Oslo Bioinformatics Workshop Week 2023

Statistical principles in machine learning for small biomedical data

Manuela Zucknick

Department of Biostatistics, University of Oslo
manuela.zucknick@medisin.uio.no

December 11, 2023

Some of the figures in this presentation are taken from "Elements of Statistical Learning" (Springer, 2009) and "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors.

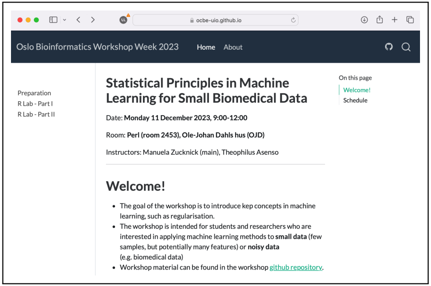
Schedule for today

Schedule

Time	Topic	Presenter
Now	Preparations	
9:00 - 10:00	(Supervised) machine learning with small data	Manuela Zucknick
	R lab 1	Manuela Zucknick
10:15 - 11:15	Overfitting, regularisation and all that	Manuela Zucknick
	R lab 2	Manuela Zucknick
11:30 - 12:00	Hierarchical models and structured penalties	Theophilus Asenso

Github, Workshop webpage and Posit Cloud project

- Github: `https://github.com/ocbe-uio/workshop-stat-highdim/`
- Workshop webpage: `https://ocbe-uio.github.io/workshop-stat-highdim/`
- Posit Cloud project: `https://posit.cloud/content/5131383/`



Some topics for this morning

Part 1

- What is supervised machine learning?
- What do we mean by small data?
- What can we do to improve ML with small data?
  - Restrict the model space → Regularisation
  - Borrow information → Include known structure in the model

Part 2

- Overfitting
- Variance vs bias
- Model selection, assessment & validation
- Prediction performance
- Resampling: Cross-validation

Further reading

James G, Witten D, Hastie T and Tibshirani R (2021), An Introduction to Statistical Learning with Applications in R, Springer, 2nd edition. `https://www.statlearning.com`

Hastie T, Tibshirani R, Friedman J (2009), The Elements of Statistical Learning, Springer, 2nd edition. `https://hastie.su.domains/ElemStatLearn/`

Holmes S, Huber W (2019), Modern Statistics for Modern Biology, Cambridge University Press. `https://www.huber.embl.de/msmb/`

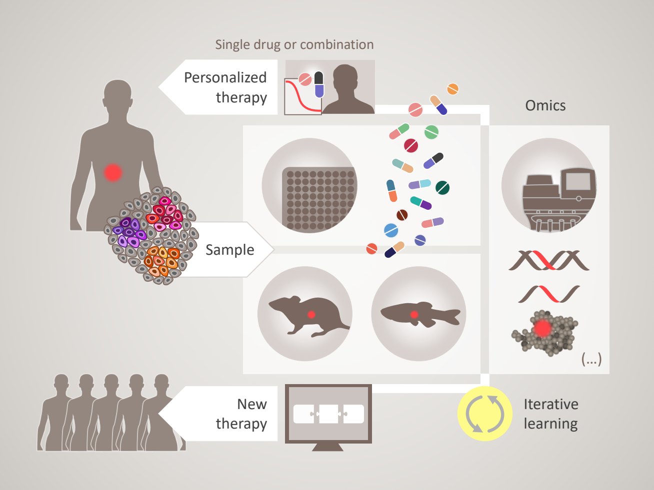
(some chapters on supervised/ unsupervised machine learning)

Introductory example:

Integrative omics for personalized cancer therapy

Personalized cancer therapy

...aims to find the best therapy for each patient based on data about the patient and tumor (e.g. genomic data).



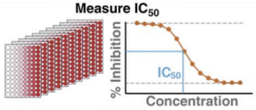
slide by Kjetil Taskén

Predict sensitivity to multiple drugs $\mathbf{Y}$ from multi-omics $\mathbf{X}$

$$\boxed{Y = XB + \epsilon}$$

- **Multivariate Y:**
  Drug dose response

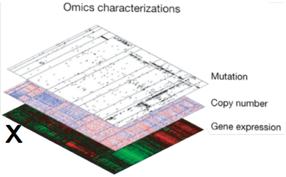
$$n \textit{ cell lines} \underbrace{}_{} \overbrace{\begin{bmatrix} | & & | \\ \mathbf{y}_{\bullet 1} & \cdots & \mathbf{y}_{\bullet m} \\ | & & | \end{bmatrix}}^{\textit{drug sensitivity}} = \mathbf{Y}$$



Source: Yang, et al. 2017

- **Heterogeneous X:**
  Integrative omics

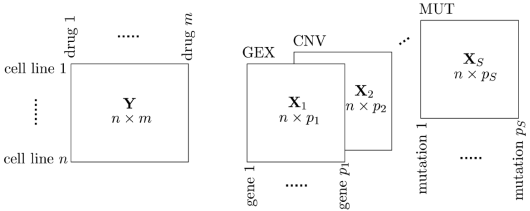
$$n \textit{ cell lines} \left[\overbrace{\mathbf{X}_1}^{\textit{gene expression}} \;\vdots\; \overbrace{\mathbf{X}_2}^{\textit{copy number}} \;\vdots\; \overbrace{\mathbf{X}_3}^{\textit{mutation}} \right] = \mathbf{X}$$



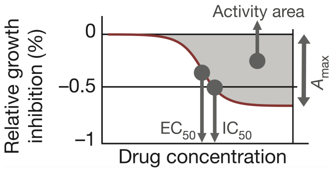
Source: TCGA, 2013

Challenges and opportunities (1)

- Small sample size
- Several types of input data **X**:
  E.g., gene expression, copy number, mutation
- Multivariate response **Y**



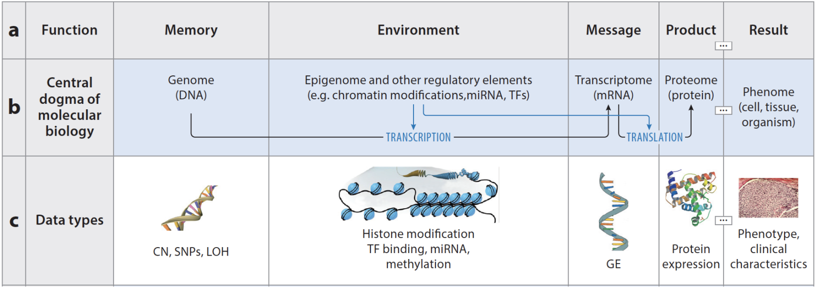
- Unclear how to define **Y**



Challenges and opportunities (2)

The data are highly structured:

1. In **Y**: relationships between drugs, e.g. due to similar chemical drug composition, same target genes/pathways
2. In **X**: relationships between molecular data sources

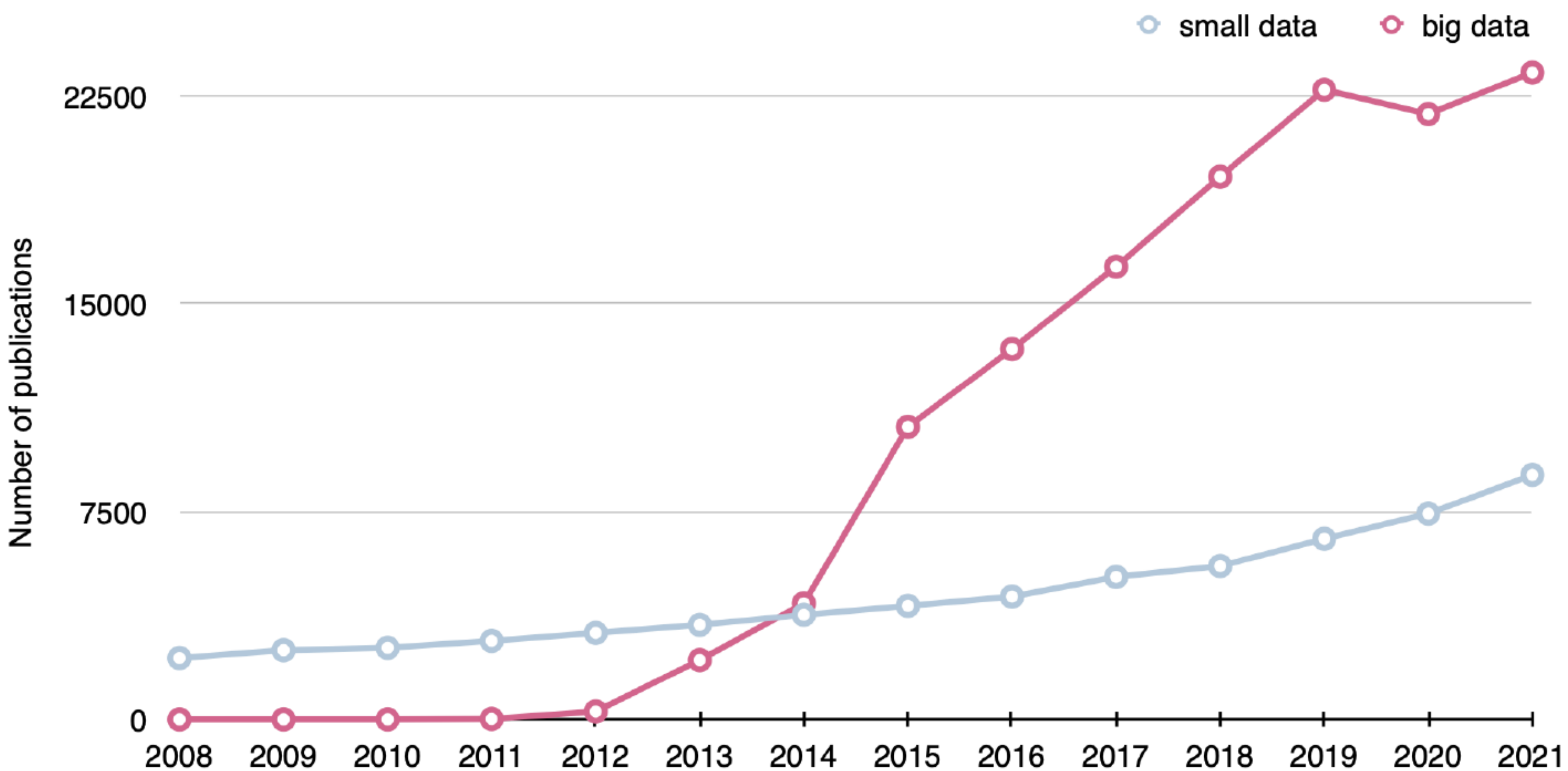


Ickstadt et al. (2018)

(Supervised) Machine Learning with Small Data

Manuela Zucknick (with slides from Maren Hackenberg)

Machine learning with small data



Machine learning with small data

- What do we mean by "small data"?
- Implications for machine learning?

- Aspects when building (multi-omic) machine learning predictors of drug response (e.g. Sammut et al. Nature 2022):

  1. Biological knowledge +
  2. Feature selection +
  3. Prioritisation of accessible data types +
  4. Machine learning algorithms

→ Develop ML methods that allow us to consider aspects 1 to 3.

What is supervised machine learning?

Supervised learning

refers to the task of inferring a functional relationship between **input data matrix $\mathbf{X}$** (e.g. gene expression array measurements) and **output data vector** Y (= response/ outcome).

The input data are used for predicting the outcome.

$$Y = f_\beta(\mathbf{X}) + \epsilon,$$

where ϵ captures measurement errors and other discrepancies, e.g. by $\epsilon \sim N(0, \sigma^2 I_n)$.

In classical statistics, this task is usually performed by (generalised) linear regression models.

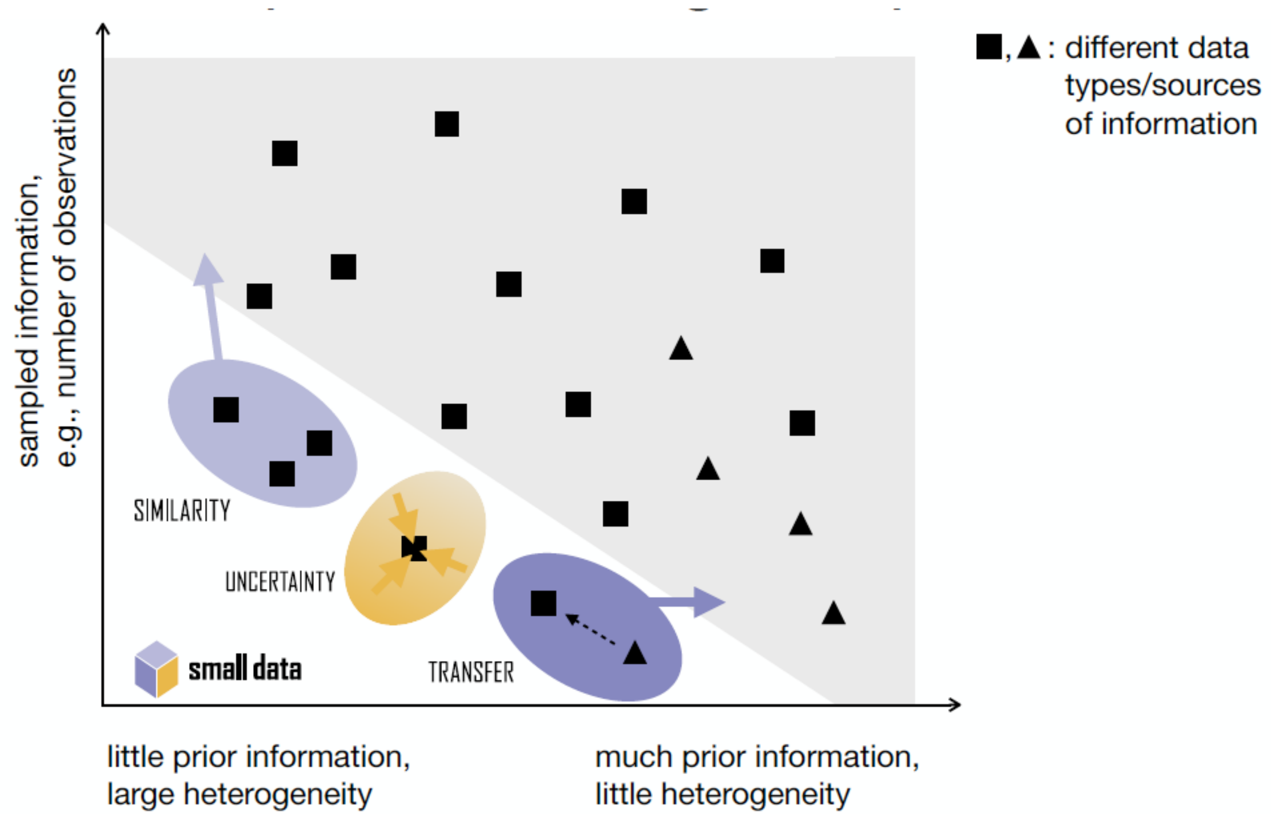
What do we mean by small data?

- Large p, small n (p>n)
- Potentially, more variables in the model than we have samples

- Classical statistical methods (e.g. linear regression) do not work:
- More parameters (e.g. regression coefficients) to estimate than observations for estimating them
- Even if all parameters can be estimated: Danger of over-fitting

- Example: Predict treatment response using gene expression data
  (n ~ 100, p ~20000)

What do we mean by small data?



What can we do?

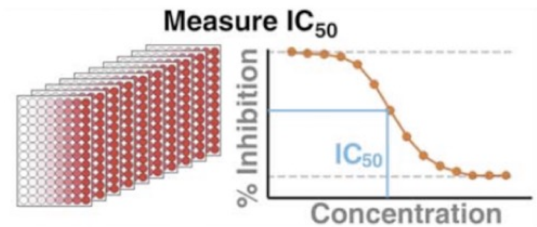
(1) Restrict the model space
(2) Borrow information across observations
(3) Increase sample size ☺

Predict sensitivity to multiple drugs **Y** from multi-omics **X**

$$\boxed{\mathbf{Y} = \mathbf{X}\mathbf{B} + \epsilon}$$

- **Multivariate Y:**
  Drug dose response

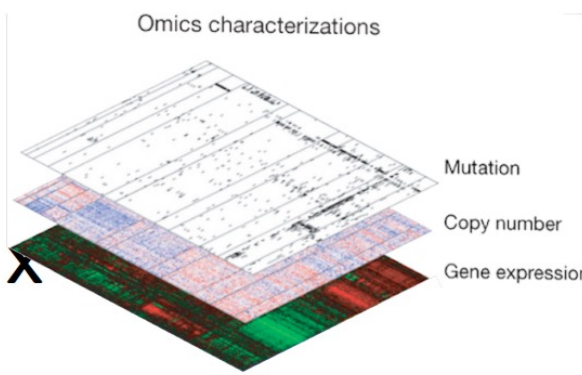
$$n \textit{ cell lines} \overbrace{\begin{bmatrix} | & & | \\ \mathbf{y}_{\bullet 1} & \cdots & \mathbf{y}_{\bullet m} \\ | & & | \end{bmatrix}}^{\textit{drug sensitivity}} = \mathbf{Y}$$



Source: Yang, et al. 2017

- **Heterogeneous X:**
  Integrative omics

$$n \textit{ cell lines} \left[\overbrace{\mathbf{X}_1}^{\textit{gene expression}} \;\vdots\; \overbrace{\mathbf{X}_2}^{\textit{copy number}} \;\vdots\; \overbrace{\mathbf{X}_3}^{\textit{mutation}} \right] = \mathbf{X}$$



Source: TCGA, 2013

What can we do?
(1) **Restrict the model space**

- (A) Careful feature engineering:
- Preselect variables by biological relevance
- Non-specific filtering, e.g. keep only variables with variance across observations larger than a threshold

- (B) Make use of known structure in the data (biological knowledge)

- (C) Use of regularisation techniques:
  - L1 and L2 penalisation
    - add a penalty term to the loss function to reduce the complexity of the model
    - Bayesian equivalents: restrictions on the prior (Bayesian variable selection)
  - Early stopping
    - train a model iteratively only until the validation error starts to decrease (boosting, neural networks)
  - Dropout regularisation
    - randomly dropping out neurons while training (neural networks) or
    - randomly dropping features when building a regression tree (random forest)

Penalised regression

- Standard regression cannot deal with $p >> n$:
  - The maximum-likelihood estimate $\hat{\beta} = \arg\max_\beta \ell(\beta)$ does not exist ($\ell =$ log-likelihood).

- **Solution:**
  Penalise the likelihood function by subtracting a penalty term and maximise penalised log-likelihood instead:

$$\hat{\beta} = \arg\max_{\beta}(\ell(\beta) - \lambda||\beta||)$$

  - λ is a penalty parameter,
  - $||\beta||$ represents the size of the regression coefficient vector,
  - The larger λ is chosen, the more the algorithm is encouraged to find a solution where $||\beta||$ is small $\rightarrow$ shrinkage.

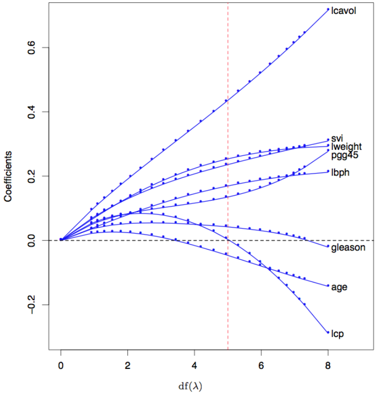
Penalised regression

- Examples for penalty terms:
  - Ridge regression (Hoerl and Kennard 1970):
    $\lambda||\beta|| := \lambda \sum_{g=1}^{p} \beta_g^2$ $\rightarrow$ **L_2 penalty**
  - Lasso regression (Tibshirani 1996):
    $\lambda||\beta|| := \lambda \sum_{g=1}^{p} |\beta_g|$ $\rightarrow$ **L_1 penalty**
  - Elastic net (Zou and Hastie 2005):
    Combination of both ridge and lasso penalty:
    $\lambda_1 \sum_{g=1}^{p} |\beta_g| + \lambda_2 \sum_{g=1}^{p} \beta_g^2$
- Advantage of lasso and elastic net:
  Both will produce a sparse solution, where only a few genes have estimate $\hat{\beta}_g \neq 0$.

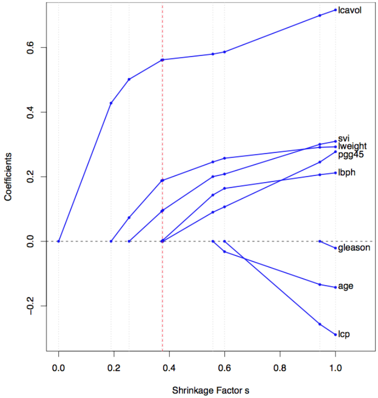
Penalised regression

Examples for coefficient paths relative to penality λ:

Ridge regresson



Lasso regression



Hastie et al. (2009), Figures 3.8 and 3.10

Penalised regression

- Ridge regression L_2: shrinks all coefficients to small, but non-zero values.
- Lasso regression L_1: shrinks some coefficients to exactly zero.
- Elastic net: mixture of the two: does shrink some coeffients to exactly zero. Keeps more variables if there is correlation.

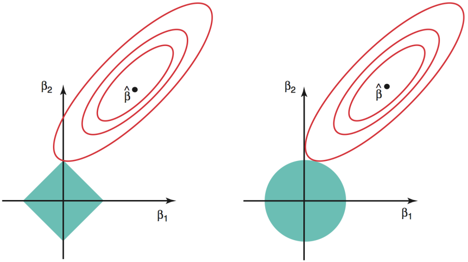


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.*

Different penalties for different types of data

Assume two data matrices **X** and **Z**:

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- Mandatory covariates: Do not penalise the parameters γ:

$$\ell_{\text{pen}}(\beta, \gamma) = \ell(\beta, \gamma) - \lambda||\beta||$$

e.g. with R packages `glmnet` or `penalized`

- Several types of molecular data sets:
Allow different penalties for β and γ:

$$\ell_{\text{pen}}(\beta, \gamma) = \ell(\beta, \gamma) - \lambda_{\beta}||\beta|| - \lambda_{\gamma}||\gamma||$$

e.g. with R packages `GRridge` (Van de Wiel *et al.*, 2016)
`http://www.few.vu.nl/~mavdwiel/grridge.html`)

Different penalties for different types of data

Assume two data matrices **X** and **Z**:

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- Several types of molecular data sets:
- Alternative: **Combine all data and use one penalty**, after scaling all features to unit variance to ensure that the data sources are treated equally.
- Example: Elastic Net models in Barretina et al. (2012)

What can we do?
(2) **Borrow information**

- Borrow information across observations in the data set
- If there is correlation, include this in your model
  - between variables (e.g. MRF prior for defining which variables to include together)
  - between samples (covariance matrix)

- Borrow information from external knowledge
  - E.g., use pathways to determine which genes should be included together

- Borrow information across data sets: transfer learning

Make use of external (biological) knowledge

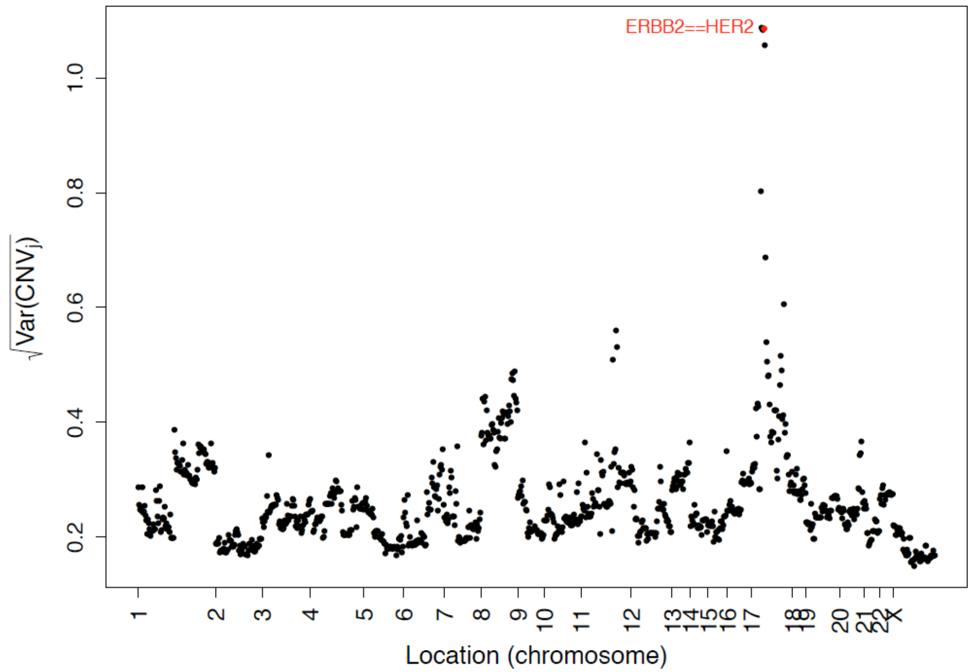
- (1) Use known relationships with one data source (CNV) to guide the variable selection in another (gene expression)

- (2) Combine the data-driven ML approach with knowledge-driven mechanistic modelling

- (3) Make use of correlations in the data
  - between input variables - to restrict the model space
  - between response variables - to borrow information

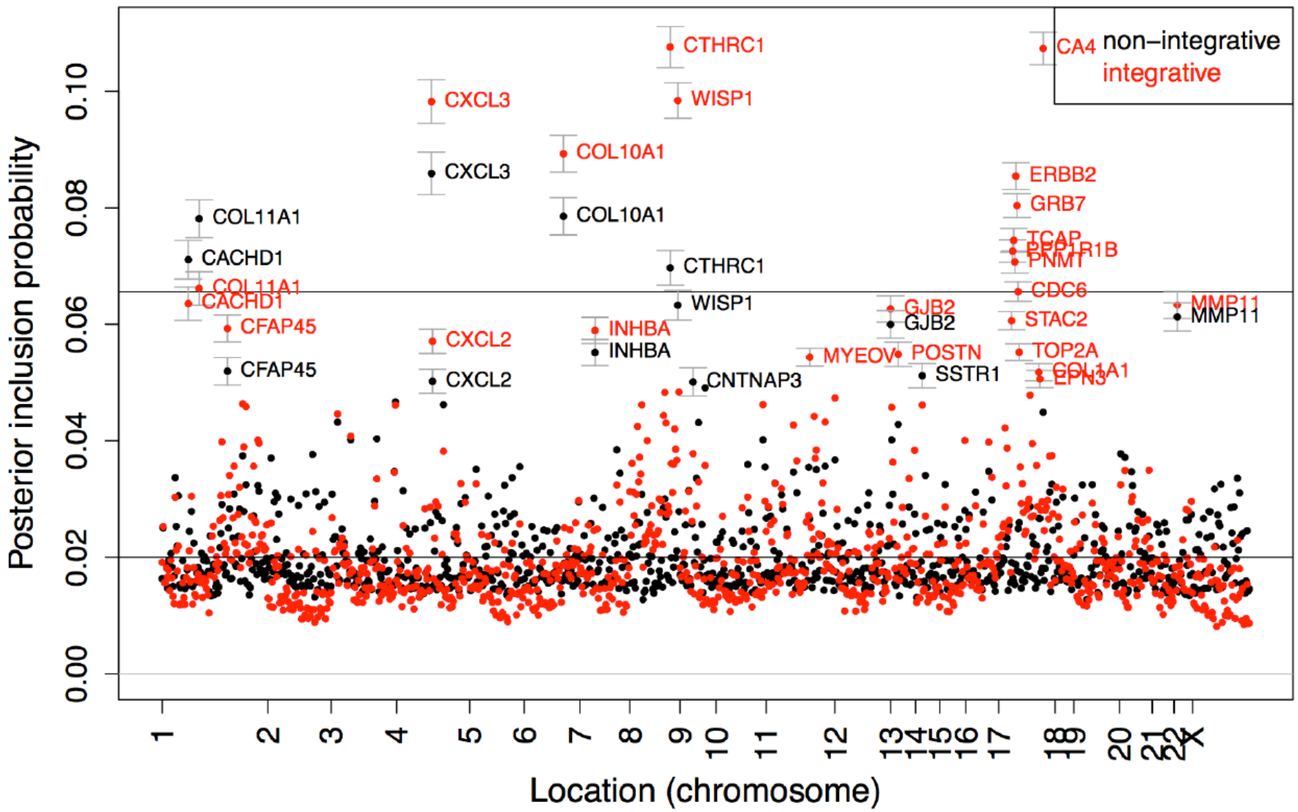
(1) Use known relationships with one data source to guide the variable selection in another

Std. dev. of CNV data of HER2-pos. breast cancer and healthy tissue samples



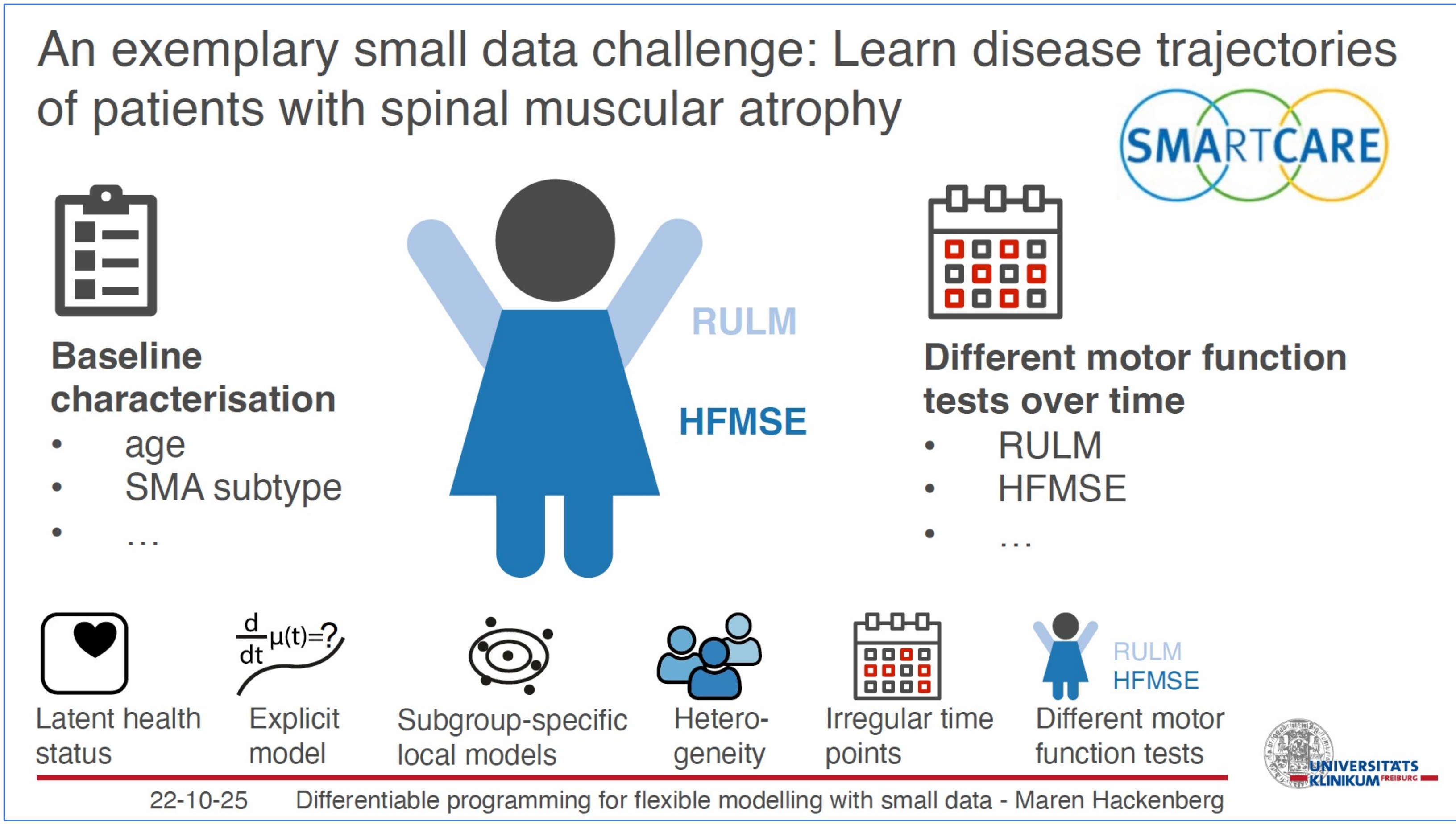
Idea: Use CNV information to weigh prior inclusion probabilities of gene expression variables in Bayesian variable selection

(1) Use known relationships with one data source to guide the variable selection in another

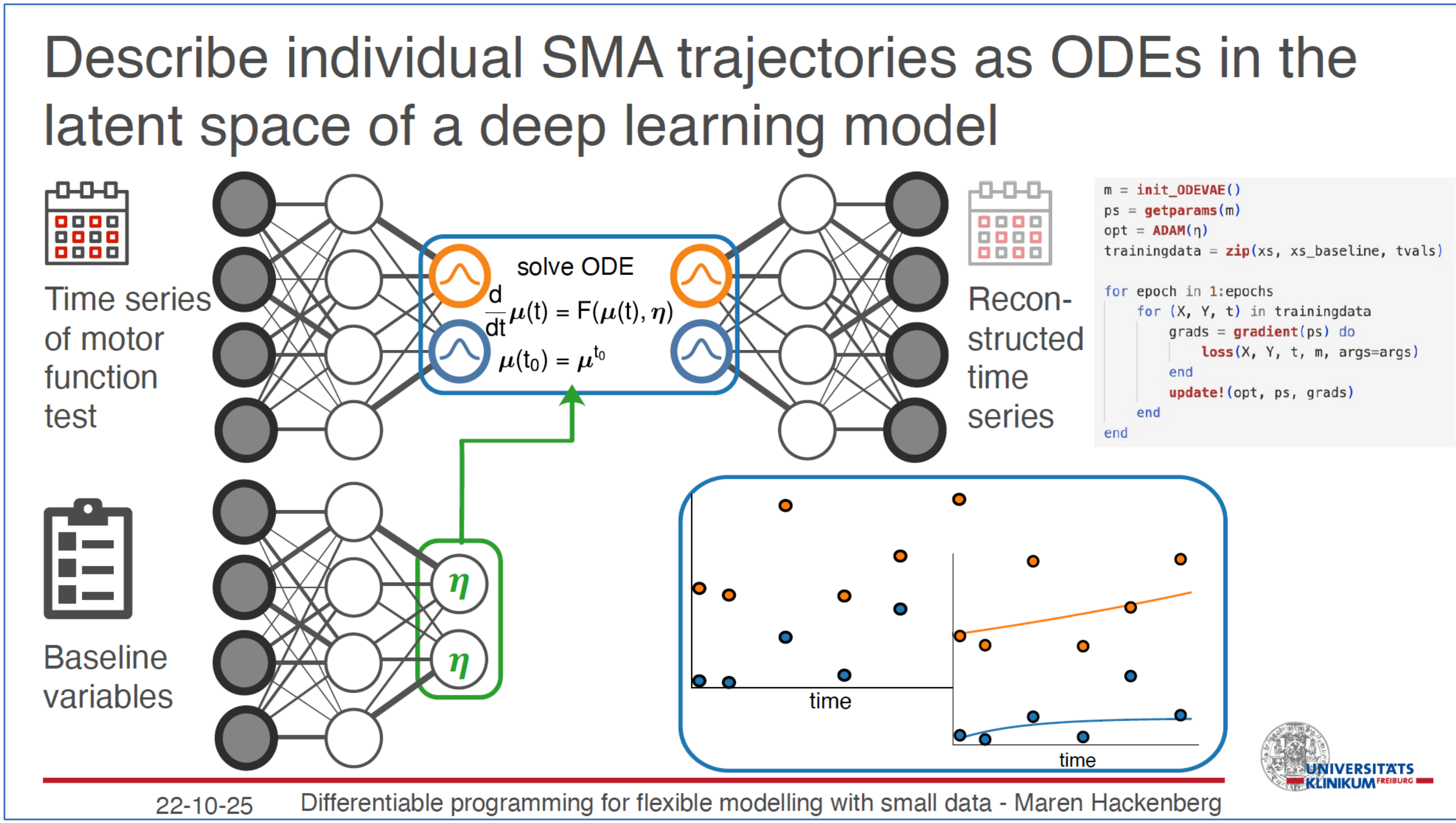


HER2 (= ERBB2) only selected in integrative analysis

(2) Combine the data-driven ML approach with knowledge-driven mechanistic modelling



(2) Combine the data-driven ML approach with knowledge-driven mechanistic modelling

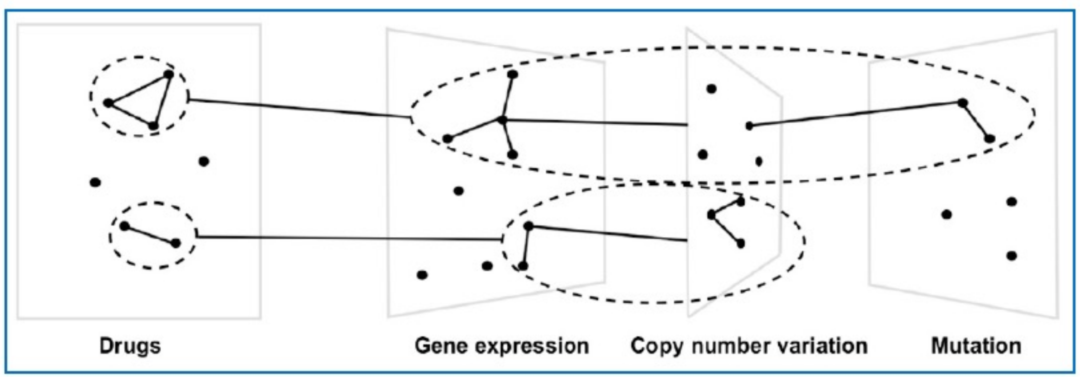


(3) Make use of correlations in the data:
between input variables - to restrict the model space

BayesSUR: An R Package for High-Dimensional Multivariate Bayesian Variable and Covariance Selection in Linear Regression

Zhi Zhao, Marco Banterle, Leonardo Bottolo, Sylvia Richardson, Alex Lewin, Manuela Zucknick — Vol. 100, Issue 11

Paper | R package (BayesSUR) | R replication code



(3) Make use of correlations in the data:
between input variables - to restrict the model space

- **Formulation of the model:**

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$
$$\text{vec}(\mathbf{U}) \sim \mathcal{N}(\mathbf{0},\ C \otimes \mathbb{I}_n)$$

$$\beta_{kj}|\gamma_{kj}, w \sim \gamma_{kj}\mathcal{N}(0,\ w) + (1-\gamma_{kj})\delta_0(\beta_{kj})$$

for each element β_{kj} in $\mathbf{B}$.

- $\mathbf{Y}$ $n \times m$ matrix of outcomes with $m \times m$ covariance matrix C,
- $\mathbf{X}$ $n \times p$ matrix of predictors for all outcomes,
- $\mathbf{B}$ $p \times m$ matrix of regression coefficients,
- $\mathbf{\Gamma} = \{\gamma_{jk}\}$ $p \times m$ binary indicator matrix for variable selection.

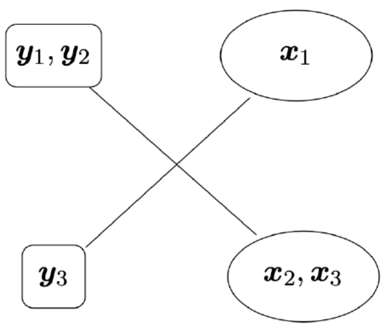
	$\gamma_{jk} \sim$ Bernoulli	$\gamma_{jk} \sim$ Hotspot	$\boldsymbol{\gamma} \sim$ MRF
$C \sim$ indep	HRR-B	HRR-H	HRR-M
$C \sim \mathcal{IW}$	dSUR-B	dSUR-H	dSUR-M
$C \sim \mathcal{HIW}_{\mathcal{G}}$	SSUR-B	SSUR-H	SSUR-M

(3) Make use of correlations in the data:
between input variables - to restrict the model space

MRF prior for pharmacogenomics

$$f(\gamma | d, e, G) \propto \exp\{d\mathbf{1}^\top \gamma + e \cdot \gamma^\top G \gamma\}$$

- d controls the model sparsity,
- e the strength of relations between responses and predictors,
- G is an adjacency matrix of the structure prior knowledge.

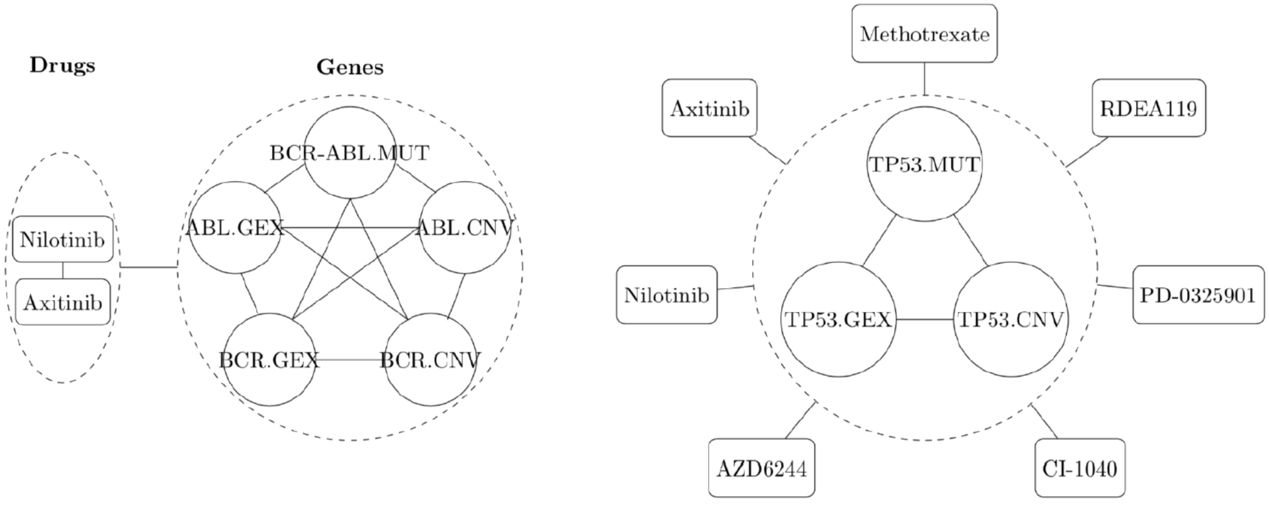


$$G = \begin{array}{c} \\ \gamma_{11} \\ \gamma_{21} \\ \gamma_{31} \\ \gamma_{12} \\ \gamma_{22} \\ \gamma_{32} \\ \gamma_{13} \\ \gamma_{23} \\ \gamma_{33} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \gamma_{11} & \gamma_{21} & \gamma_{31} & \gamma_{12} & \gamma_{22} & \gamma_{32} & \gamma_{13} & \gamma_{23} & \gamma_{33} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

(3) Make use of correlations in the data:
between input variables - to restrict the model space

Application to Genomics of Drug Sensitivity in Cancer data

- Same data as before, but now only use $m = 7$ cancer drugs



(3) Make use of correlations in the data:
between input variables - to restrict the model space

Results ($\boldsymbol{\Gamma}$): Which covariates are important?

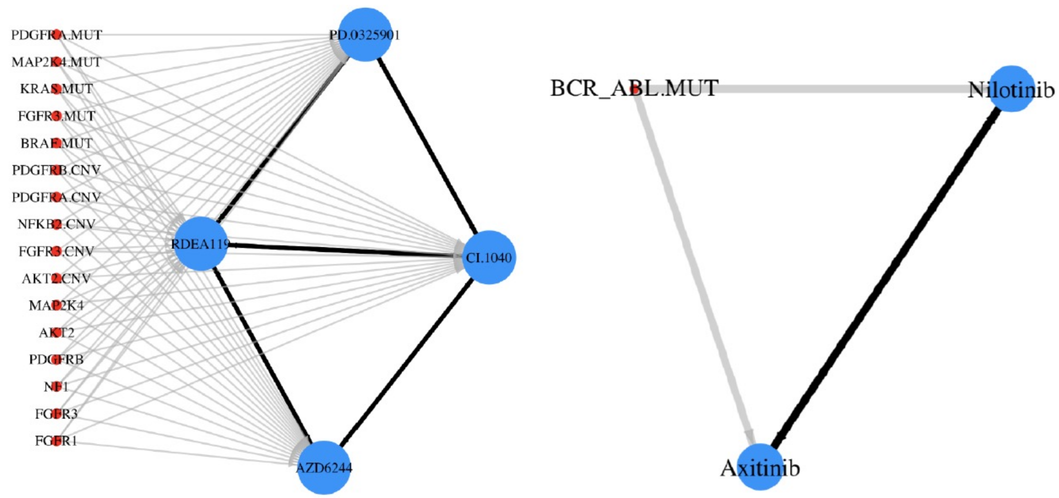


Fig: Important covariates related to the MEK inhibitors (left) or Bcr-Abl inhibitors (right) based on threshold for posterior marginal inclusion probabilities (mPIP > 0.5).

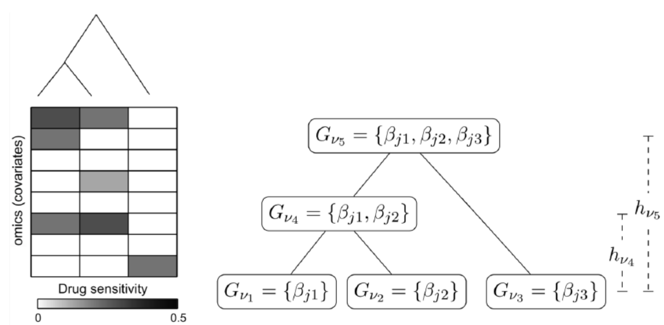
(3) Make use of correlations in the data:
between response variables - to borrow information

(Multi-response) Tree-guided group lasso (Kim & Xing 2012)

- Include dependencies between columns of **Y** in a group lasso
- Extension to IPF-tree lasso

$$\textbf{Tree lasso:}\ \text{pen}(\mathbf{B}) = \lambda \sum_{j=1}^{p} \sum_{\nu \in \{V_{\text{int}}, V_{\text{leaf}}\}} \omega_\nu \|\boldsymbol{\beta}_j^{G_\nu}\|_{\ell_2}$$

$$\textbf{IPF-tree lasso:}\ \text{pen}(\mathbf{B}) = \sum_{s} \lambda_s \left(\sum_{j_s} \sum_{\nu \in \{V_{\text{int}}, V_{\text{leaf}}\}} \omega_\nu \|\boldsymbol{\beta}_{j_s}^{G_\nu}\|_{\ell_2} \right)$$



(a) (b)

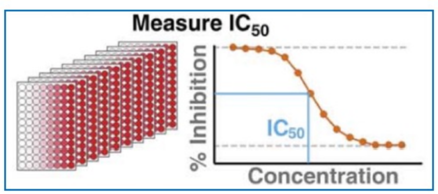
(3) Make use of correlations in the data:
between response variables - to borrow information

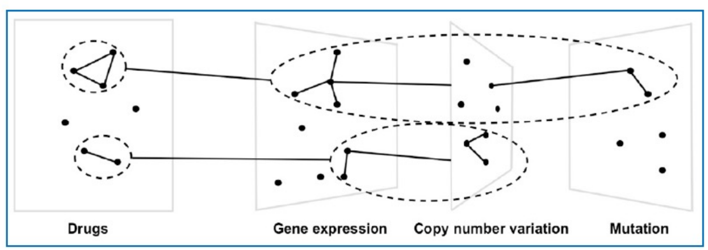
Drug screens for precision cancer medicine:
How to predict the drugs’ effect with data on drugs and tumour?

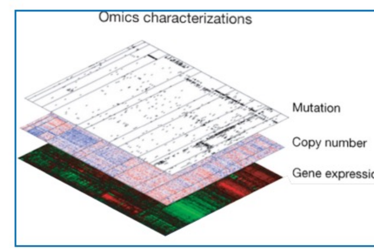
Original Article | Open Access

Structured penalized regression for drug sensitivity prediction

Zhi Zhao, Manuela Zucknick





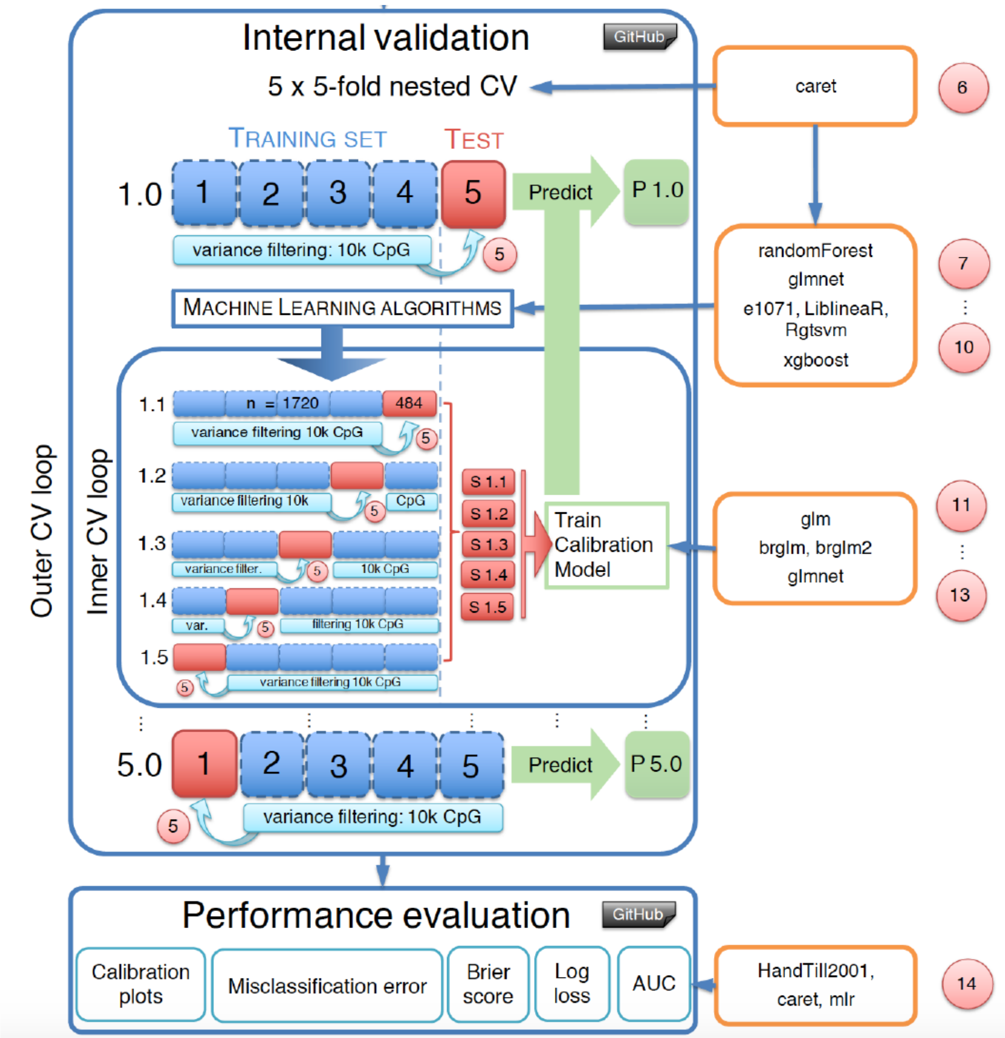


Model validation is crucial with small data

- Careful and correctly set up the model validation framework is even more important with small data
- To avoid over-fitting when selecting tuning parameters or selecting models
- To avoid being too optimistic when estimating prediction error

- Learning curve: How many samples are needed in the training set to approach optimal model training?
- Nested cross-validation
- .632+ bootstrapping vs .632+ subsampling

Nested cross-validation



from: Maros et al. (2020)

Learning curve:
How many samples are needed in the training data?