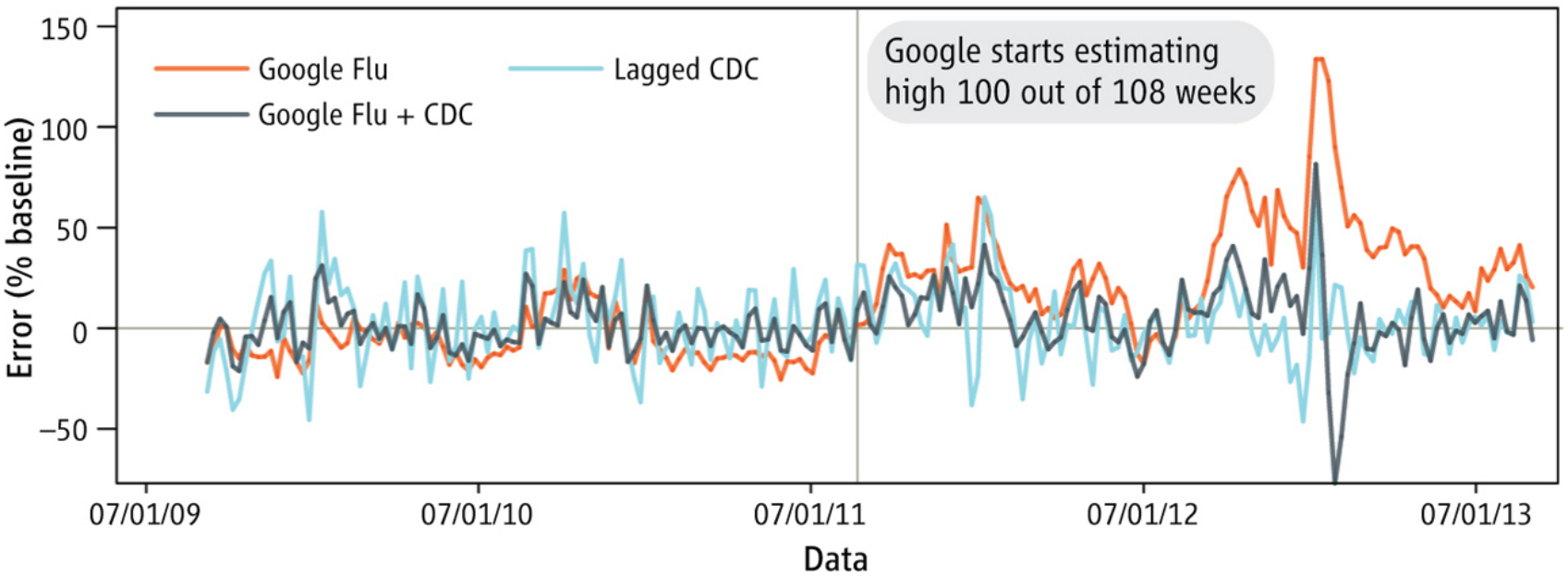
the full text of the article at NEJM.org. Some readers were confused by our incomplete description of part of the data analysis and our use of the term "validation set." We acknowledge that our use of "validation set" was not appropriate in the context of the search for a neoantigen signature, since information from both data sets was used to derive the results. Here,

In the article, we did not use "validation set" in the conventional way that the term is typically used in biomarker studies — namely, as an entirely independent data set in which findings from the discovery set are either confirmed or refuted. Rather, the term was carried over from

# Google Flu Trends

D. Lazer, R. Kennedy, G. King and A. Vespignani. **The Parable of Google Flu: Traps in Big Data Analysis.** Science 2014

- **p = 50 million** search terms
- **n = 1152** data points




- Predictive search terms include keywords related to high school basketball.

# Measuring stability

- **Similarity between two sets of selected features: Jaccard's index**

$$J(\mathcal{S}, \mathcal{S}') = \frac{|\mathcal{S} \cap \mathcal{S}'|}{|\mathcal{S} \cup \mathcal{S}'|}$$

- **Kuncheva's consistency index** [Kun07]

$$I_C(\mathcal{S}, \mathcal{S}') = \frac{\text{Observed}(|\mathcal{S} \cap \mathcal{S}'|) - \text{Expected}(|\mathcal{S} \cap \mathcal{S}'|)}{\text{Maximum}(|\mathcal{S} \cap \mathcal{S}'|) - \text{Expected}(|\mathcal{S} \cap \mathcal{S}'|)}$$

$$\text{Maximum}(|\mathcal{S} \cap \mathcal{S}'|) = \min(|\mathcal{S}|, |\mathcal{S}'|) \qquad \text{Expected}(|\mathcal{S} \cap \mathcal{S}'|) = \frac{|\mathcal{S}||\mathcal{S}'|}{p}$$

- **Similarity between $k$ sets of selected features:**

$$I_C(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k) = \frac{k(k-1)}{2} \sum_{i=1}^{k} \sum_{j=i+1}^{k} I_C(\mathcal{S}_i, \mathcal{S}_j)$$

# 2. Reducing p: Regularization

# Dimensionality reduction

- **Feature selection:** Keep relevant features only
  - **Filter approaches:** Apply a statistical test to assign a score to each feature.
  - **Wrapper approaches:** Use a greedy search to find the "best" set of features for a given predictive model.
  - **Embedded approaches:** Fit a **sparse** model, i.e. that is encouraged to not use all the features.
- **Feature extraction:** Project the data on a new space
  - Creates new features, which makes interpretability harder.
  - **Matrice factorization techniques:** PCA, factorial analysis, NMF, kPCA.
  - **Manifold learning:** Multidimensional scaling, t-SNE.
  - **Autoencoders.**

# Integrating prior knowledge

- **Regularization**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}}$$

# Integrating prior knowledge

- **Regularization**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}}$$

# Integrating prior knowledge

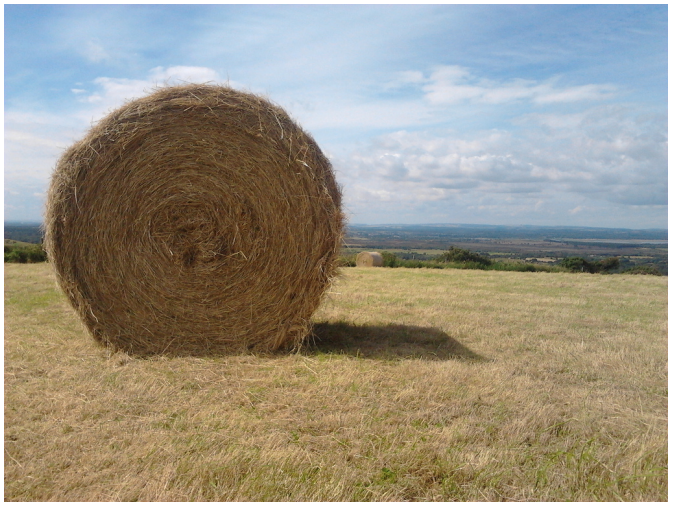
- **Regularization**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \ldots, w_j)}_{\text{regularizer}}$$

# Integrating prior knowledge

- **Regularization**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \ldots, w_j)}_{\text{regularizer}}$$



# Integrating prior knowledge

- **Regularization**

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \ldots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.

# Integrating prior knowledge

- **Regularization**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \ldots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.
- **Lasso**

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p} |w_j|}_{\text{sparsity}}$$

# Integrating prior knowledge

- **Regularization**

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \ldots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.
- **Lasso**

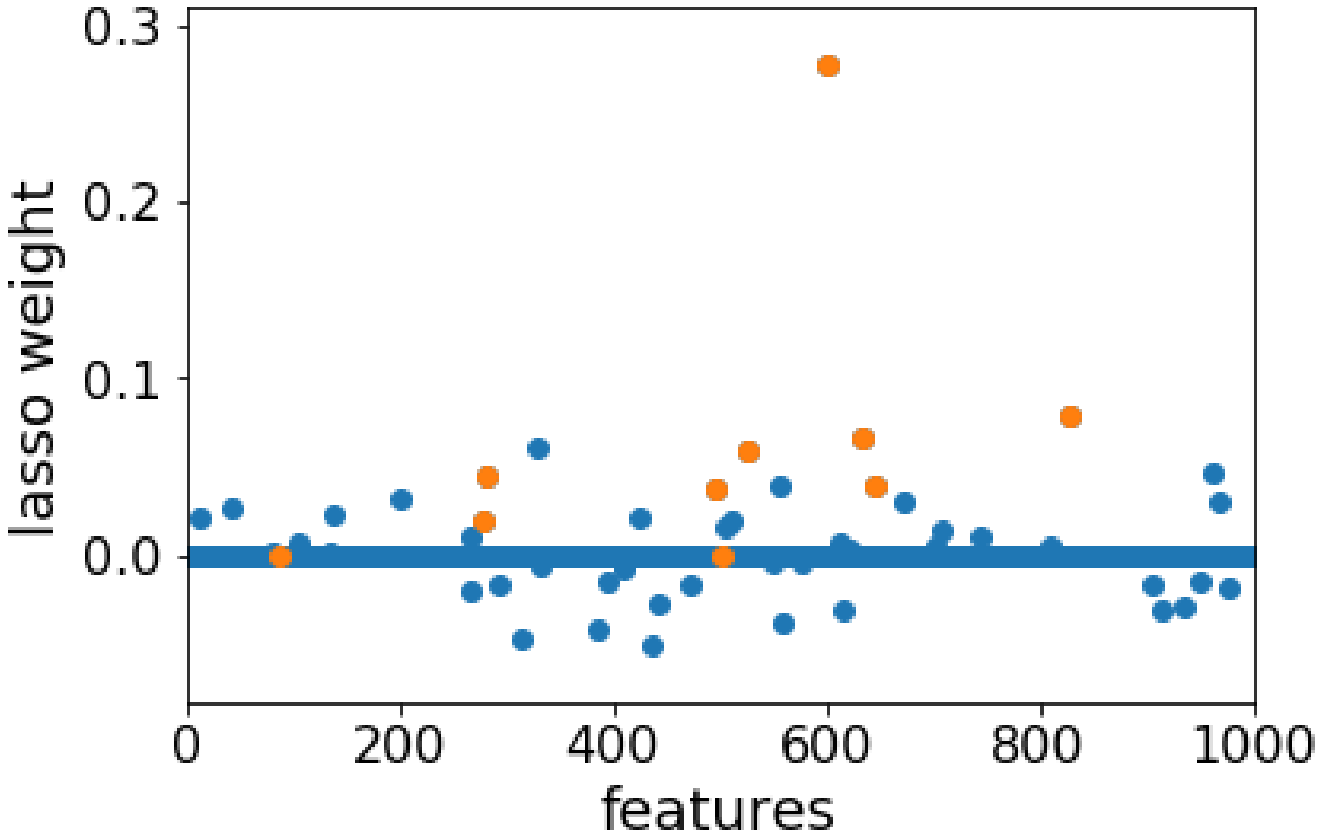
$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p} |w_j|}_{\text{sparsity}}$$

- **Sparsity:** many features are assigned a weight of 0.

  They can be removed from the model.

# Simulation

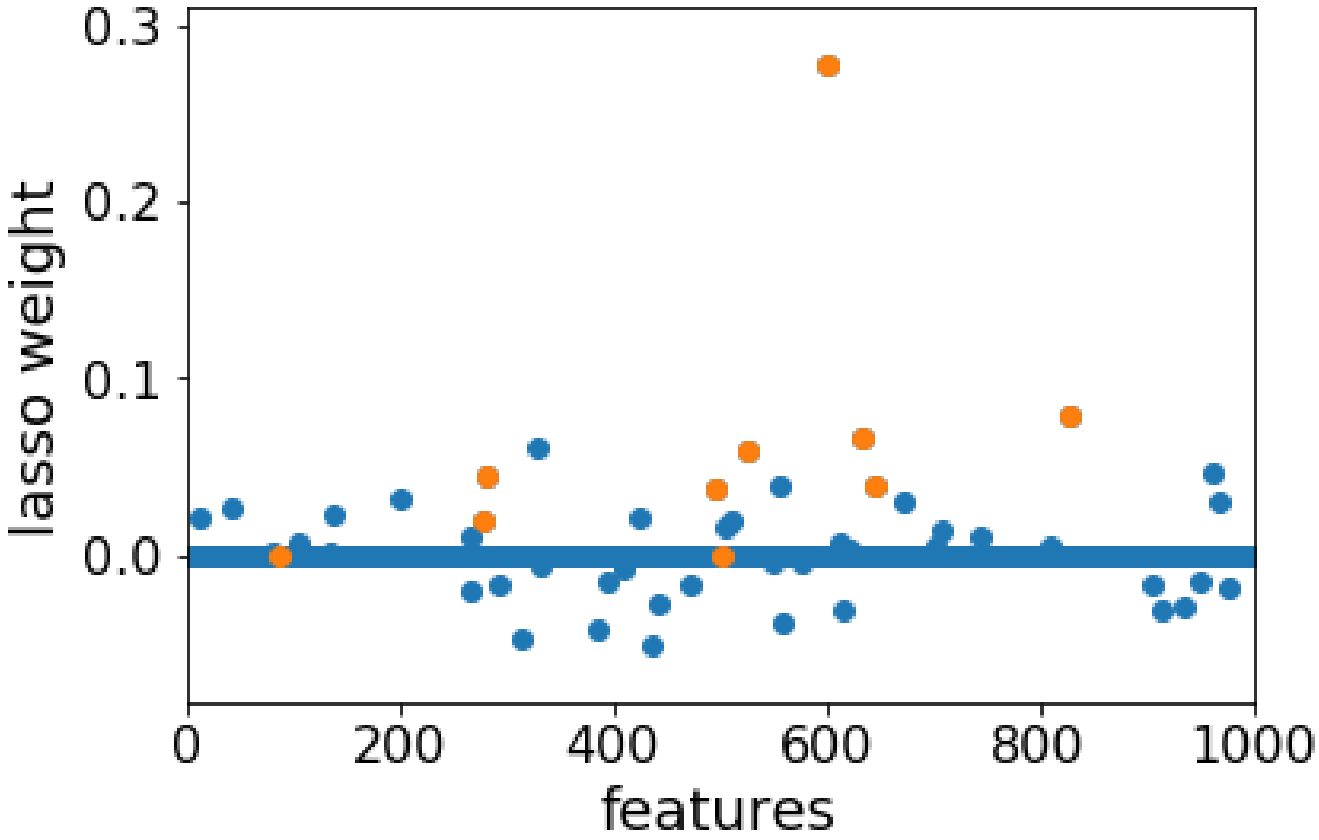
## Lasso regression:




– 55 features have non-zero weights.

# Simulation

## Lasso regression:




- 55 features have non-zero weights.
- RMSE on test set: 0.19.

# But what about **p-values?**

# But what about **p-values?**

- **Do you really want a p-value?**

# But what about **p-values?**

- **Do you really want a p-value?**
  - **What a p-value is: the probability to observe a value as extreme** as the one you obtain, under the null.

# But what about **p-values?**

- **Do you really want a p-value?**
  - **What a p-value is: the probability to observe a value as extreme** as the one you obtain, under the null.
  - **What a p-value is not: all-powerful magic,** nor **biological evidence.**

# But what about **p-values?**

- **Do you really want a p-value?**
  - **What a p-value is: the probability to observe a value as extreme** as the one you obtain, under the null.
  - **What a p-value is not: all-powerful magic,** nor **biological evidence.**

    "The p-value was never intended to be a substitute for scientific reasoning." [WL+16]

# But what about **p-values?**

- **Do you really want a p-value?**
  - **What a p-value is: the probability to observe a value as extreme** as the one you obtain, under the null.
  - **What a p-value is not: all-powerful magic,** nor **biological evidence.**

    "The p-value was never intended to be a substitute for scientific reasoning." [WL+16]

[Ioa05; Nuz14; Hea+15; Hol18]

# But what about **p-values?**

- **Do you really want a p-value?**
  - **What a p-value is: the probability to observe a value as extreme** as the one you obtain, under the null.
  - **What a p-value is not: all-powerful magic,** nor **biological evidence.**

    "The p-value was never intended to be a substitute for scientific reasoning." [WL+16]

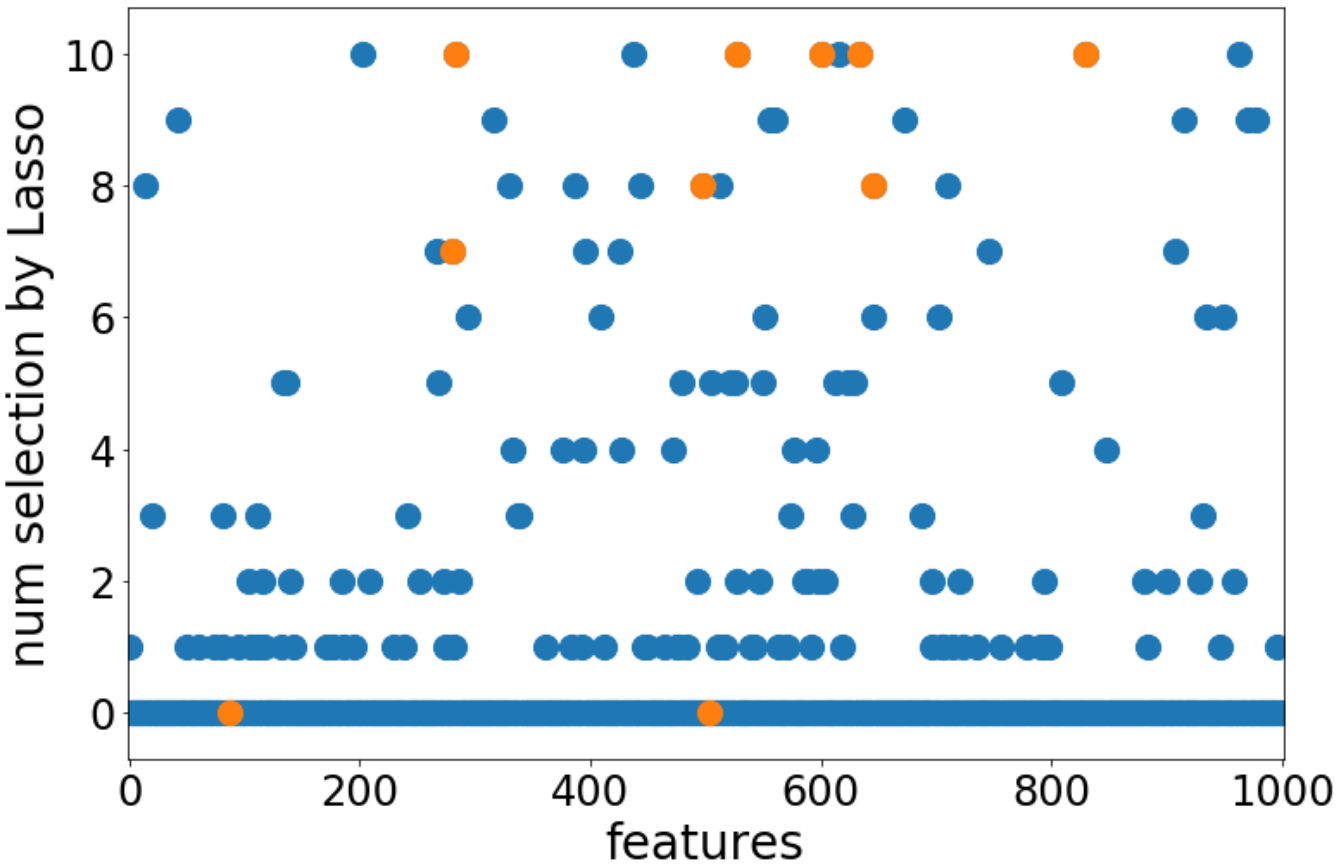
  [Ioa05; Nuz14; Hea+15; Hol18]

- For the **lasso**, possible but computationally intensive

  R. Lockhart et al. **A significance test for the lasso.** Annals of Stats 2014.

# Stability

10-fold cross-validation: how often was each feature selected?




**Consistency index:** 0.58.

# Elastic Net

- **Lasso**

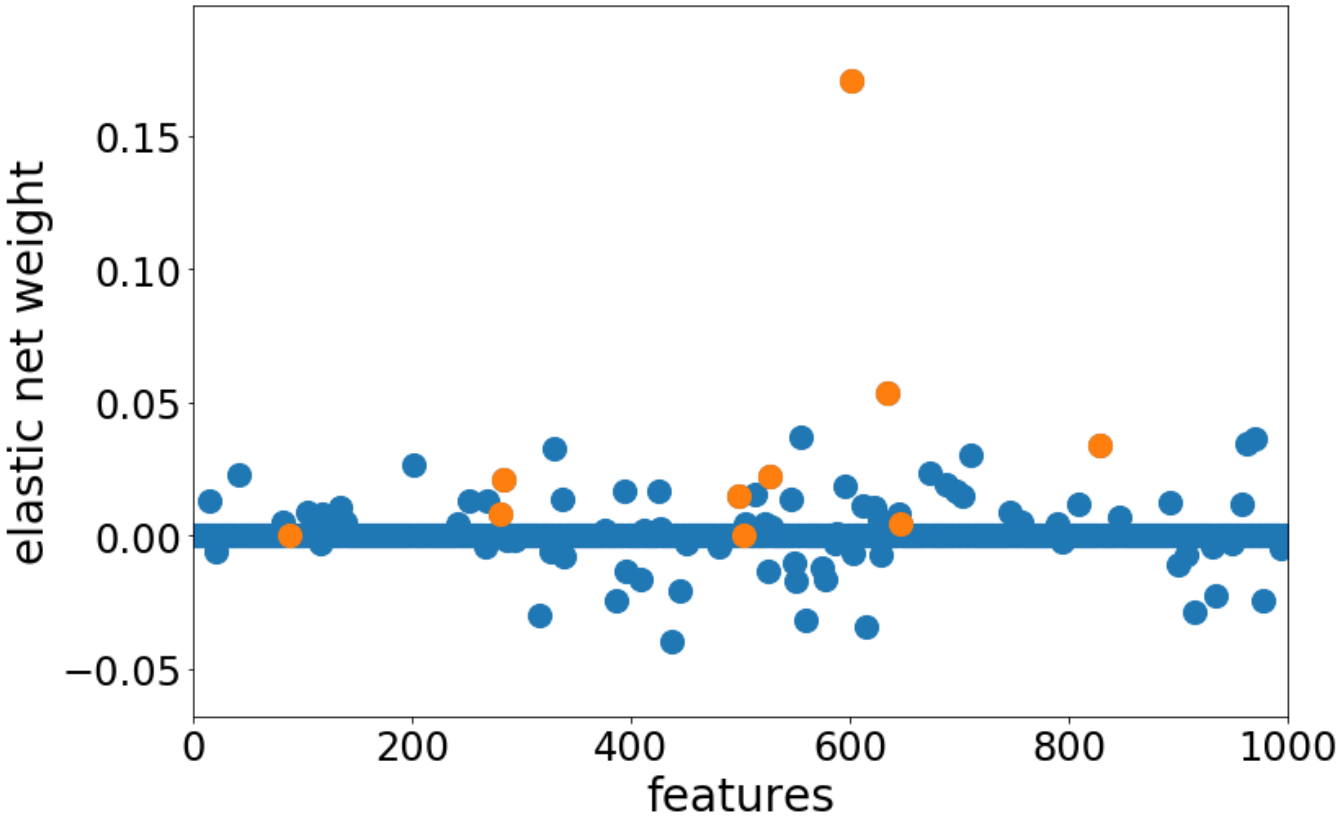
$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p}|w_j|}_{\text{sparsity}}$$

- **Elastic Net:** [ZH05]

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \left( \underbrace{\alpha \sum_{j=1}^{p}|w_j|}_{\text{sparsity}} + (1-\alpha) \underbrace{\sum_{j=1}^{p} w_j^2}_{\text{weight decay}} \right)$$

# Simulation

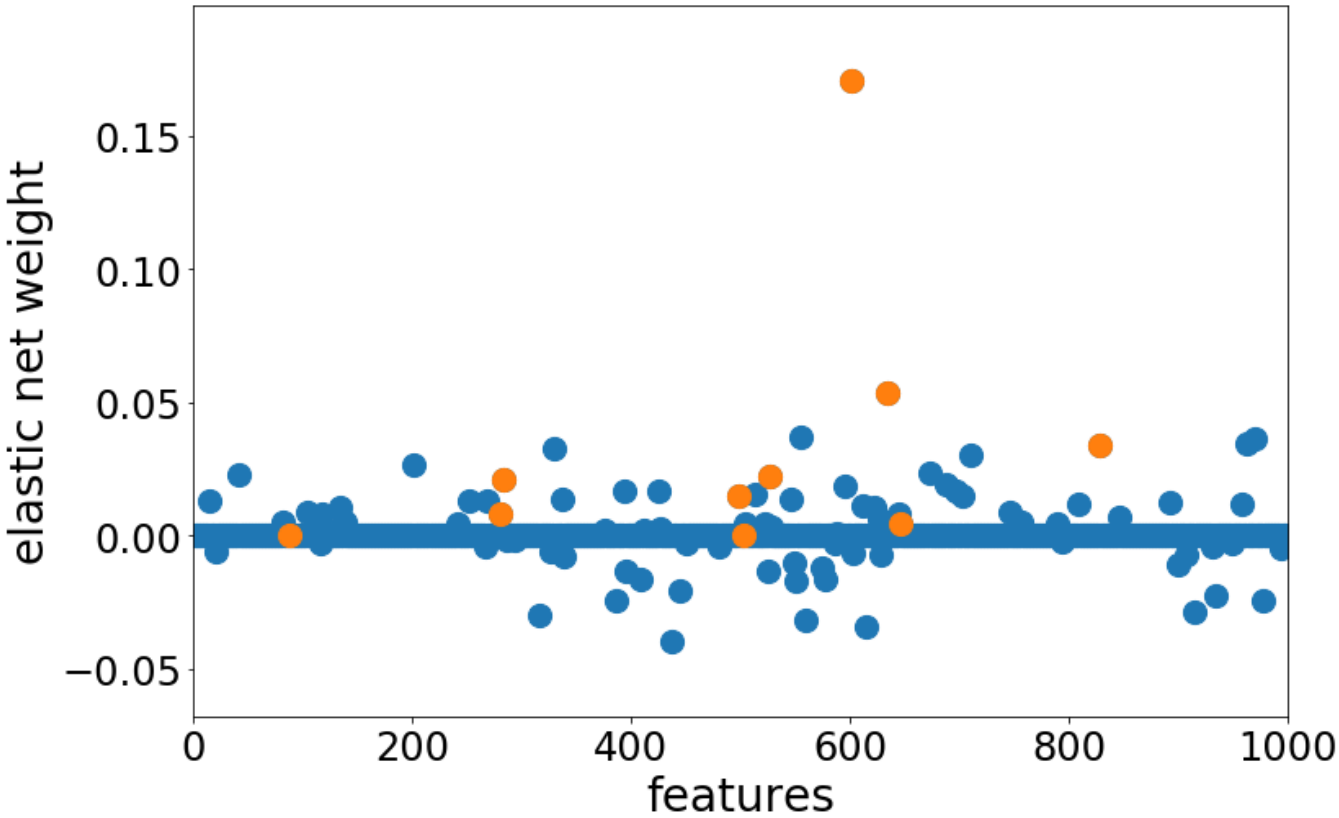
## Elastic net:




– 92 features have non-zero weights.

# Simulation

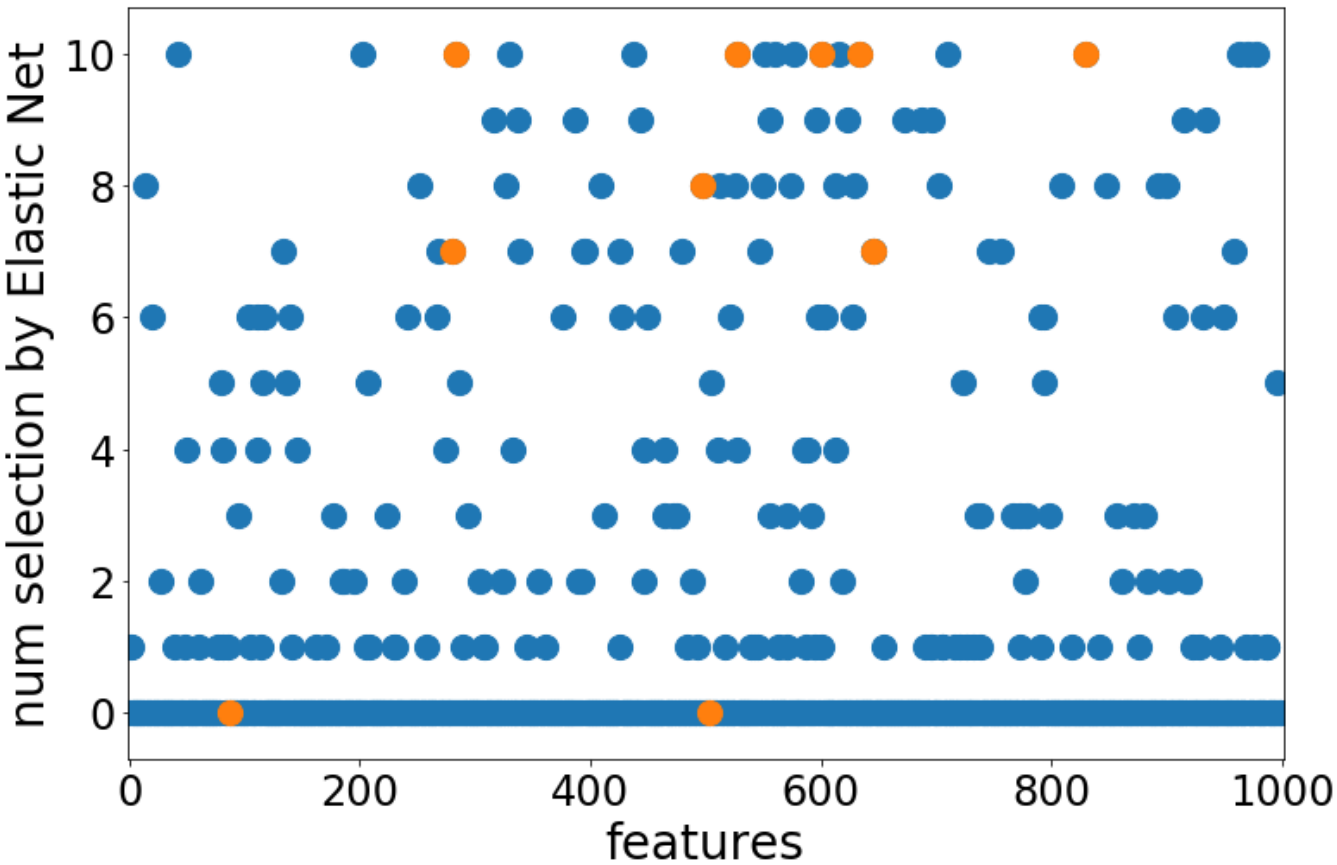
## Elastic net:




- 92 features have non-zero weights.
- RMSE on test set: 0.20.

# Stability

10-fold cross-validation: how often was each feature selected?




**Consistency index:** 0.62.

# Stability selection with the lasso

- Repeat:
  - randomly **subsample the training data** $I \subset \{1, 2, \ldots, n\}$
  - randomly **scale the regularizer** $s_j \in \{s, 1\}$ (Bernouilli; $0 < s < 1$)
  - fit a lasso

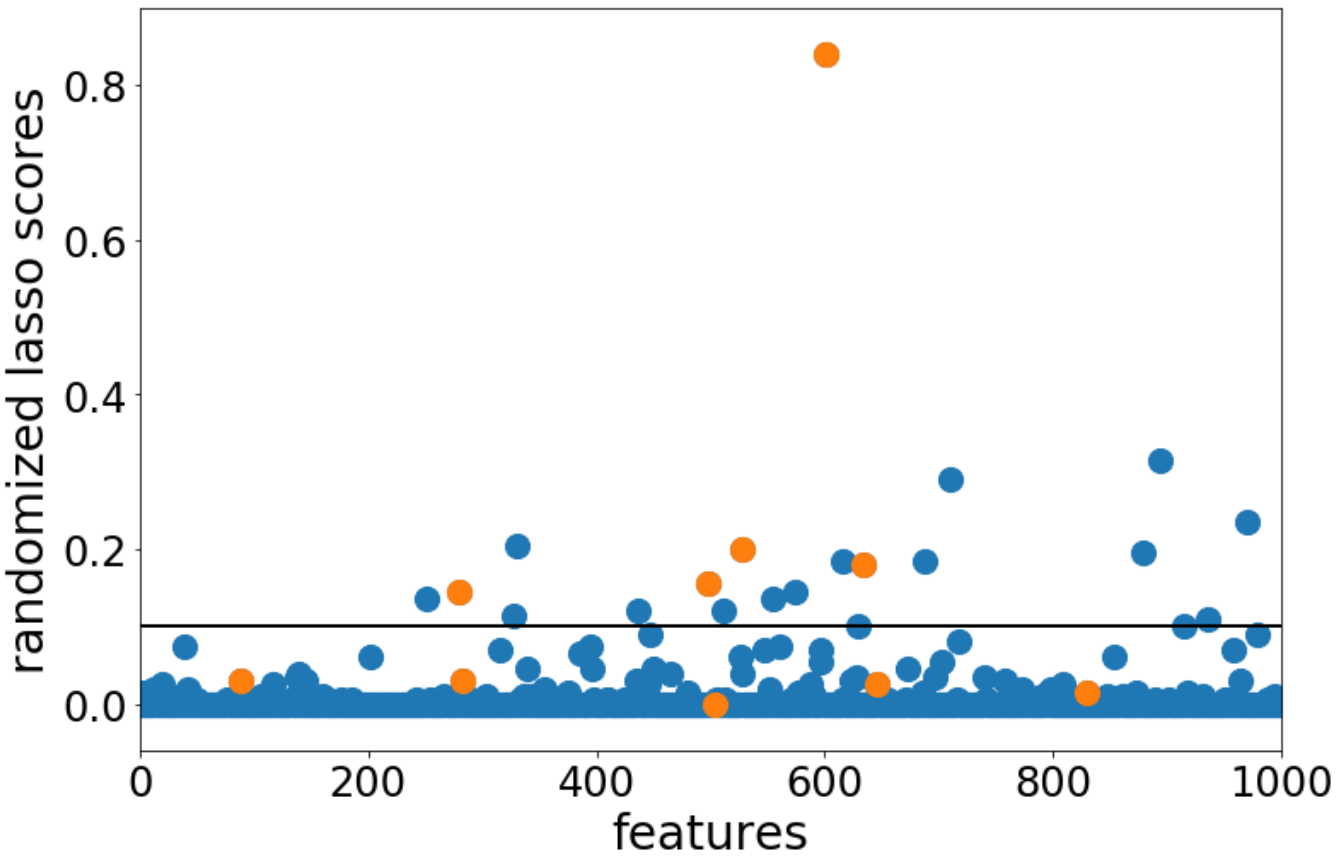
$$\underset{\mathbf{w} \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{|I|} \sum_{i \in I} \left( y_i - \sum_{j=1}^{p} w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p} \frac{|w_j|}{s_j}}_{\text{sparsity}}$$

- Select features that appear in **multiple models.**

[MB10; SS13; AL11; DP15; NB16]

# Simulation

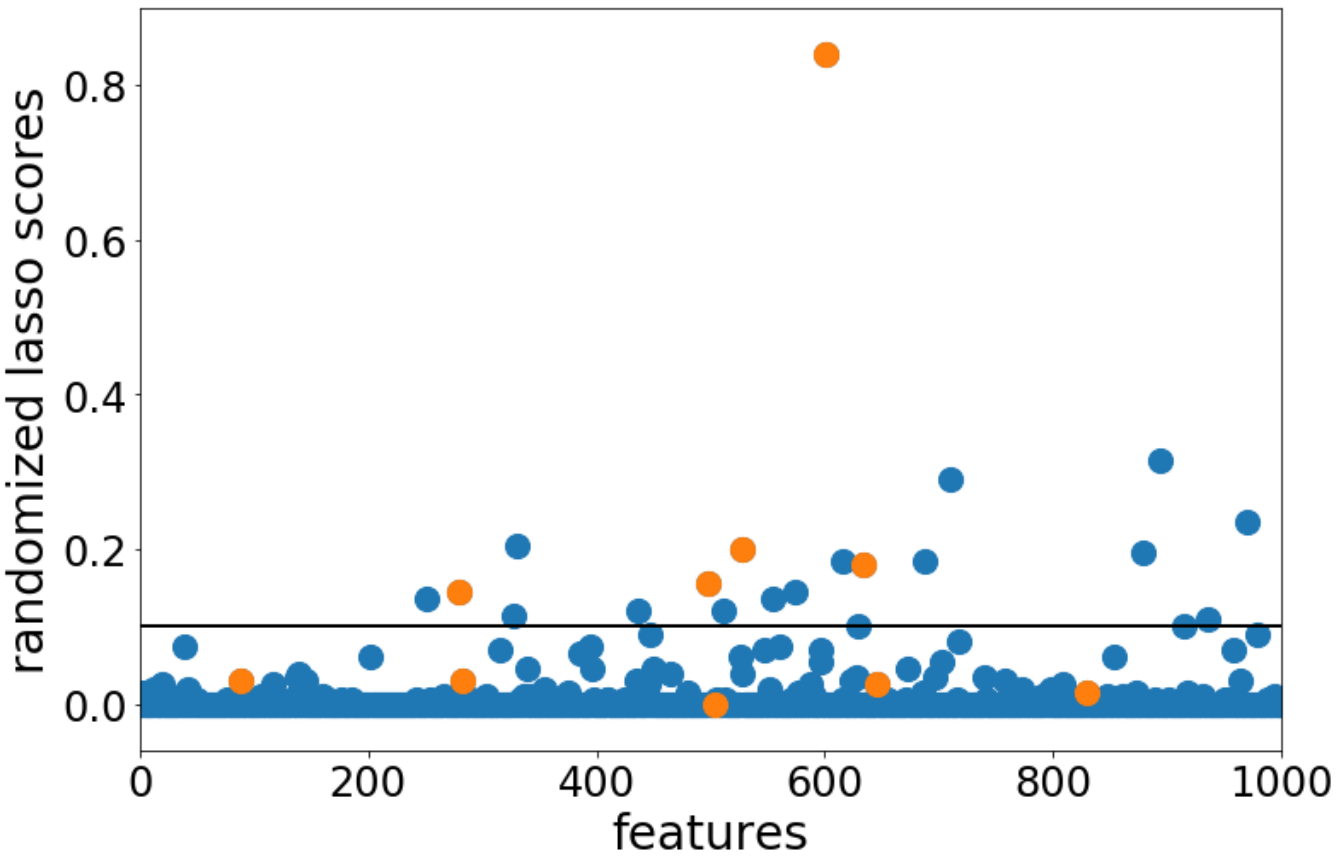
## Stability selection lasso




19 features are selected in more than 10% of the folds.

# Simulation

## Stability selection lasso




19 features are selected in more than 10% of the folds.

- RMSE on test set: 0.16.

# Integrating prior knowledge

**Use prior knowledge as a constraint on the selected features**

- **Consistant** with previously established knowledge
- Increases **interpretability** and **statistical power.**

Prior knowledge can be represented as **structure:**

- Linear structure of the DNA
- **Groups:** e.g. pathways
- **Networks:** molecular, 3D structure.

# Integrating prior group knowledge

- **Group lasso**

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p}|w_j|}_{\text{sparsity}} + \eta \underbrace{\sum_{g\in\mathcal{G}}||\mathbf{w}_g||}_{\text{grouping}}$$

- **Overlapping group lasso** for when a feature can belong to multiple groups

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p}|w_j|}_{\text{sparsity}} + \eta \underbrace{\sum_{g\in\mathcal{G}}\sum_{\mathbf{x}\in\mathcal{V}_\mathcal{G},\sum_{g\in\mathcal{G}}\mathbf{x}_g=\mathbf{w}}||\mathbf{x}_g||}_{\text{grouping}}$$

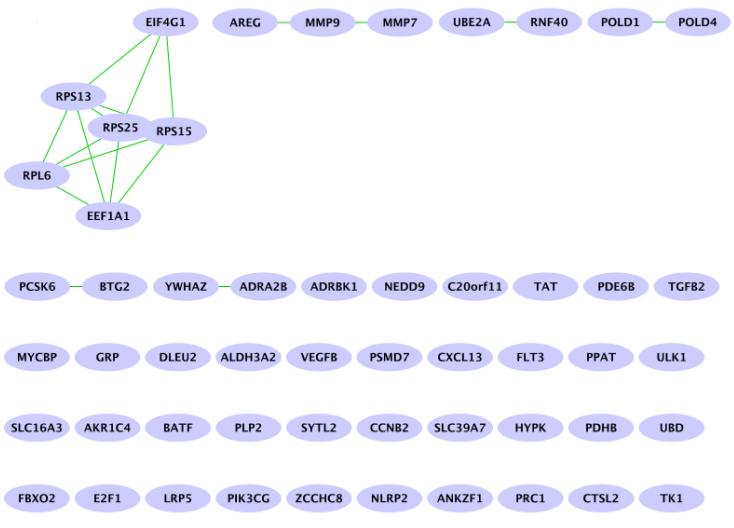
# Gene selection with the Graph lasso

icml2009

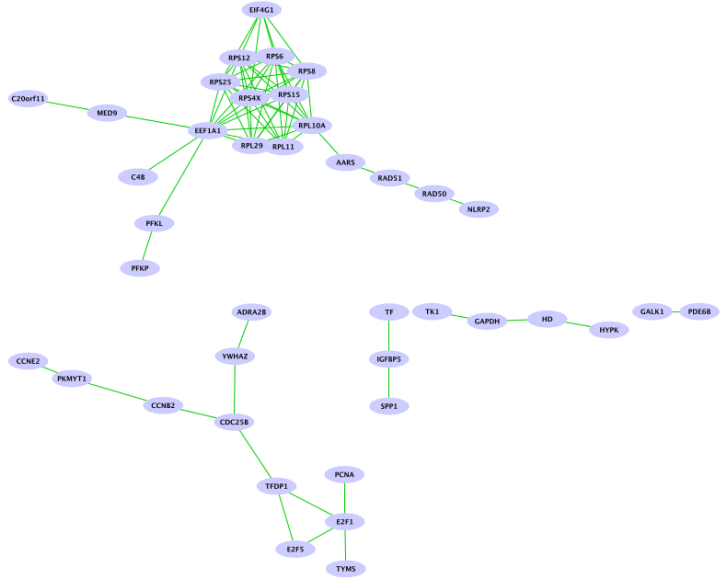
**Group Lasso with Overlaps and Graph Lasso**

*Laurent Jacob, Guillaume Obozinski and Jean-Philippe Vert*

## Lasso



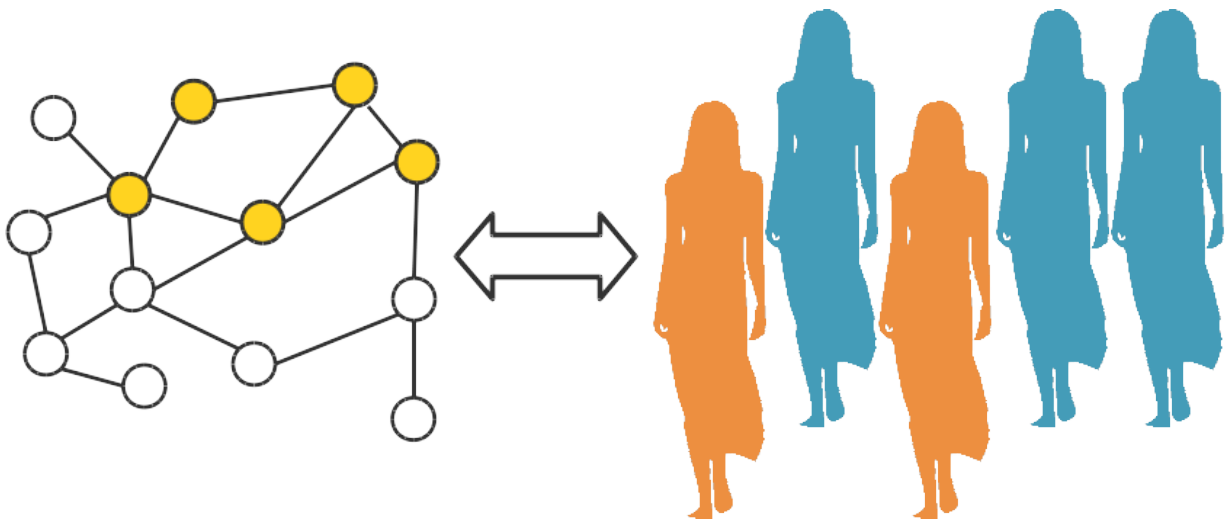

## Graph Lasso




Ref: [JOV09]

# Network-guided biomarker discovery

Goal: Find a **set of explanatory features** compatible with a **given network** structure.



C.-A. Azencott, Network-guided biomarker discovery, LNCS 2016

# Integrating prior network knowledge

- **Network-constrained lasso** [LL08]

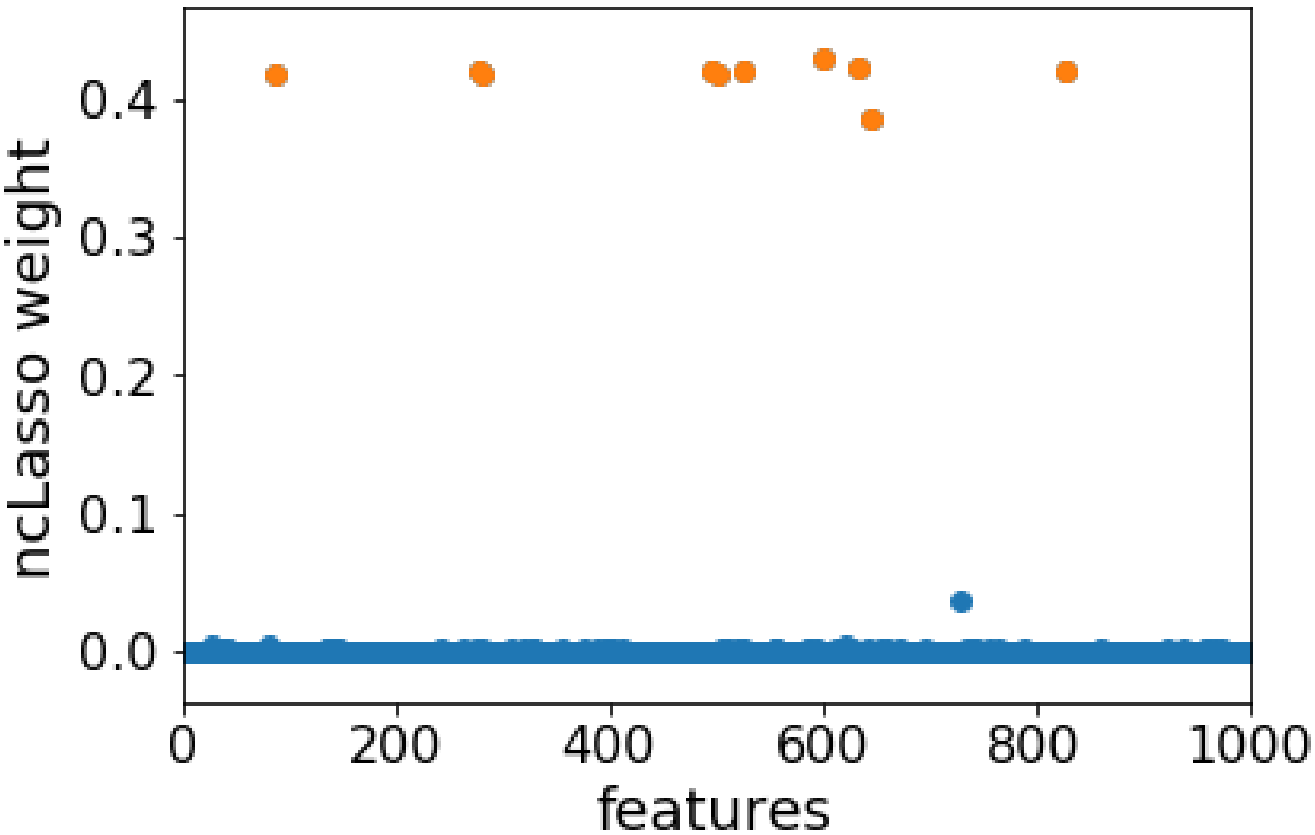
$$\underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \underbrace{||\mathbf{y}-\mathbf{w}X||_2^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^{p}|w_j|}_{\text{sparsity}} + \eta \underbrace{\sum_{j=1}^{p}\sum_{k=1}^{p} w_j L_{jk} w_k}_{\text{connectivity}}$$

- **Graph Laplacian** $L$ ensures $w$ varies **smoothly** on the network.

$$L_{jk} = \begin{cases} 1 & \text{if } j = k \\ -W_{jk}/\sqrt{d_j d_j} & \text{if } j \neq k \\ 0 & \text{otherwise.} \end{cases}$$

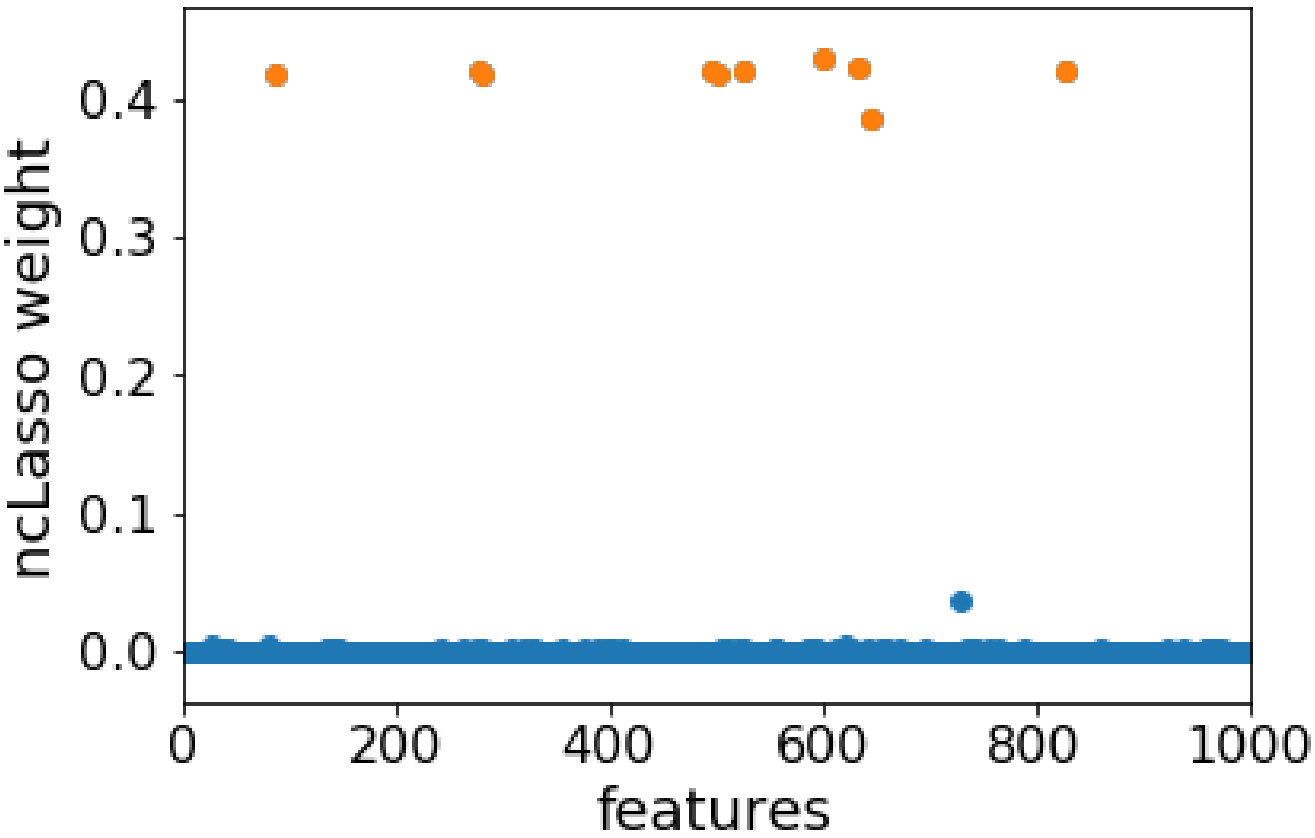
# Simulation

## Network-constrained lasso



# Simulation

## Network-constrained lasso




- RMSE on test set: 0.14.

# Integrating prior network knowledge

- **Regularized relevance** Set $\mathcal{V}$ of $p$ variables.

$$\arg\max_{\mathcal{S}\subseteq\mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{\Omega(\mathcal{S})}_{\text{regularizer}}$$

- **Network-regularized relevance**

$$\arg\max_{\mathcal{S}\subseteq\mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{|\mathcal{S}|}_{\text{sparsity}} - \eta \underbrace{\sum_{j\in\mathcal{S}}\sum_{k\notin\mathcal{S}} W_{jk}}_{\text{connectivity}}$$

# Integrating prior network knowledge

- **Regularized relevance** Set $\mathcal{V}$ of $p$ variables.

$$\arg\max_{\mathcal{S}\subseteq\mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{\Omega(\mathcal{S})}_{\text{regularizer}}$$

- **Network-regularized relevance**

$$\arg\max_{\mathcal{S}\subseteq\mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{|\mathcal{S}|}_{\text{sparsity}} - \eta \underbrace{\sum_{j\in\mathcal{S}}\sum_{k\notin\mathcal{S}} W_{jk}}_{\text{connectivity}}$$

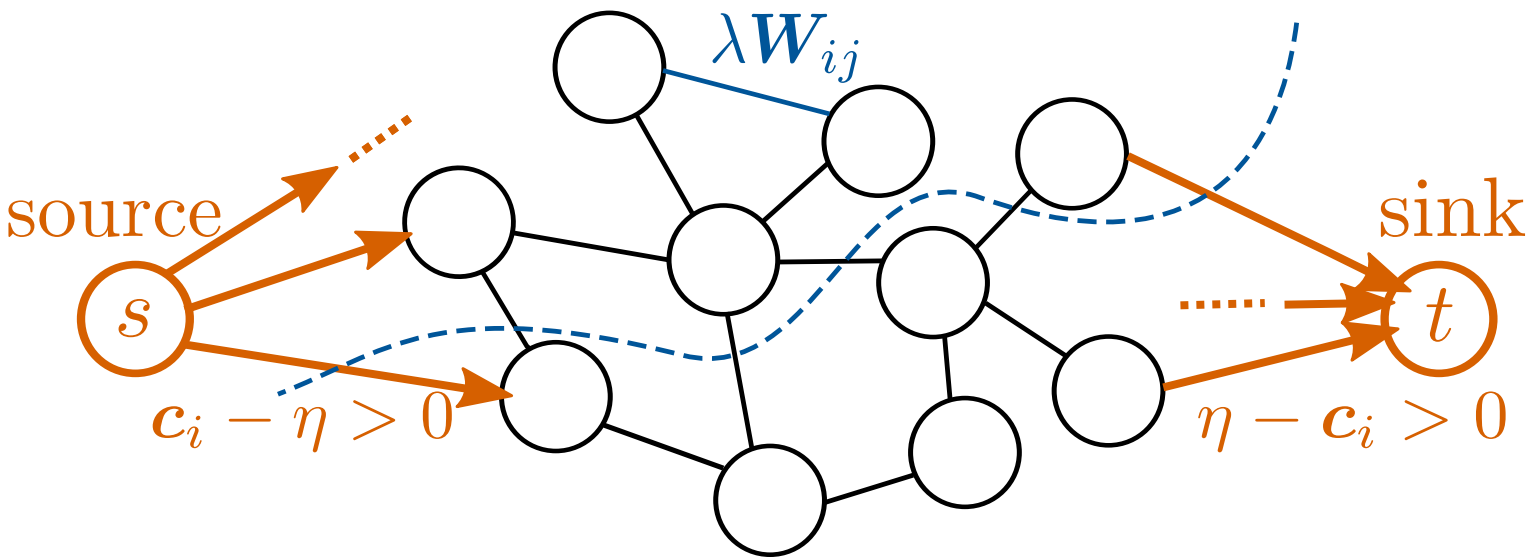
**SConES: S**electing **Con**nected **E**xplanatory **S**NPs.

C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara and K. Borgwardt (2013) **Efficient network-guided multi-locus association mapping with graph cuts**, Bioinformatics 29 (13), i171–i179 doi:10.1093/bioinformatics/btt238 [Aze+13]
https://github.com/chagaz/scones Bioconductor/martini

# Minimum cut reformulation

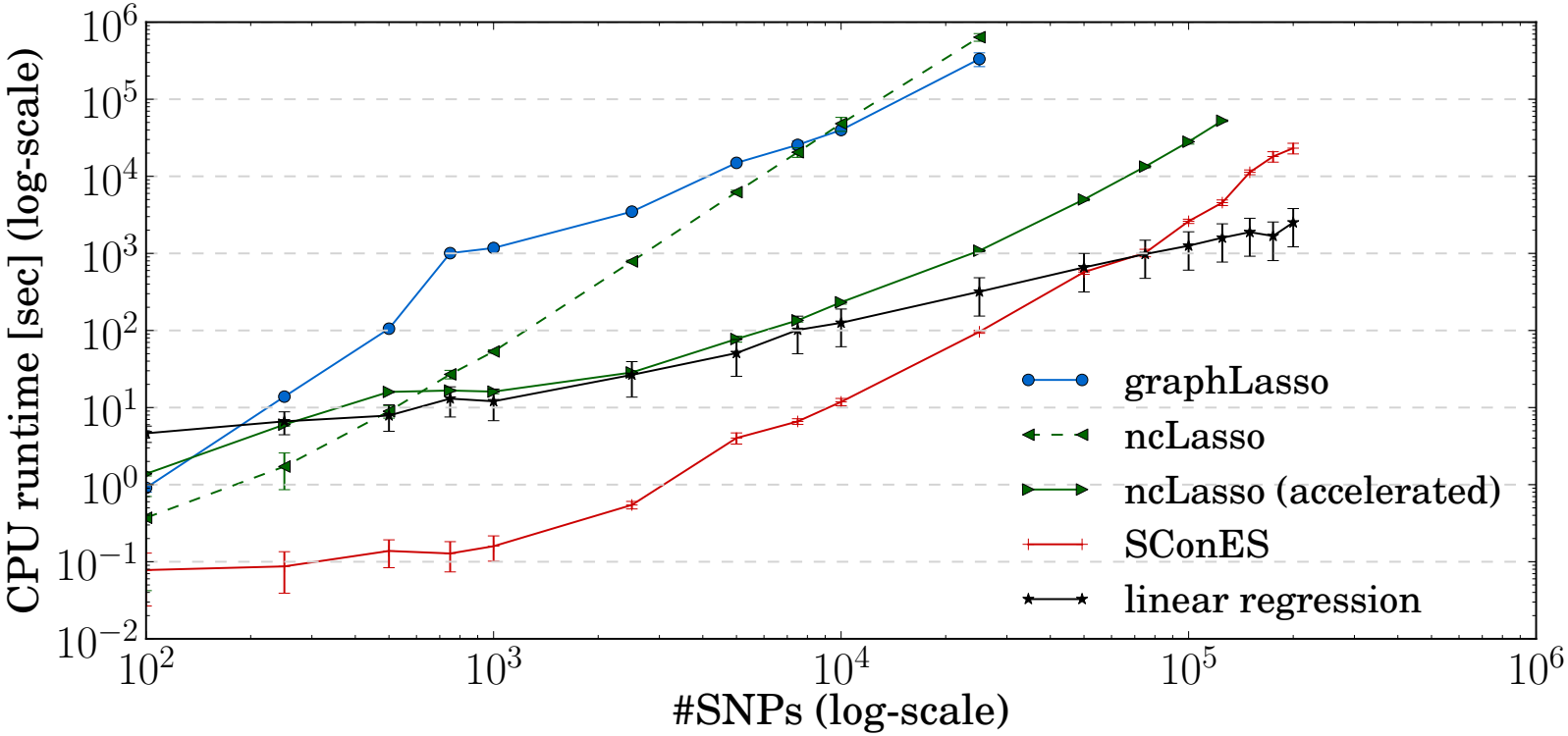
The graph-regularized maximization of score $Q(*)$ is equivalent to a $s/t$-min-cut for a graph with adjacency matrix $\mathbf{A}$ and two additional nodes $s$ and $t$, where $\mathbf{A}_{ij} = \lambda \mathbf{W}_{ij}$ for $1 \leq i, j \leq p$ and the weights of the edges adjacent to nodes $s$ and $t$ are defined as

$$\mathbf{A}_{si} = \begin{cases} c_i - \eta & \text{if } c_i > \eta \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{A}_{it} = \begin{cases} \eta - c_i & \text{if } c_i < \eta \\ 0 & \text{otherwise} \end{cases}.$$




**SConES: S**electing **Con**nected **E**xplanatory **S**NPs.

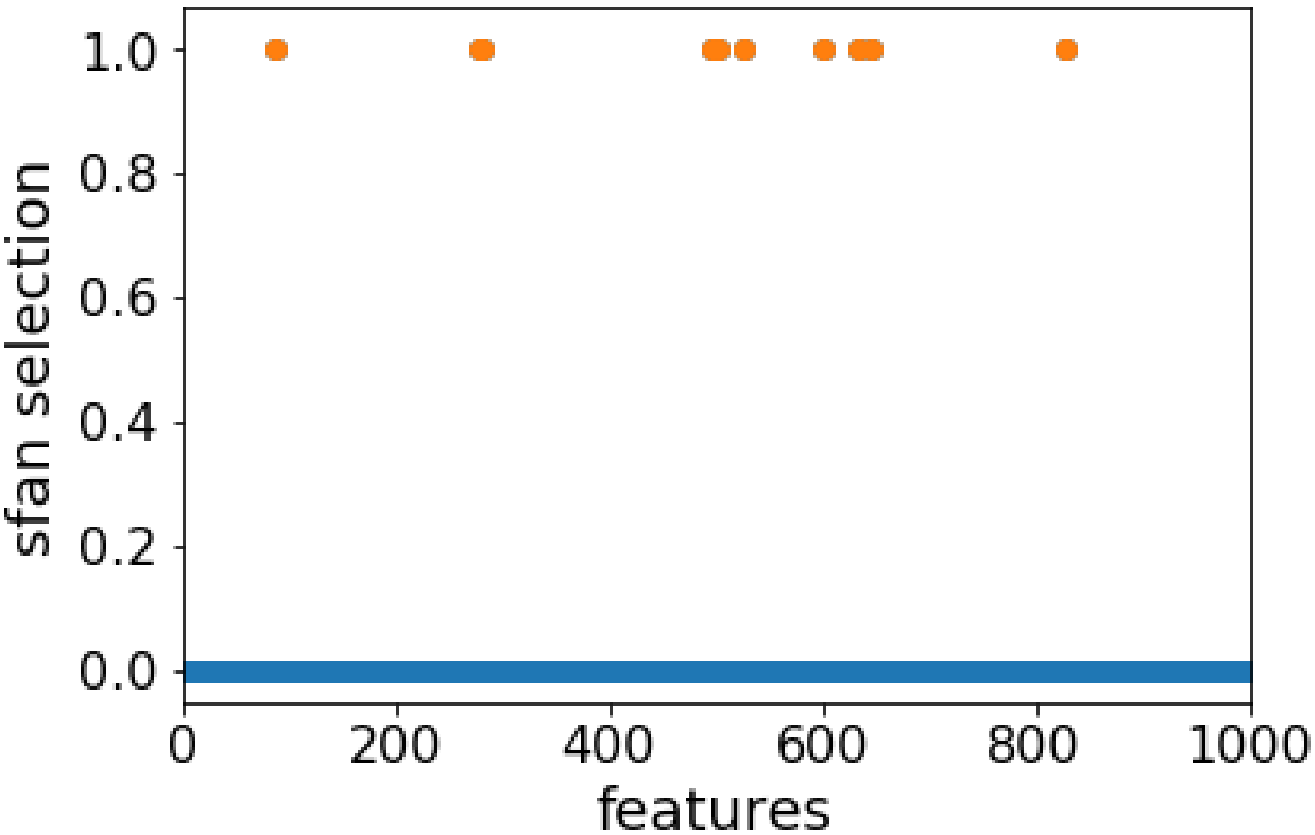
# Runtime




$n = 200$ exponential random network (2 % density)

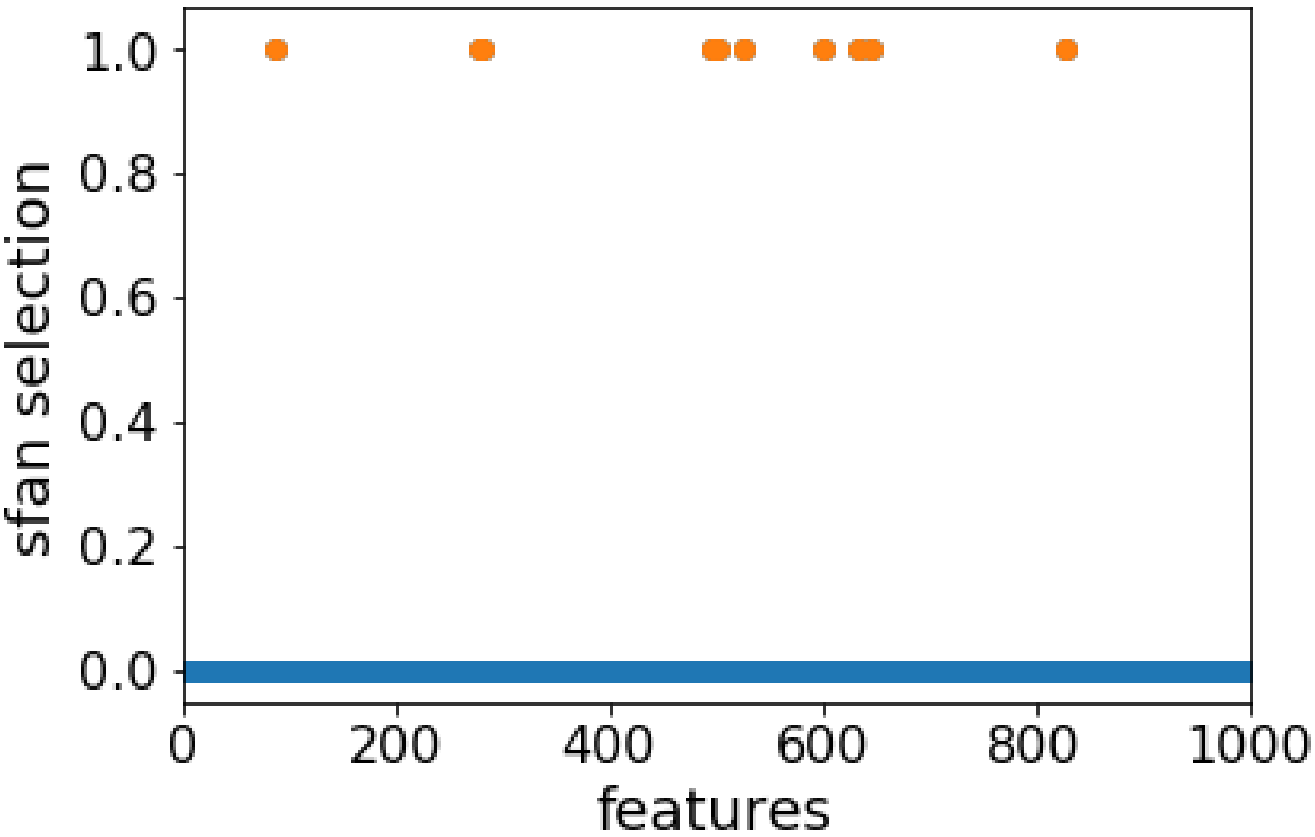
# Simulation

## SConES



# Simulation

## SConES

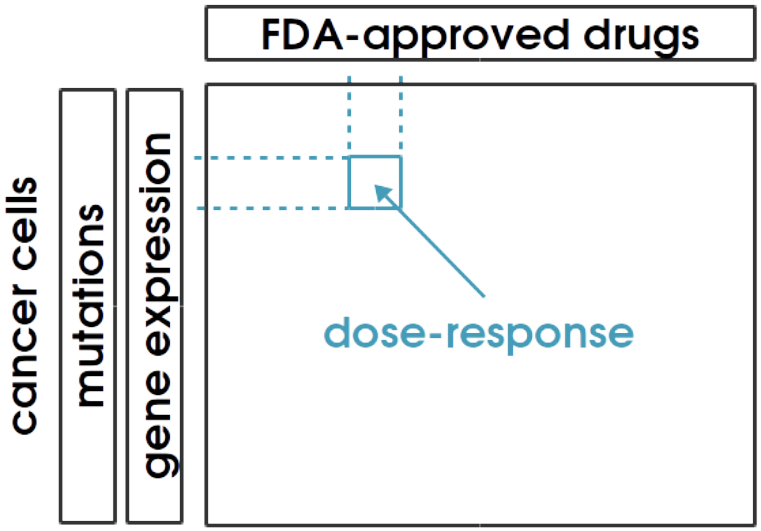



– RMSE on test set: 0.10.

# 3. Increasing n: Multitask approaches

# Multitask machine learning

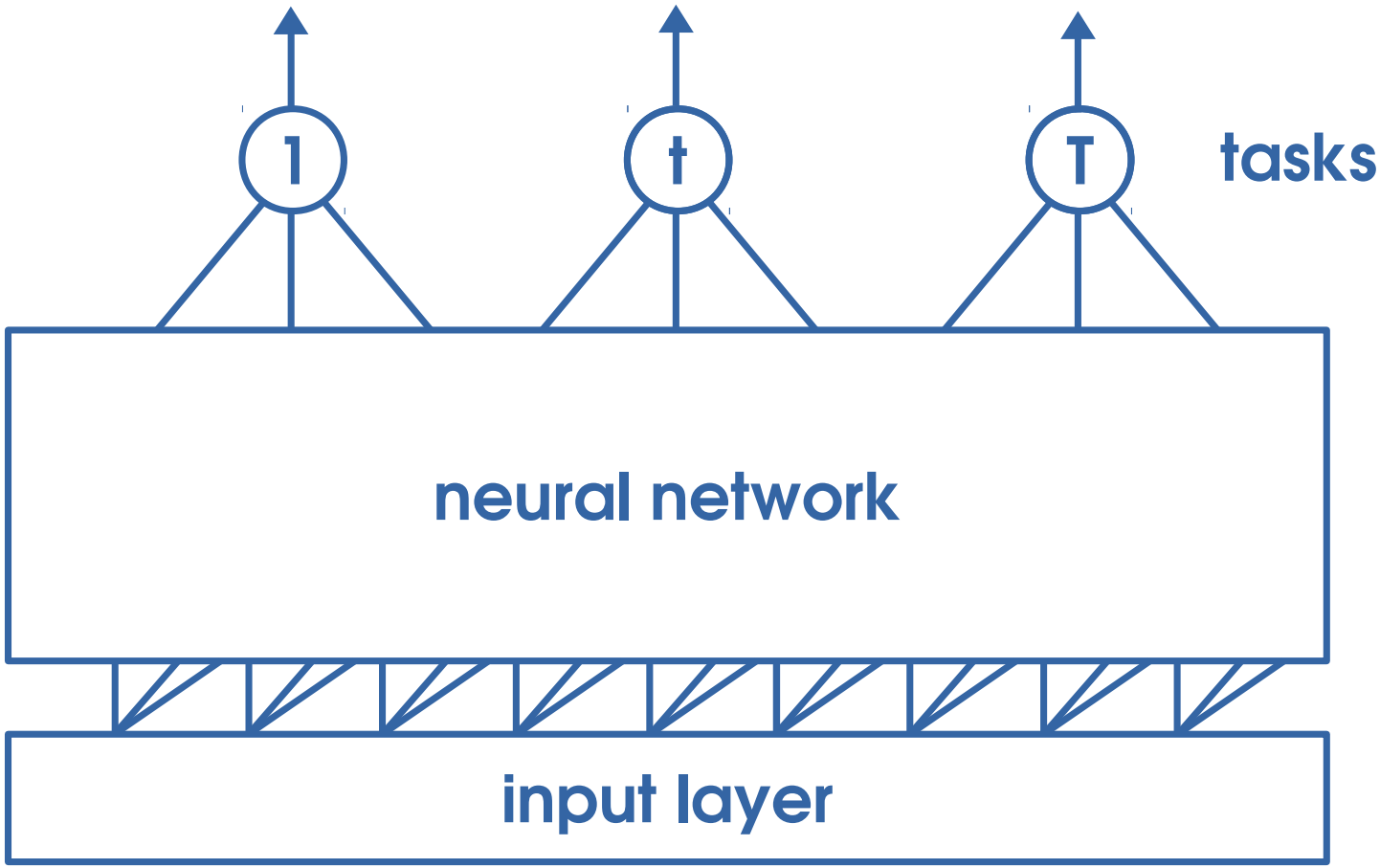
- Solve **multiple related tasks** simultaneously
  - QSAR for multiple related assays
  - **Pharmacogenetics / toxicogenetics / chemogenomics:** predict the response of a cell line to multiple drugs [Edu+15; Dre16; Ior+16]




  - x = a cell line, described by **multiomics features**
  - $y_t$ = **dose-response** to a given drug
  - t = **drug**

# Multitask neural networks

One output unit per task [Car97]



# Multitask linear regression

Group regularization [OTJ06]

linear predictor for task t

weight of feature j for task t

feature j of sample i

$$f_t(x_i) = \sum_{j=1}^{p} w_{tj} x_{ij}$$

$$\min \sum_{t=1}^{T} \boxed{\frac{1}{n_T} \sum_{i=1}^{n_T} (y_i - f_t(x_i))^2} + \lambda \boxed{\sum_{j=1}^{p} \sum_{t=1}^{T} w_{tj}^2}$$

least squares error

**regularization term:** same non-zero weights across tasks

- **Same features selected across tasks**
- The more **similar** the tasks, the more **similar** the features?

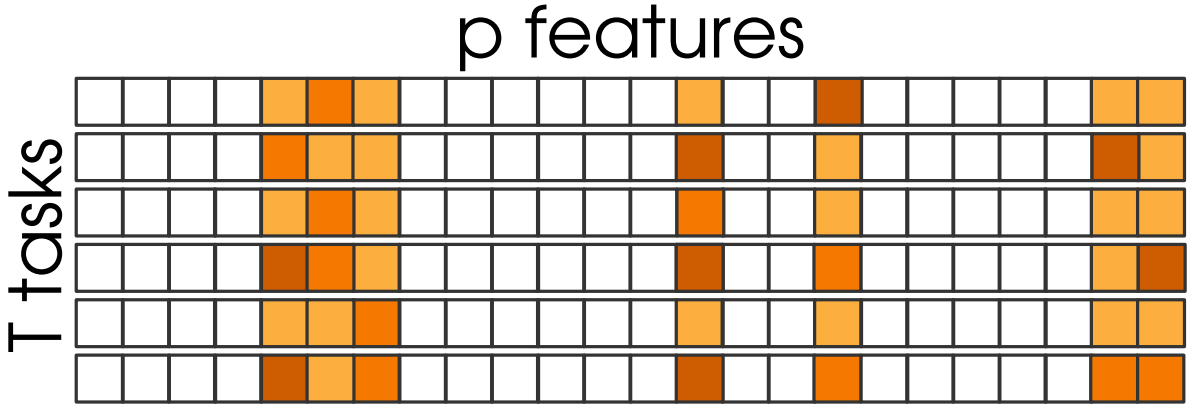
V. Bellón, C.-A. Azencott, and V. Stoven (2016) **Multitask feature selection with task descriptors.** PSB

# Multiplicative Multitask Lasso

- **Multitask Lasso** [OTJ06]

$$\underset{\beta \in \mathbb{R}^{T \times p}}{\arg\min} \underbrace{\sum_{t=1}^{T} \frac{1}{n_t} \sum_{m=1}^{n_t} \left( y_{tm} - \sum_{j=1}^{p} \beta_{tj} x_{tmj} \right)^2}_{\text{loss}} + \underbrace{\lambda \sum_{j=1}^{p} ||\beta_j^2||_2}_{\text{task sharing}}$$

  - Selects the **same features across tasks**
  - Controls weights magnitude ($\ell_2$ shrinkage).

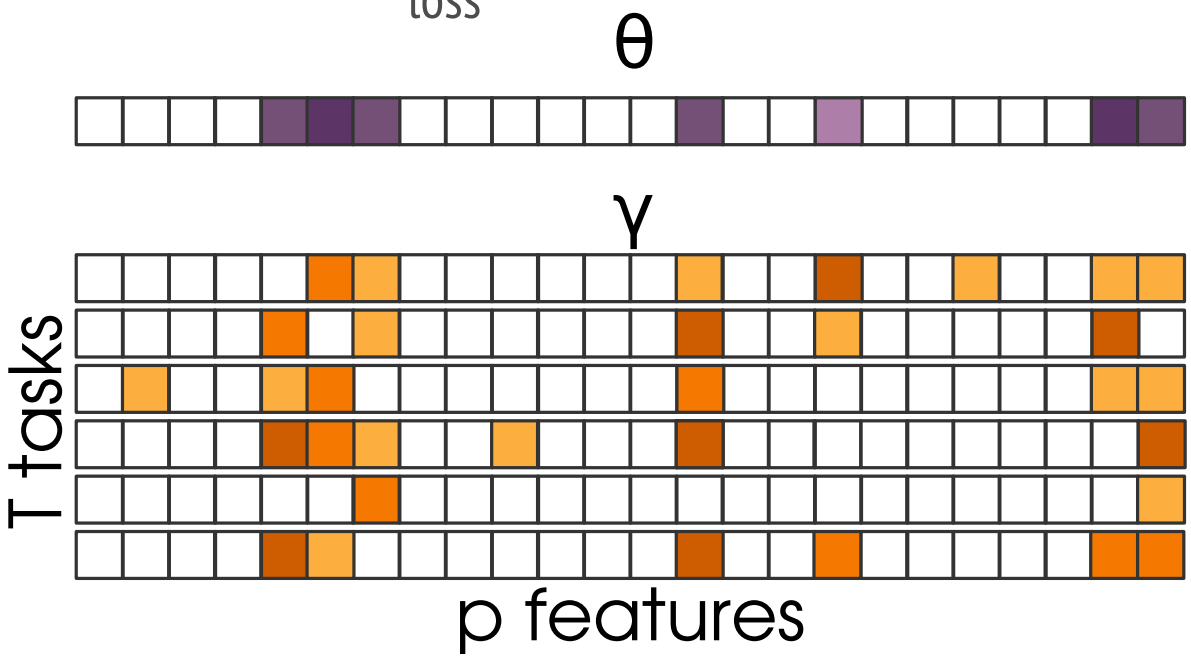


# Multiplicative Multitask Lasso

– **Multilevel Multitask Lasso** [SL12]

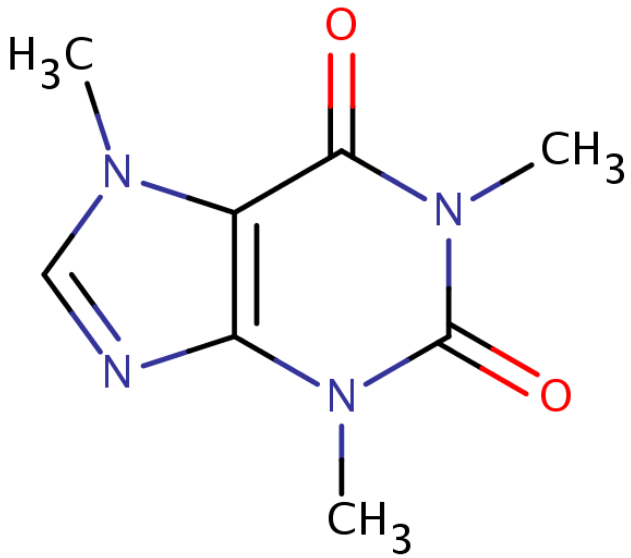
Modulate a sparse, global, task-independent effect $\theta_j$ with a task-specific $\gamma_{tj}$.

$$\underset{\boldsymbol{\theta}\in\mathbb{R}_+^p,\boldsymbol{\gamma}\in\mathbb{R}^{T\times p}}{\arg\min} \underbrace{\sum_{t=1}^{T}\frac{1}{n_t}\sum_{m=1}^{n_t}\left(y_{tm}-\sum_{j=1}^{p}\theta_j\gamma_{tj}x_{tmj}\right)^2}_{\text{loss}} + \underbrace{\lambda_1\,||\boldsymbol{\theta}||_1}_{\text{sparsity}} + \underbrace{\lambda_2\sum_{j=1}^{p}\sum_{t=1}^{T}|\gamma_{tj}|}_{\text{task sharing}}$$



# Task relatedness

- **Tasks that are "more similar" should share more features**
- Large literature on **chemical compounds representation** and **similarity**
  - Using the **molecular graph**
  - 3D structure, physico-chemical descriptors, etc.




C.-A. Azencott et al. (2007) **One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties** JCIM

# Expert knowledge descriptors

- **DRAGON** descriptors
  - 3224 molecular features
  - atom types, functional group counts, topological descriptors
- **PowerMV**
  - 1000+ descriptors
  - atom-based, fragment-based, real-valued descriptors
  - includes: partial charges, electronegativities, estimated solubility

**Hard to define/compute and potentially incomplete.**

# Molecular fingerprints

- Define **feature vectors** that record the presence/absence (or number of occurrences) of pre-determined molecular features in a compound.

$$\phi(A) = (\phi_s(A))_{s \text{ feature}}$$

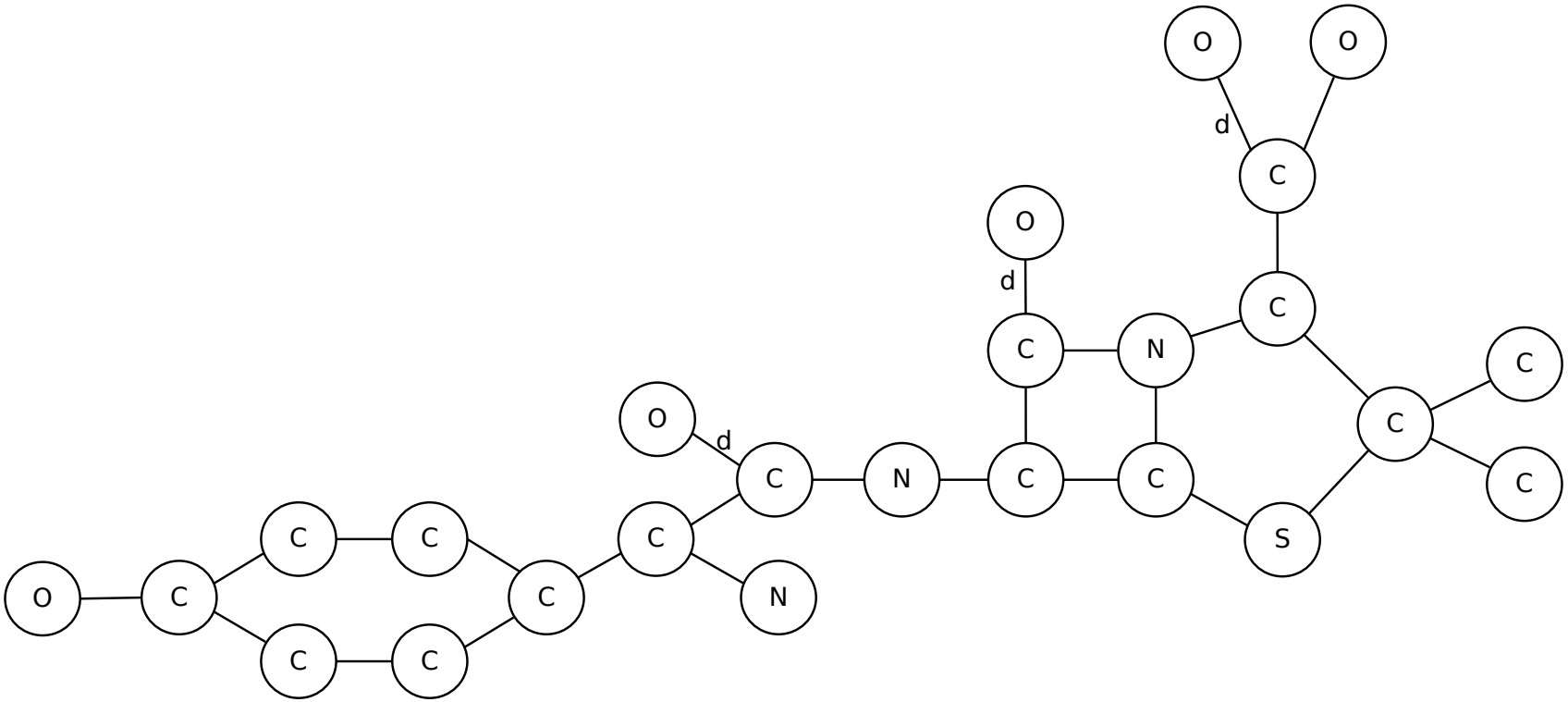
where

$$\phi_s(A) = \begin{cases} 1 & \text{if } s \text{ occurs in } A \\ 0 & \text{otherwise} \end{cases}$$

# Examples of molecular fingerprints

- **MACCS keys**: answers to a set of true/false questions about a chemical structure
  - "Are there fewer than 3 oxygen atoms?"
  - "Is there at least one halogen atom present?"
  - . . .
- **PubChem Substructure Fingerprints** (aka **CACTVS Subgraph Keys)**: presence/absence of molecular substructures, e.g.:
  - individual atoms of a given type
  - rings of a given size containing given atoms
  - more general patterns involving several bonds and atom types

# The molecular graph I




undirected labeled graph

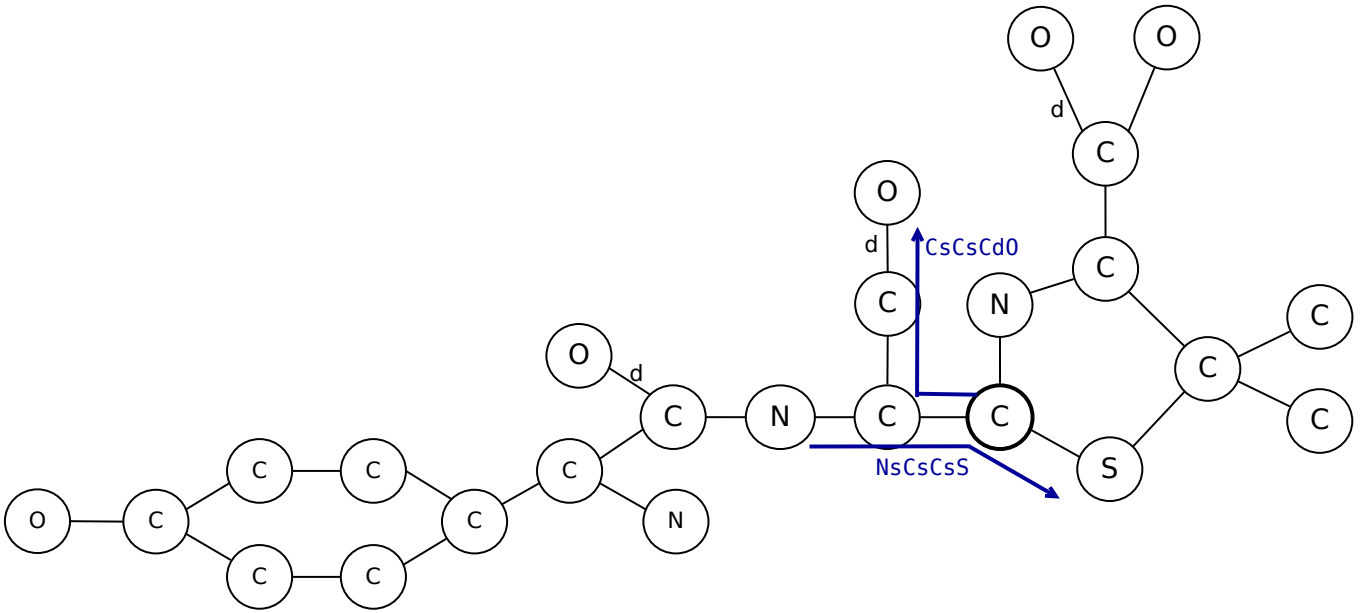
# The molecular graph II

Examples of labels:

- The **element**
- The **binding affinity**
  E.g. XSCORE: polar, hydrophobic, hydrogen-bond donating, hydrogen-bond accepting, other.
- The **element-hybridization state**
  E.g. Csp3, Nsp2.
- SYBYL: 53 labels
  E.g. sp3 carbon, trigonal planar nitrogen, halogen, sulfoxide sulfur.

# Paths fingerprints

- **Path fingerprints:** Labeled **paths** (or **walks**)

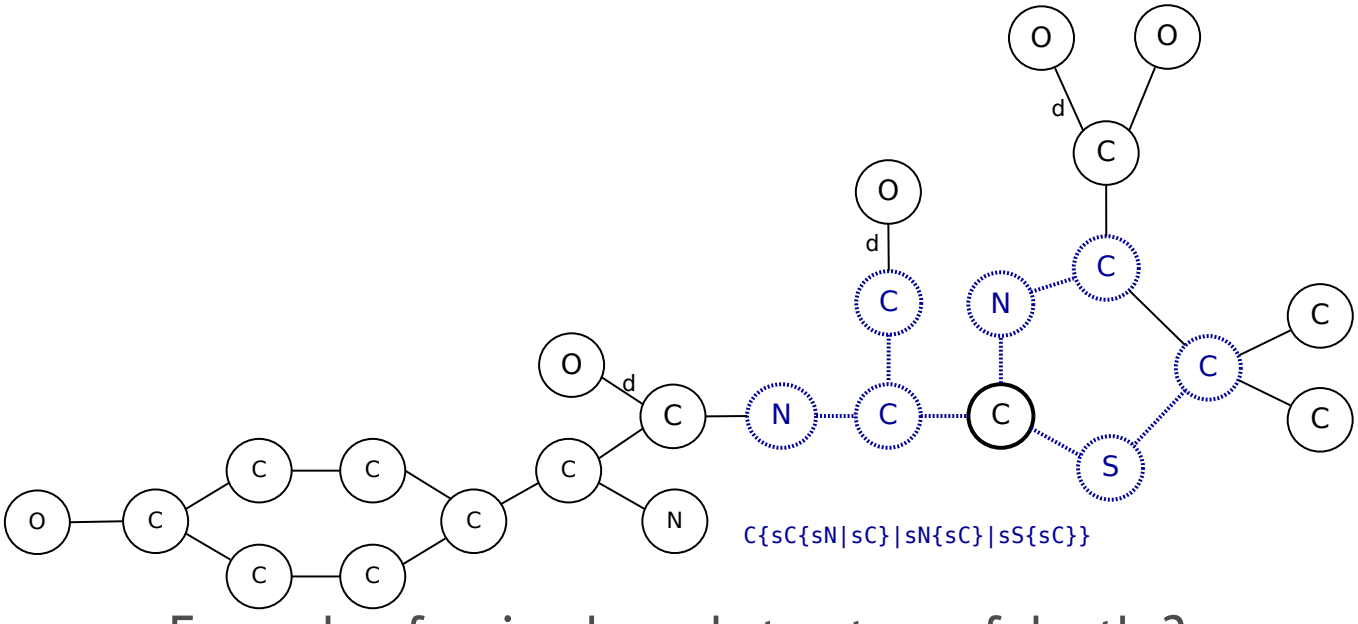



Some sub-paths of length 3

- Each entry of the fingerprint (feature vector) records the presence/absence of a given path in the molecular graph.

# Circular fingerprints

- Labeled **sub-trees**: **Extended-Connectivity** (or **Circular**) features




Example of a circular substructure of depth 2

- Each entry of the fingerprint (feature vector) records the presence/absence of a given tree (EC pattern) in the molecular graph.

# Similarity between two fingerprints

- Idea: Compare the number of entries that are **common** to the two fingerprints.
- **Binary** fingerprints $\mathcal{X} = \{0,1\}^p$ : **Tanimoto similarity**

$$s(x,z) = \frac{\sum_{j=1}^{p}(x_j \text{ AND } z_j)}{\sum_{j=1}^{p}(x_j \text{ OR } z_j)}$$

- **Count** fingerpritns $\mathcal{X} = \mathbb{N}^p$ : **Minmax similarity**

$$s(x,z) = \frac{\sum_{j=1}^{p}\min(x_j,z_j)}{\sum_{j=1}^{p}\max(x_j,z_j)}$$

- For a binary fingerprint, Tanimoto = MinMax.
- Tanimoto and MinMax are **kernels** and can be computed efficiently!

# Limitations

These representations do not model

- **accessible surface**
- **3D configuration**
  - Information can be difficult to get (use simulations)
  - Multiple conformations
- **stereoisomers**

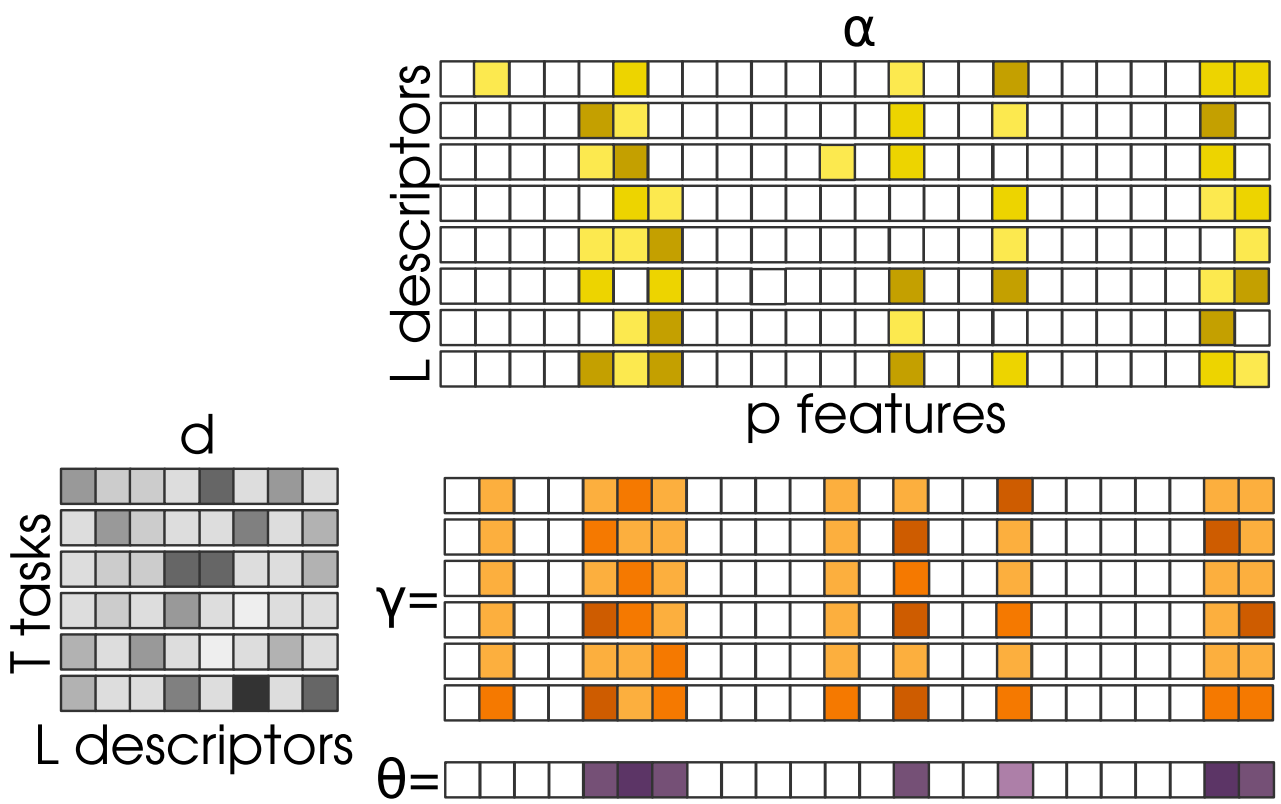
# What about large molecules?

- **Drug-like** compounds: typically about 22 heavy atoms
- What about proteins?
- **Sequence-based** representations
  - **Local Alignment kernel:** based on the Smith-Waterman alignment algorithm [Sai+04]
  - **String kernels:** [Les+04]
    - use all possible sequences of amino acids of fixed length k
    - soft penalization of mismatches
- Use **hierarchy** from the Enzyme Commission [JV08]
- Use the **3D binding pocket** information [Kel+06]

# Multiplicative Multitask Lasso with Task Descriptors

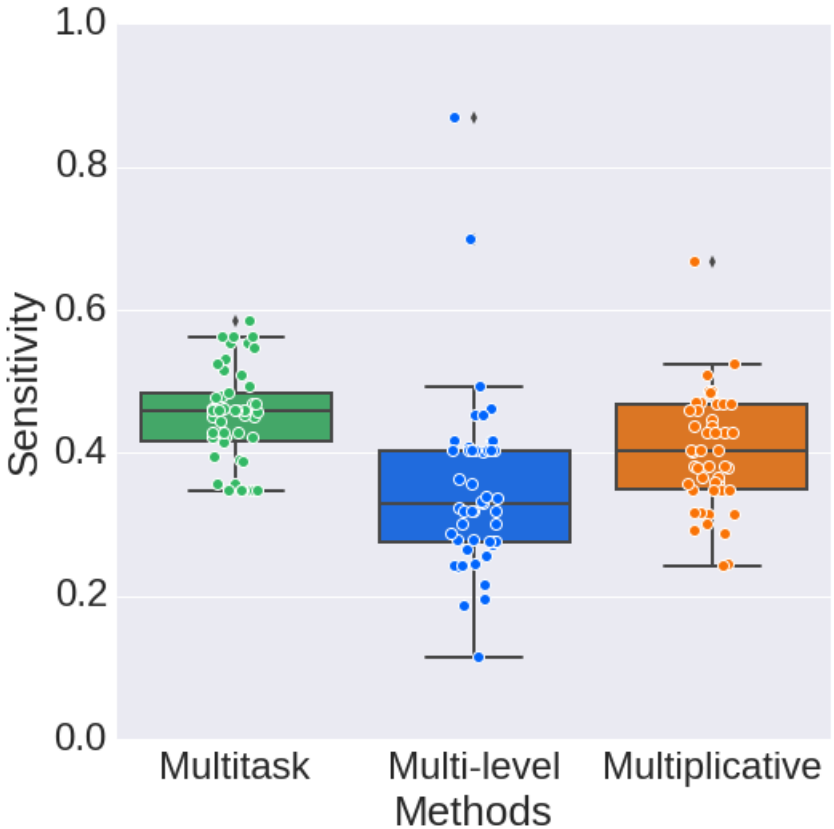
Tasks are described by $\boldsymbol{d} \in \mathbb{R}^{T \times L}$ Hypothesis: $\gamma_{tj} = \sum_{l=1}^{L} \alpha_{jl} d_{tl}$.

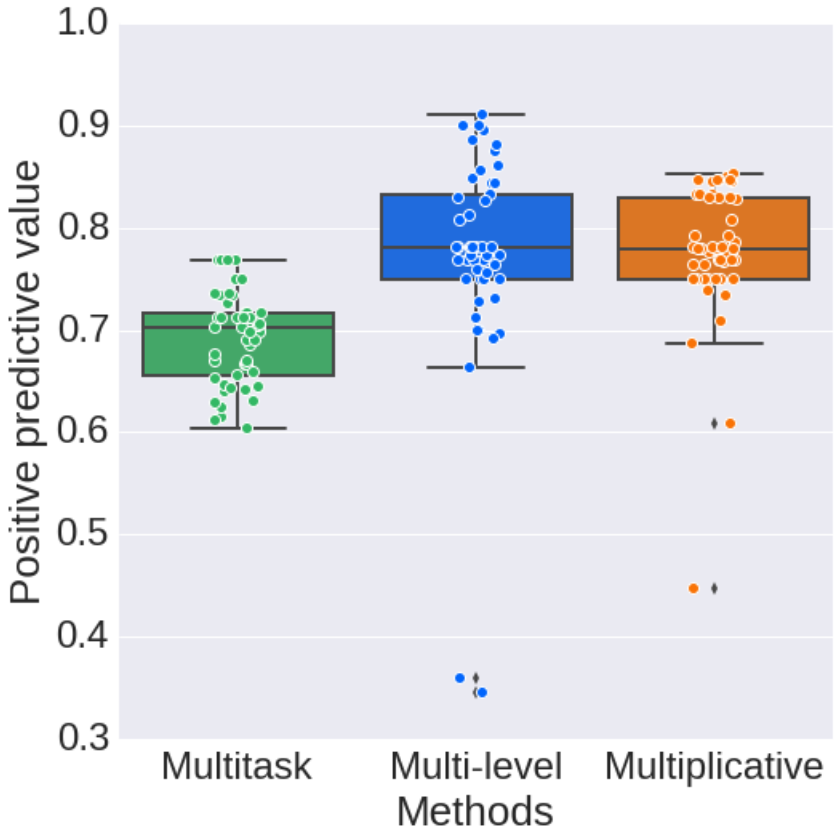
$$\underset{\boldsymbol{\theta} \in \mathbb{R}_+^p, \alpha \in \mathbb{R}^{p \times L}}{\arg\min} \underbrace{\sum_{t=1}^{T} \frac{1}{n_t} \sum_{m=1}^{n_t} \left( y_{tm} - \sum_{j=1}^{p} \theta_j \left( \sum_{l=1}^{L} \alpha_{jl} d_{tl} \right) x_{tmj} \right)^2}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{j=1}^{p} \sum_{l=1}^{L} |\alpha_{jl}|}_{\text{task sharing}}$$



# Performance of MMLD

- On **simulations:**

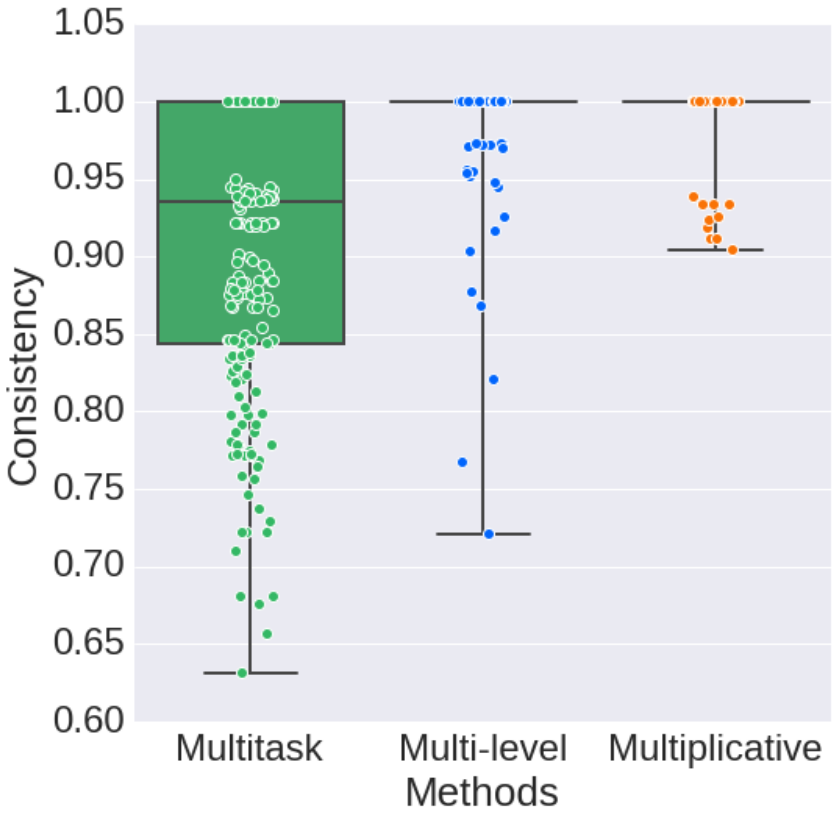


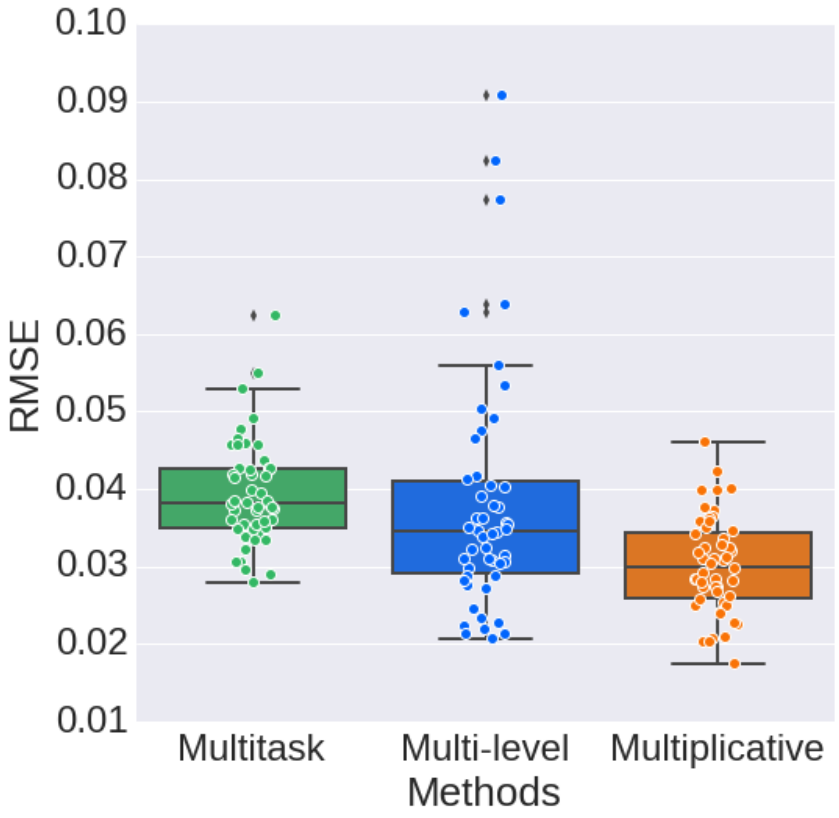




- **Sparser** solution
- Better **recovery of true features** (higher PPV)

# Performance of MMLD

- On **simulations:**

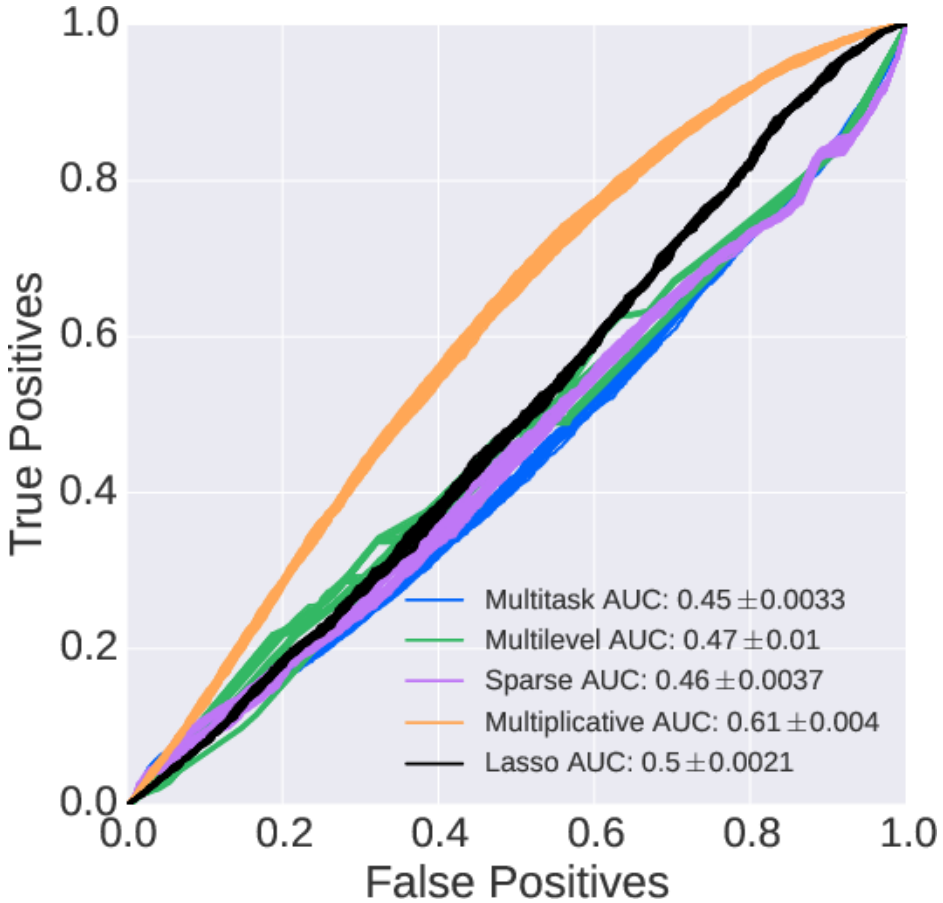



- Improved **stability** (consistency)
- Better **predictivity** (RMSE).

# Performance of MMLD

– Making predictions for tasks for which you have **no data.**




V. Bellón, V. Stoven, and C.-A. Azencott (2016) **Multitask feature selection with task descriptors**, PSB.

https://github.com/vmolina/MultitaskDescriptor

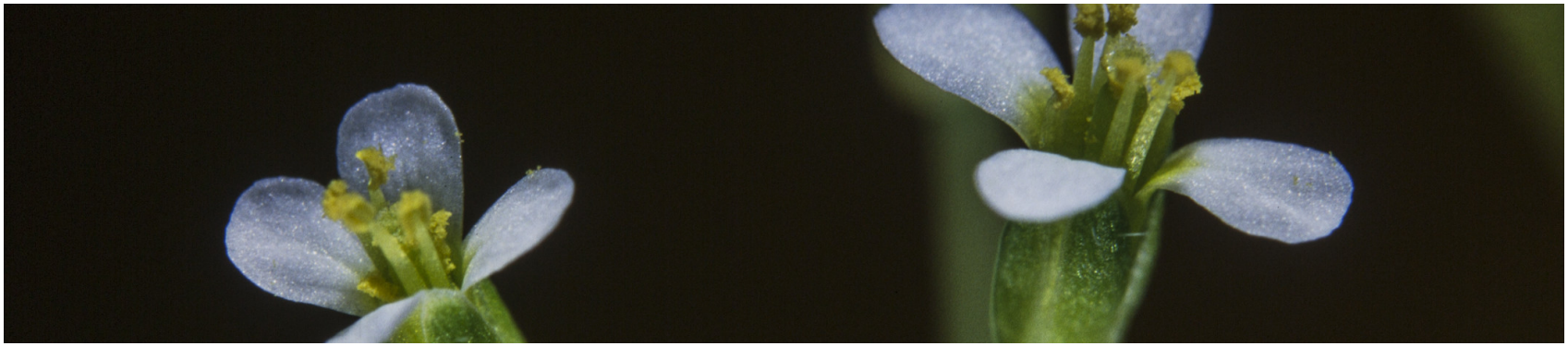
# Randomized MMLD

- **Stability:** Selecting the same features on multiple related data sets (or multiple subsets of the same data set)
- Crucial for **interpretability**
- **Randomized Lasso:** Summarize many repetitions on subsamples of the data in a consensus feature selection [MB10].

  $\Rightarrow$ **Randomized MMLD**

# Experimental results

*Arabidopsis thaliana* genotypes



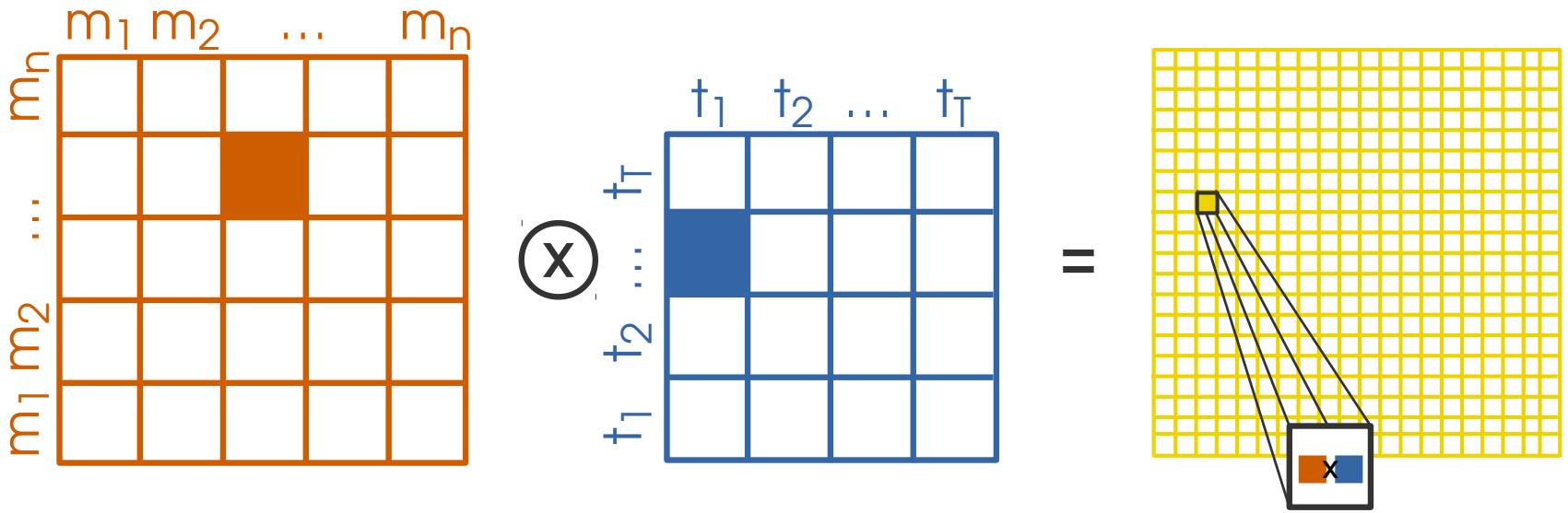
- 6 flowering time phenotypes (= tasks)
- $\sim$ 180 samples per task, $\sim$ 280 000 SNPs
- **candidate SNPs** = SNPs from genes known to be related to flowering time.

| | # Selected SNPS | Consistency | Consistency on candidate SNPs |
|---|---|---|---|
| Lasso | $871 \pm 12$ | $0.21 \pm 0.04$ | $0.01 \pm 0.01$ |
| MMLD | $616 \pm 71$ | $0.18 \pm 0.03$ | $0.01 \pm 0.07$ |
| Randomized MMLD | $2295 \pm 118$ | $0.37 \pm 0.02$ | $0.45 \pm 0.13$ |

# Multitask kernel methods

## Kronecker kernel:

$K((mol1, task1), (mol2, task2)) = K_{mol}(mol1, mol2) \times K_{task}(task1, task2)$




## Computational issues:

- n molecules, T tasks $\Rightarrow$ nT observations
- Can be solved by using **mostly nearby observations**

  B. Playe, C.-A. Azencott, and V. Stoven (2017) **Efficient multi-task chemogenomics for drug specificity prediction.** bioRxiv

# Conclusion

- Applying **machine learning** to **genomics data** is challenging because of the **dimensionality of the data**
- "Large p, small n" problem
- Reducing p: **interpretable dimensionality reduction**

  $\Rightarrow$ **regularization**
- Increasing n: **multi-task** approaches.

# Current ML challenges

- **Robustness/stability**

  Recovering the same biomarkers when the data changes slightly.

- **Complex interaction patterns**

  - Limited to additive or quadrative effects
  - Little existing work on complex ML models.

- **Statistical significance**

  - Computing p-values
  - Correcting for multiple hypotheses
  - Controlling false discovery rate.

# 4. A few hot topics around ML and health

# Further challenges: Privacy

- More data $\rightarrow$ Data sharing $\rightarrow$ **ethical** concerns
- How to learn from **privacy-protected** patient data?

C.-A. Azencott. **Machine learning and genomics: precision medicine vs. patient privacy.** Phil Trans Roy Soc A 2018.

# Further challenges: Heterogeneity

- Multiple relevant **data sources** and **types**
- Multiple (unknown) **populations** of samples.

**Tumor heterogeneity**

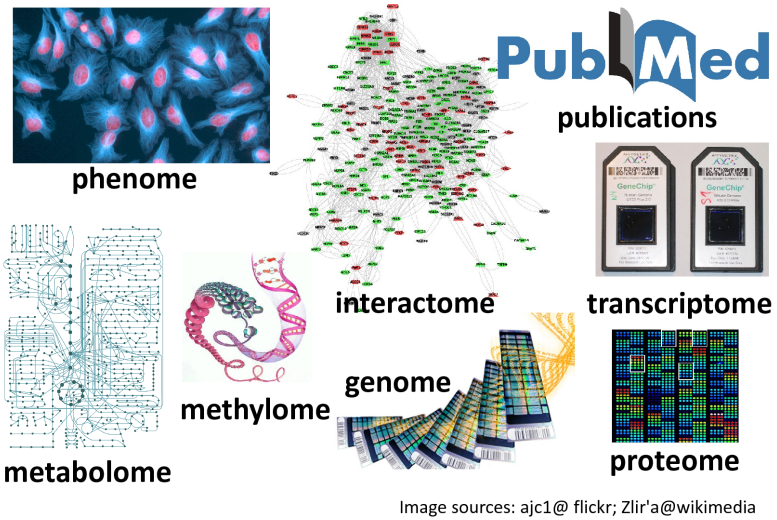
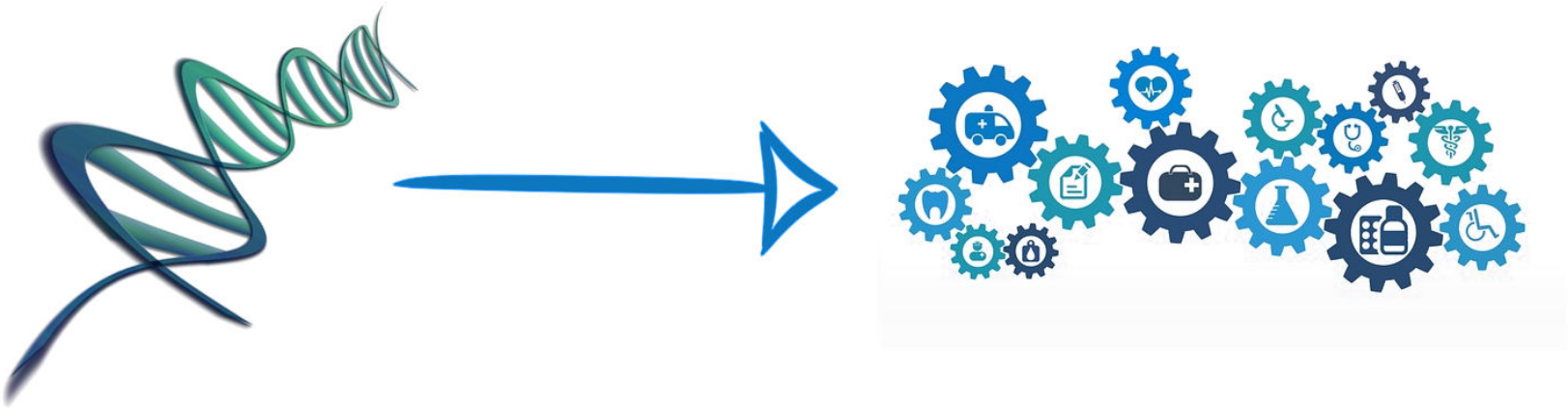



Image sources: ajc1@ flickr; Zlir'a@wikimedia

**Heterogeneous data sources**

[GBG16]

# Further challenges: Risk prediction



- State of the art: **Polygenic Risk Scores**

  **Linear** combination of SNPs with high p-values (**summary statistics**)
  Weighted by log odd ratios / univariate linear regression coefficients.

- **Complex models** slow to adopt – **reliability?** [Oks+14; SS17]

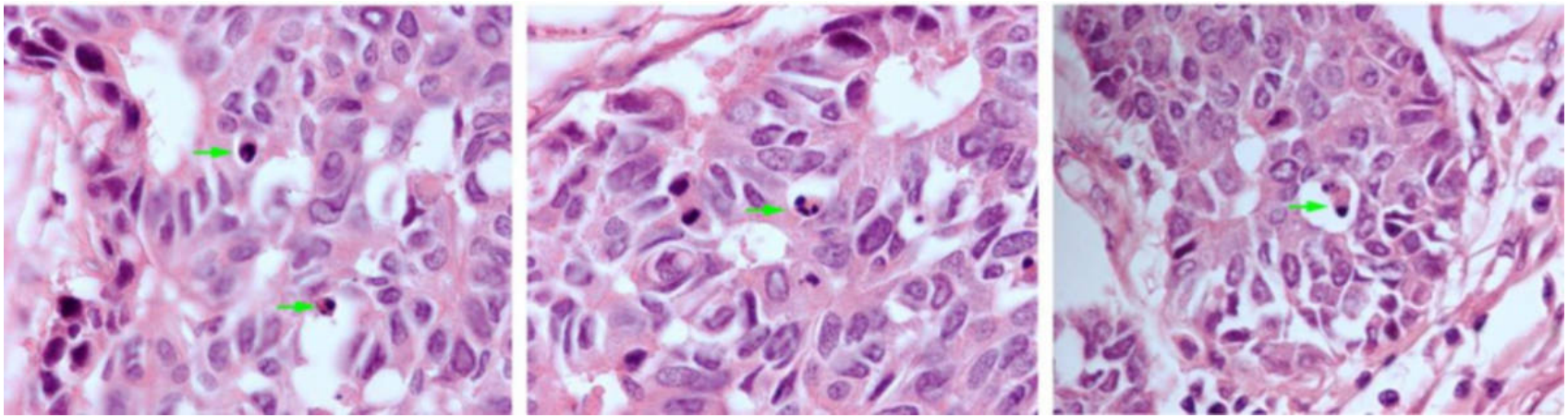
# Further challenges: Bioimage informatics

High-throughput **molecular** and **cellular** images

- **Subcellular location** analysis
- **High-content screening**
- Segmentation, tracking, registration.

**BioImage Informatics** `http://bioimageinformatics.org/`

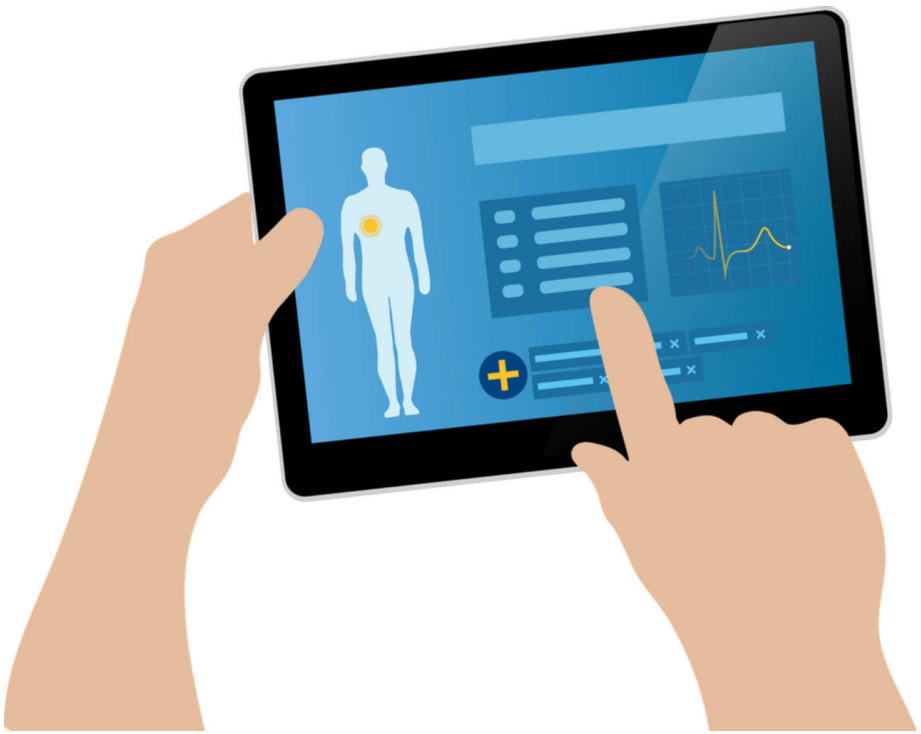
[Est+17; Che+18]



**Detecting cells undergoing apoptosis**

# Further challenges: Electronic health records

- **Clinical notes:** incomplete, imbalanced, time series
- Combine **text** + **images** + **genetics**
- Assisting **evidence-based medicine**



[Mio+16; Raj+18]

# Molecules

- **Representation learning:**
  - Use deep learning to **learn better representations** of molecular graphs [LPB13; Duv+15; Col+17a; HYL17]
  - But not only deep learning
    S. Rensi and R. B. Altman (2017) **Shallow Representation Learning via Kernel PCA Improves QSAR Modelability.** JCIM 57 (8), 1859–1967
- Use **(deep) generative models** to come up with new chemicals (still very much exploratory) [Kad+17; CKK17]
- **Deep learning**
  - **ML Docking:** Protein-ligand docking from 3D representations [Rag+16]
  - Y. Xu et al. (2017) **Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships.** JCIM Article ASAP.

# A few starting places

## Data and Challenges

- **DREAM Challenges:** Crowdsourcing challenges for biology and medicine `http://dreamchallenges.org/`
- **Epidemium:** Cancer research through data challenges `http://www.epidemium.cc/`
- **MIMIC:** Deidentified electronic health records `https://mimic.physionet.org/`
- **BioImage Informatics Challenges** `https://bii.eecs.wsu.edu/challenges/`

# References


[AB06] Christopher P. Adams and Van V. Brantner. "Estimating the cost of new drug development: is it really $802 million?" In: Health Aff 25.2 (2006), pp. 420–428.

[AL11] David H. Alexander and Kenneth Lange. "Stability selection for genome-wide association". In: Genetic Epidemiology 35.7 (2011), pp. 722–728.

[AR15] Samuel J. Aronson and Heidi L. Rehm. "Building the foundation for genomics in precision medicine". In: Nature 526.7573 (2015), pp. 336–342.

[Aze+07] Chloé-Agathe Azencott et al. "One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties". In: Journal of Chemical Information and Modeling 47.3 (2007), pp. 965–974.

[Aze+13] Chloé-Agathe Azencott et al. "Efficient network-guided multi-locus association mapping with graph cuts". In: Bioinformatics 29.13 (2013), pp. i171–i179.

[Aze16] Chloé-Agathe Azencott. "Network-guided biomarker discovery". In: Machine Learning for Health Informatics. Lecture Notes in Computer Science 9605. Springer International Publishing, 2016.

[Baj08] Jürgen Bajorath. "Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery". In: Informa Healthcare 3.12 (2008), pp. 1371–1376.

[Baj15] Jürgen Bajorath. "Computer-aided drug discovery". In: F1000Research 4 (2015).

[Baj17] Jürgen Bajorath. "Computational scaffold hopping: cornerstone for the future of drug design?" In: Future Medicinal Chemistry 9.7 (2017), pp. 629–631.

[Bar+12] Fredrik BarrenÃďs et al. “Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms”. In: Genome Biology 13 (2012), R46.

[Bel17] Victor Bellon. “Prédiction personalisée des effets secondaires indésirables de médicaments”. PhD thesis. 2017.

[BMG96] Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. “The art and practice of structure-based drug design: A molecular modeling perspective”. In: Med. Res. Rev. 16.1 (1996), pp. 3–50.

[Bot+17] V. Botu et al. “Machine learning force fields: construction, validation, and outlook”. In: The Journal of Physical Chemistry C 121.1 (2017), pp. 511–522.

[BR09] Lorenz C. Blum and Jean-Louis Reymond. “970 million druglike small molecules for virtual screening in the chemical universe database GDB-13”. In: Journal of the American Chemical Society 131.25 (2009), pp. 8732–8733.

[BR15] Venkatesh Botu and Rampi Ramprasad. “Ab-initio molecular dynamics acceleration scheme with an adaptive machine learning framework”. In: (2015). arXiv: 1410.3353.

[BSA16] Victor Bellon, VÃľronique Stoven, and ChloÃĺ-Agathe Azencott. “Multitask feature selection with task descriptors”. In: Pacific Symposium on Biocomputing. Vol. 21. 2016, pp. 261–272.

[CA17] HÃľctor Climente and ChloÃĺ-Agathe Azencott. martini: GWAS incorporating networks in R. 2017. URL: https://bioconductor.org/packages/devel/bioc/html/martini.html.

[Cam+11] Aurel Cami et al. “Predicting adverse drug events using pharmacological network models”. In: Science Translational Medicine 3.114 (2011), 114ra127–114ra127.

[Car97] Rich Caruana. "Multitask Learning". In: Machine Learning 28.1 (1997), pp. 41–75.

[CH+08] Gerda Claeskens, Nils Lid Hjort, et al. "Model selection and model averaging". In: Cambridge Books (2008).

[Che+18] Po-Hsuan Cameron Chen et al. An Augmented Reality Microscope for Real-time Automated Detection of Cancer. Tech. rep. Google AI Healthcare, 2018.

[Chi+17] Travers Ching et al. "Opportunities and obstacles for deep learning in biology and medicine". In: bioRxiv (2017), p. 142760.

[Chm+17] Stefan Chmiela et al. "Machine learning of accurate energy-conserving molecular force fields". In: Science Advances 3.5 (2017), e1603015.

[CKB10] Jonathan H. Chen, Matthew A. Kayala, and Pierre Baldi. "Reaction Explorer: Towards a knowledge map of organic chemistry to support dynamic assessment and personalized instruction". In: Enhancing Learning with Online Resources, Social Networking, and Digital Libraries. Vol. 1060. 1060. 2010, pp. 191–209.

[CKK17] Mehdi Cherti, Balazs Kegl, and Akin Kazakci. "De novo drug design with deep generative models : an empirical study". In: ICLR. 2017.

[Cla+08] Robert Clarke et al. "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data". In: Nature Reviews Cancer 8.1 (2008), pp. 37–49.

[Col+17a] Connor W. Coley et al. "Convolutional embedding of attributed molecular graphs for physical property prediction". In: J. Chem. Inf. Model. 57.8 (2017), pp. 1757–1772.

[Col+17b] Connor W. Coley et al. “Prediction of organic reaction outcomes using machine learning”. In: ACS Cent. Sci. 3.5 (2017), pp. 434–443.

[Cow+17] Lenore Cowen et al. “Network propagation: a universal amplifier of genetic associations”. In: Nature Reviews Genetics 18.9 (2017), pp. 551–562.

[CZ14] Feixiong Cheng and Zhongming Zhao. “Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties”. In: J Am Med Inform Assoc 21 (2014), e278–e286.

[DGH16] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. “Innovation in the pharmaceutical industry: new estimates of R&D costs”. In: Journal of health economics 47 (2016), pp. 20–33.

[DHG03] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. “The price of innovation: new estimates of drug development costs”. In: Journal of Health Economics 22.2 (2003), pp. 151–185.

[DJS14] George E. Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. “Multi-task neural networks for QSAR predictions”. In: (2014). arXiv: 1406.1231.

[DL16] Oleg T. Devinyak and Roman B. Lesyk. “5-Year trends in QSAR and its machine learning methods”. In: Curr Comput Aided Drug Des 12.4 (2016), pp. 265–271.

[DP15] Nicoletta DessÃň and Barbara Pes. “Stability in Biomarker Discovery: Does Ensemble Feature Selection Really Help?” In: Current Approaches in Applied Artificial Intelligence. Lecture Notes in Computer Science 9101. DOI: 10.1007/978-3-319-19066-2_19. 2015, pp. 191–200.

[Dre16] Liam Drew. “Pharmacogenetics: the right drug for you”. In: Nature 537.7619 (2016), S60–S62.

[Duv+15] David K Duvenaud et al. "Convolutional networks on graphs for learning molecular fingerprints". In: Advances in Neural Information Processing Systems 28. 2015, pp. 2224–2232.

[ED+05] Liat Ein-Dor et al. "Outcome signature genes in breast cancer: is there a unique set?" In: Bioinformatics 21.2 (2005), pp. 171–178.

[Edu+15] Federica Eduati et al. "Prediction of human population responses to toxic compounds by a collaborative competition". In: Nature Biotechnology 33.9 (2015), pp. 933–940.

[Eri+17] Spencer S. Ericksen et al. "Machine learning consensus scoring improves performance across targets in structure-based virtual screening". In: J. Chem. Inf. Model. 57.7 (2017), pp. 1579–1590.

[Est+17] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: Nature 542.7639 (2017), p. 115.

[EW11] Sean Ekins and Antony J. Williams. "Finding promiscuous old drugs for new uses". In: Pharmaceutical Research 28.8 (2011), pp. 1785–1791.

[Fra05] Simon Frantz. "Drug discovery: Playing dirty". In: Nature 437.7061 (2005), pp. 942–943.

[Fur13] Laura I. Furlong. "Human diseases through the lens of network biology". In: Trends in Genetics 29.3 (2013), pp. 150–159.

[Gas16] Johann Gasteiger. "Chemoinformatics: achievements and challenges, a personal view". In: Molecules 21.2 (2016), p. 151.

[GBG16] Laura Gay, Ann-Marie Baker, and Trevor A Graham. "Tumour cell heterogeneity". In: F1000Research 5 (2016).

[Gig+15] Sébastien Giguère et al. "Machine Learning Assisted Design of Highly Active Peptides for Drug Discovery". In: PLOS Computational Biology 11.4 (2015), e1004074.

[Gra+15] G'Sell Max Grazier et al. "Sequential selection procedures and false discovery rate control". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78.2 (2015), pp. 423–444.

[GVB10] Hanna Geppert, Martin Vogt, and Jürgen Bajorath. "Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation". In: Journal of Chemical Information and Modeling (2010).

[GVB13] Levi A. Garraway, Jaap Verweij, and Karla V. Ballman. "Precision oncology: an overview". In: J. Clin. Oncol. 31.15 (2013), pp. 1803–1805.

[Hea+15] Megan L Head et al. "The extent and consequences of p-hacking in science". In: PLoS biology 13.3 (2015), e1002106.

[HGV11] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures". In: PLoS ONE 6.12 (2011), e28210.

[Hol18] Susan Holmes. "Statistical proof? The problem of irreproducibility." In: Bulletin (New Series) of the American Mathematical Society 55.1 (2018).

[Hop09] Andrew L. Hopkins. "Drug discovery: Predicting promiscuity". In: Nature 462.7270 (2009), pp. 167–168.

[HYL17] William L. Hamilton, Rex Ying, and Jure Leskovec. "Representation learning on graphs: methods and applications". In: arxiv:1709.05584 [cs] (2017). arXiv: 1709.05584.

[Ioa05] John PA Ioannidis. "Why most published research findings are false". In: PLoS medicine 2.8 (2005), e124.

[Ior+16] Francesco Iorio et al. "A landscape of pharmacogenomic interactions in cancer". In: Cell 166.3 (2016), pp. 740–754.

[JOV09] L. Jacob, G. Obozinski, and J.-P. Vert. "Group lasso with overlap and graph lasso". In: ICML. 2009, pp. 433–440.

[JV07] Laurent Jacob and Jean-Philippe Vert. "Kernel methods for in silico chemogenomics". In: arXiv:0709.3931 (2007).

[JV08] Laurent Jacob and Jean-Philippe Vert. "Protein-ligand interaction prediction: an improved chemogenomics approach". In: Bioinformatics 24.19 (2008), pp. 2149–2156.

[Kad+17] Artur Kadurin et al. "druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico". In: Mol. Pharmaceutics 14.9 (2017), pp. 3098–3104.

[Kel+06] Esther Kellenberger et al. "sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank". In: J. Chem. Inf. Model. 46.2 (2006), pp. 717–727.

[Kun07] Ludmila I. Kuncheva. "A stability index for feature selection". In: Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications. AIAP'07. Anaheim, CA, USA: ACTA Press, 2007, pp. 390–395.

[Laz+14] David Lazer et al. "The parable of Google Flu: traps in big data analysis". In: Science 343.6176 (2014), pp. 1203–1205.

[Les+04] Christina S. Leslie et al. "Mismatch string kernels for discriminative protein classification". In: Bioinformatics 20.4 (2004), pp. 467–476.

[Li11] C. Li. "Personalized medicine – the promised land: are we there yet?" In: Clinical Genetics 79.5 (2011), pp. 403–412.

[LJ12] Yvonne Y Li and Steven JM Jones. "Drug repositioning for personalized medicine". In: Genome Medicine 4.3 (2012), p. 27.

[LL08] C. Li and H. Li. "Network-constrained regularization and variable selection for analysis of genomic data". In: Bioinformatics 24.9 (2008), pp. 1175–1182.

[LL10] Caiyan Li and Hongzhe Li. "Variable selection and regression analysis for graph-structured covariates with an application to genomics". In: The annals of applied statistics 4.3 (2010), pp. 1498–1516.

[Loc+14] Richard Lockhart et al. "A significance test for the lasso". In: Annals of statistics 42.2 (2014), p. 413.

[Lou+12] Eugen Lounkine et al. "Large-scale prediction and testing of drug activity on side-effect targets". In: Nature 486.7403 (2012), pp. 361–367.

[LPB13] Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. "Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules". In: J. Chem. Inf. Model. 53.7 (2013), pp. 1563–1575.

[LPC98] J. Lazarou, B. H. Pomeranz, and P. N. Corey. "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies". In: JAMA 279.15 (1998), pp. 1200–1205.

[Man+09] Teri A. Manolio et al. "Finding the missing heritability of complex diseases". In: Nature 461.7265 (2009), pp. 747–753.

[Man10] Teri A. Manolio. "Genome-wide association studies and assessment of the risk of disease". In: N. Engl. J. Med. 363.2 (2010), pp. 166–176.

[Man13] Teri A Manolio. "Bringing genome-wide association findings into clinical use". In: Nature Reviews Genetics 14.8 (2013), pp. 549–558.

[Mar+17] Eric J. Martin et al. "Profile-QSAR 2.0: Kinase virtual screening accuracy comparable to four-concentration ic50s for realistically novel compounds". In: J. Chem. Inf. Model. 57.8 (2017), pp. 2077–2088.

[MB10] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: J. R. Stat. Soc. 72.4 (2010), pp. 417–473.

[Mig+12] A. Miguel et al. "Frequency of adverse drug reactions in hospitalized patients: a systematic review and meta-analysis." In: Pharmacoepidemiol Drug Saf 21.11 (2012), pp. 1139–1154.

[Mio+16] Riccardo Miotto et al. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In: Scientific reports 6 (2016), p. 26094.

[NB16] Sarah Nogueira and Gavin Brown. "Measuring the stability of feature selection". In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science 9852. Springer International Publishing, 2016, pp. 442–457.

[Nos16] Nicola Nosengo. "Can you teach old drugs new tricks?" In: Nature News 534.7607 (2016), p. 314.

[Nuz14] Regina Nuzzo. "Scientific method: statistical errors". In: Nature News 506.7487 (2014), p. 150.

[Oks+14] Sebastian Okser et al. "Regularized machine learning in the genetic prediction of complex traits". In: PLoS genetics 10.11 (2014), e1004754.

[OTJ06] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. Multi-task feature selection. Tech. rep. UC Berkeley, 2006.

[PAS18] Benoit Playe, Chloe-Agathe Azencott, and Veronique Stoven. "Efficient multi-task chemogenomics for drug specificity prediction". In: bioRxiv (2018), p. 193391.

[Pen07] Elizabeth Pennisi. "Human Genetic Variation". In: Science 318.5858 (2007), pp. 1842–1843.

[PM17] Joshua Pottel and Nicolas Moitessier. "Customizable generation of synthetically accessible, local chemical subspaces". In: J. Chem. Inf. Model. 57.3 (2017), pp. 454–467.

[RA17] Stefano E. Rensi and Russ B. Altman. "Shallow representation learning via kernel PCA improves QSAR modelability". In: J. Chem. Inf. Model. 57.8 (2017), pp. 1859–1867.

[Rag+16] Matthew Ragoza et al. "Protein-ligand scoring with convolutional neural networks". In: (2016). arXiv: 1612.02751.

[Raj+18] Alvin Rajkomar et al. "Scalable and accurate deep learning with electronic health records". In: NPJ Digital Medicine 1.1 (2018), p. 18.

[Ral+05] Liva Ralaivola et al. "Graph kernels for chemical informatics". In: Neural Networks 18.8 (2005), pp. 1093–1110.

[Raz82] Razinger. "Extended connectivity in chemical graphs". In: Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 61.6 (1982), pp. 581–586.

[Rey15] Jean-Louis Reymond. "The Chemical Space Project". In: Acc. Chem. Res. 48.3 (2015), pp. 722–730.

[Sad+16] Peter Sadowski et al. "Synergies between quantum mechanics and machine learning in reaction prediction". In: J. Chem. Inf. Model. 56.11 (2016), pp. 2125–2128.

[Sai+04] Hiroto Saigo et al. "Protein homology detection using string alignment kernels". In: Bioinformatics 20.11 (2004), pp. 1682–1689.

[Sch+17] Kristof T. Schütt et al. "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions". In: (2017). arXiv: 1706.08566.

[Sch15] Nicholas J. Schork. "Personalized medicine: Time for one-person trials". In: Nature News 520.7549 (2015), p. 609.

[SL12] Grzegorz Swirszcz and Aurelie C. Lozano. "Multi-level Lasso for Sparse Multi-task Regression". In: Proceedings of the 29th International Conference on Machine Learning (ICML-12). 2012, pp. 361–368.

[Sny+14] Alexandra Snyder et al. "Genetic basis for clinical response to CTLA-4 blockade in melanoma". In: New England Journal of Medicine 371.23 (2014), pp. 2189–2199.

[SS13] Rajen D. Shah and Richard J. Samworth. "Variable selection with error control: another look at stability selection". In: Journal of the Royal Statistical Society 75.1 (2013), pp. 55–80.

[SS17] Hon-Cheong So and Pak C Sham. "Improving polygenic risk prediction from summary statistics by an empirical Bayes approach". In: Scientific Reports 7 (2017), p. 41262.

[Sug+14] Mahito Sugiyama et al. "Multi-task feature selection on multiple networks via maximum flows". In: SIAM ICDM. 2014, pp. 199–207.

[Swa+09] S. Joshua Swamidass et al. "Influence Relevance Voting: An accurate and interpretable virtual high throughput screening method". In: Journal of Chemical Information and Modeling 49.4 (2009), pp. 756–766.

[Sør+01] Therese Sørlie et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". In: Proceedings of the National Academy of Sciences 98.19 (2001), pp. 10869–10874.

[TB16] Alexander Tropsha and Jürgen Bajorath. "Computational methods for drug discovery and design". In: J. Med. Chem. 59.1 (2016).

[Tib94] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: J. R. Stat. Soc. 58 (1994), pp. 267–288.

[Var17] Alexandre Varnek. Tutorials in Chemoinformatics. 2017.

[VDD11] David Venet, Jacques E. Dumont, and Vincent Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In: PLoS Computational Biology 7.10 (2011), e1002240.

[Vis+17] Peter M Visscher et al. "10 years of GWAS discovery: biology, function, and translation". In: The American Journal of Human Genetics 101.1 (2017), pp. 5–22.

[VV+02] Laura J Van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: nature 415.6871 (2002), p. 530.

[WDAG16] Jennifer N. Wei, David Duvenaud, and AlÃąn Aspuru-Guzik. "Neural networks for the prediction of organic chemistry reactions". In: ACS Cent. Sci. 2.10 (2016), pp. 725–732.

[Wei88] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: Journal of Chemical Information and Computer Sciences 28.1 (1988), pp. 31–36.

[WL+11] M. C. Wu, S. Lee, et al. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test". In: AJHG 89.1 (2011), pp. 82–93.

[WL+16] Ronald L Wasserstein, Nicole A Lazar, et al. "The ASAâĂŹs statement on p-values: context, process, and purpose". In: The American Statistician 70.2 (2016), pp. 129–133.

[WWW89] David Weininger, Arthur Weininger, and Joseph L. Weininger. "SMILES. 2. Algorithm for generation of unique SMILES notation". In: Journal of Chemical Information and Computer Sciences 29.2 (1989), pp. 97–101.

[XS00] Jun Xu and James Stevenson. "Drug-like Index:âĂĽ A new approach to measure drug-like compounds and their diversity". In: J. Chem. Inf. Comput. Sci. 40.5 (2000), pp. 1177–1187.

[Xu+17] Yuting Xu et al. "Demystifying multitask deep neural networks for quantitative structure-activity relationships". In: J. Chem. Inf. Model. (2017).

[Yad17] Maneesh K. Yadav. "On the synthesis of machine learning and automated reasoning for an artificial synthetic organic chemist". In: New J. Chem. 41.4 (2017), pp. 1411–1416.

[YL06] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006), pp. 49–67.

[YPK12] Yoshihiro Yamanishi, Edouard Pauwels, and Masaki Kotera. "Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces". In: J. Chem. Inf. Model. 52.12 (2012), pp. 3284–3292.

[ZH05] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005), pp. 301–320.