StableMate:

A statistical method to select stable predictors in omics data

Yidi Deng (Melbourne Integrative Genomics)

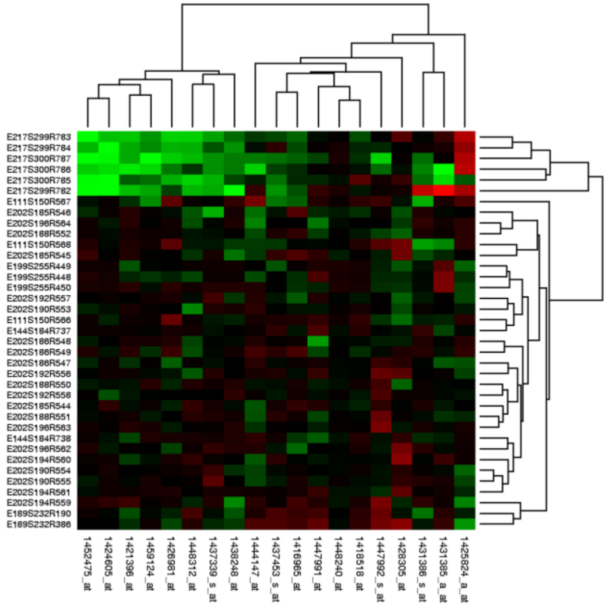
Supervised by:
Prof. Kim-Anh Lê Cao
Dr. Jarny Choi
Dr. Jiadong Mao

THE UNIVERSITY OF
MELBOURNE

Motivation

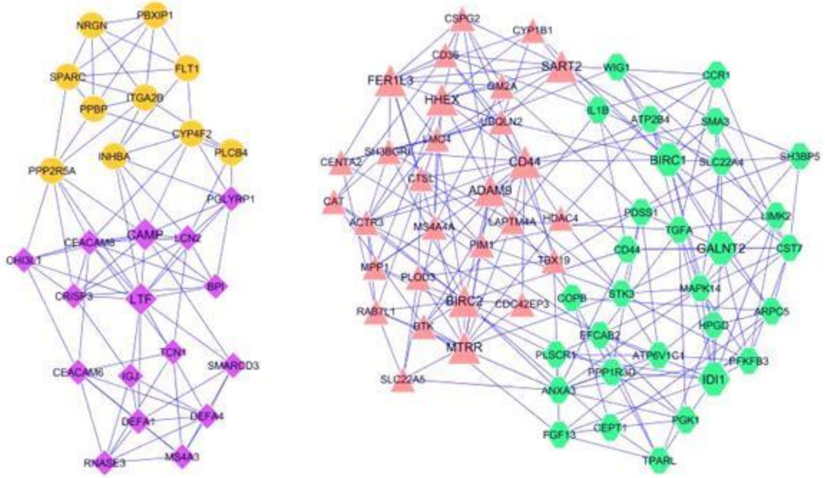
Infer biological relationship from statistical association

Example: gene regulatory network



Gene expression data

Retrieved from https://en.wikipedia.org/wiki/Gene_expression_profiling



Network

Retrieved from https://www.ese.wustl.edu/~nehorai/research/genomic/grn.html

Limitation of current methods

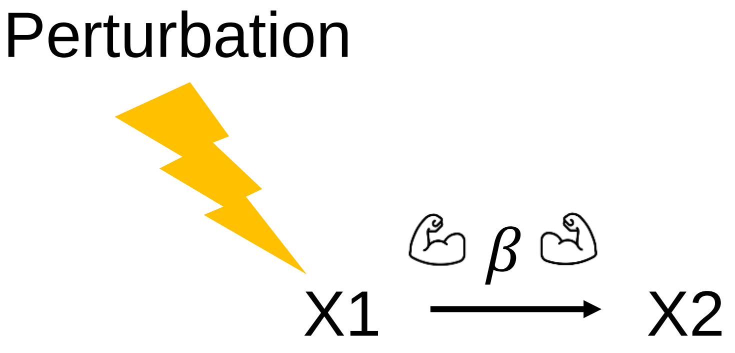
1. Lack of interpretability:

   Statistical association $\xrightarrow{?}$ Biological hypothesis for validation

2. Lack of generalizability:

   Study1 $\xrightarrow{?}$ Study2

Stable association



1. Robust to perturbation

   Generalizable

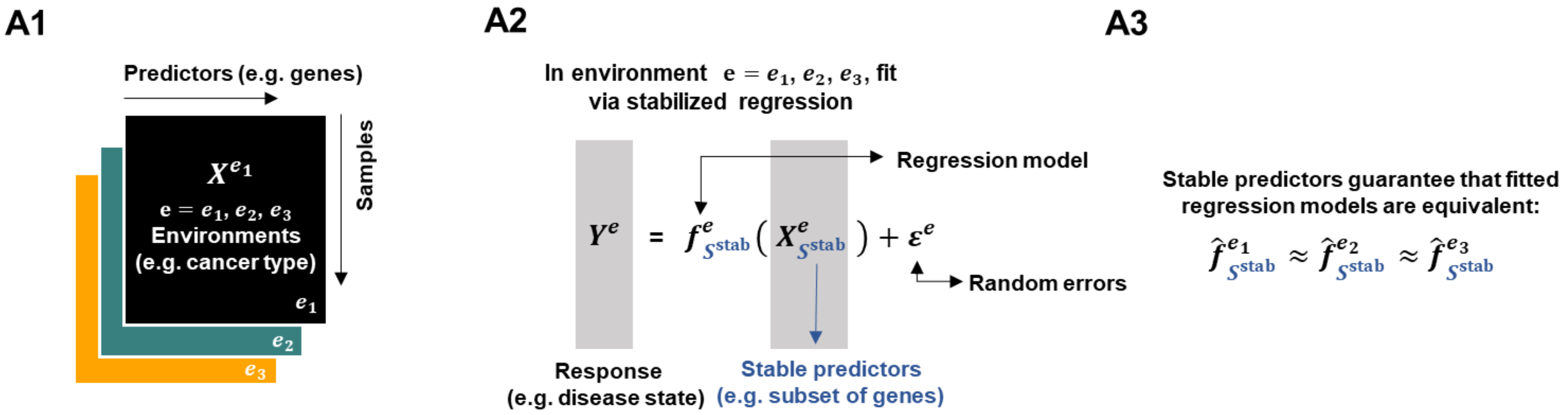
2. Causal implication

   Interpretable

Bühlmann, P. (2020). Invariance, causality and robustness.

Stabilized Regression
(Pfister et al. 2021)

Goal of Stabilized Regression (SR)



A1

Predictors (e.g. genes)

Samples

X^{e_1}

$e = e_1, e_2, e_3$
Environments
(e.g. cancer type)

e_1 e_2 e_3

A2

In environment $e = e_1, e_2, e_3$, fit
via stabilized regression

$$Y^e = f^e_{S^{stab}}\left(X^e_{S^{stab}}\right) + \varepsilon^e$$

Regression model

Random errors

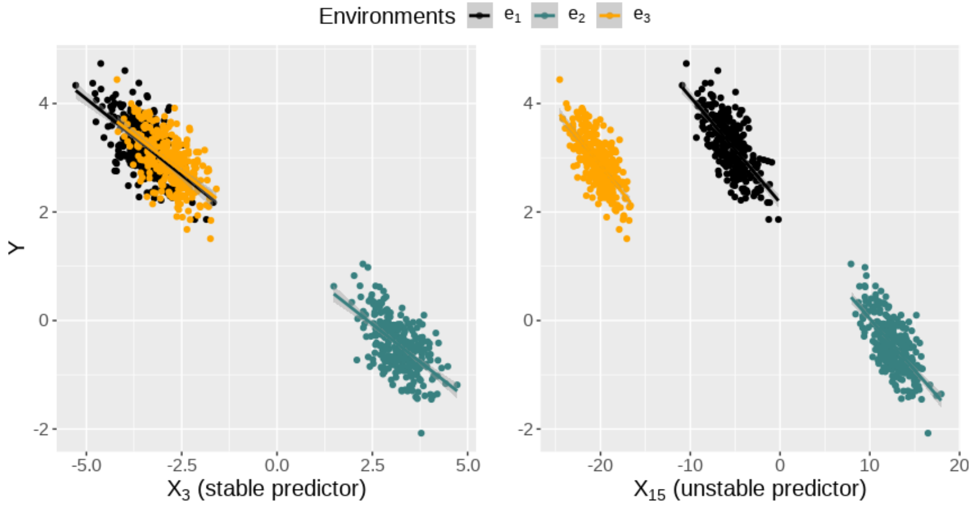
Response
(e.g. disease state)

Stable predictors
(e.g. subset of genes)

A3

Stable predictors guarantee that fitted regression models are equivalent:

$$\hat{f}^{e_1}_{S^{stab}} \approx \hat{f}^{e_2}_{S^{stab}} \approx \hat{f}^{e_3}_{S^{stab}}$$

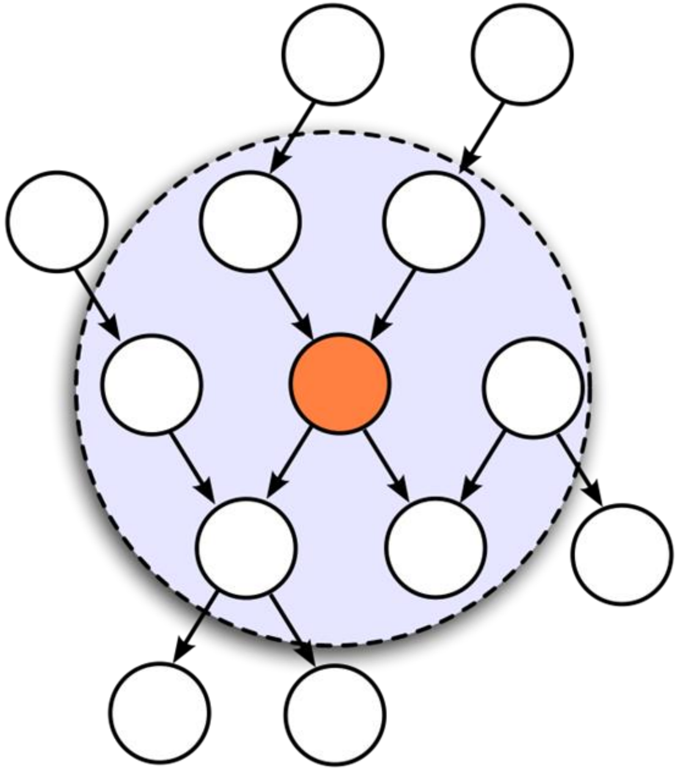


Infer generalizable functional dependency on the response

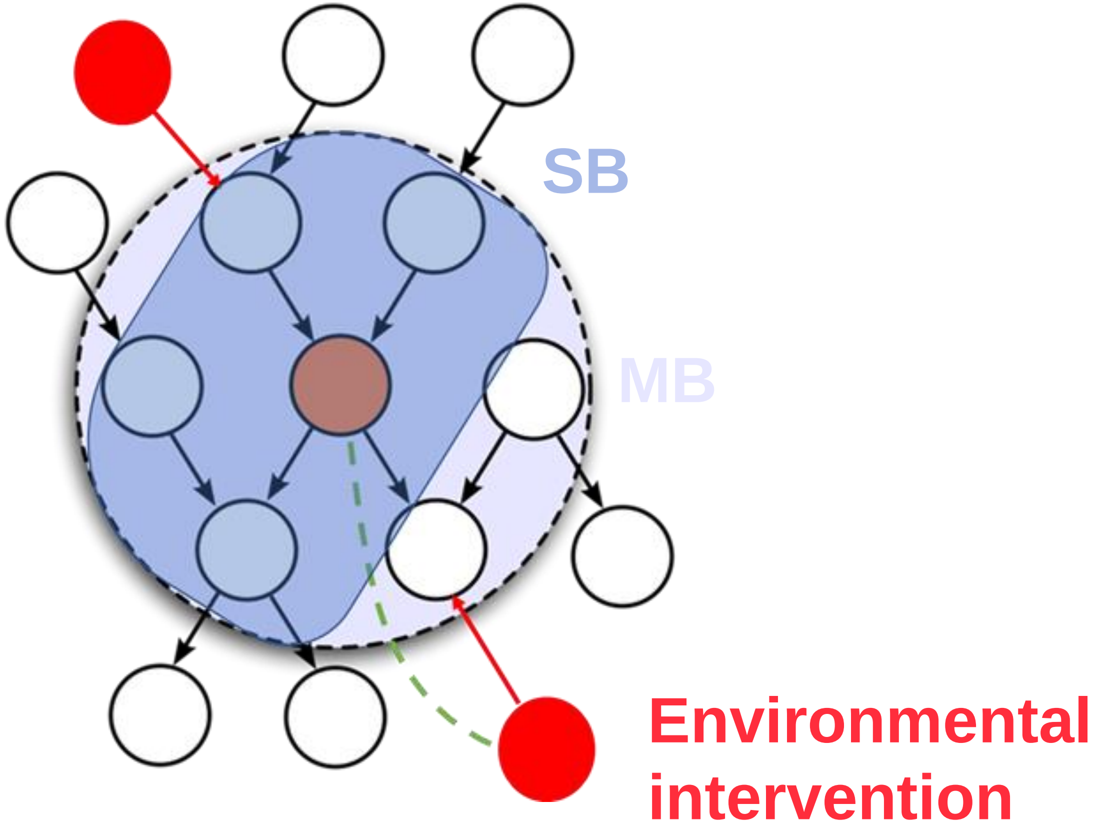
Pfister, N., (2021). Stabilizing variable selection and regression

Markov Blanket and Stable Blanket

MB: The most predictive set



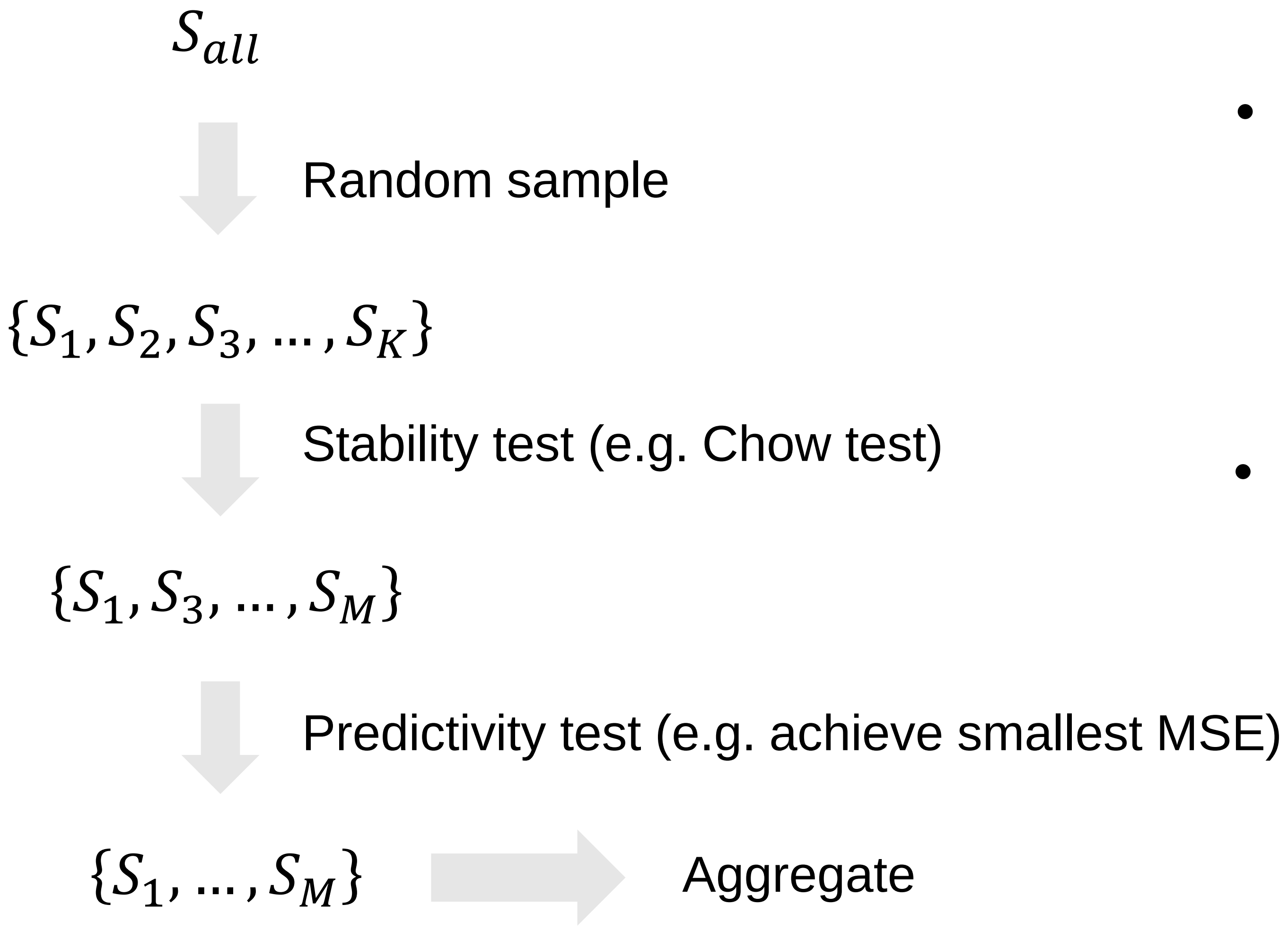
SB: The most predictive stable set



PA <= SB <= MB

SR algorithm

Given a response, predictors and environments



S_{all}

Random sample

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

Stability test (e.g. Chow test)

$\{S_1, S_3, \ldots, S_M\}$

Predictivity test (e.g. achieve smallest MSE)

$\{S_1, \ldots, S_M\}$

Aggregate

- Inaccurate:
  Hard to generate enough subsets to test.

- Solution:
  Over sample
  Pre-filtering

StableMate

(Deng et al. 2023)

Deng, Y., Mao, J., Choi, J., & Lê Cao, K. A. (2023). StableMate: a statistical method to select stable predictors in omics data. bioRxiv, 2023-09.

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}

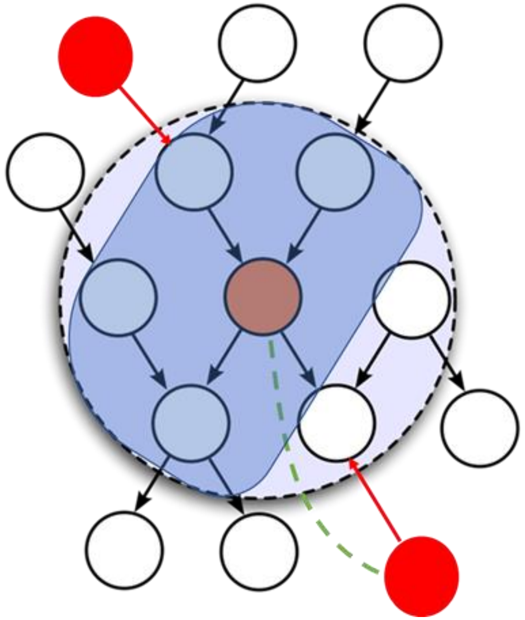
ST2: select predictive sets

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

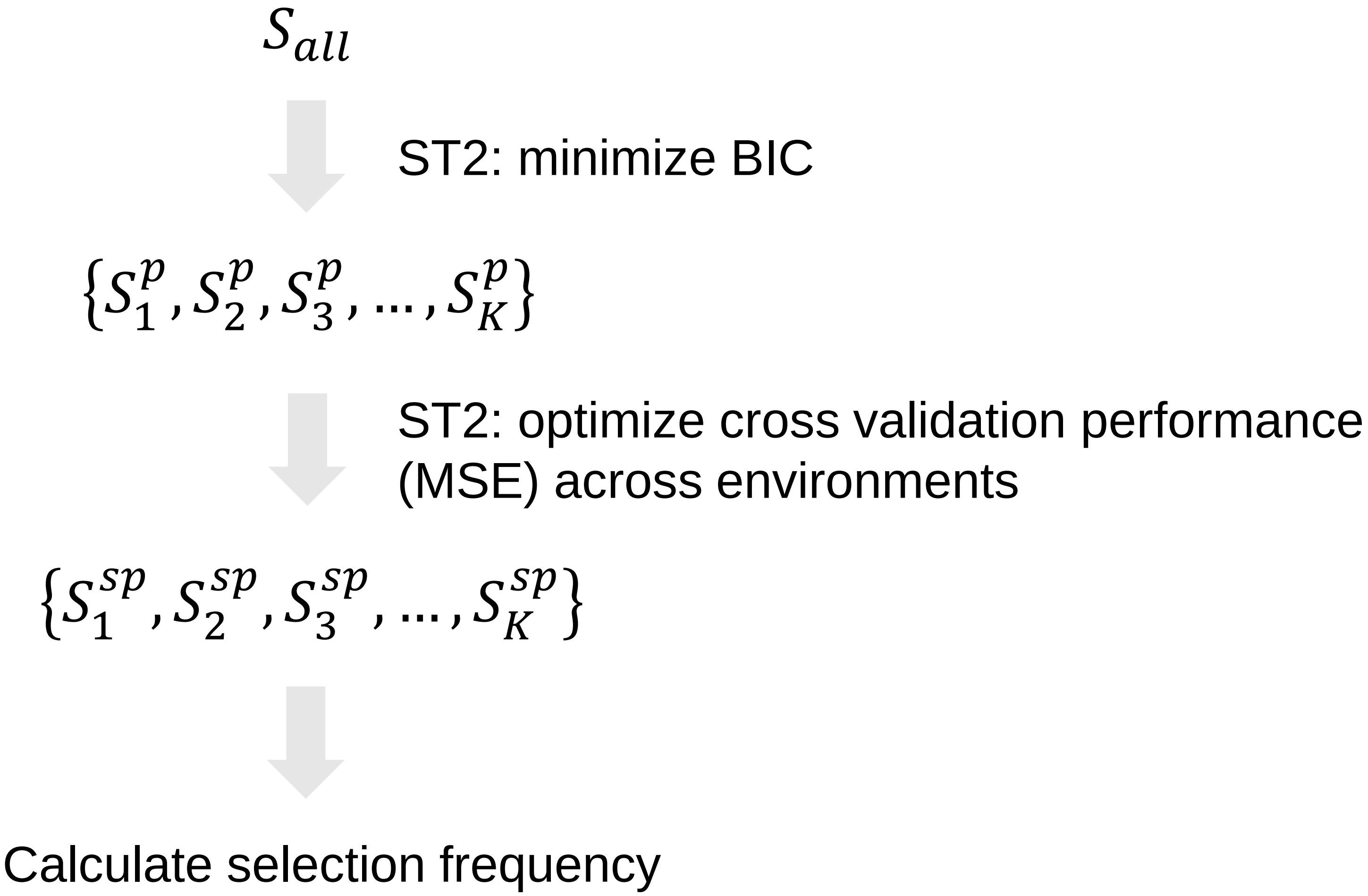
ST2: select stable and predictive sets $S_k^{sp} \subseteq S_k^p$

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

- SB must be the subset of MB



Objectives



S_{all}

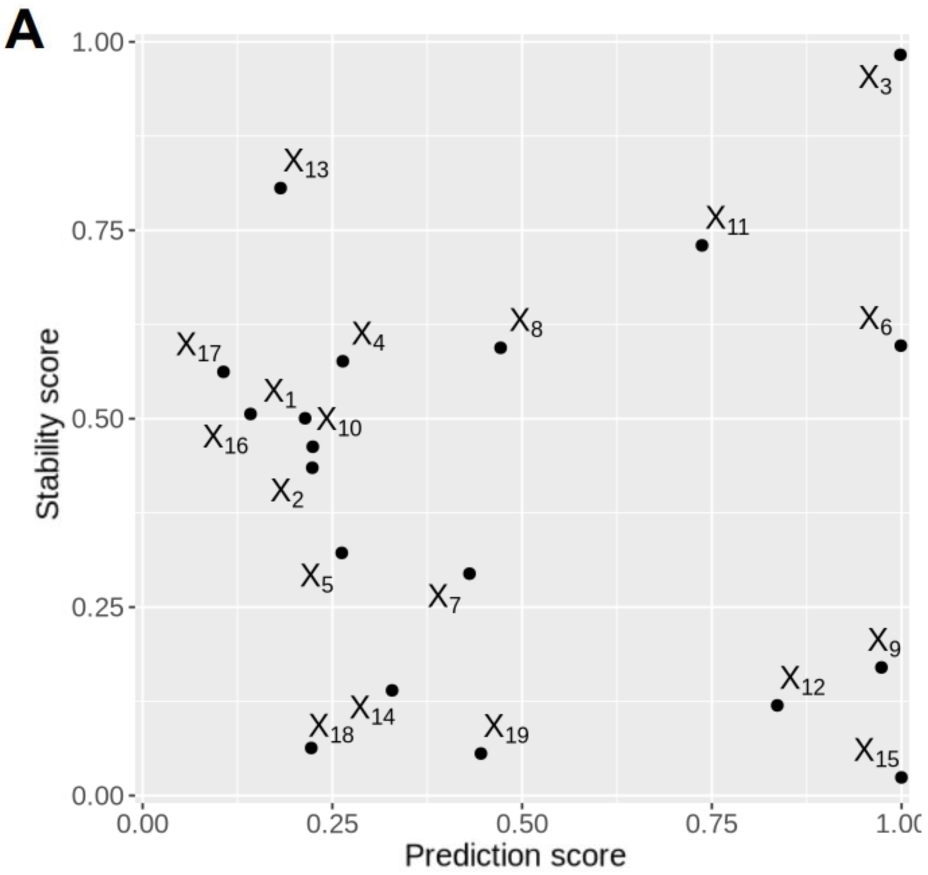
ST2: minimize BIC

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

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

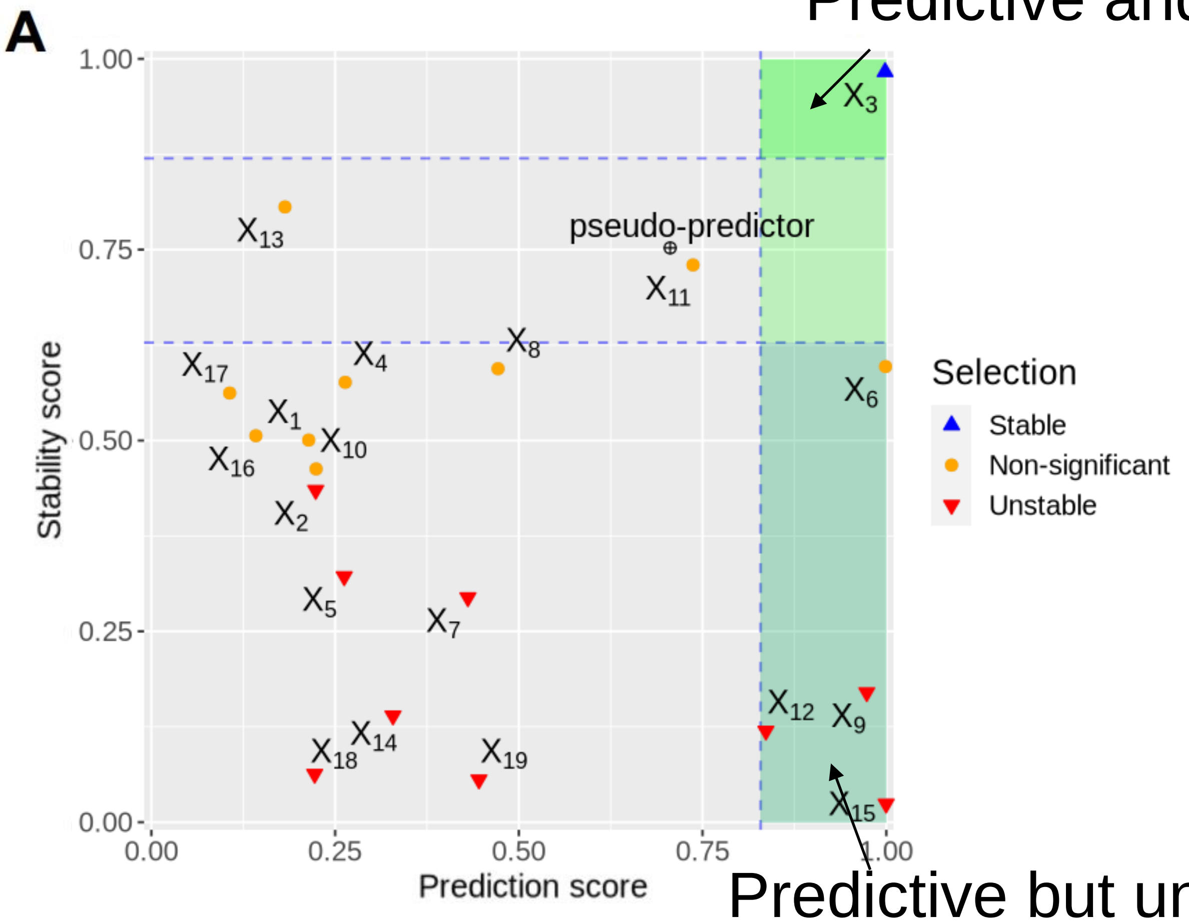
S_{all} ↙ Add a pseudo-predictor (Can be selected but doesn’t influence model fitting).

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

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

Calculate selection frequency

Predictive and stable



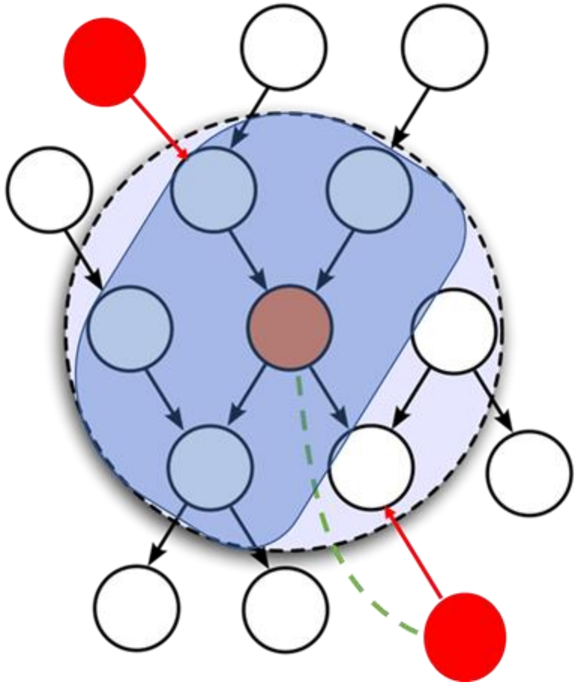
Predictive but unstable

THE UNIVERSITY OF
MELBOURNE

Simulation study

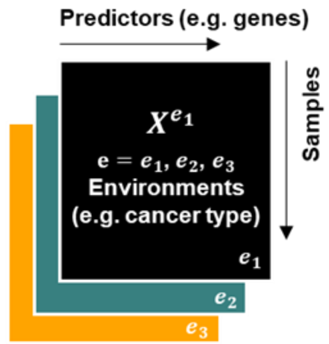
Setting

Simulate a structural causal model (SCM) in different environments



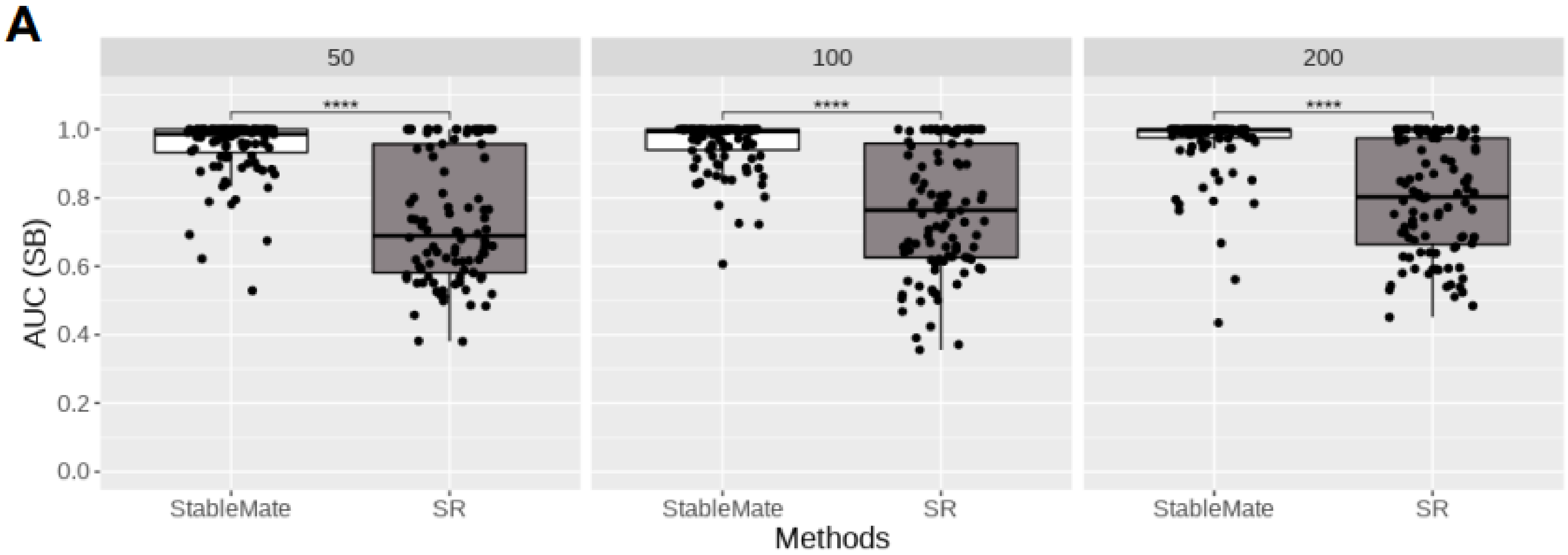
Differs by exogenous perturbation

- Generate data according to the SCM (Three training, one testing environment)



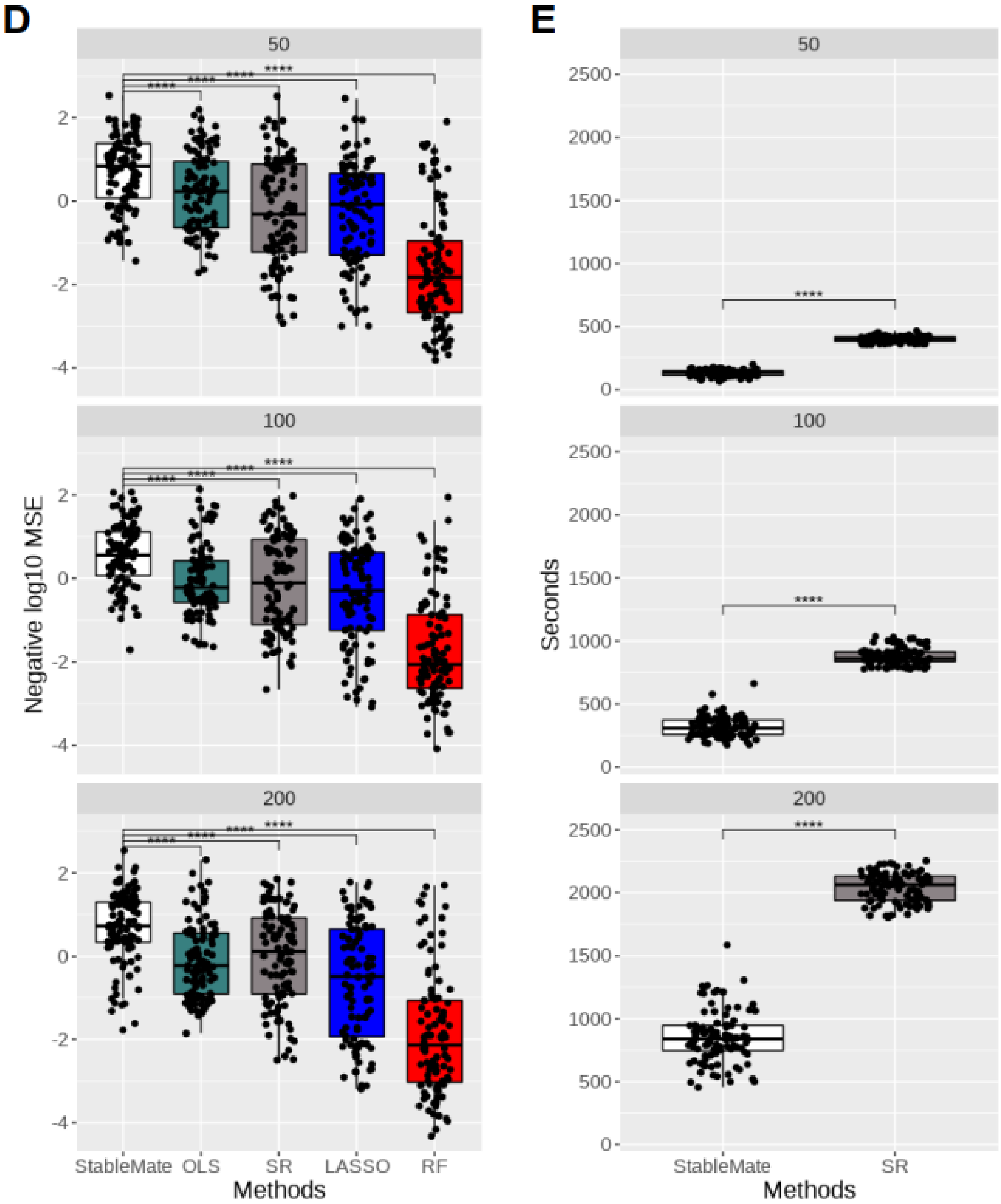
- Mask the SCM

- Identify SB from observational data



StableMate makes better selections

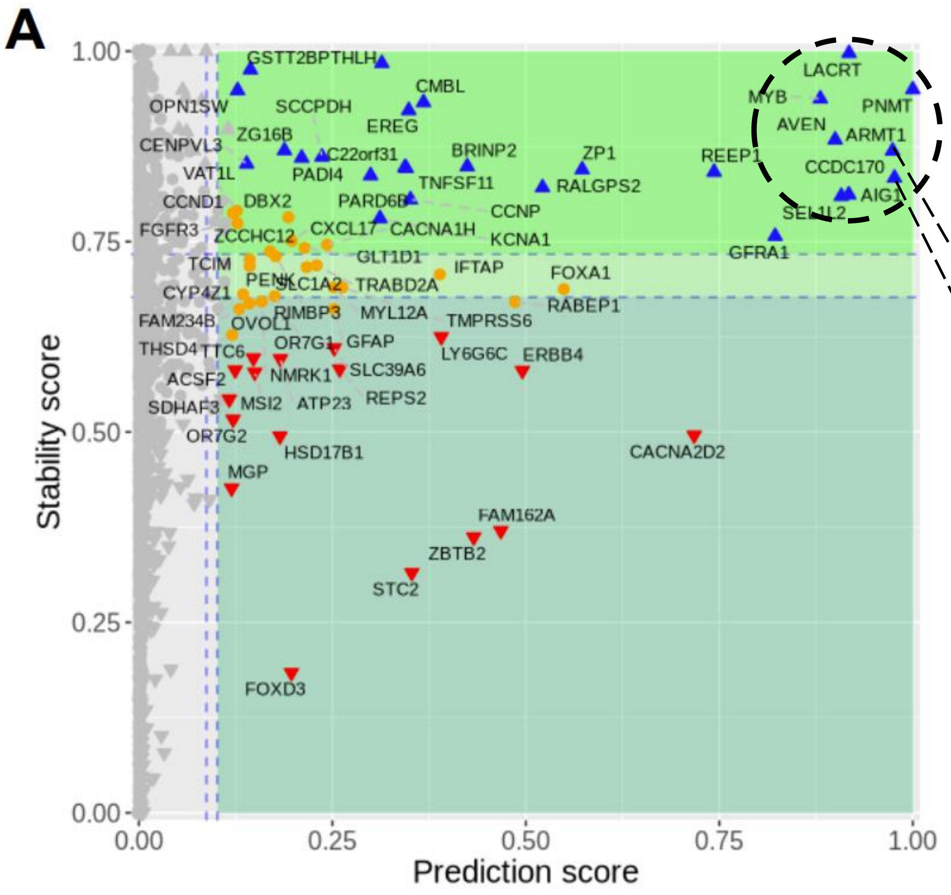
StableMate extrapolate better, with greatly reduced running time



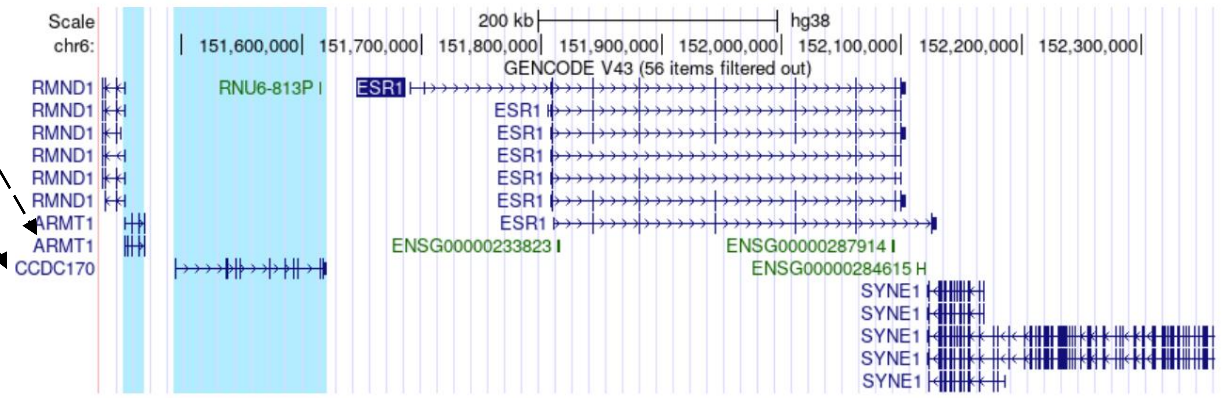
A case study on breast cancer RNA-seq data with BRCA gene mutation

- Source: TCGA (The Cancer Genome Atlas Program) consortium
- Data: RNA-seq (gene expression, a count matrix)
- Response: ESR1 (estrogen receptor 1) gene expression
- Environment: disease status (113 normal or 778 ER+ samples)

ESR1 vs other genes



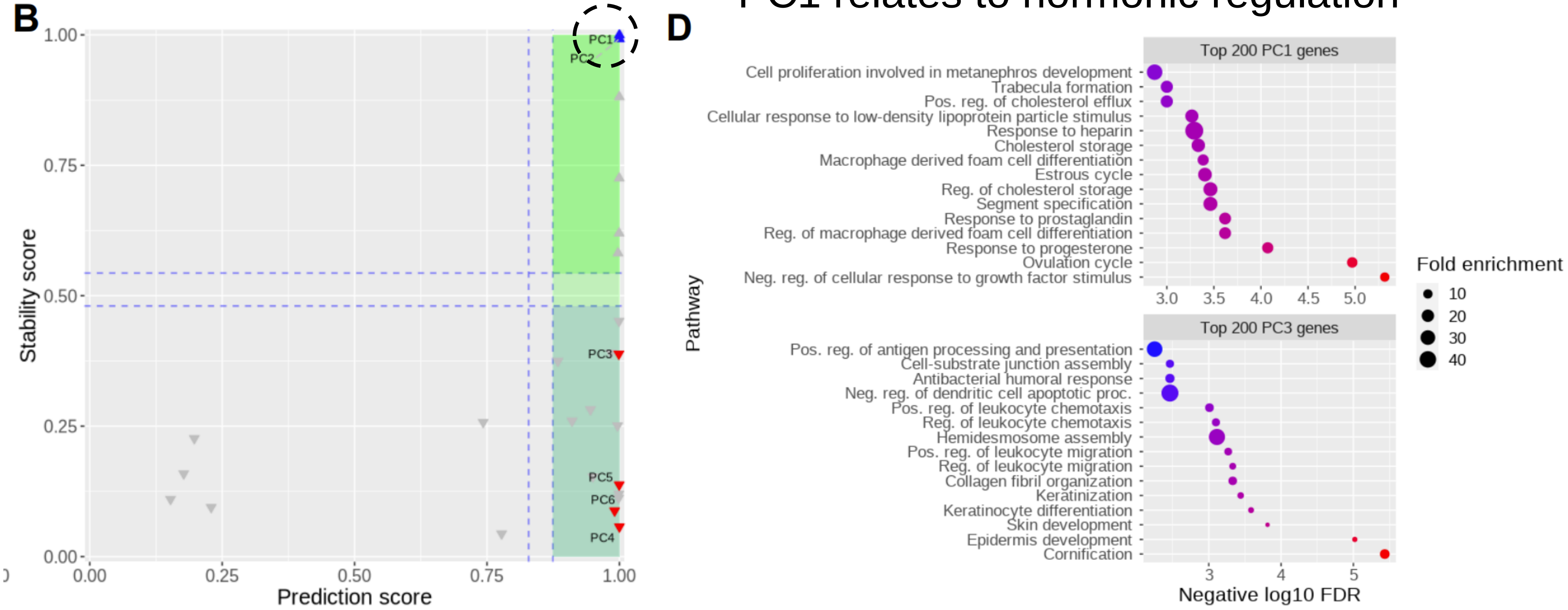
ARMT1 and CCDC170 may subject to the same transcriptional regulation as ESR1



ESR1 vs principal components

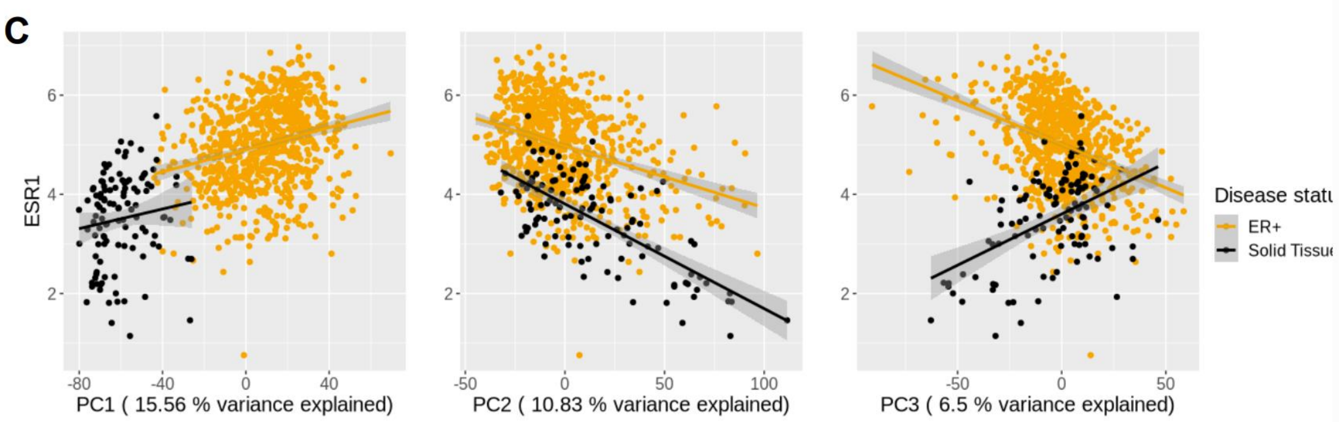
(Estrogen receptor 1)

PC1 relates to hormonic regulation



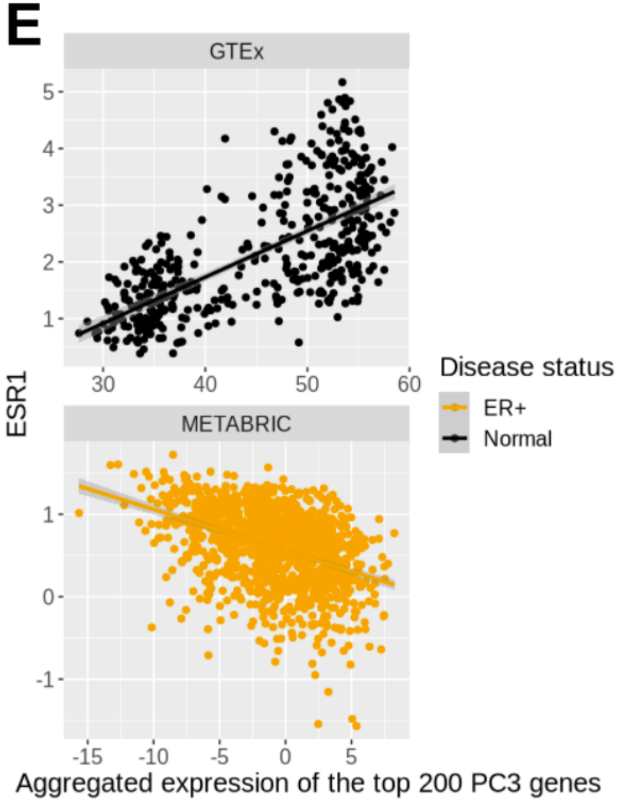
PC3 relates to epidermis development

ESR1 vs principal components



Consistent pattern

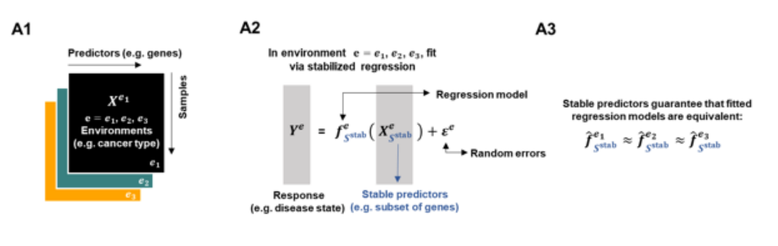
Inconsistent pattern



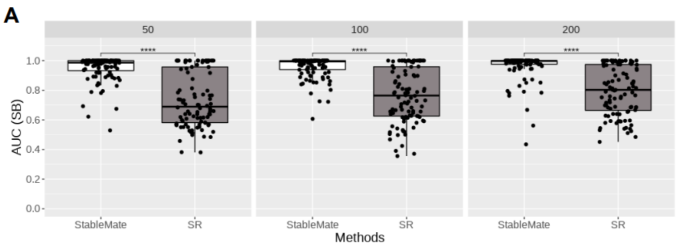
StableMate reveals interesting functional heterogeneity to be investigated.

External validation on two consortium studies (GTEX, METABRIC)

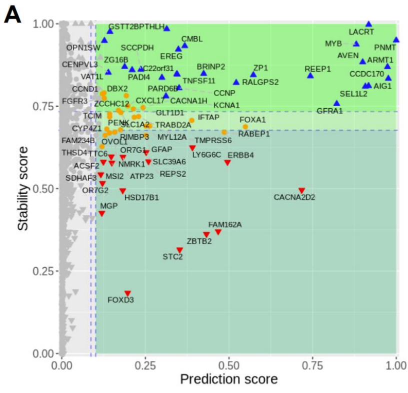
Summary


A1
Predictors (e.g. genes)
Samples
X^{e_1}
$e = e_1, e_2, e_3$
Environments
(e.g. cancer type)
e_1
e_2
e_3
A2
In environment $e = e_1, e_2, e_3$, fit
via stabilized regression
Regression model
$Y^e = f^e_{S^{stab}}(X^e_{S^{stab}}) + \varepsilon^e$
Random errors
Response
(e.g. disease state)
Stable predictors
(e.g. subset of genes)
A3
Stable predictors guarantee that fitted
regression models are equivalent:
$\hat{f}^{e_1}_{S^{stab}} \approx \hat{f}^{e_2}_{S^{stab}} \approx \hat{f}^{e_3}_{S^{stab}}$

- StableMate is a method for selecting consistent and inconsistent functional dependencies across heterogeneous datasets.



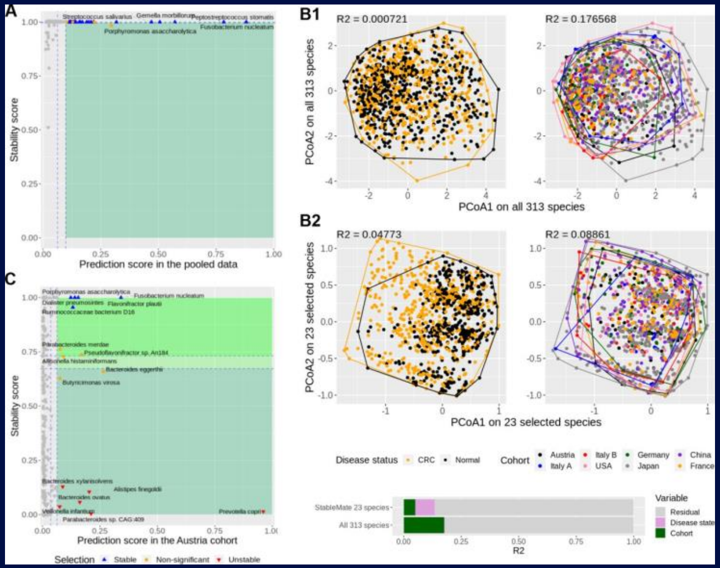
- StableMate outperform Stabilized Regression.



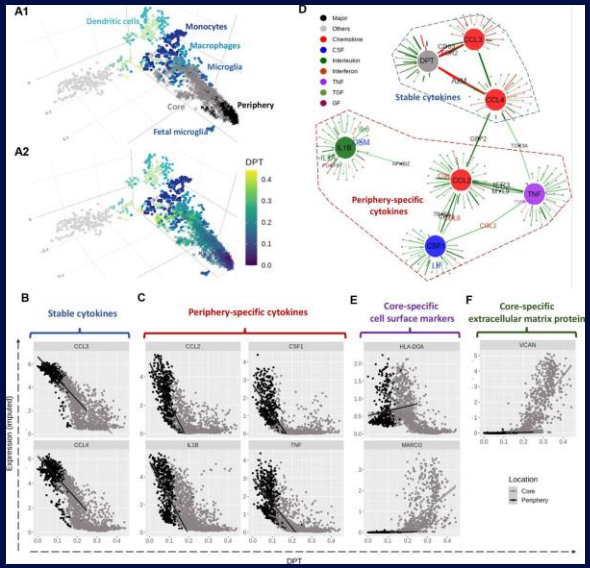
- StableMate makes interpretable inference of biological relationships via variable selection

What else?

Metagenomic data: predict colon cancer incidence with fecal microbiome abundance



scRNA-seq data: Characterizing cell identity transition of glioblastoma tumor-infiltrating microglia



Deng, Y., Mao, J., Choi, J., & Lê Cao, K. A. (2023). StableMate: a statistical method to select stable predictors in omics data. bioRxiv, 2023-09.