

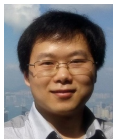
# Machine learning with small data: Examples from pharmacogenomic screens for personalised medicine

Manuela Zucknick

Oslo Centre for Biostatistics and Epidemiology, University of Oslo  
`manuela.zucknick@medisin.uio.no`

Hans van Houwelingen Symposium, Utrecht, 15-06-2023

Zhi Zhao



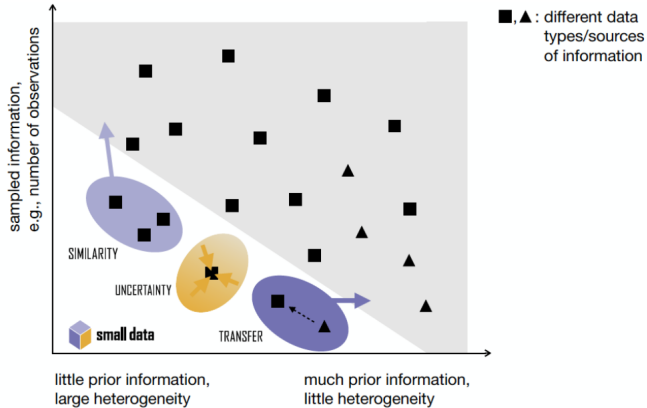
Theo Asenso

## 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.

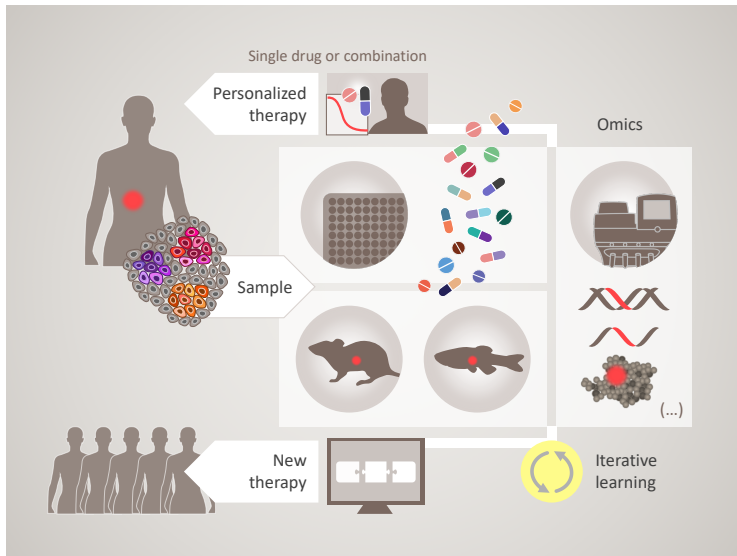
# What are small data in ML and what can we do?



by Maren Hackenberg

1. Increase sample size :-)
2. Borrow information across observations (incl. between data sets)
3. Restrict the model space

# Pharmacogenomic screens for personalised medicine





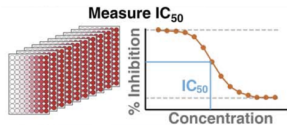
# Predict sensitivity to multiple drugs $Y$ from multi-omics $X$

$$Y = XB + \epsilon$$

- Drug dose response

*drug sensitivity*

$$n \text{ cell lines} \begin{bmatrix} | & & | \\ \mathbf{y}_{\bullet 1} & \dots & \mathbf{y}_{\bullet m} \\ | & & | \end{bmatrix} = \mathbf{Y}$$

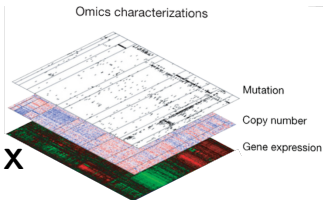


Source: Yang, et al. 2017

- Integrative omics

*gene expression copy number mutation*

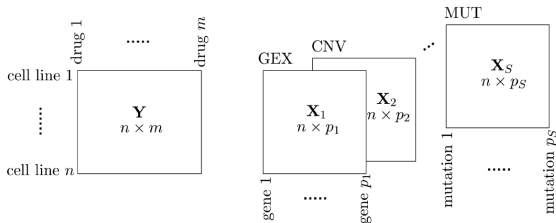
$$n \text{ cell lines} \begin{bmatrix} \mathbf{X}_1 & | & \mathbf{X}_2 & | & \mathbf{X}_3 \\ \vdots & & \vdots & & \vdots \end{bmatrix} = \mathbf{X}$$



Source: TCGA, 2013

# Challenges and opportunities (1)


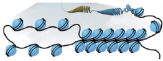



- Small sample size
- Heterogeneous populations (tumours in different tissues)
- Several types of input data  $\mathbf{X}$ :  
E.g., gene expression, copy number, mutation
- Multivariate response  $\mathbf{Y}$



## 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

a	Function	Memory	Environment	Message	Product	Result
b	Central dogma of molecular biology	Genome (DNA)	Epigenome and other regulatory elements (e.g. chromatin modifications, miRNA, TFs)	Transcriptome (mRNA)	Proteome (protein)	Phenome (cell, tissue, organism)
c	Data types	 CN, SNPs, LOH	 Histone modification TF binding, miRNA, methylation	 GE	 Protein expression	 Phenotype, clinical characteristics

Ickstadt et al. (2018)

## Drug screens for precision cancer medicine: Predict sensitivity to multiple drugs $Y$ from multi-omics $X$

### High-dimensional multi-response regression with variable selection with

- Correlated responses (drugs with same target or similar mechanism of action)
- Non-i.i.d. observations (cell lines representing different cancer types)
- Several related input data sets (multi-omics)

1. Penalised regressions – with structured (tailored) penalty terms
2. Bayesian variable selection models – with structured selection priors

## Structured penalized regression for drug sensitivity prediction

ROYAL  
STATISTICAL  
SOCIETY  
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society  
**Applied Statistics**  
Series C

Original Article |  Open Access |   

### Structured penalized regression for drug sensitivity prediction

Zhi Zhao  Manuela Zucknick

- Different penalties for different data sources
- Group lasso for the coefficients corresponding to correlated responses (tree structure or any overlapping groups)
- <https://github.com/zhizuio/IPFStructPenalty> and <https://github.com/zhizuio/mixlasso>

## Multi-response penalised linear regression

Objective function:

$$\min_{\beta_0, \mathbf{B}} \left\{ \frac{1}{2mn} \|\mathbf{Y} - \mathbf{1}_n \beta_0^T - \mathbf{X}\mathbf{B}\|_F^2 + \text{pen}(\mathbf{B}) \right\}$$

Standard penalised regression assigns the same penalty to all data sources, and treats columns of  $\mathbf{Y}$  as independent:

- **Lasso:**  $\text{pen}(\mathbf{B}) = \lambda \|\mathbf{B}\|_{\ell_1}$
- **Elastic-net:**  $\text{pen}(\mathbf{B}) = \lambda(\alpha \|\mathbf{B}\|_{\ell_1} + \frac{1}{2}(1 - \alpha) \|\mathbf{B}\|_{\ell_2}^2)$

## Integrative LASSO with Penalty Factors (Boulesteix et al. 2017)

- Allow different penalties for different data sources
- Extensions of IPF-lasso to multi-response regression and to the elastic net

$$\text{IPF-lasso: } \text{pen}(\mathbf{B}) = \sum_s \lambda_s \|\mathbf{B}_s\|_{\ell_1}$$

$$\text{IPF-sEN: } \text{pen}(\mathbf{B}) = \sum_s \lambda_s \left( \alpha \|\mathbf{B}_s\|_{\ell_1} + \frac{1}{2} (1 - \alpha) \|\mathbf{B}_s\|_{\ell_2}^2 \right)$$

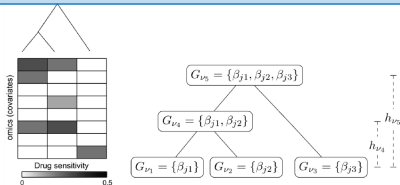
$$\text{IPF-EN: } \text{pen}(\mathbf{B}) = \sum_s \lambda_s \left( \alpha_s \|\mathbf{B}_s\|_{\ell_1} + \frac{1}{2} (1 - \alpha_s) \|\mathbf{B}_s\|_{\ell_2}^2 \right)$$

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

- Include dependencies between columns of  $\mathbf{Y}$  in a group lasso
- Extension to IPF-tree lasso

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

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





# Drug screens for precision cancer medicine: Predict sensitivity to multiple drugs Y from multi-omics X

iScience

CellPress  
OPEN ACCESS

Article

Tissue-specific identification of multi-omics features for pan-cancer drug response prediction

Zhi Zhao,<sup>1,2</sup> Shixiong Wang,<sup>1</sup> Manuela Zucknick,<sup>2,\*</sup> and Tero Aittokallio<sup>1,2,3,4,\*</sup>

- Include random effects, e.g. for different cancer sub-types (*What to do with V?*)
- Improved optimization of penalty parameters  
(*Smoothing proximal gradient with a proxy for the random effect covariance V*)
- Allows for missing values in the responses
- Drug Set Enrichment Analysis (R package “EnrichIntersect”)

# Drug screens for precision cancer medicine: Predict sensitivity to multiple drugs $Y$ from multi-omics $X$

An ADMM approach for multi-response regression with overlapping groups and interaction effects

Theophilus Quachie Asenso<sup>a\*</sup>, Manuela Zucknick<sup>a</sup>

<sup>a</sup> Oslo Center for Epidemiology and Biostatistics, Institute of Basic Medical Sciences, University of Oslo

- Pliable lasso for interactions, e.g. with the tissue types (Tibshirani & Friedman 2020)
- Incorporate pliable lasso in tree-lasso type multi-response regression
- Penalty parameter optimisation with ADMM, alternating direction method of multipliers (Boyd et al, 2011)

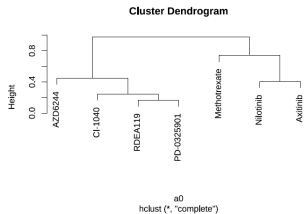
$$\hat{Y}_d = \beta_{0d}\mathbf{1} + Z\theta_{0d} + \sum_j X_j\beta_{jd}\mathbf{1} + \sum_j (X_j \circ Z)\theta_{jd}, \quad (24)$$

Objective function for a general multi-response pliable lasso:

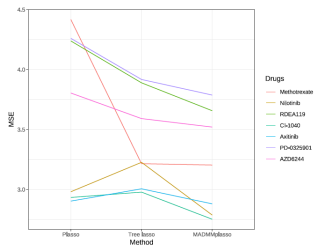
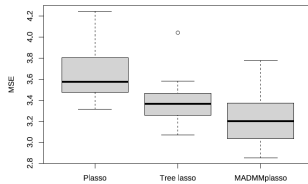
$$\min_{B \in \mathbb{R}^{p \times (1+K) \times D}} \frac{1}{2N} \|Y - \hat{Y}\|_F^2 + \sum_{d=1}^D \left[ (1 - \alpha)\lambda \sum_{j=1}^p (\|B_{jd}\|_2 + \|B_{j(-1)d}\|_2) + \alpha\lambda \sum_{j=1}^p \|B_{j(-1)d}\|_1 \right]. \quad (27)$$

Objective function for multi-response pliable lasso with tree-guided structure:

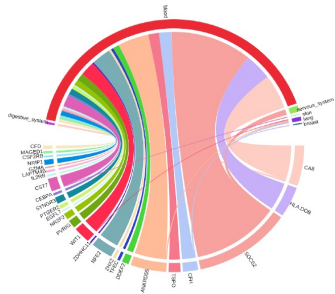
$$\min_{B \in \mathbb{R}^{p \times (1+K) \times D}} \frac{1}{2N} \|Y - \hat{Y}\|_F^2 + \lambda_1 \sum_{j=1}^p \sum_{m \in M_{\text{int}}} w_m \|B_j^{\mathcal{G}_m}\|_2 + \lambda_2 \sum_{j=1}^p \sum_{m \in M_{\text{leaf}}} w_m \|B_j^{\mathcal{G}_m}\|_2 + \sum_d \left[ (1 - \alpha)\lambda_3 \sum_{j=1}^p (\|B_{jd}\|_2 + \|B_{j(-1)d}\|_2) + \alpha\lambda_3 \sum_{j=1}^p \|B_{j(-1)d}\|_1 \right]. \quad (28)$$



(a)



(c)



Zhao et al. (Journal of Statistical Software, 2021).

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

(BayesSUR = Bayesian Seemingly Unrelated Regression)

`https://CRAN.R-project.org/package=BayesSUR`.  
R package version 2.1-3.

Joint work with Zhi Zhao, Marco Banterle, Alex Lewin, Leonardo Bottolo, Sylvia Richardson.

# Bayesian seemingly unrelated regression for variable and covariance selection (Bottolo et al. 2021; Zhao et al. 2021)

- **Matrix formulation of the model:**

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U},$$

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

- $\mathbf{Y}$   $n \times m$  matrix of outcomes with  $m \times m$  covariance matrix  $\mathbf{C}$ ,
  - $\mathbf{X}$   $n \times p$  matrix of predictors for all outcomes,
  - $\mathbf{B}$   $p \times m$  matrix of regression coefficients.
- In addition: Variable selection indicator matrix  $\mathbf{\Gamma}$

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

## We can introduce structure/ sparsity in two places:

### 1. Prior for variable selection indicator $\gamma$ .

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

- Binary latent indicator matrix  $\Gamma = \{\gamma_{jk}\}$  for variable selection
- Spike-and-slab prior on vectorised  $\beta = \text{vec}(\mathbf{B})$  and  $\gamma = \text{vec}(\Gamma)$
- and  $w \sim \mathcal{IG}(a_w, b_w)$  and  $\delta_0(\cdot)$  is the Dirac delta function.

### 2. Prior for covariance matrix: $C \sim \mathcal{HIW}_{\mathcal{G}}$ with further hyper-prior on graph $\mathcal{G}$ (Bottolo et al. 2021)

- Graph  $\mathcal{G}$  encodes conditional dependence between responses. Sparse  $\mathcal{G}$  implies sparse precision matrix  $C^{-1}$ .
- Sparse Seemingly Unrelated Regression (SSUR)

## Options for covariance matrix structure (Bottolo et al. 2021)

- **Diagonal:** Hierarchical Related Regression (Richardson et al. 2011)

$$C = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \sigma_s^2 \end{pmatrix}$$

Independent inverse Gamma priors  $\sigma_k^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$

- **Dense:** dense Seemingly Unrelated Regressions (dSUR)  
Inverse Wishart prior  $C \sim \mathcal{IW}(\nu, \tau \mathbb{I}_s)$
- **Sparse:** Sparse Seemingly Unrelated Regressions (SSUR)  
Hyper-inverse Wishart prior  $C \sim \mathcal{HIW}_G(\nu, \tau \mathbb{I}_s)$



## Options for variable selection ( $j = 1, \dots, p; k = 1, \dots, m$ )

- Independent Bernoulli prior:

$$\gamma_{jk} | \omega_{jk} \sim \text{Ber}(\omega_{jk}), \quad \text{with } \omega_j \sim \text{Beta}(a_\omega, b_\omega).$$

- Hotspot prior: (Bottolo et al. 2021)

$$\begin{aligned} \gamma_{jk} | \omega_{jk} &\sim \text{Ber}(\omega_{jk}), \quad \text{with } \omega_{jk} = o_k \times \pi_j, \\ o_k &\sim \text{Beta}(a_o, b_o), \pi_j \sim \text{Gamma}(a_\pi, b_\pi). \end{aligned}$$

- Markov Random Field (MRF) prior: (e.g. Chekouo et al. 2015.)

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



## Multivariate Bayesian structured variable selection for pharmacogenomic studies

Zhao, Banterle, Lewin, Zucknick  
(arXiv.2101.05899, update coming soon)

## SSUR model with MRF prior and random intercepts

$$\mathbf{Y} = \mathbf{Z}\mathbf{B}_0 + \mathbf{X}\mathbf{B} + \mathbf{U},$$

$$\beta_{0,tj}|w_0 \sim \mathcal{N}(0, w_0),$$

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

$$w_0 \sim \mathcal{IG}(a_{w_0}, b_{w_0}),$$

$$w \sim \mathcal{IG}(a_w, b_w),$$

$$\gamma|d, e, G \propto \exp\{d\mathbb{1}^\top \gamma + e\gamma^\top G\gamma\},$$

$$\text{vec}\{\mathbf{U}\} \sim \mathcal{N}(\mathbf{0}, \Psi \otimes \mathbb{I}_n),$$

$$\Psi \sim \mathcal{HIW}_G(\nu, \tau\mathbb{I}_m),$$

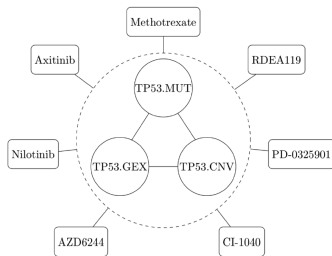
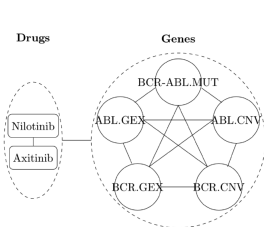
$$\tau \sim \mathcal{Gamma}(a_\tau, b_\tau),$$

# Application to Genomics of Drug Sensitivity in Cancer data

(Garnett et al., 2012)

- Large-scale pharmacogenomic study with  $n=498$  cell lines and  $m=97$  drugs. We illustrate the model with  $m = 7$  drugs.
- Outcome data:  $\log(IC_{50})$  from dose-response experiments
- Random draws of 80% cell lines as training data and 20% as validation data.
- Input data:
  - cancer type ( $p_0 = 13$ ) → included as **random intercept effects**,
  - mRNA expression ( $p_1 = 2602$ ),
  - copy numbers ( $p_2 = 426$ ) and
  - DNA mutations ( $p_3 = 68$ )

- **MRF prior to include structure**, with edges between:
  - **drugs**: Group1 ("RDEA119", "PD-0325901", "CI-1040" and "AZD6244"); Group2 ("Nilotinib", "Axitinib")
  - **genes** in MAPK/ERK pathway (targets of Group1)
  - **genes** in the Bcr-Abl fusion gene (targets of Group2)
  - **genes** of MAPK/ERK pathway and Group1
  - **genes** of the Bcr-Abl fusion gene and Group2
  - **each gene feature** in different data sources (GEX, CNV, MUT)



$$\underbrace{G_y}_{2 \text{ drugs}} \otimes \underbrace{G_x}_{5 \text{ features}} - \mathbb{I}_{10} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} - \mathbb{I}_{10}$$

$$\underbrace{G_y}_{7 \text{ drugs}} \otimes \underbrace{G_x}_{3 \text{ features}} - \mathbb{I}_{21} = \mathbb{I}_7 \otimes \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} - \mathbb{I}_{21}$$

# Results ( $\Gamma$ ): Variable selection more stable with MRF prior

TABLE 4  
*GDSC data application: Number of identified genomic features corresponding to each drug by the SSUR-Ber and SSUR-MRF models.*

	Nilotinib	Axitinib	RDEA119	PD-0325901	CI-1040	AZD6244	Methotrexate
<b>SSUR-Ber</b>							
Feature set I	5	5	2	3	1	0	3
Feature set II	1	2	3	1	1	2	2
Feature set III	8	11	8	4	8	10	8
<b>SSUR-MRF</b>							
Feature set I	1	2	42	41	40	40	0
Feature set II	9	10	56	56	56	57	9
Feature set III	39	38	87	86	86	89	41

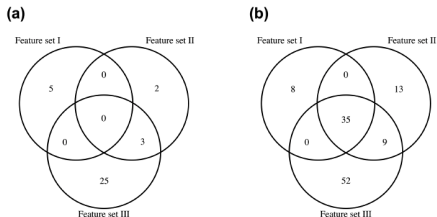


Fig 8: GDSC data application: A Venn diagram for the numbers of identified features for the MAPK inhibitors by SSUR-Ber (panel (a)) and SSUR-MRF (panel (b)) models and overlaps between the models fitted with feature sets I, II, and III.

# Results ( $\mathcal{G}$ ): Residual covariance structure between drugs

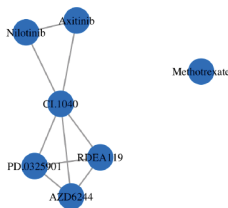


Fig 7: GDSC data application: Estimated residual structure between the seven drugs by the SSUR-MRF model based on features set III with  $\hat{\mathcal{G}}$  thresholded at 0.5.

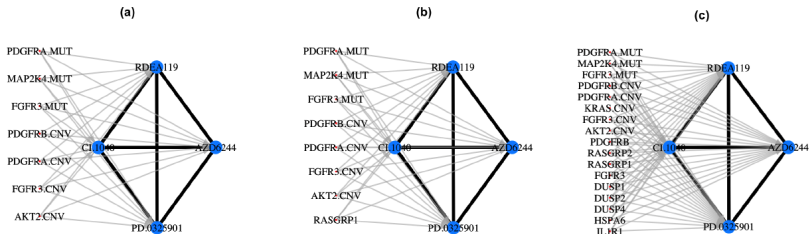


Fig 9: GDSC data application: Estimated network between the MAPK inhibitors and identified target genes based on  $\hat{\mathcal{G}}$  and  $\hat{\mathcal{I}}$  thresholded at 0.5 by SSUR-MRF corresponding to feature set I, II and III respectively.



# Simulation setup

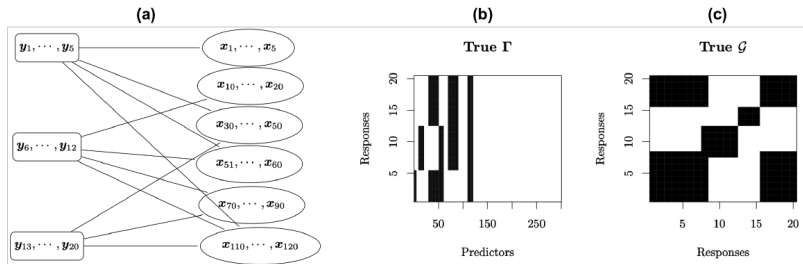


Fig 2: Simulation scenarios: True relationships between response variables and predictors. (a) Network structure between  $\mathbf{Y}$  and  $\mathbf{X}$ ; (b) latent indicator variable  $\Gamma$  for the associations between  $\mathbf{Y}$  and  $\mathbf{X}$  in the SUR model; (c) additional structure  $\mathcal{G}$  between response variables not explained by  $\mathbf{XB}$ . Black indicates a true relation between the response variables and predictors.

# Simulation: MRF prior can improve variable selection

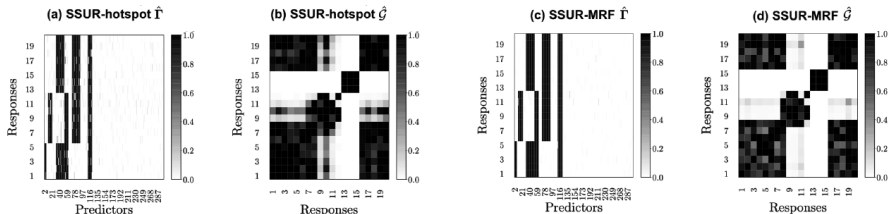


Fig 3: Results for simulation scenario 1: Posterior mean of  $\Gamma$  and  $\mathcal{G}$  by models SSUR-hotspot (panels (a) and (b)) and SSUR-MRF (panels (c) and (d))

TABLE 1

*Results for simulation scenario 1: Accuracy of variable selection and prediction performance of models SSUR-hotspot and SSUR-MRF prior*

	accuracy	sensitivity	specificity	RMSE	RMSPE
SSUR-hotspot	0.988	0.936	0.999	0.800	0.693
SSUR-MRF	0.989	0.998	0.986	0.643	0.412

# Simulation: Results robust to mis-specified MRF prior

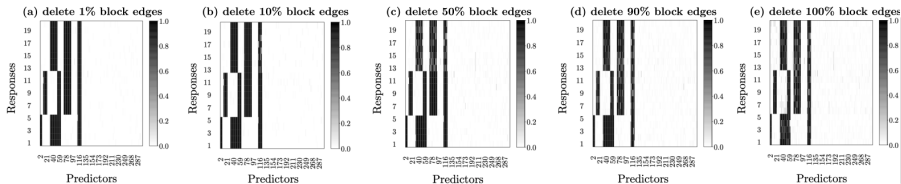


Fig 4: Results for simulation scenario 1: Sensitivity analysis for case 2, i.e. when blocks of edges are deleted (i.e. delete edges non-uniformly).

# Simulation: Random intercepts for e.g. tissue effects

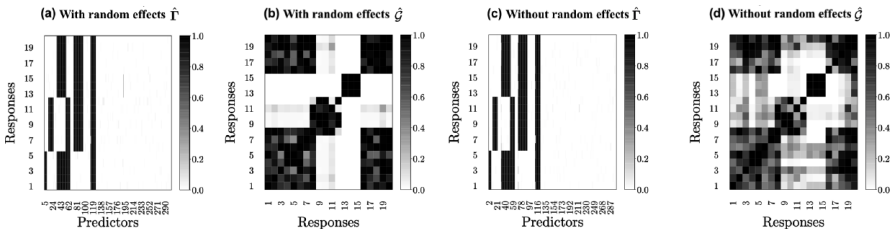


Fig 5: Results for simulation scenario 2: Posterior mean of  $\Gamma$  and  $\mathcal{G}$  by the SSUR-MRF with random effects based on the simulated data from scenario 2

## Thank you!

- **BigInsight:** <https://www.biginsight.no>
- **PerCaThe:**  
<https://www.uio.no/english/research/strategic-research-areas/life-science/research/convergence-environments/percathe/>
- **RESCUER:** <https://www.rescuer.uio.no>
- **PINpOINT:**  
<https://digitallifenorway.org/research/pinpoint/>
- **Scientia Fellows Programme**, Faculty of Medicine, UiO  
<https://www.med.uio.no/english/research/scientia-fellows/>

# References

-  Ickstadt K, Schäfer M, Zucknick M (2018). Toward integrative Bayesian analysis in molecular biology. *ARSIA*. 5:141–167.
-  Bottolo L, Banterle M, Richardson S, Ala-Korpela M, Järvelin MR, Lewin A (2021). A computationally efficient Bayesian Seemingly Unrelated Regressions model for high-dimensional quantitative trait loci discovery. *JRSSC* 70(4):886–908.
-  Zhao Z, Banterle M, Bottolo L, Richardson S, Lewin A, Zucknick M (2021). BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in linear regression. *JSS* 100(11):1–32. <https://CRAN.R-project.org/package=BayesSUR>.
-  Zhao Z, Banterle M, Lewin A, Zucknick M (2022). Structured Bayesian variable selection for multiple correlated response variables and high-dimensional predictors. arXiv.2101.05899. Updated version coming soon to arXiv.