Utiliser l'IA pour la recherche sur le cancer

Promesses et défis de l'apprentissage statistique pour la génomique

Chloé-Agathe Azencott

Centre for Computational Biology (CBIO)
Mines Paris – PSL, Institut Curie & INSERM U1331
PR[AI]RIE-PSAI

21 mai 2026

http://cazencott.info chloe-agathe.azencott@minesparis.psl.eu @cazencott@lipn.info

MINES PARIS | PSL

PR[AI]RIE Paris School

institut Curie

Inserm

Machine learning

- Learn/build/define a **statistical model** using **data**
- **Model:** a function of input variables

$$f : \mathbb{R}^p \to \mathbb{R}$$
$$\vec{x} \mapsto \dots$$

```
def model(x):
    ...
    return ...
```

Supervised machine learning problems

- **Supervised** machine learning: learn a **predictive model**

Example: cancer detection from liquid biopsies

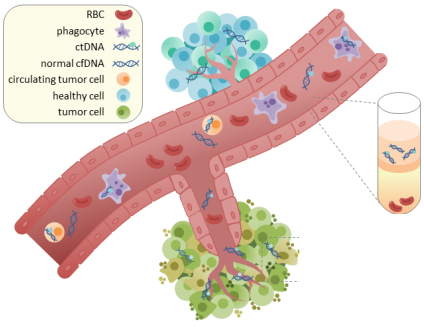
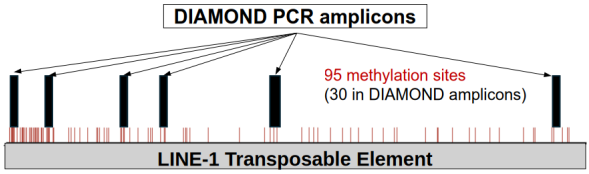


Image by Racheljunewong via Wikimedia Commons



ML task: predict **cancer vs healthy** status from the methylation levels of 30 LINE-1 CpG sites

M. Michel et al. **Noninvasive multicancer detection using DNA hypomethylation of LINE-1 retrotransposons** Clin Can Res 2024

Example: cancer detection from liquid biopsies

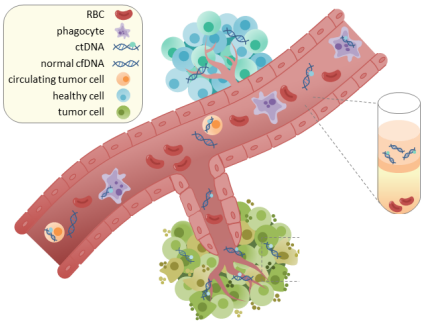
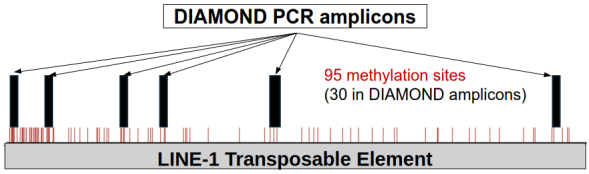


Image by Racheljunewong via Wikimedia Commons



	Discovery	Validation
Early ovarian	95% (n=18)	56% (n=45)
Metastatic colorectal	100% (n=75)	95%(n=17)
Sensitivity at 99% specificity		

ML task: predict **cancer vs healthy** status from the methylation levels of 30 LINE-1 CpG sites

M. Michel et al. **Noninvasive multicancer detection using DNA hypomethylation of LINE-1 retrotransposons** Clin Can Res 2024

Example: cancer detection from liquid biopsies

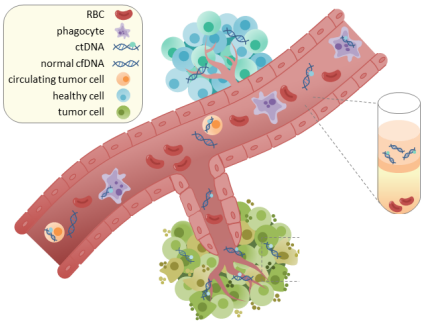
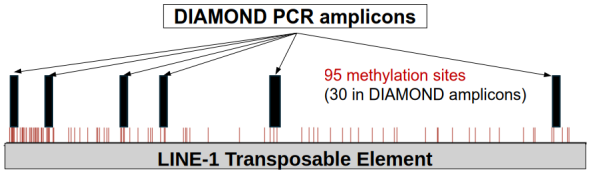


Image by Racheljunewong via Wikimedia Commons



	Discovery	Validation
Early ovarian	95% (n=18)	56% (n=45)
Metastatic colorectal	100% (n=75)	95%(n=17)

Sensitivity at 99% specificity

ML task: predict **cancer vs healthy** status from the methylation levels of 30 LINE-1 CpG sites

M. Michel et al. **Noninvasive multicancer detection using DNA hypomethylation of LINE-1 retrotransposons** Clin Can Res 2024

Example: cancer detection from liquid biopsies

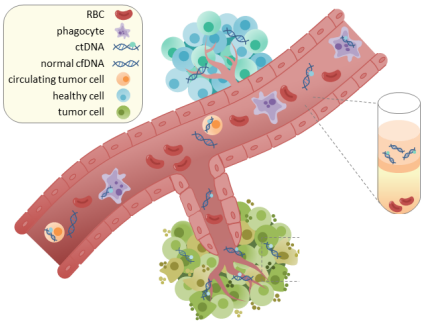
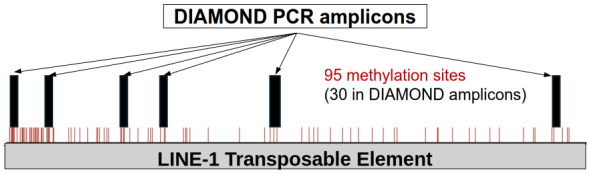


Image by Racheljunewong via Wikimedia Commons



	Discovery	Validation
Early ovarian	95% (n=18)	56% (n=45)
Metastatic colorectal	100% (n=75)	95%(n=17)
Sensitivity at 99% specificity		

ML task: predict **cancer vs healthy** status from the methylation levels of 30 LINE-1 CpG sites

M. Michel et al. **Noninvasive multicancer detection using DNA hypomethylation of LINE-1 retrotransposons** Clin Can Res 2024

Example: cancer detection from liquid biopsies

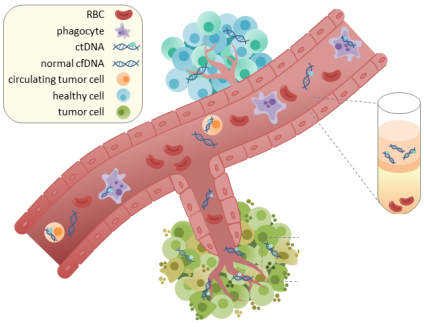
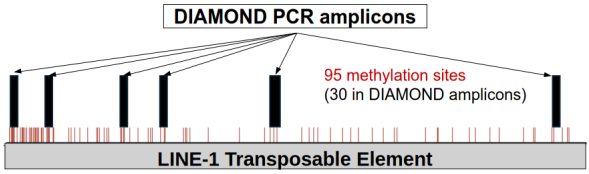


Image by Racheljunewong via Wikimedia Commons



	Discovery	Validation
Early ovarian	95% (n=18)	56% (n=45)
Metastatic colorectal	100% (n=75)	95%(n=17)
Sensitivity at 99% specificity		
Early ovarian (AUC)	1.00 (n=18)	0.96 (n=45)

ML task: predict **cancer vs healthy** status from the methylation levels of 30 LINE-1 CpG sites

M. Michel et al. **Noninvasive multicancer detection using DNA hypomethylation of LINE-1 retrotransposons** Clin Can Res 2024

Machine learning works best …

- when the data is **really big**
  - ImageNet: 14 million images
  - Llama4 training set: 30 trillion tokens

Machine learning works best …

- when the data is **really big**
  - ImageNet: 14 million images
  - Llama4 training set: 30 trillion tokens
- when the **nature of the data** is well understood

  $\Rightarrow$ good representations/modeling/architecture

Machine learning works best …

- when the data is **really big**
  - ImageNet: 14 million images
  - Llama4 training set: 30 trillion tokens
- when the **nature of the data** is well understood

  $\Rightarrow$ good representations/modeling/architecture
- when the **nature of the problem** is well understood

  humans can do it

Machine learning works best …

- when the data is **really big**
  - ImageNet: 14 million images
  - Llama4 training set: 30 trillion tokens
- when the **nature of the data** is well understood

  $\Rightarrow$ good representations/modeling/architecture
- when the **nature of the problem** is well understood

  humans can do it
- for **making predictions** rather than **explaining how** they were made

Machine learning works best …

- when the data is **really big**
  - ImageNet: 14 million images
  - Llama4 training set: 30 trillion tokens
- when the **nature of the data** is well understood

  $\Rightarrow$ good representations/modeling/architecture
- when the **nature of the problem** is well understood

  humans can do it
- for **making predictions** rather than **explaining how** they were made

Genomics does not fit this picture very well!

Talk outline

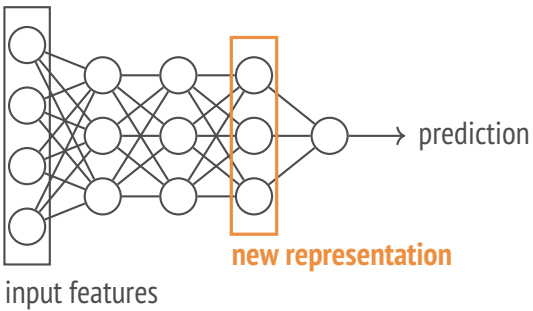
I. Good representations of genomic data

II. Identifying relevant genomic features

I. Good representations of genomic data

Representation learning

- **Good representations** = features from which learning is "easy"
- If we cannot **handcraft** good features using domain knowledge, can we **learn** them?



Foundation models

- **LOTS** of **broad** data
  - LAION-5B: 5.85 billion image-text pairs
  - GPT-3 was trained on 570 GB of text
- **self-supervision:**
  - Masked language modeling, next sentence prediction
  - Reconstructing a blurred, partially erased or scrambled image
- **Fine-tuning:** learned representations can then be used for any downstream task

Foundation models in genomics

- Pre-training = masked language modeling

- **NucleotideTransformer** [DT+24]
  - trained on 4k genomes (300B 6bp tokens)
  - 50M to 2.5B parameters
  - 12 kbp context length
  - trained on 16 $\times$8 A100 GPUs ($\sim$ 20 000 €)
  - Try it out: https://hclimente.eu/blog/hf-transformers/

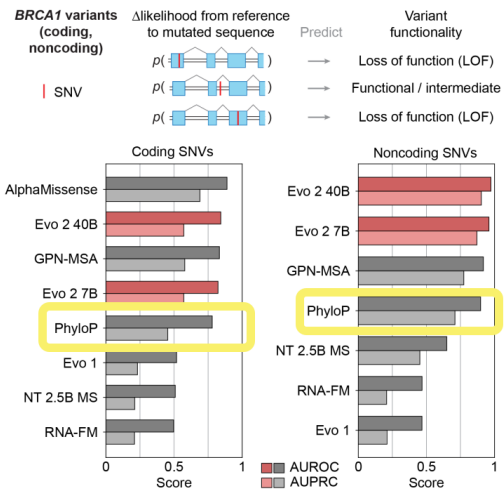
- **Evo2** [Bri+25]
  - trained on up 8.8 Tbp (1 token = 1bp)
  - 7 to 40 B parameters
  - 1 million bp context length
  - training took 2.25 $\times 10^{24}$ FLOPS (on par with Llama 3.1)

Variant pathogenicity prediction

- **Evo2** predicts BRAC1 variant pathogenicity
  - without training!
  - “unnatural” sequence = pathogenic

Variant pathogenicity prediction

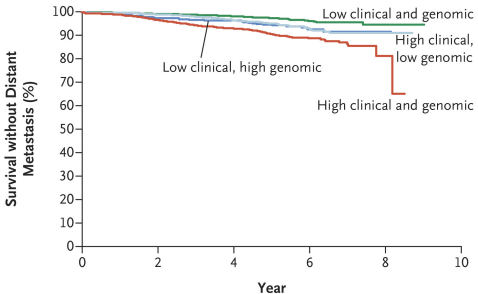
- **Evo2** predicts BRAC1 variant pathogenicity
  - without training!
  - “unnatural” sequence = pathogenic
- So does **PhyloP** [Pol+09]
  - conservation score
  - number of parameters: 2
- much better than **NucleotideTransformer**



II. Identifying relevant genomic features

Selecting features in genomic data

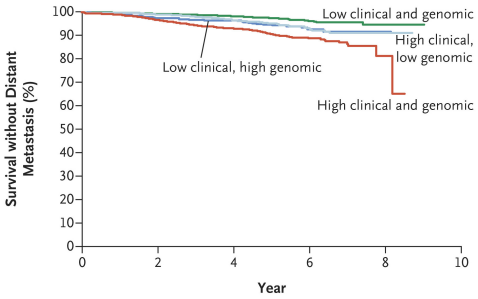
- **Goal:** Identify **genomic features** that can be used to predict a phentoype
  - **biomarker** or **signature** discovery



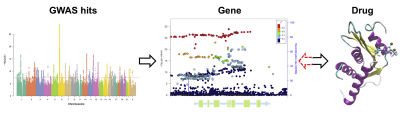
F. Cardoso et al., **70-gene signature as an aid to treatment decisions in early-stage breast cancer** NEJM 2016

Selecting features in genomic data

- **Goal:** Identify **genomic features** that can be used to predict a phentoype
  - **biomarker** or **signature** discovery
  - hypotheses on the underlying **biological mechanisms**



F. Cardoso et al., **70-gene signature as an aid to treatment decisions in early-stage breast cancer** NEJM 2016



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-CI-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

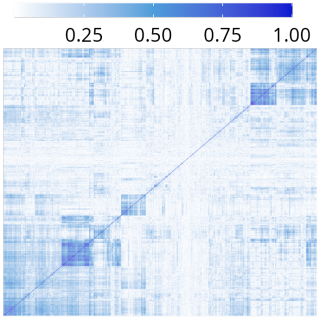
P. Visscher et al., **10 years of GWAS discovery: biology, function and translation** AJHG 2017

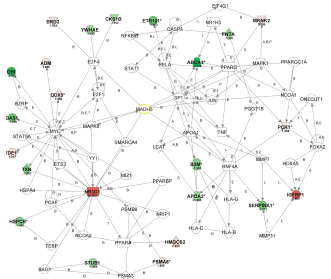
Selecting features in genomic data

- **Challenges:**
  - Orders of magnitude **more features than samples** ($p \gg n$)
    - 20 000 **transcripts** or 500 000 **Single Nucleotide Poymorphisms** for a few hundreds/thousands samples

Selecting features in genomic data

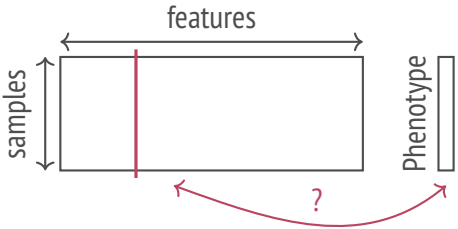
- **Challenges:**
  - Orders of magnitude **more features than samples** ($p \gg n$)
    - 20 000 **transcripts** or 500 000 **Single Nucleotide Poymorphisms** for a few hundreds/thousands samples
  - **Correlations** between features

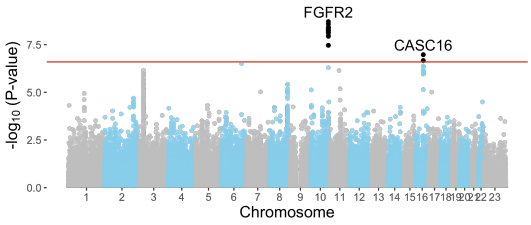


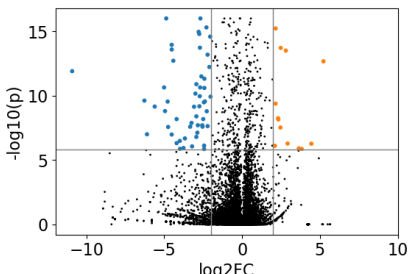


P. Lee et al., **Glucocorticoid Receptor-Dependent Gene Regulatory Networks** AJHG 2005

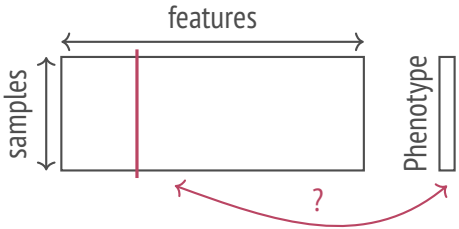
Classical statistical testing

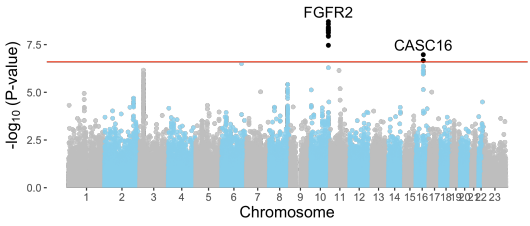




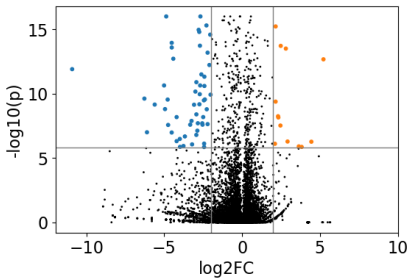


Classical statistical testing





- ☹ Lack of **power**
- ☹ Does not account for **other features** (joint/conditional effects)
- ☹ Potentially high **false discovery rate** [Hej+24]

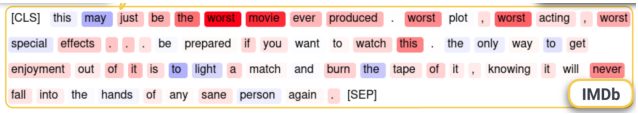


Explainable AI

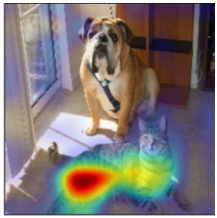
- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest

Explainable AI

- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest
- permutation importance, LINE, SHAP, gradient-based attribution



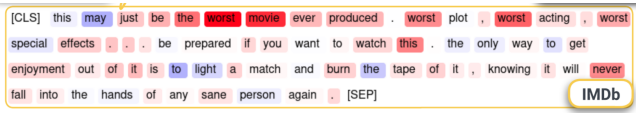
Integrated gradients



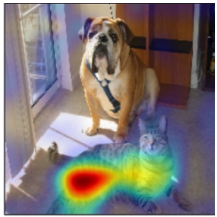
Grad-CAM

Explainable AI

- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest
- permutation importance, LINE, SHAP, gradient-based attribution



Integrated gradients



Grad-CAM

- How **reliable** are those? No measure of **uncertainty!**

Stability

– **Stability:** Selecting the same features on subsets of the data

☹ **Gene signatures** are notoriously unstable [VDD11]

Stability

- **Stability:** Selecting the same features on subsets of the data

☹ **Gene signatures** are notoriously unstable [VDD11]

- **Stability selection** procedures [Bac08; MB10; SS13]

  Keep the features that are **selected the most often** on **multiple subsamples** of the data.

Stability

- **Stability:** Selecting the same features on subsets of the data

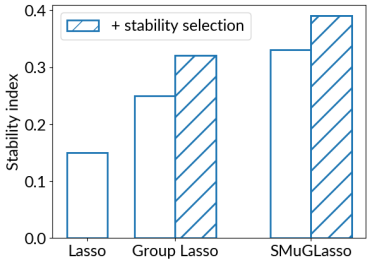
☹ **Gene signatures** are notoriously unstable [VDD11]

- **Stability selection** procedures [Bac08; MB10; SS13]

  Keep the features that are **selected the most often** on **multiple subsamples** of the data.

- **Structured sparsity** to incorporate **prior knowledge**

  E.g. Select features at the level of **predefined groups** or **connected in a given network** [Aze+13; Aze16; NA25]



A. Nouira & C.-A. Azencott. **Sparse multitask group lasso for genome-wide association studies**. PLoS Comp Bio 2025

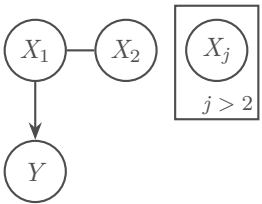
Post-selection inference

- How can one obtain **p-values** or **confidence intervals** on the parameters of models obtained from complex machine learning procedures?

- **Post-Selection Inference:** perform inference that **accounts for selection events** [Lee+16]
  - ☹ Model-specific
  - ☺ Possible for many **structured sparsity** approaches [Lee+16; Yan+16] and **kernel-based approaches** [Sli+19]
  - ☹ Computationally intensive

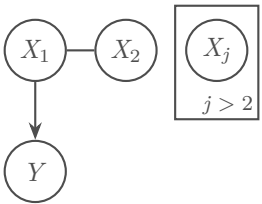
- **Conditional Feature Importance** [Str+08; RLNT26]
  - ☺ Model-agnostic
  - ☹ Does not currently scale to thousands of features

Correlation between features



Features that are **correlated** with important features **appear important** as well.

Correlation between features



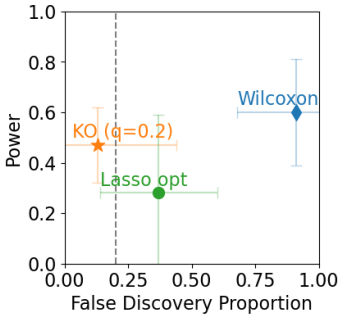
Features that are **correlated** with important features **appear important** as well.

Statistical knockoffs [BC15; Can+18]

- Create **knockoffs** of the original features:
  same correlation structure, no relationship with the phenotype
- Compare importance of a feature with importance of its knockoff
- Threshold on the difference can be set to **control the false discovery rate**

Applying knockoffs to transcriptomics data

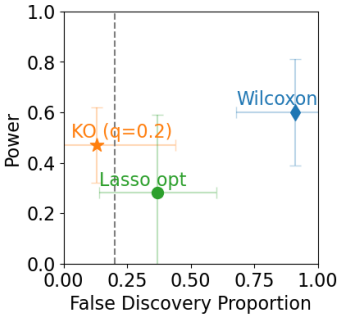
- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso



J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** Briefings in Bioinformatics 2026

Applying knockoffs to transcriptomics data

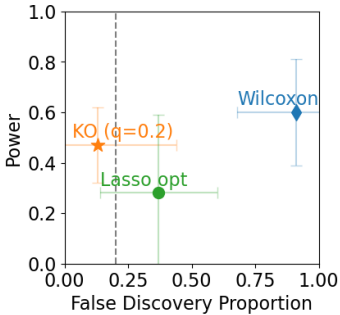
- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso
  - ☹ The KO framework fails if the relationship between phenotype and features is **non-linear**



J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** Briefings in Bioinformatics 2026

Applying knockoffs to transcriptomics data

- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso
  - ☹ The KO framework fails if the relationship between phenotype and features is **non-linear**



- On **real** outcomes:
  - ☹ The KO framework is **overly conservative**

cohort	CRUKPAP	AEGIS	BC HER2	BC ER	BC PgR
discoveries (q=0.5)	3	7	13	1	0

J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** Briefings in Bioinformatics 2026

Take-home messages

- The **performance** of a machine learning algorithm should always be analyzed in light of
  - the real use case
  - a cost-benefit analysis
  - comparison to state-of-the-art tools

Take-home messages

- The **performance** of a machine learning algorithm should always be analyzed in light of
  - the real use case
  - a cost-benefit analysis
  - comparison to state-of-the-art tools
- **Prediction** is easier than **explainability**
  - **XAI** usually comes with **no guarantees**
  - Incorporating **prior knowledge** can make XAI more **stable**
  - **post-selection inference** and **conditional feature importance** can give **p-values** or **confidence intervals** on model parameters
  - **statistical knockoffs** can help **control false discovery rate**

Thanks

CBIO: Youmna Ayadi, **Asma Nouira** (now ASNR), **Julie Cartier**, Héctor Climente-González (now Isomorphic Labs), Florian Massip, **Lotfi Slim** (now NVIDIA Genomics), Jean-Philippe Vert (now Owkins and Bioptimus).

Institut Curie: **Marc Michel** (now Sancare), Charlotte Proudhon (now IRSET).

Funding: European Research Council (MLFPM2018, H2020-MSCA-ITN-2018 813533); Agence Nationale pour la Recherche (SCAPHE, PRAIRIE)

References I

[Aze+13] Chloé-Agathe Azencott et al. "Efficient network-guided multi-locus association mapping with graph cuts". In: *Bioinformatics* 29.13 (2013), pp. i171i179. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt238.

[Aze16] Chloé-Agathe Azencott. "Network-guided biomarker discovery". In: *Machine Learning for Health Informatics*. Springer International Publishing, 2016, pp. 319336. DOI: 10.1007/978-3-319-50478-0_16.

[Bac08] Francis R Bach. "Bolasso: model consistent lasso estimation through the bootstrap". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 33–40.

[BC15] Rina Foygel Barber and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015). ISSN: 0090-5364. DOI: 10.1214/15-aos1337.

[Ber+87] H. C. P. Berbee et al. "Hit-and-run algorithms for the identification of nonredundant linear inequalities". In: *Mathematical Programming* 37.2 (1987), pp. 184207. ISSN: 1436-4646. DOI: 10.1007/bf02591694. URL: http://dx.doi.org/10.1007/BF02591694.

[Bla+23] Alexandre Blain et al. "False Discovery Proportion control for aggregated Knockoffs". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 78193–78204. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/f6712d5191d2501dfc7024389f7bfcdd-Paper-Conference.pdf.

References II

[Bla+25] Alexandre Blain et al. *When Knockoffs fail: diagnosing and fixing non-exchangeability of Knockoffs.* 2025. arXiv: 2407.06892 [stat.ME]. URL: https://arxiv.org/abs/2407.06892.

[Bri+25] Garyk Brixi et al. "Genome modeling and design across all domains of life with Evo 2". In: (2025). DOI: 10.1101/2025.02.18.638918.

[Can+18] Emmanuel Candès et al. "Panning for Gold: Model-X Knockoffs for High Dimensional Controlled Variable Selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551577. ISSN: 1467-9868. DOI: 10.1111/rssb.12265.

[Car+25] Julie Cartier et al. "Statistical knockoffs improve biomarker discovery from transcriptomic data". In: (2025). DOI: 10.1101/2025.07.04.663147.

[CG+21] Héctor Climente-González et al. "Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer". In: *PLOS Computational Biology* 17.3 (2021), e1008819. DOI: 10.1371/journal.pcbi.1008819.

[CGAY23] Héctor Climente-González, Chloé-Agathe Azencott, and Makoto Yamada. "A network-guided protocol to discover susceptibility genes in genome-wide association studies using stability selection". In: *STAR Protocols* 4.1 (2023). DOI: 10.1016/j.xpro.2022.101998.

References III

[DAN15] Alia Dehman, Christophe Ambroise, and Pierre Neuvial. "Performance of a blockwise approach in variable selection using linkage disequilibrium information". In: *BMC Bioinformatics* 16.1 (2015). DOI: 10.1186/s12859-015-0556-6.

[DT+24] Hugo Dalla-Torre et al. "Nucleotide Transformer: building and evaluating robust foundation models for human genomics". In: *Nature Methods* 22.2 (2024), pp. 287297. DOI: 10.1038/s41592-024-02523-z.

[Gre+05] Arthur Gretton et al. "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: *Algorithmic Learning Theory*. Springer Berlin Heidelberg, 2005, pp. 6377. ISBN: 9783540316961. DOI: 10.1007/11564089_7. URL: http://dx.doi.org/10.1007/11564089_7.

[Hej+24] Boris P. Hejblum et al. "Neglecting the impact of normalization in semi-synthetic RNA-seq data simulations generates artificial false positives". In: *Genome Biology* 25.1 (2024). DOI: 10.1186/s13059-024-03231-9.

[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.

[Hun+10] David J. Hunter et al. *Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) Oncoarray Genotypes*. 2010. URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1.

[ISO20] ISO/IEC JTC 1/SC 7. *TR 29119-11:2020 Part 11: Guidelines on the testing of AI-based systems*. International Organization for Standardization and International Electrotechnical Commission. 2020.

References IV

[JOV09] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. "Group lasso with overlap and graph lasso". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML 09. ACM, 2009, pp. 433440. DOI: 10.1145/1553374.1553431.

[KSX09] Seyoung Kim, Kyung-Ah Sohn, and Eric P. Xing. "A multivariate regression approach to association analysis of a quantitative trait network". In: *Bioinformatics* 25.12 (2009), pp. i204i212. DOI: 10.1093/bioinformatics/btp218.

[Lee+16] Jason D. Lee et al. "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3 (2016). ISSN: 0090-5364. DOI: 10.1214/15-aos1371. URL: http://dx.doi.org/10.1214/15-AOS1371.

[LL07] Chun Li and Mingyao Li. "GWAsimulator: a rapid whole-genome simulation program". In: *Bioinformatics* 24.1 (2007), pp. 140142. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm549.

[LL08] Caiyan Li and Hongzhe Li. "Network-constrained regularization and variable selection for analysis of genomic data". In: *Bioinformatics* 24.9 (2008), pp. 11751182. DOI: 10.1093/bioinformatics/btn081.

[LT15] Joshua R. Loftus and Jonathan E. Taylor. *Selective inference in regression models with groups of variables.* 2015. arXiv: 1511.01478 [stat.ME]. URL: https://arxiv.org/abs/1511.01478.

[MB10] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010). ISSN: 1467-9868.

References V

[Mol25] Christoph Molnar. *Interpretable machine learning: A guide for making black box models explainable*. 2025.

[NA] Asma Nouira and Chloé-Agathe Azencott. "Multitask group Lasso for Genome Wide association Studies in diverse populations". In: *Pacific Symposium on Biocomputing 2022*, pp. 163–174. DOI: 10.1142/9789811250477_0016.

[NA25] Asma Nouira and Chloé-Agathe Azencott. "Sparse multitask group Lasso for genome-wide association studies". In: *PLOS Computational Biology* 21.9 (2025), e1012734. DOI: 10.1371/journal.pcbi.1012734.

[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.

[OTJ09] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2009), pp. 231252. DOI: 10.1007/s11222-008-9111-x.

[Pol+09] Katherine S. Pollard et al. "Detection of nonneutral substitution rates on mammalian phylogenies". In: *Genome Research* 20.1 (2009), pp. 110121. DOI: 10.1101/gr.097857.109.

[RLNT26] Angel Reyero-Lobo, Pierre Neuvial, and Bertrand Thirion. *Conditional Feature Importance revisited: Double Robustness, Efficiency and Inference*. 2026. arXiv: 2501.17520 [stat.ME]. URL: https://arxiv.org/abs/2501.17520.

References VI

[RTT17] Stephen Reid, Jonathan Taylor, and Robert Tibshirani. "A General Framework for Estimation and Inference From Clusters of Features". In: *Journal of the American Statistical Association* 113.521 (2017), pp. 280293. ISSN: 1537-274X. DOI: 10.1080/01621459.2016.1246368. URL: http://dx.doi.org/10.1080/01621459.2016.1246368.

[Sli+19] Lotfi Slim et al. "kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5857–5865. URL: https://proceedings.mlr.press/v97/slim19a.html.

[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: Series B (Statistical Methodology)* 75.1 (2013), pp. 55–80.

[Str+08] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1 (2008), p. 307. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-307.

[Sug+14] Mahito Sugiyama et al. "Multi-task feature selection on multiple networks via maximum flows". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. 2014, pp. 199207. DOI: 10.1137/1.9781611973440.23.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267288. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

References VII

[Toz+22] Veronica Tozzo et al. "Where do we stand in regularization for life science studies?" In: *Journal of Computational Biology* 29.3 (2022), pp. 213232. DOI: 10.1089/cmb.2019.0371.

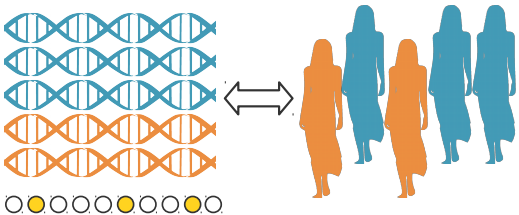
[VDD11] David Venet, Jacques E. Dumont, and Vincent Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In: *PLoS Computational Biology* 7.10 (2011), e1002240.

[Wu+11] MichaelăC. Wu et al. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test". In: *The American Journal of Human Genetics* 89.1 (2011), pp. 8293. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2011.05.029. URL: http://dx.doi.org/10.1016/j.ajhg.2011.05.029.

[Yan+16] Fan Yang et al. "Selective inference for group-sparse linear models". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 24772485. ISBN: 9781510838819.

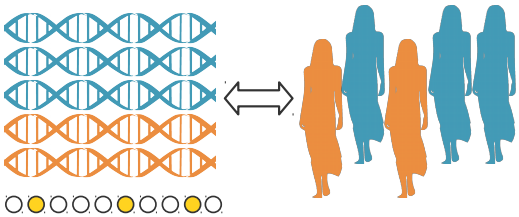
[YL05] Ming Yuan and Yi Lin. "Model Selection and Estimation in Regression with Grouped Variables". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.1 (2005), pp. 4967. DOI: 10.1111/j.1467-9868.2005.00532.x.

Genotype-to-phenotype studies



Which genomic features explain the phenotype?

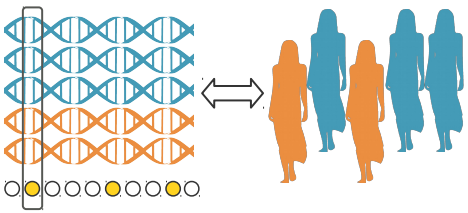
Genotype-to-phenotype studies



Which genomic features explain the phenotype?

– Typically **fewer samples** than **genomic features** (gene expressions, SNPs, etc)

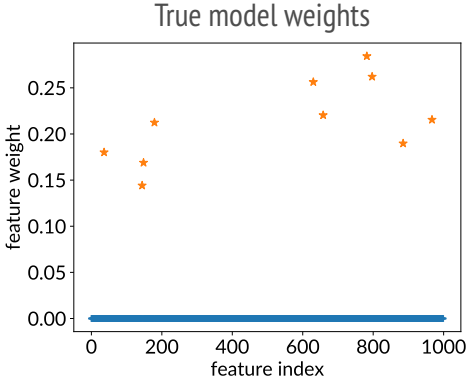
State of the art: Statistical tests



Perform a **statistical test of association** between **each feature** and the phenotype.

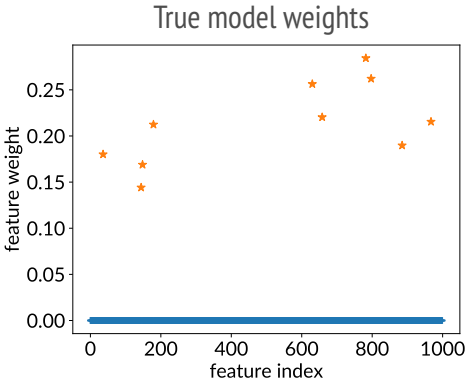
Simulation

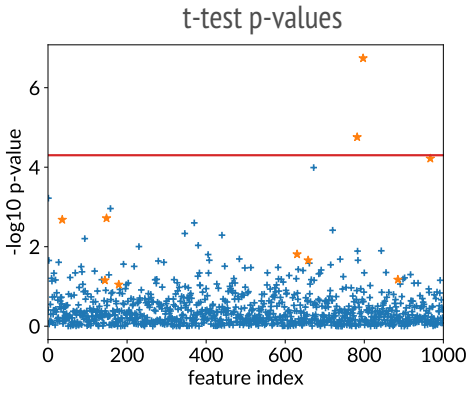
- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$



Simulation

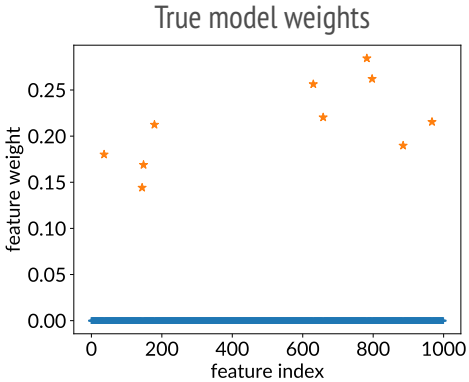
- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$





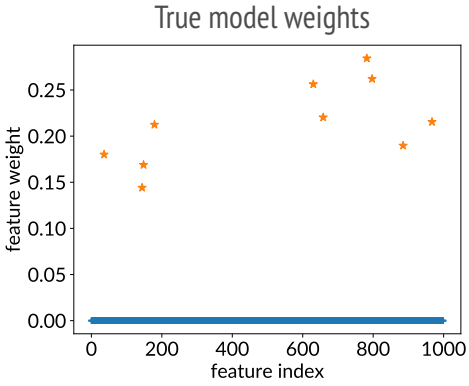
Simulation: linear regression

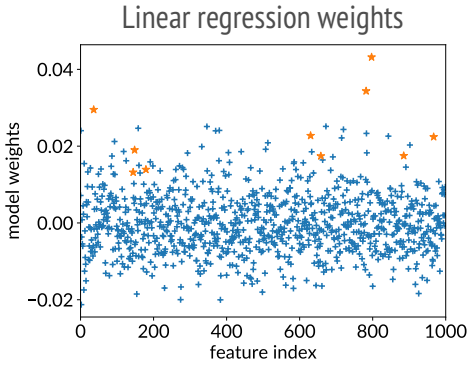
- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$



Simulation: linear regression

- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$





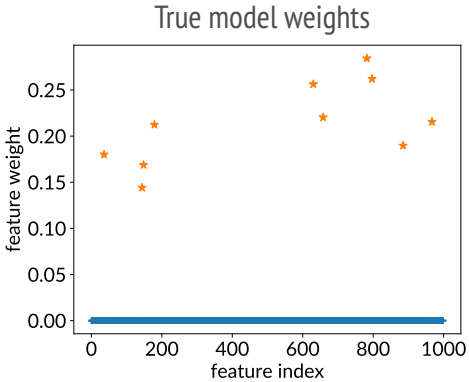
Regularization

- **Empirical risk minimization:** find a model with minimal error on the training data
- **Regularization:** force the model to respect some additional constraints
  - **Weight decay**: don't allow the model parameters to take large values
    (or ridge/Tikhonov/ℓ_2 regularization)
  - **Sparsity**: don't allow too many of the model parameters to have non-zero values
    E.g.: Lasso (or ℓ_1 regularization)

V. Tozzo et al. **Where do we stand in regularization for life science studies?** Journal of Computational Biology 2022

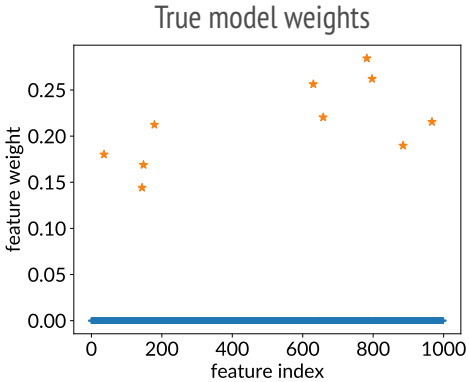
Simulation: Lasso

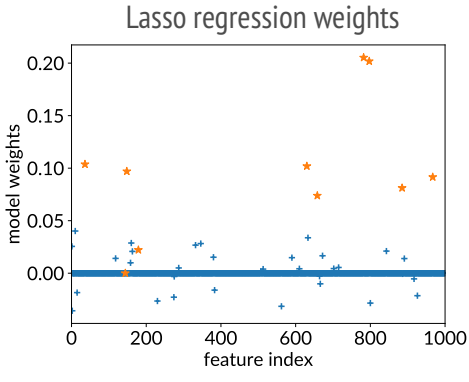
- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$



Simulation: Lasso

- 100 **samples** 1 000 **features** 10 of which **influence** the phenotype
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$



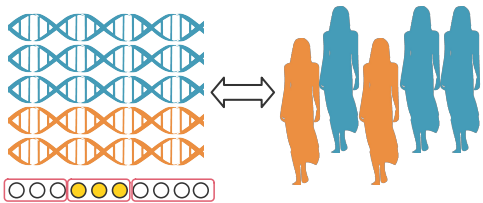


Regularization to integrate prior biological knowledge

- **Goals:**
  - Make the model **consistant** with previously established knowledge
  - Help find a **good model**
  - Increase **interpretability**
- Prior biological knowledge has **structure:**
  - **Groups:** genes belonging to the same pathway / regulated by the same transcription factor; SNPs belonging to the same LD block
  - **Graphs:** biological networks

V. Tozzo et al. **Where do we stand in regularization for life science studies?** Journal of Computational Biology 2022

Group-based regularization



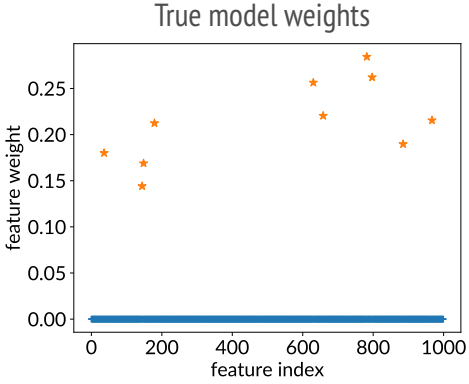
Variants of the **Lasso** encourage the sparsity pattern to **respect a given groups structure**: features that belong to the same provided group will tend to be selected together

- Group Lasso [YL05]
- Overlapping Group Lasso [JOV09]

Simulation: Group Lasso

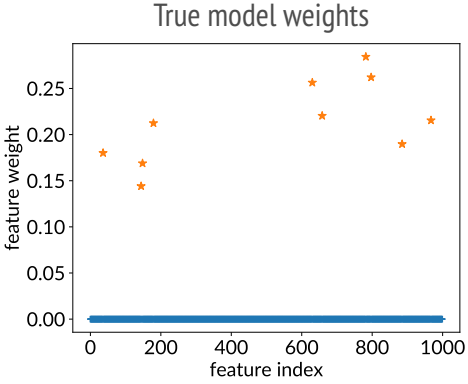
- 100 **samples**
  1 000 **features** 10 of which **influence** the phenotype and **form two of the provided groups**

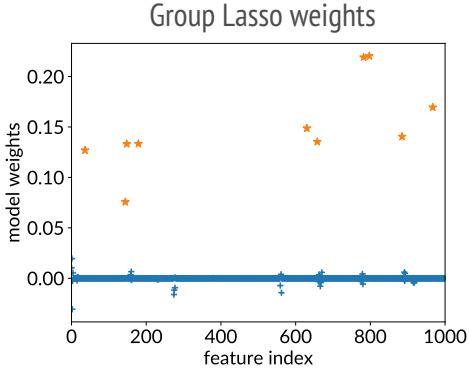
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$



Simulation: Group Lasso

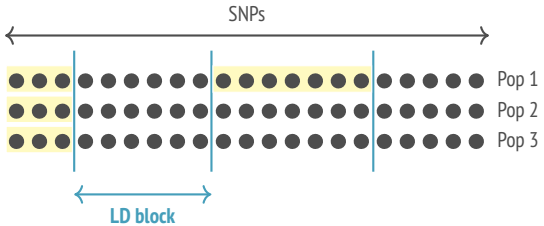
- 100 **samples**
  1 000 **features** 10 of which **influence** the phenotype and **form two of the provided groups**
- $y = \sum_{j=1}^{1000} w_j x_j + \varepsilon$





SMuGLasso for GWAS in diverse populations

- **G = Group:** Group SNPs by **linkage disequilibrium blocks** [DAN15]
- Split samples by genetically homogeneous population (PCA + clustering) $\rightarrow$ **tasks**
  - **Mu = Multitask:** same blocks are selected across tasks [OTJ09]
  - **S = Sparse:** some blocks are task-specific

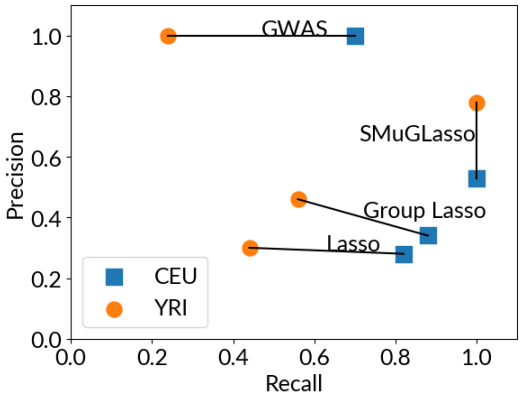


A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso has better recall than other methods

Simulation with GWAsimulator [LL07]

- 2 populations from HapMap3:
  - **CEU** (1300 cases, 1700 controls)
  - **YRI** (400 cases, 600 controls)
- 50 000 SNPs
  - 200 disease-causing SNPs
  - 50 **CEU-specific** SNPs
  - 50 **YRI-specific** SNPs

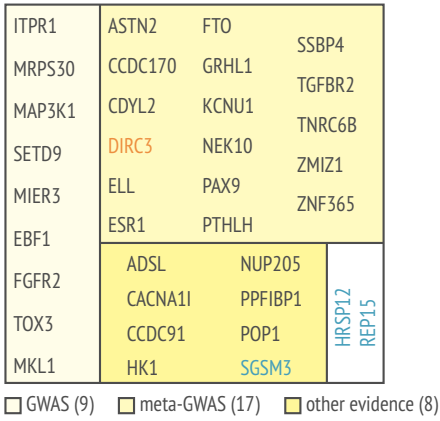


A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso identifies disease genes

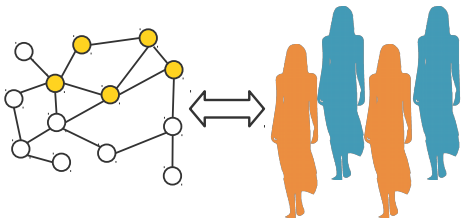
DRIVE dataset [Hun+10]

- 13 846 breast cancer cases, 14 435 controls
- 312 237 SNPs after quality control
- PCA + kmeans $\rightarrow$ 2 populations:
  - **Pop1** (USA, Australia, Denmark)
  - **Pop2** (USA, Cameroon, Nigeria, Uganda)



A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

Graph-based regularization



Variants of the **Lasso** encourage the sparsity pattern to **respect the structure of a given graph**: features that are connected on the provided graph will tend to be selected together

- Network-constrained Lasso [LL08]
- Graph Lasso [JOV09]
- Graph-guided fused Lasso [KSX09]

H. Climente-González et al. **A network-guided protocol to discover susceptibility genes in genome-wide association studies**. STAR Prot 2023
H. Climente-González et al. **Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer**. PLoS Comp Bio 2021
C.-A. Azencott. **Network-guided biomarker discovery**. Lecture Notes in Computer Science 2016

Empirical risk minimization

– The idea behind (most) supervised machine learning algorithms:

  Find a model f in the **hypothesis space** $\mathcal{F}$ that **minimizes** the **empirical risk**.

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, f(\vec{x}_i))}_{\text{loss}}$$

Empirical risk minimization

- The idea behind (most) supervised machine learning algorithms:

  Find a model f in the **hypothesis space** $\mathcal{F}$ that **minimizes** the **empirical risk**.

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, f(\vec{x}_i))}_{\text{loss}}$$

- Examples of **losses:**
  - For a regression problem, the **quadratic loss**

$$\mathcal{L}(y, f(\vec{x})) = (y - f(\vec{x}))^2$$

  - For a binary classification problem, the **logistic loss**

$$\mathcal{L}(y, f(\vec{x})) = -y \log(f(\vec{x})) - (1 - y) \log(1 - f(\vec{x}))$$

Regularized empirical risk minimization

- Idea: impose **a priori constraints** on the solution of the empirical risk minimization problem
- **Parametric** models: $\mathcal{F} = \{f_{\vec{w}}; \vec{w} \in \mathbb{R}^d\}$

$$\min_{\vec{w} \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f_{\vec{w}}(\vec{x}_i))}_{\text{loss}} + \lambda \underbrace{\Omega(\vec{w})}_{\text{regularizer}}$$

Example: Lasso

- **Linear model:** $f_{\vec{w}}(\vec{x}) = \langle \vec{w}, \vec{x} \rangle = w_0 + \sum_{j=1}^{p} w_j x_j$
- Regularized empirical risk minimization

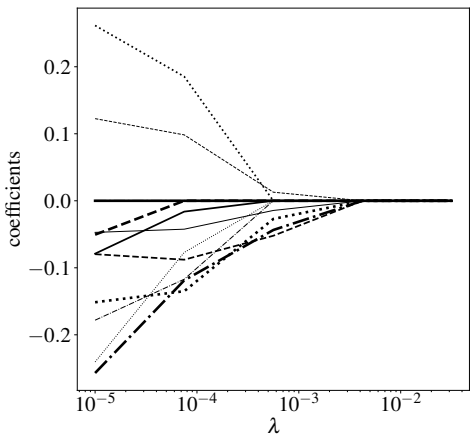
$$\min_{\vec{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \langle \vec{w}, \vec{x_i} \rangle)}_{\text{loss}} + \lambda \underbrace{\Omega(\vec{w})}_{\text{regularizer}}$$

- **Prior knowledge / a priori constraints:** few features are relevant.
- **Lasso:** $\Omega(\vec{w}) = ||\vec{w}||_1 = \sum_{j=0}^{p} |w_j|$ [Tib96]
- **Sparsity:** many features are assigned a weight of 0. They can be removed from the model.

Regularization coefficient λ

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, \langle \vec{w}, \vec{x}_i\rangle)}_{\text{loss}} + \lambda \underbrace{\Omega(\vec{w})}_{\text{regularizer}}$$

- λ controls the amount of regularization
- Typically set by **grid search** + **cross-validation:** {number of folds} $\times$ {number of values on the grid} experiments
- For the lasso, efficient ways to get the entire regularization path $\{\vec{w}_\lambda$ for $\lambda \in \{\lambda_{\text{min}}, \ldots, \lambda_{\text{max}}\}\}$



Group-based regularization

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, \langle \vec{w}, \vec{x}_i\rangle)}_{\text{loss}} + \lambda \underbrace{\Omega_{\text{group}}(\vec{w})}_{\text{group-level regularizer}}$$

- Given a way of grouping the p features in G groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$, each of size p_g, define Ω_{group} to encourage **the selection of only a few groups**

- **Group Lasso** [YL05]

$$\Omega_{\text{group}}(\vec{w}) = \sum_{g=1}^{G} \sqrt{p_g} \sum_{j\in\mathcal{G}_g} w_j^2$$

- **Overlapping Group Lasso** [JOV09]

Multitask regularization

$$\min_{\vec{w}\in\mathbb{R}^p} \sum_{t=1}^{T} \underbrace{\frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_i^{(t)}, \langle \vec{w}^{(t)}, \vec{x}_i^{(t)} \rangle)}_{\text{loss}} + \lambda \underbrace{\Omega_{\text{tasks}}(\vec{w}^{(1)}, \ldots, \vec{w}^{(T)})}_{\text{task regularizer}}$$

- Given T **related tasks,** define Ω_{tasks} so as to solve the T empirical risk minimization problems in such a way that **the same features are selected across tasks.**

- **Multitask Lasso** [OTJ09]

$$\Omega_{\text{tasks}}(\vec{w}^{(1)}, \ldots, \vec{w}^{(T)}) = \sum_{t=1}^{T} \sum_{j=1}^{p} \left(w_j^{(t)} \right)^2$$

MuGLasso

Multitask **G**roup Lasso:

- **multitask group**-level sparsity
- the same groups are selected for all tasks

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_i^{(t)}, \langle \vec{w}^{(t)}, \vec{x}_i^{(t)} \rangle)}_{\text{loss}} + \lambda \underbrace{\sum_{g=1}^{G} \sqrt{p_g} \sum_{j\in\mathcal{G}_g} \sum_{t=1}^{T} \left(w_j^{(t)} \right)^2}_{\text{mutltitask group-level sparsity}}$$

- If $T = 1 \rightarrow$ group Lasso
- If $G = p$ and $\mathcal{G}_1, \ldots, \mathcal{G}_p = \{1\}, \ldots, \{p\} \rightarrow$ multitask Lasso

A. Nouira & C.-A. Azencott. **Multitask group Lasso for genome-wide association studies in diverse populations**. PSB 2022

SMuGLasso

Sparse **Mu**ltitask **G**roup Lasso:

- the same groups are selected for all tasks
- among those, some groups can be task-specific

$$\min_{\vec{w}\in\mathbb{R}^p} \sum_{t=1}^{T} \underbrace{\frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}(y_i^{(t)}, \langle \vec{w}^{(t)}, \vec{x}_i^{(t)} \rangle)}_{\text{loss}} + \lambda \underbrace{\sum_{g=1}^{G} \sqrt{p_g} \sum_{j\in\mathcal{G}_g} \sum_{t=1}^{T} \left(w_j^{(t)}\right)^2}_{\text{mutltitask group-level sparsity}} + \lambda_2 \underbrace{\sum_{g=1}^{G} \sqrt{p_g} \sum_{j\in\mathcal{G}_g} \sum_{t=1}^{T} \left|w_j^{(t)}\right|}_{\text{task-level sparsity}}$$

A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

Graph-based regularization

$$\min_{\vec{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \langle \vec{w}, \vec{x}_i \rangle)}_{\text{loss}} + \lambda_s \underbrace{||\vec{w}||_1}_{\text{sparsity}} + \lambda_g \underbrace{\Omega_{\text{graph}}(\vec{w})}_{\text{connectivity}}$$

– Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of p nodes over the features, define Ω_{graph} to encourage the sparsity pattern to **respect the structure of** $\mathcal{G}$.

Graph-based regularization

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i,\langle\vec{w},\vec{x}_i\rangle)}_{\text{loss}} + \lambda_s \underbrace{||\vec{w}||_1}_{\text{sparsity}} + \lambda_g \underbrace{\Omega_{\text{graph}}(\vec{w})}_{\text{connectivity}}$$

- Given a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ of p nodes over the features, define Ω_{graph} to encourage the sparsity pattern to **respect the structure of $\mathcal{G}$**.
- **Graph-fused Lasso** [KSX09]

$$\Omega_{\text{graph}}(\vec{w}) = \sum_{(v_j,v_k)\in\mathcal{E}} |w_j - w_k|$$

- **Network-constrained Lasso** [LL08]

$$\Omega_{\text{graph}}(\vec{w}) = \vec{w}^\top L\vec{w} = \sum_{(v_j,v_k)\in\mathcal{E}} A_{jk}(w_j - w_k)^2$$

Graph-based regularization

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i,\langle\vec{w},\vec{x}_i\rangle)}_{\text{loss}} + \lambda_s \underbrace{||\vec{w}||_1}_{\text{sparsity}} + \lambda_g \underbrace{\Omega_{\text{graph}}(\vec{w})}_{\text{connectivity}}$$

- Given a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ of p nodes over the features, define Ω_{graph} to encourage the sparsity pattern to **respect the structure of** $\mathcal{G}$.
- **Graph Lasso:** overlapping group lasso with edges as groups [JOV09]

$$\Omega_{\text{graph}}(\vec{w}) = \inf_{(\vec{\beta}_1,\ldots,\vec{\beta}_{\mathcal{E}}):\vec{w}=\sum_{k=1}^{|\mathcal{E}|}\beta_k} \sum_{k=1}^{|\mathcal{E}|}||\beta_k||_2^2 \qquad \vec{\beta}_k\in\mathbb{R}^p \text{ s.t. } \vec{\beta}_{kj}\neq 0 \text{ iff node } j \text{ in edge } k$$

Network-constrained Lasso

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i,\langle\vec{w},\vec{x}_i\rangle)}_{\text{loss}} + \lambda_s \underbrace{||\vec{w}||_1}_{\text{sparsity}} + \lambda_g \underbrace{\vec{w}^\top L\vec{w}}_{\text{connectivity}}$$

Can be solved as a **Lasso on transformed data**

$$X^* = \frac{1}{\sqrt{1+\lambda_g}}\begin{pmatrix} X \\ \sqrt{\lambda_g}S^\top \end{pmatrix} \in \mathbb{R}^{(n+m)\times p} \qquad \vec{y}^* = \begin{pmatrix} \vec{y} \\ 0 \end{pmatrix} \in \mathcal{Y}^{n+m}$$

where $S \in \mathbb{R}^{m\times p}$ such that $L = SS^\top$

with regularization parameter $\frac{\lambda_s}{\sqrt{1+\lambda_g}}$ and then $\vec{w} = \sqrt{1+\lambda_g}\vec{w}^*$

Network-constrained Lasso

$$\min_{\vec{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i,\langle\vec{w},\vec{x_i}\rangle)}_{\text{loss}} + \lambda_s \underbrace{||\vec{w}||_1}_{\text{sparsity}} + \lambda_g \underbrace{\vec{w}^\top L\vec{w}}_{\text{connectivity}}$$

Can be solved as a **Lasso on transformed data**

$$X^* = \frac{1}{\sqrt{1+\lambda_g}}\begin{pmatrix} X \\ \sqrt{\lambda_g}S^\top \end{pmatrix} \in \mathbb{R}^{(n+m)\times p} \qquad \vec{y}^* = \begin{pmatrix} \vec{y} \\ 0 \end{pmatrix} \in \mathcal{Y}^{n+m}$$

where $S \in \mathbb{R}^{m\times p}$ such that $L = SS^\top$

with regularization parameter $\frac{\lambda_s}{\sqrt{1+\lambda_g}}$ and then $\vec{w} = \sqrt{1+\lambda_g}\vec{w}^*$

- Defining S:
  - Option 1 ($m = p$) and $S = U\Lambda^{1/2}$ with $L = U\Lambda U^\top \rightarrow$ runtime issues ☹
  - Option 2 ($m = |\mathcal{E}|$) and S is the incidence matrix of $\mathcal{G} \rightarrow$ memory issues ☹

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test**

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test**
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test** relies on the hypothesis of normally-distributed residuals $Y \sim \mathcal{N}(\mu(X), \sigma^2)$
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test** relies on the hypothesis of normally-distributed residuals $Y \sim \mathcal{N}(\mu(X), \sigma^2)$

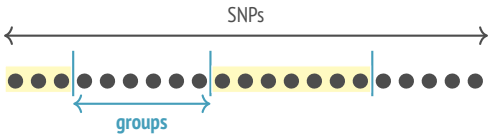
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!

    Y must be compatible with the observed set of selected features
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

    i.e. characterize $Y | \widehat{S}(Y) = S$

Group-based feature selection with kernels

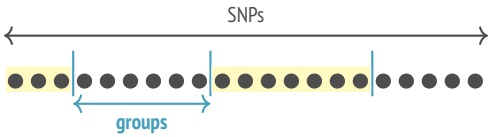
- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$



- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$

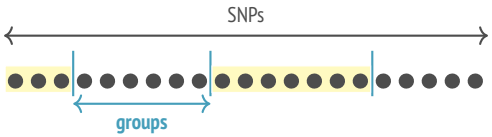


- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]

L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$

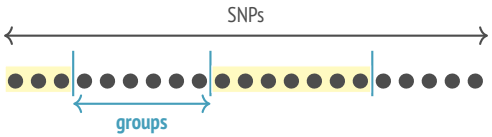


- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]
- **Group selection procedure:** select the kernels with the highest scores

L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$



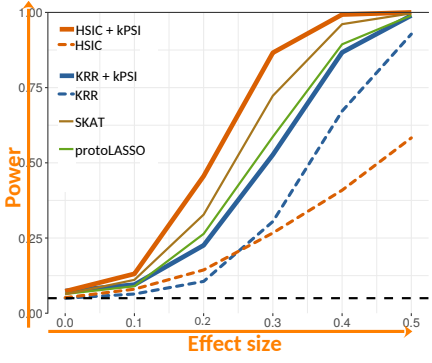
- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]
- **Group selection procedure:** select the kernels with the highest scores

  Post-selection inference needed

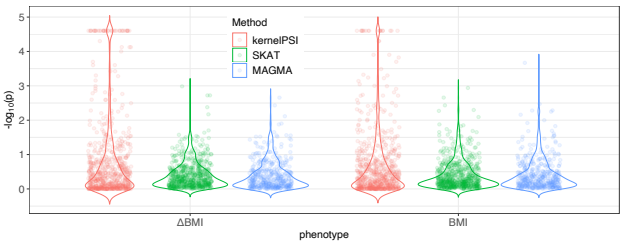
L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

kernelPSI

– **kernelPSI:** characterize the selection event and sample efficiently from the resulting distribution



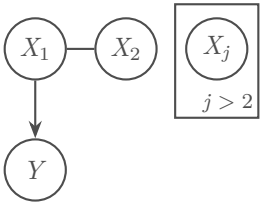
Simulation



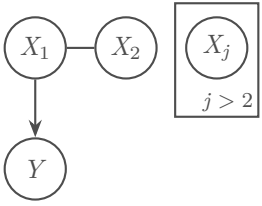
On UKBiobank data

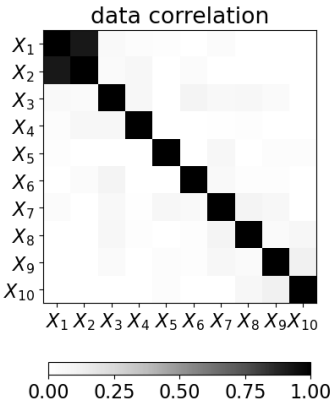
L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

Challenge: correlation between variables

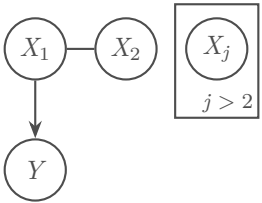


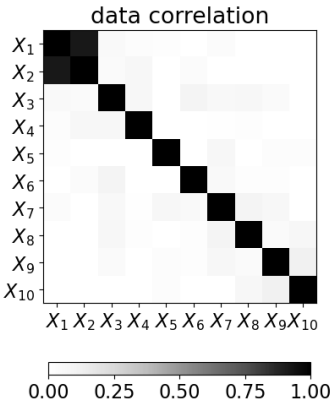
Challenge: correlation between variables

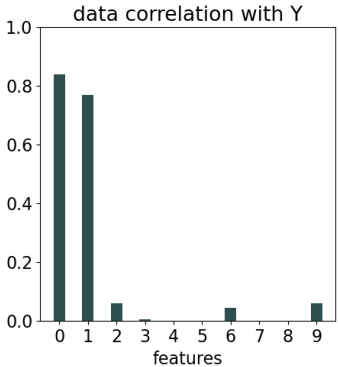




Challenge: correlation between variables

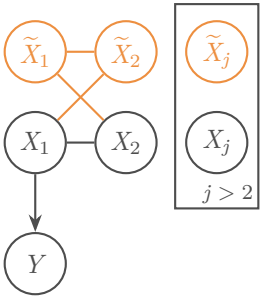


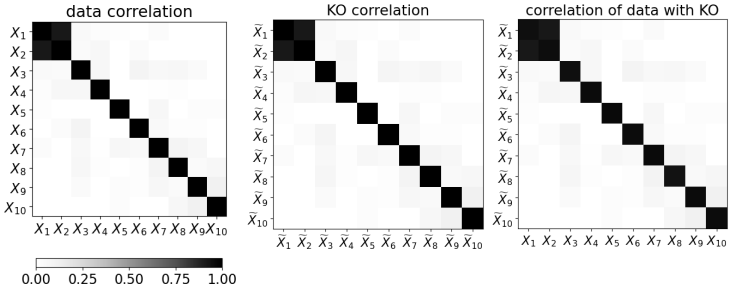




Features that are **correlated** with features associated with the phenotype **appear associated** with the phenotype

Knockoffs

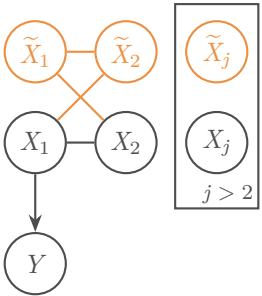


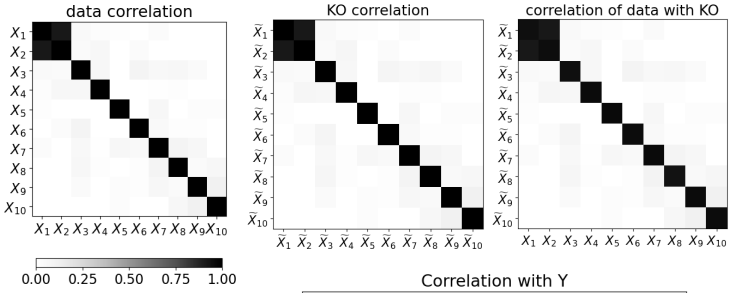


- $\widetilde{X}$ built from X only: $\widetilde{X} \perp Y | X$
- $\widetilde{X}$ has the **same covariance structure** as X
- $\widetilde{X}$ as **different** from X as possible

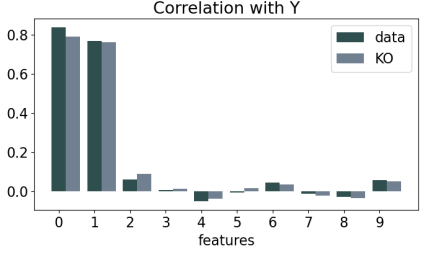
R. Barber & E. Candès **Controlling the false discovery rate via knockoffs**
Ann. Stat. 2015

Knockoffs



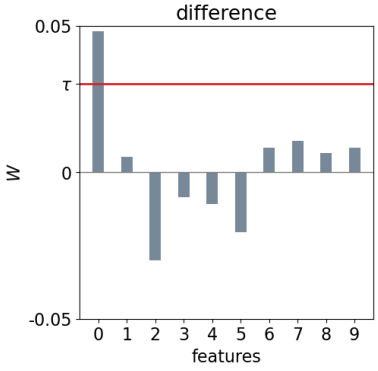


- $\widetilde{X}$ built from X only: $\widetilde{X} \perp Y | X$
- $\widetilde{X}$ has the **same covariance structure** as X
- $\widetilde{X}$ as **different** from X as possible



R. Barber & E. Candès **Controlling the false discovery rate via knockoffs**
Ann. Stat. 2015

Knockoff Statistics



- $W_j = Q(X_j, Y) - Q(\widetilde{X}_j, Y)$
- Example: **Lasso Coefficient-Difference** [Can+18]
- $\tau = \min\left\{t > 0 \colon \frac{|\{j \colon W_j \leq -t\}|}{|\{j \colon W_j \geq t\}|} \leq q\right\}$ guarantees **control of the target FDR** q

R. Barber & E. Candès et al. **Controlling the false discovery rate via knockoffs** Ann. Stat. 2015
E. Candès et al. **Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection** J. R. Stat. Soc. B 2018