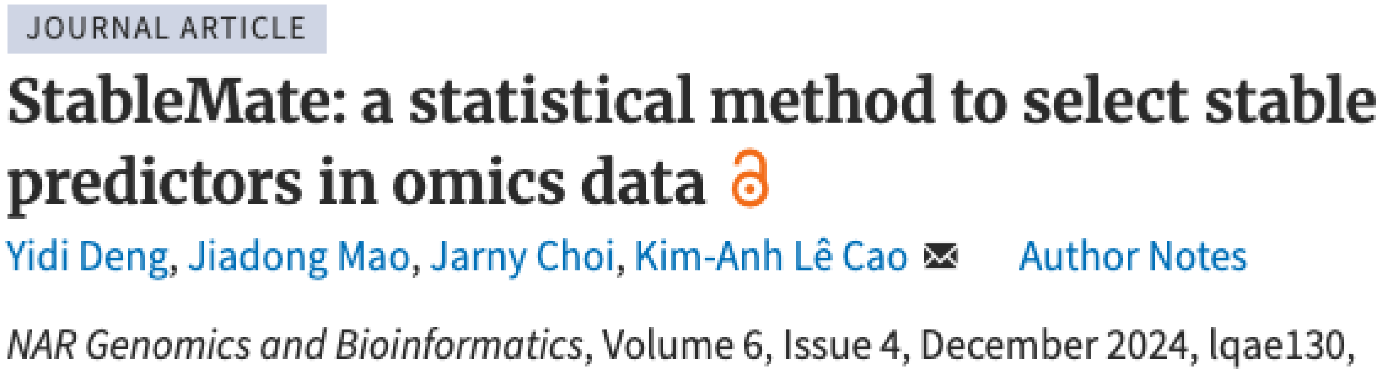
Embracing instability

Assessing stability of biological variables via *stabilised regression*



JOURNAL ARTICLE

StableMate: a statistical method to select stable predictors in omics data

Yidi Deng, Jiadong Mao, Jarny Choi, Kim-Anh Lê Cao ✉ Author Notes

NAR Genomics and Bioinformatics, Volume 6, Issue 4, December 2024, lqae130,

Omics data: big and complex

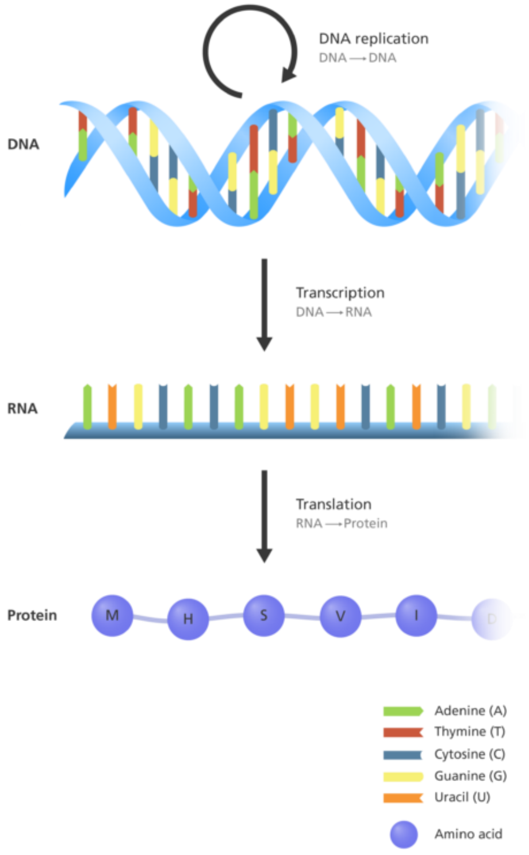
- ‘Omics’: measure many molecular activities in one go
  - eg genomics, transcriptomics, epigenomics, proteomics
  - Epigenomics: DNA methylation, ATAC-seq, histone modification, high-C, …
  - Eg >20,000 gene expression, >100,000 ATAC peaks

- Complexity 1: many variables

- Complexity 2: high resolution
  - Single-cell & spatial omics
  - Not clear what is a ‘sample’

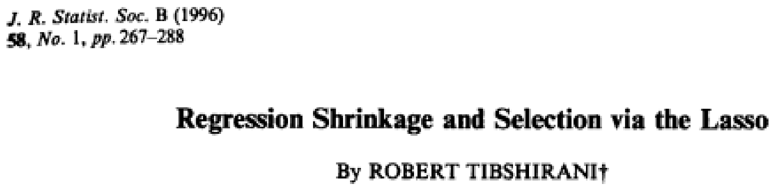
- Combined: big and complex
  - High-dimensional
  - Sparse (a lot of zeros)
  - Heterogeneous (cell types, experimental batches)

Laura Olivares Boldú / Wellcome Connecting Science



Feature selection: old and new paradigms

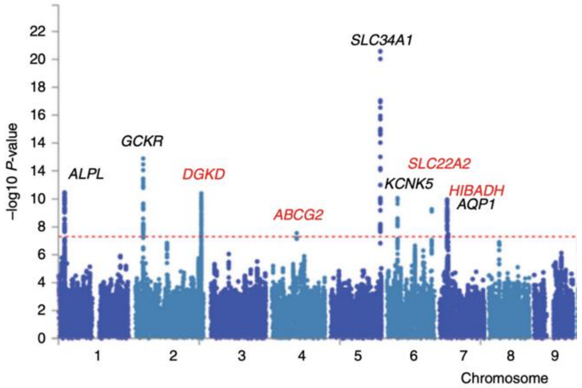
- Need more tools for
- Old paradigm: sparsity assumption, ie only a few variables, eg genes, are doing the biology, and hopefully they are independent
- But sometimes a lot of variable are doing some biology together, each doing a little bit, eg GWAS, ATAC-seq
- Lasso becomes unstable: sensitive to small perturbation in data
- But instability can be useful: ensemble learning



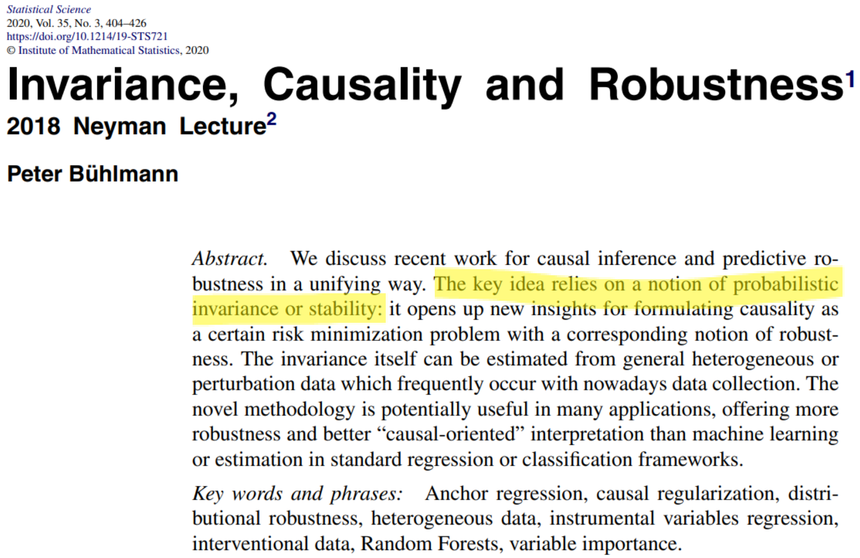
J. R. Statist. Soc. B (1996)
58, No. 1, pp. 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†



GWAS [wikipedia]



Statistical Science
2020, Vol. 35, No. 3, 404–426
https://doi.org/10.1214/19-STS721
© Institute of Mathematical Statistics, 2020

Invariance, Causality and Robustness[1]

2018 Neyman Lecture[2]

Peter Bühlmann

Abstract. We discuss recent work for causal inference and predictive robustness in a unifying way. The key idea relies on a notion of probabilistic invariance or stability: it opens up new insights for formulating causality as a certain risk minimization problem with a corresponding notion of robustness. The invariance itself can be estimated from general heterogeneous or perturbation data which frequently occur with nowadays data collection. The novel methodology is potentially useful in many applications, offering more robustness and better "causal-oriented" interpretation than machine learning or estimation in standard regression or classification frameworks.

Key words and phrases: Anchor regression, causal regularization, distributional robustness, heterogeneous data, instrumental variables regression, interventional data, Random Forests, variable importance.

Predictivity & Stability

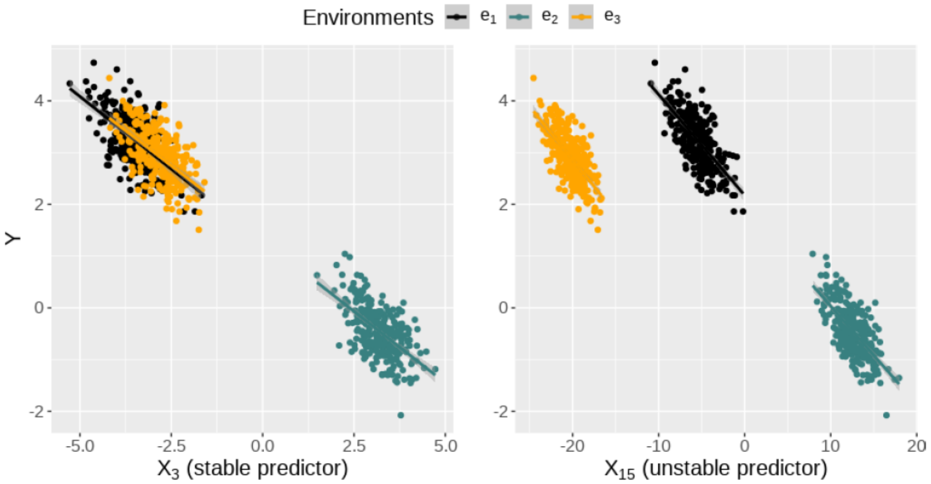
- Predictive: including X helps increase prediction of Y

- Stable: X is invariantly predictive across conditions

- Conditions:
  - Biological: ethnicity, cancer type, cell type
  - Technical: experimental batches

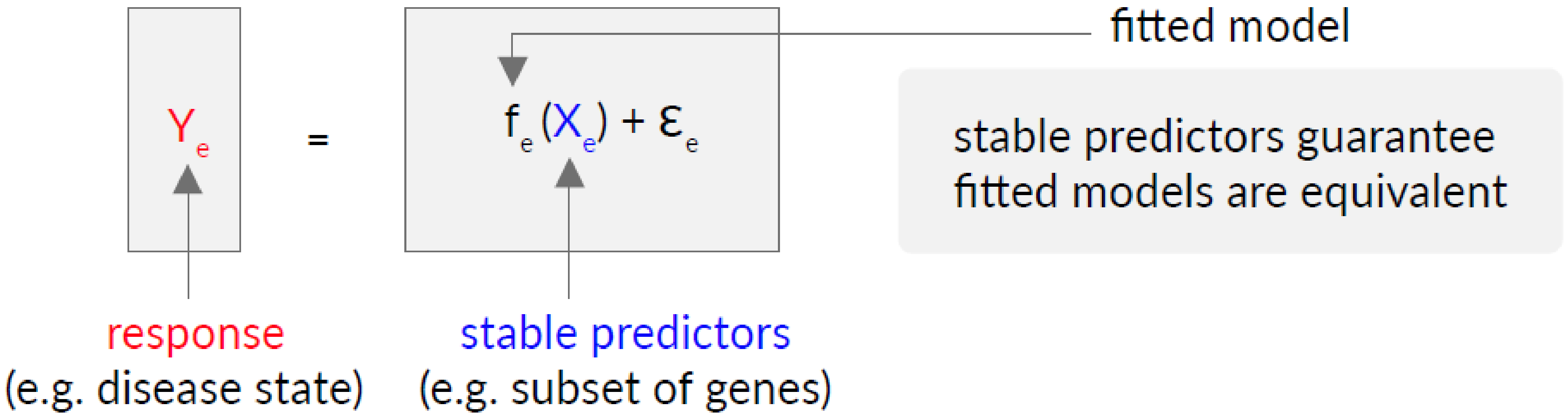
- Is instability bad?
  - When condition is purely technical, want to find stable predictors
  - When condition is biological, unstable ones also interesting
  - Eg ‘pan-cancer’ and ‘breast-cancer specific’
  - Let’s call unstable ones ‘condition-specific’

- Two types of questions:
  - Can find potentially causal variables, stable across technical conditions (perturbations)
  - Can find universal and condition-specific predictors



StableMate model

In each environment, e1, e2, e3
fit via stablised regression



$$Y_e = f_e(X_e) + \varepsilon_e$$

fitted model

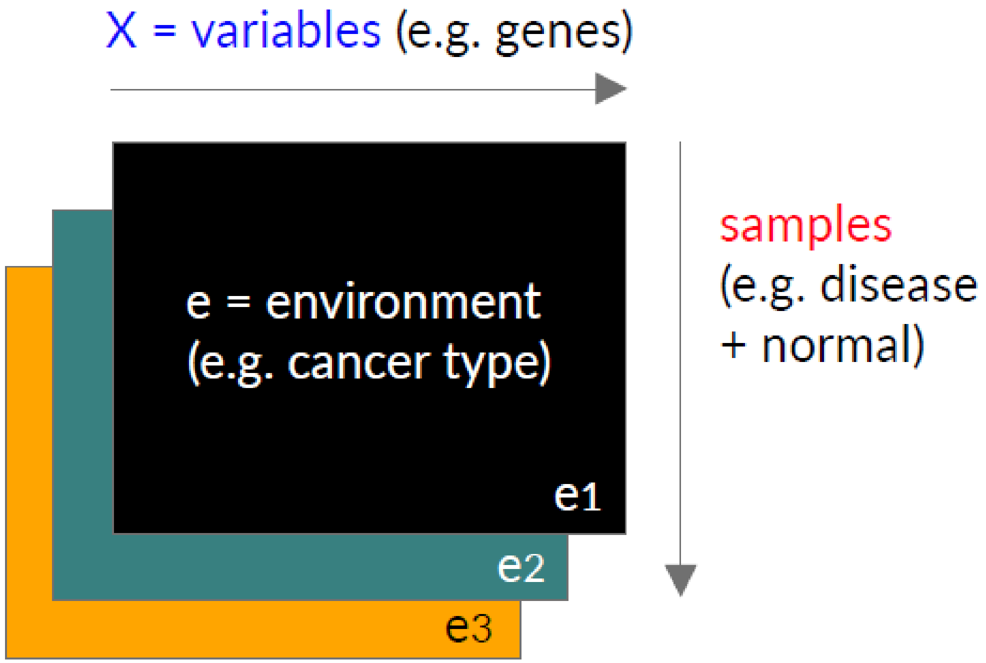
stable predictors guarantee fitted models are equivalent

response
(e.g. disease state)

stable predictors
(e.g. subset of genes)

Original stabilised regression (Pfister et al., 2021) is slow and cannot deal with too many variables.
We have greatly improved computational efficiency.

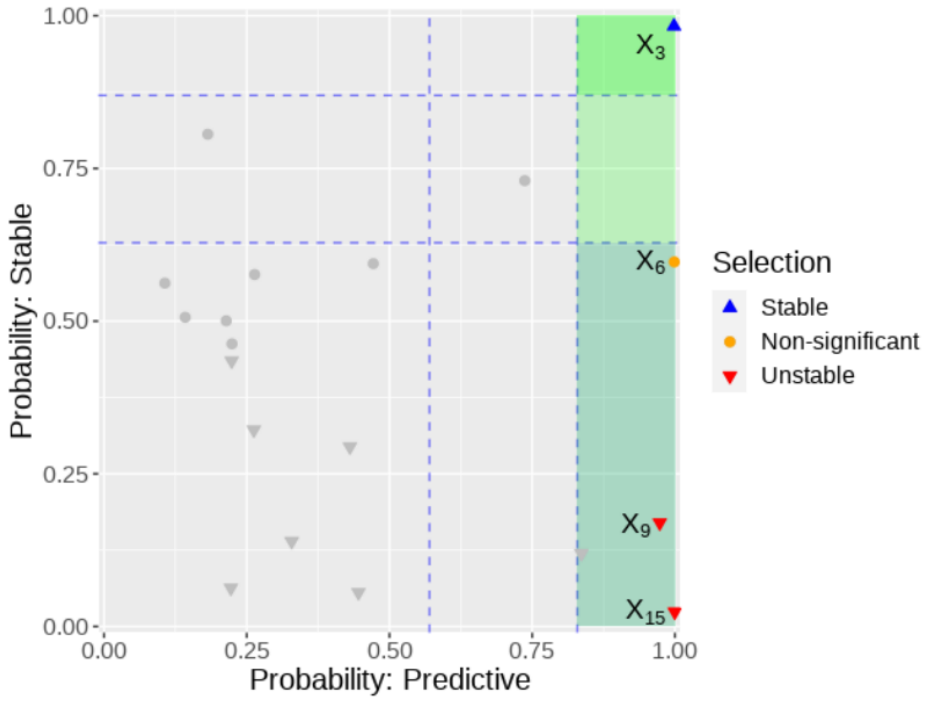
Input & output



X = variables (e.g. genes)

samples (e.g. disease + normal)

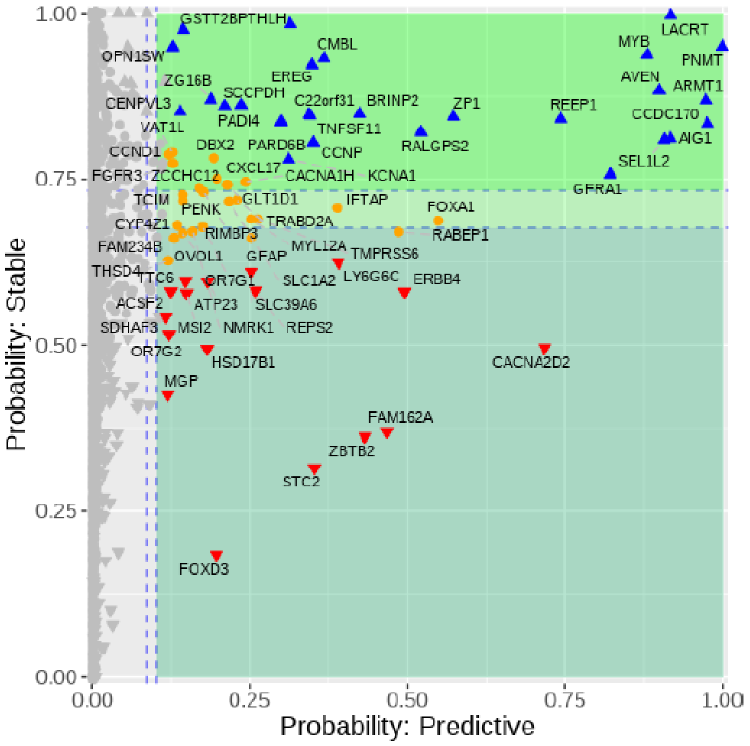
e = environment (e.g. cancer type)



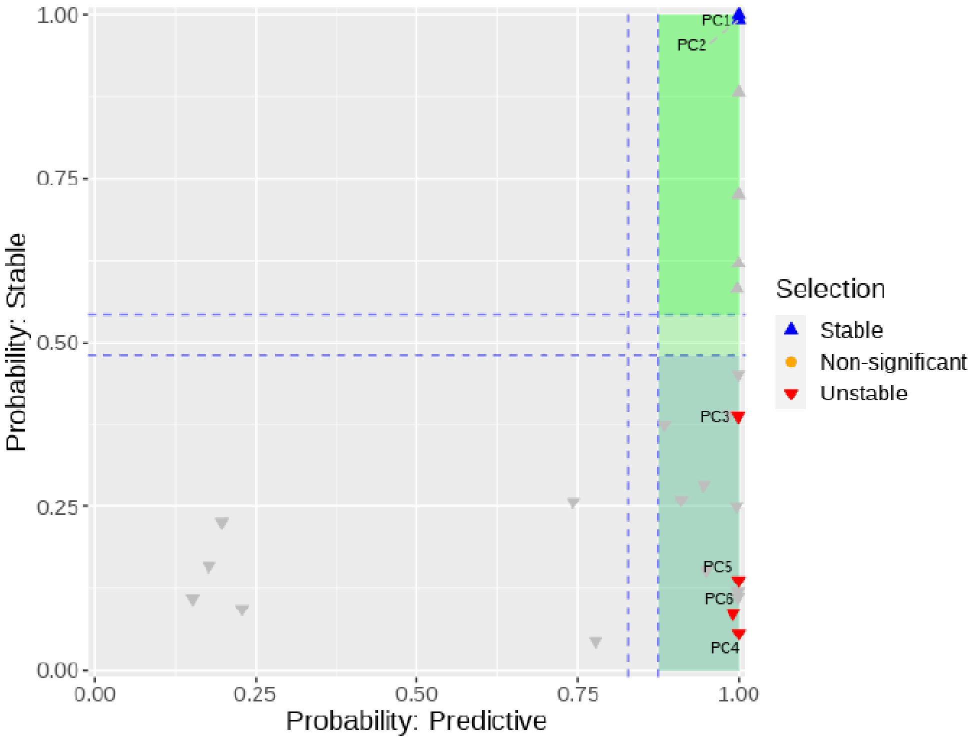
Breast cancer ESR1 regulation

- Data source: TCGA (The Cancer Genome Atlas) breast cancer (BRCA) data
- Data format: bulk RNA-seq (gene expression, count matrix)
- Response: ESR1 (estrogen receptor 1) gene expression
- Predictors: expression of other genes
- Conditions: disease status (113 healthy vs 778 ER+ BRCA patients)
- Biological background:
    - ESR1 is a biomarker for ER+ BRCA
    - Mutation in ESR1 may lead to increased tumour growth & drug resistance
    - Important to know how other genes **regulate** ESR1
    - StableMate helps understand how other genes regulate ESR1 in cancer vs healthy

Results

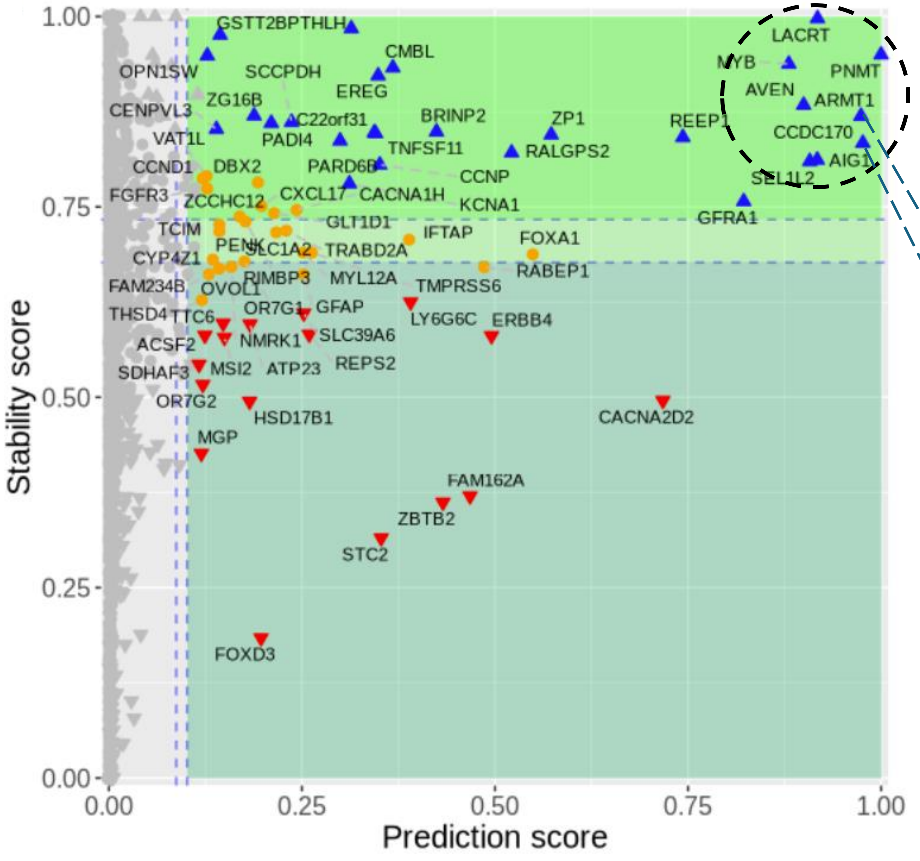


X = other genes

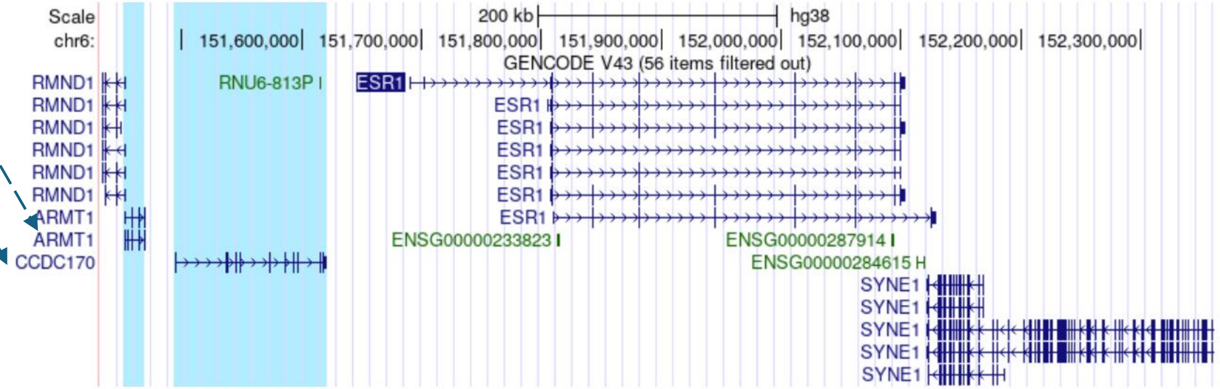


X = principal components of other genes

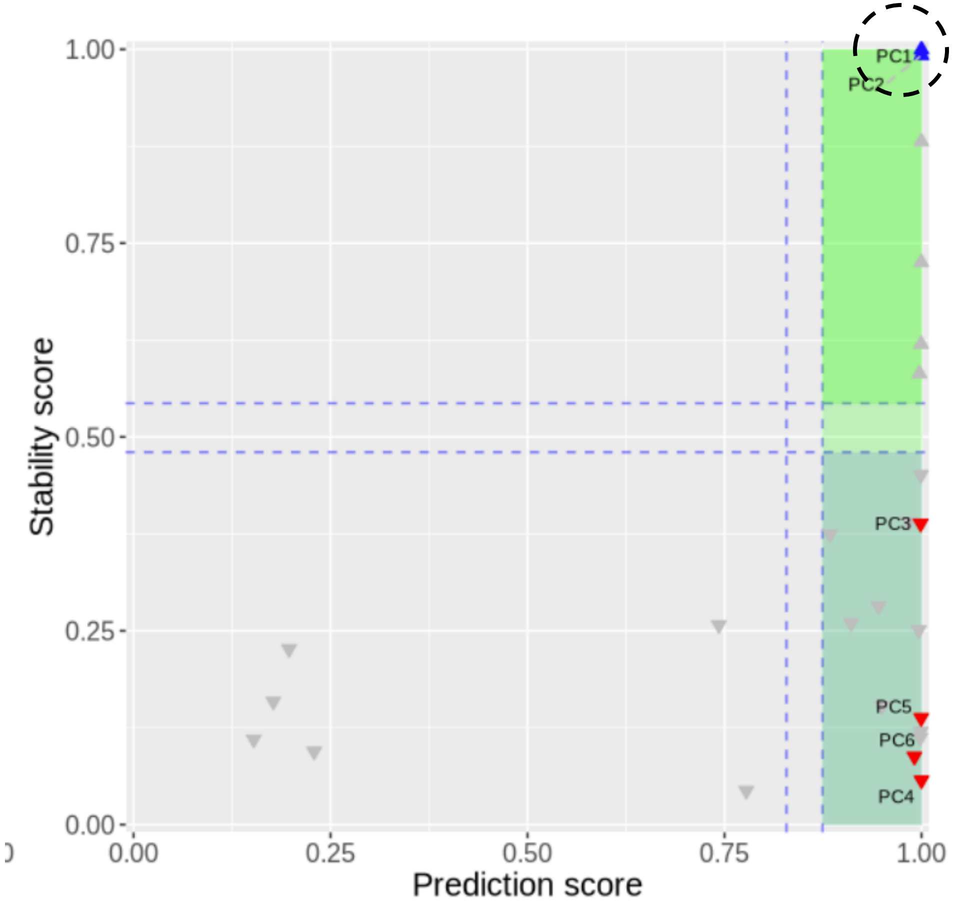
Interpretation: gene as X



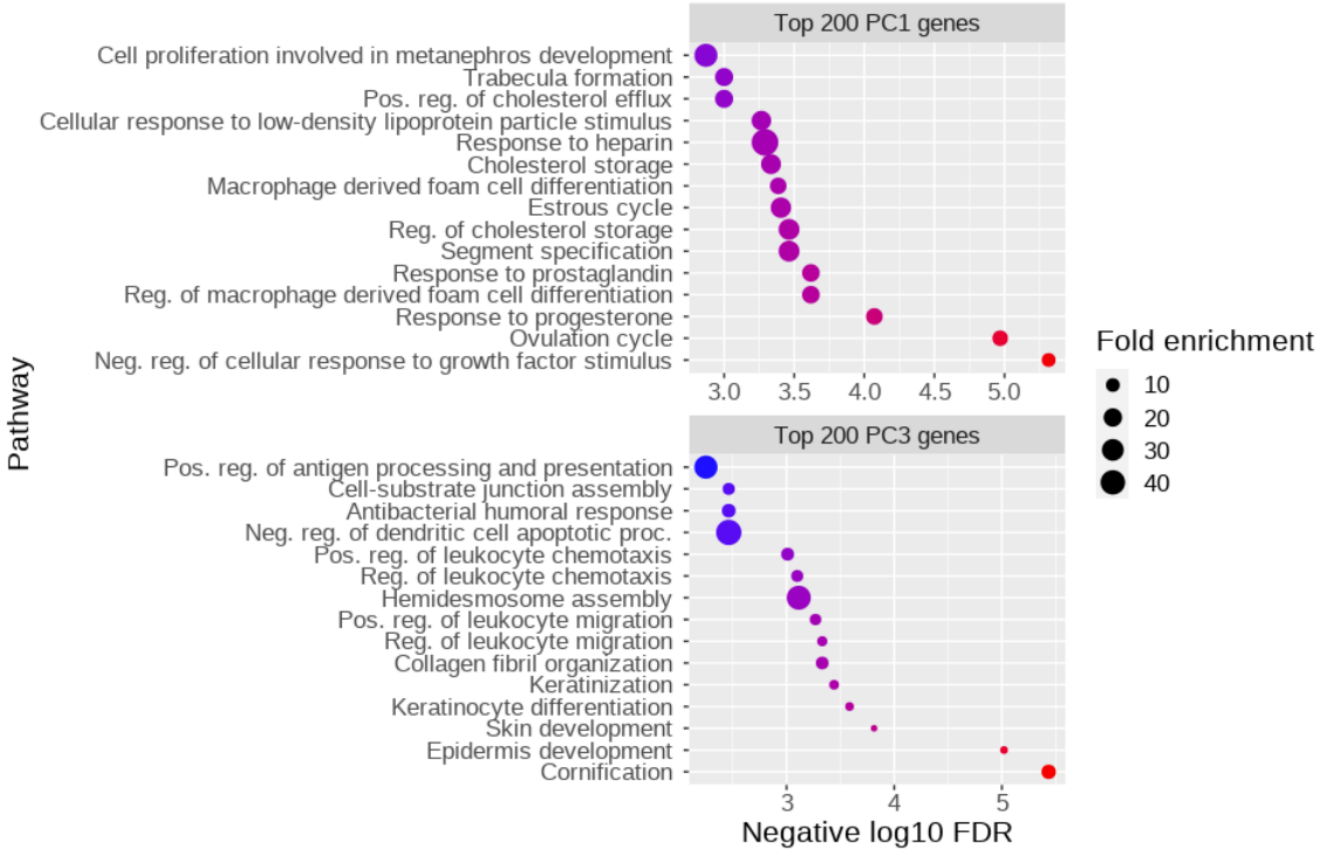
ARMT1 and CCDC170 may subject to the same transcriptional regulation as ESR1



Interpretation: PC as X

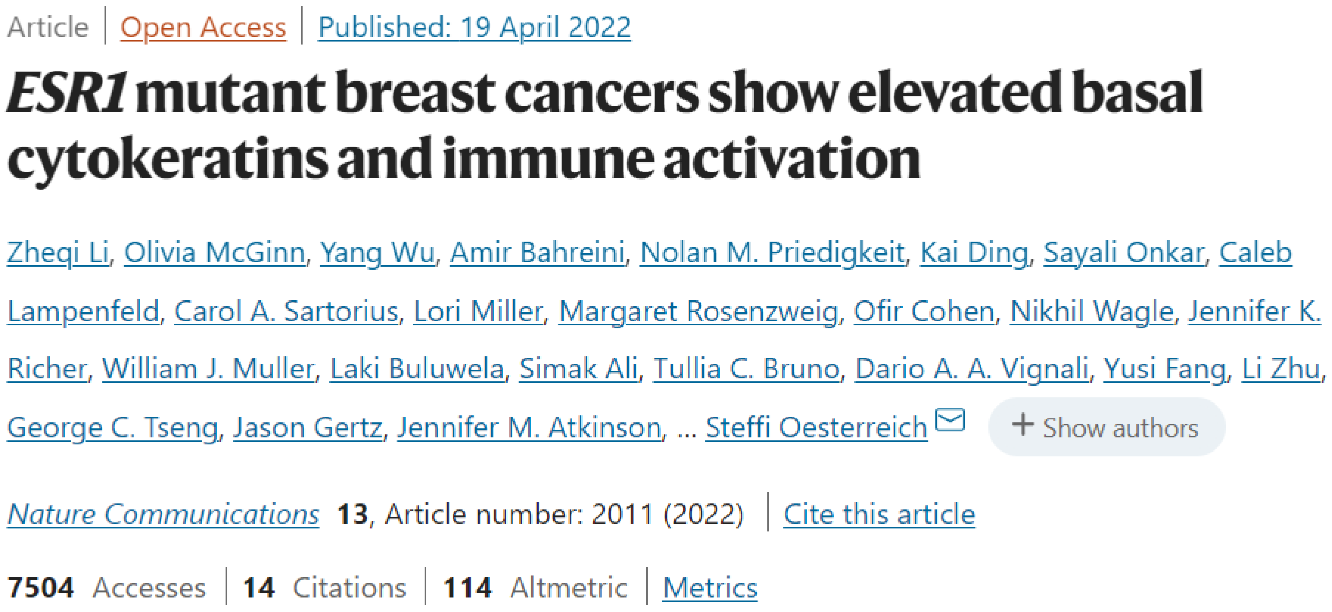


PC1 relates to hormonic regulation



PC3 relates to epidermis development

Does this makes sense?



Article | Open Access | Published: 19 April 2022

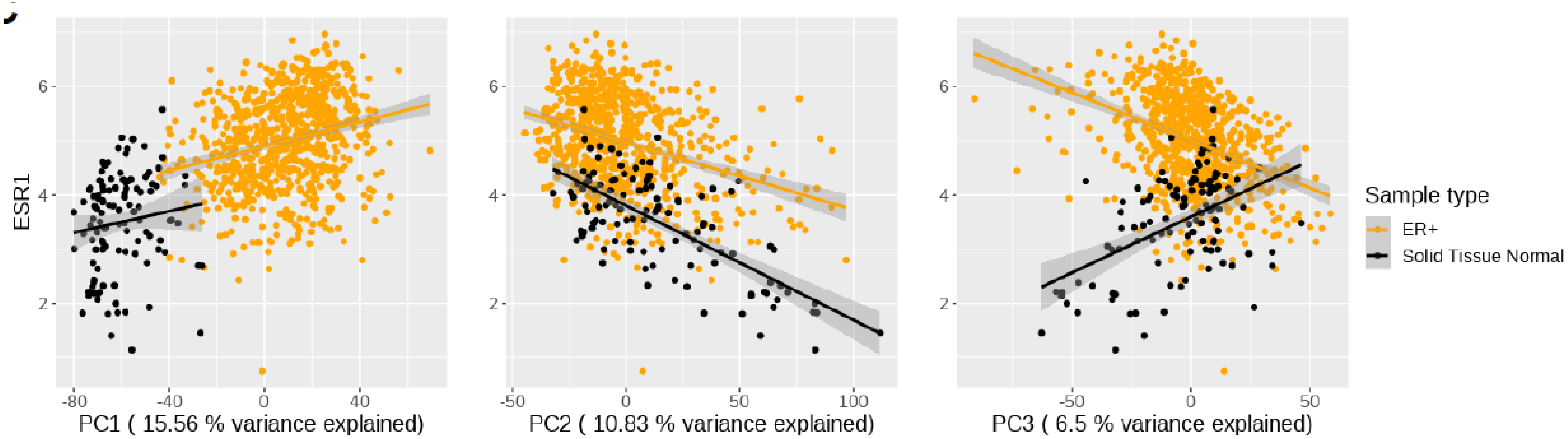
ESR1 mutant breast cancers show elevated basal cytokeratins and immune activation

Zheqi Li, Olivia McGinn, Yang Wu, Amir Bahreini, Nolan M. Priedigkeit, Kai Ding, Sayali Onkar, Caleb Lampenfeld, Carol A. Sartorius, Lori Miller, Margaret Rosenzweig, Ofir Cohen, Nikhil Wagle, Jennifer K. Richer, William J. Muller, Laki Buluwela, Simak Ali, Tullia C. Bruno, Dario A. A. Vignali, Yusi Fang, Li Zhu, George C. Tseng, Jason Gertz, Jennifer M. Atkinson, ... Steffi Oesterreich ✉ + Show authors

Nature Communications **13**, Article number: 2011 (2022) | Cite this article

7504 Accesses | **14** Citations | **114** Altmetric | Metrics

PC3 shows cancer specific regulation



Thoughts

- Stabilised regression is very natural for problems in public health
- Stability of polygenic risk scores?
- Factor model: PCA and PLS

Appendix

StableMate algorithm

based on stochastic stepwise (ST2, Xin et al, 2012) variable selection

Classic

1. Fit regression model.
2. Add or remove one variable per step.
3. Stop until no improvement.

ST2

1. Fit regression model.
2. Randomly subsample some predictor sets.
3. Add or remove one set per step
4. Stop until no improvement.

Xin, L., & Zhu, M. (2012). Stochastic stepwise ensembles for variable selection.

StableMate algorithm

S_{all}

ST2: select predictive sets

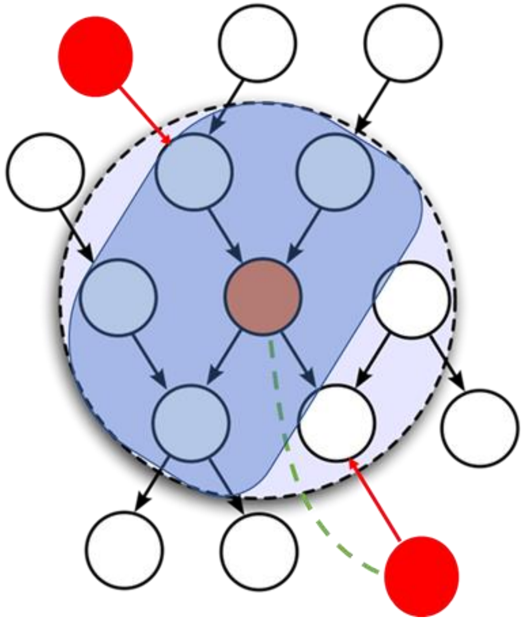
$\{S_1^p, S_2^p, S_3^p, \dots, S_K^p\}$

ST2: select stable and predictive sets

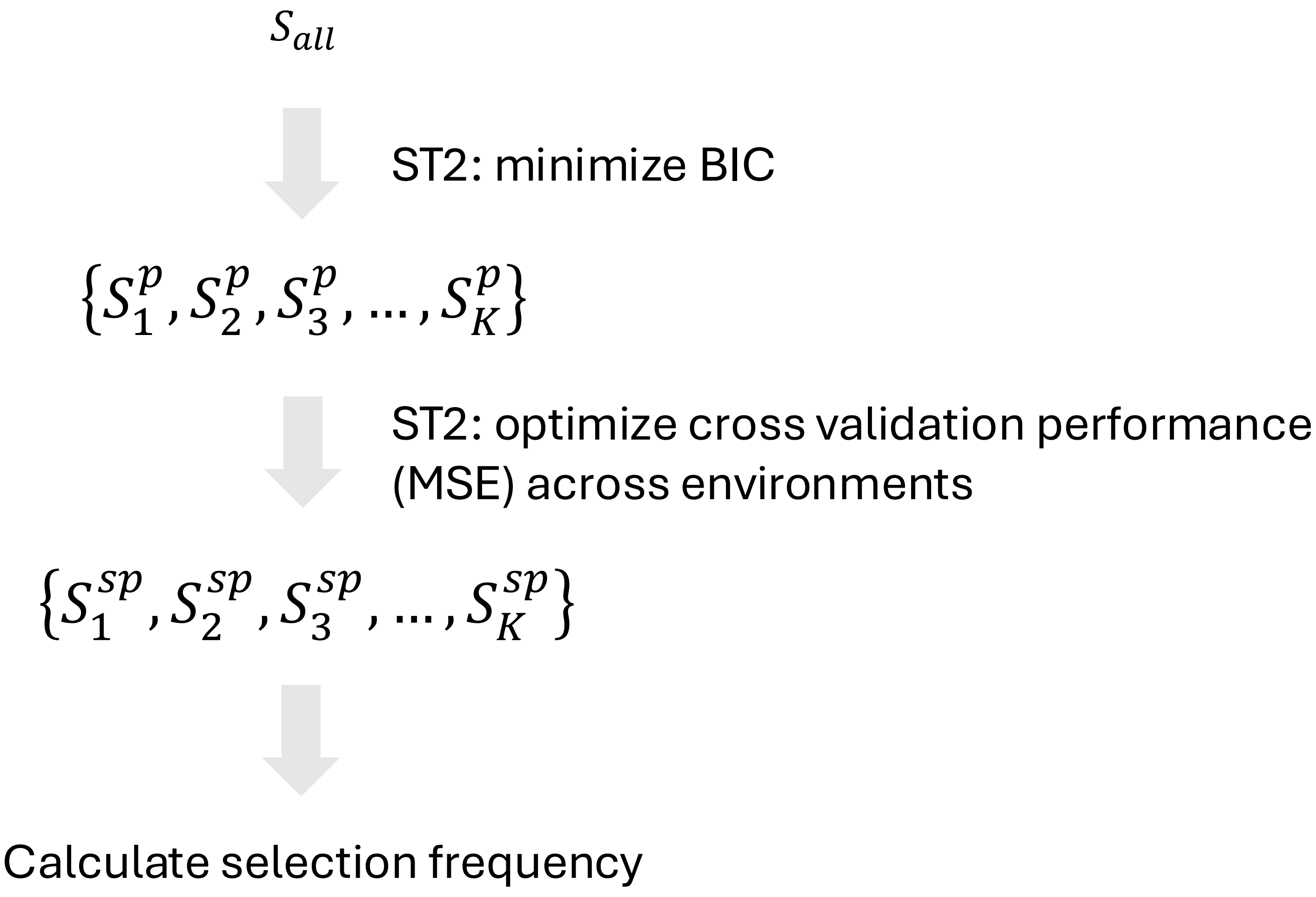
$\{S_1^{sp}, S_2^{sp}, S_3^{sp}, \dots, S_K^{sp}\}$

- SB must be the subset of MB

$S_k^{sp} \subseteq S_k^p$



Objectives



S_{all}

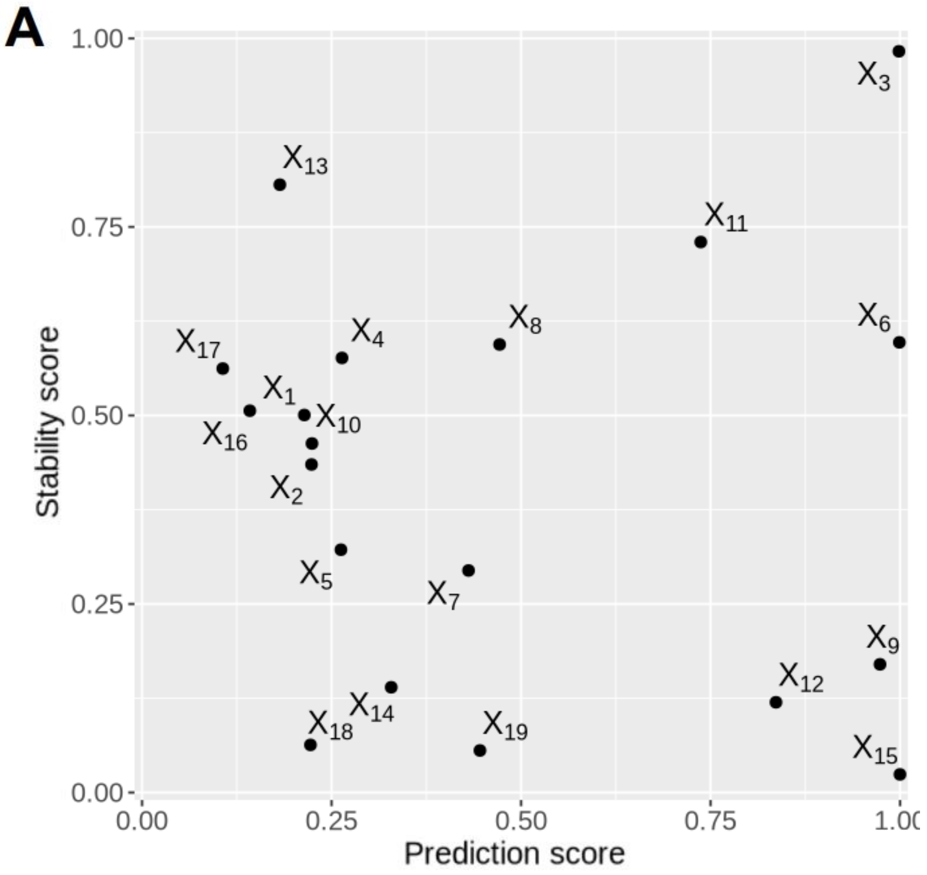
ST2: minimize BIC

$\{S_1^p, S_2^p, S_3^p, \dots, S_K^p\}$

ST2: optimize cross validation performance (MSE) across environments

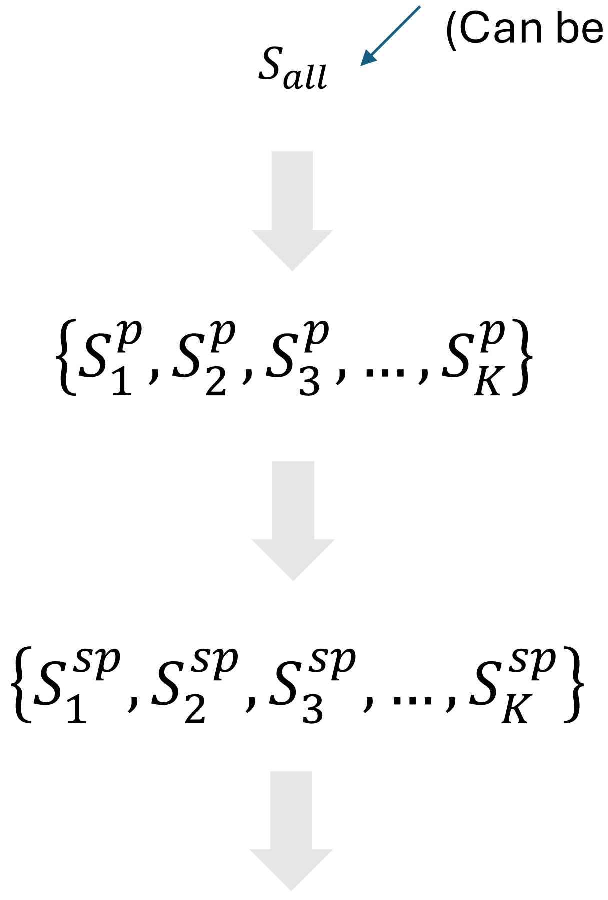
$\{S_1^{sp}, S_2^{sp}, S_3^{sp}, \dots, S_K^{sp}\}$

Calculate selection frequency



Make selection

Add a pseudo-predictor
(Can be selected but doesn’t influence model fitting).

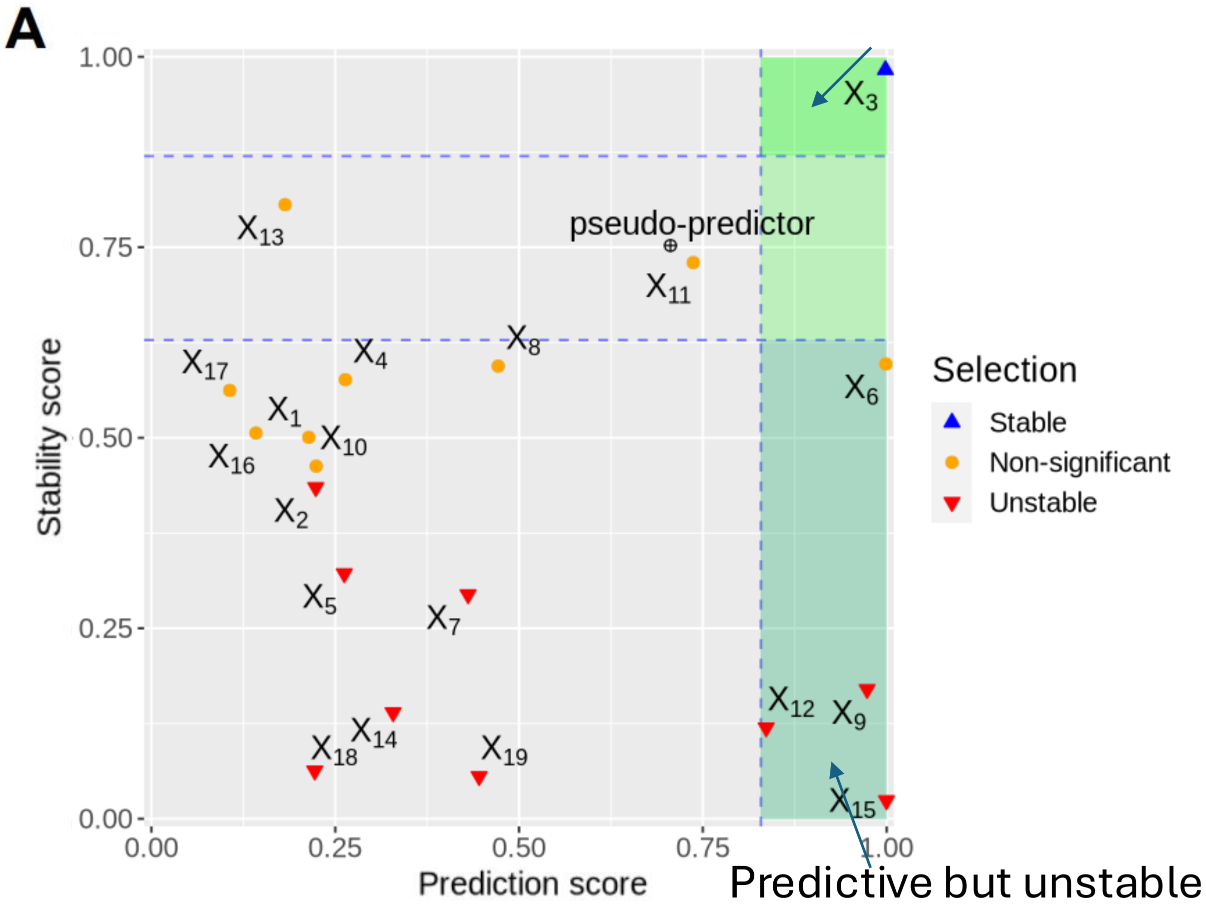


S_{all}

$\{S_1^p, S_2^p, S_3^p, \ldots, S_K^p\}$

$\{S_1^{sp}, S_2^{sp}, S_3^{sp}, \ldots, S_K^{sp}\}$

Calculate selection frequency



Predictive and stable

Predictive but unstable