# G2PSR :

# Bayesian Genome-to-Phenome Sparse Regression

IA et Santé : Approches interdisciplinaires

30/06/2022

Marie Deprez

# A system biology approach … toward multi-omics data integration

**Rationale**: A complex biological system is a combination of many simpler processes, which combination provides functional outputs of greater expressiveness than the sum of its part.

The analysis of multiple biological layers through system level approaches has already been reported to provide great insights in the study of complex diseases.

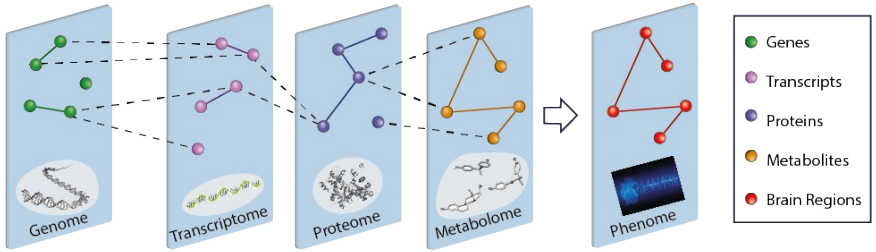



Fig: The interplay between biological layer

**Challenges**

- Large data dimensionality
- Limited effect size of biological features
- Wide heterogeneity in data type/properties
- Computational bottleneck
- Lack of flexibility to integrate both known and unknown interaction between omic layers

**Interdisciplinary approach using Deep learning methods :**

**Bayesian Sparse Regression**

# A system biology approach … toward multi-omics data integration

**Rationale**: A complex biological system is a combination of many simpler processes, which combination provides functional outputs of greater expressiveness than the sum of its part.

The analysis of multiple biological layers through system level approaches has already been reported to provide great insights in the study of complex diseases.

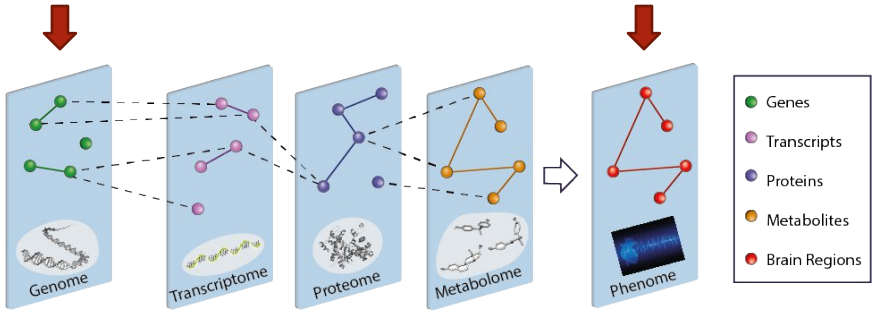



Fig: The interplay between biological layer

**Challenges**

- Large data dimensionality
- Limited effect size of biological features
- Wide heterogeneity in data type/properties
- Computational bottleneck
- Lack of flexibility to integrate both known and unknown interaction between omic layers

**Interdisciplinary approach using Deep learning methods :**

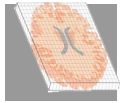
**Bayesian Sparse Regression**

# Genome-to-Phenome neural network with biologically inspired constraint

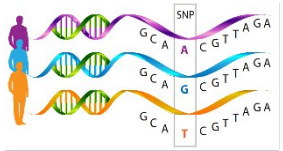
**Rationale**: Predict multivariate phenotypic features from genetic data.

- **Phenotypic data:** clinical data, medical evaluation, ...

→ Check-up evaluation
(aptitude evaluation, cognition score ...)
→ Physical examination
(Systolic and Diastolic Blood Pressure)
→ Laboratory measures
(creatinine, hemoglobin, glucose ...)
→ Diagnostic procedures
(Stress test, imaging results)



- **Genetic data:** Single Nucleotide Polymorphisme



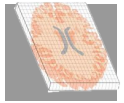

**Possible values :**
0 – Homozygous reference
1 -- Heterozygous
2 – Homozygous alternative

# Genome-to-Phenome neural network with biologically inspired constraint

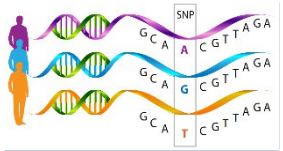
**Rationale**: Predict multivariate phenotypic features from genetic data.

- **Phenotypic data:** clinical data, medical evaluation, ...

→ Check-up evaluation
(aptitude evaluation, cognition score ...)
→ Physical examination
(Systolic and Diastolic Blood Pressure)
→ Laboratory measures
(creatinine, hemoglobin, glucose ...)
→ Diagnostic procedures
(Stress test, imaging results)



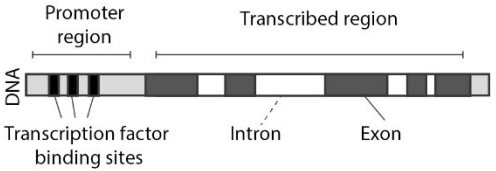
- **Genetic data:** Single Nucleotide Polymorphisme




**Possible values :**
0 – Homozygous reference
1 -- Heterozygous
2 – Homozygous alternative

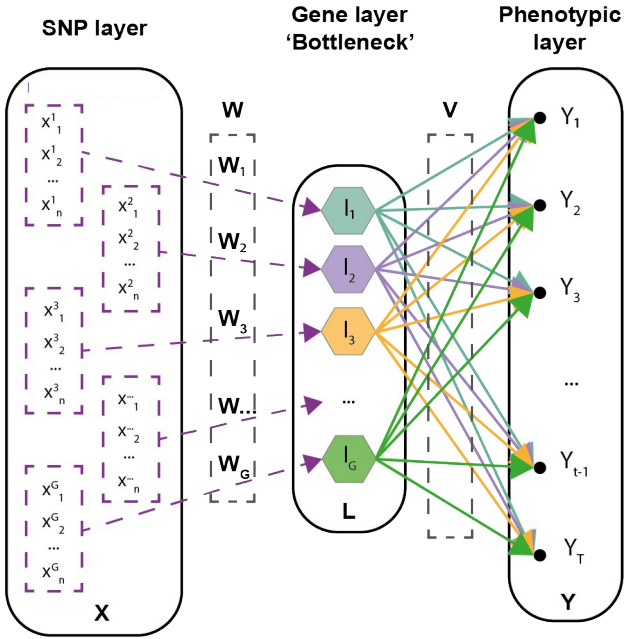
**Constraint**: Functional and structural relationship between SNPs within the transcribed or the regulation range of their asociated gene.

→ **SNPs-to-Gene matrix** (n x g) with n the number of SNPs and g the number of genes containing the SNPs.




Schematic gene structure

# Bayesian Genome-to-Phenome Sparse Regression




Biologically constrained Bayesian G2PSR

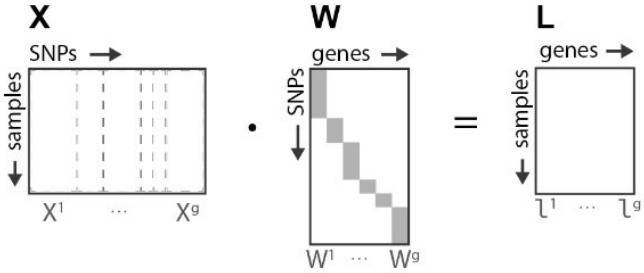
## G2PSR generative model

$\boldsymbol{X^g}$ All SNPs associated to gene *g*

$\boldsymbol{W^g}$ Linear transformation mapping SNPs to their respective gene g.

$$w_i = \mathbf{0}\, if\ i \notin A^{(g)}$$

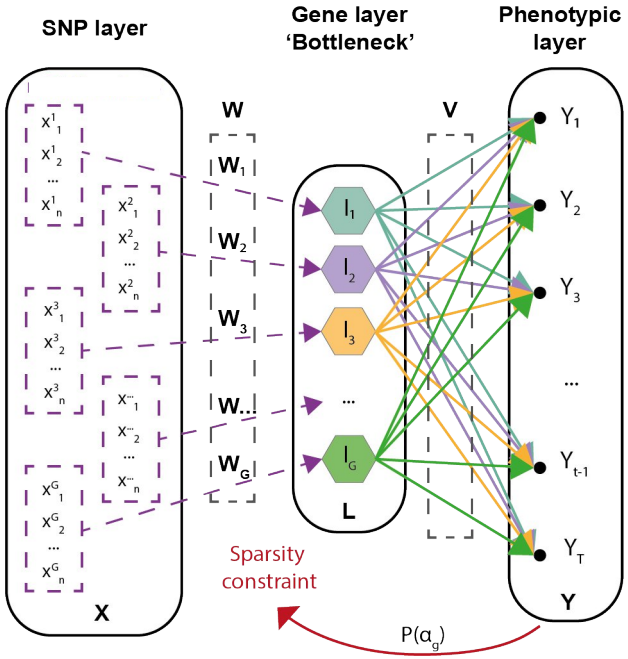
$\boldsymbol{L = X.W} = (l^1, \ldots, l^g)$ Intermediate/gene layer




$\boldsymbol{Y} = N(\mathbf{L}, \mathbf{V}, \boldsymbol{\Sigma})$ Reconstruction of the phenotypic features Y

$\boldsymbol{\Sigma}$ Variance of the observational noise

# Bayesian Genome-to-Phenome Sparse Regression




Biologically constrained Bayesian G2PSR

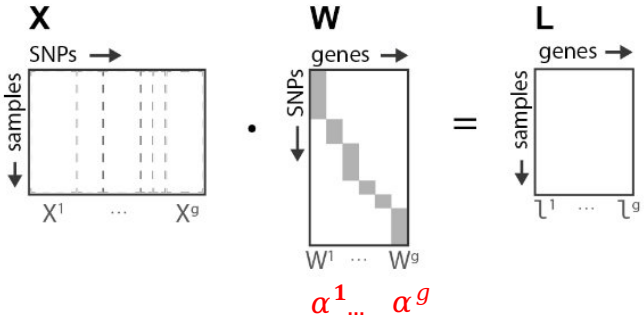
## G2PSR group-sparsity constraint

$X^g$ All SNPs associated to gene *g*

$W^g$ Linear transformation mapping SNPs to their respective gene g.

Introduction of variational dropout
as a group-wise regularisation technique.

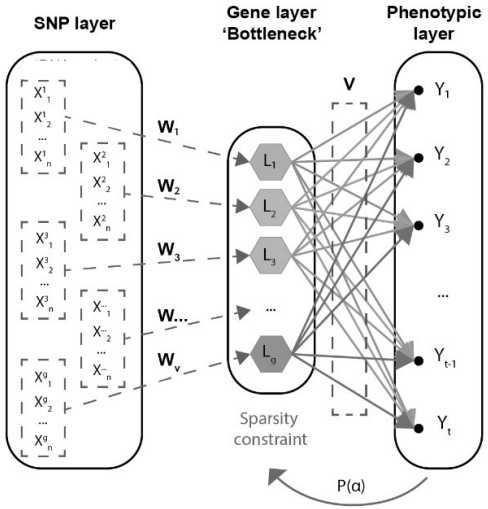
$$w_i^g = N(\mu_i^g, \alpha^g \cdot \mu_i^{g2})$$




$\alpha^1 \dots \alpha^g$

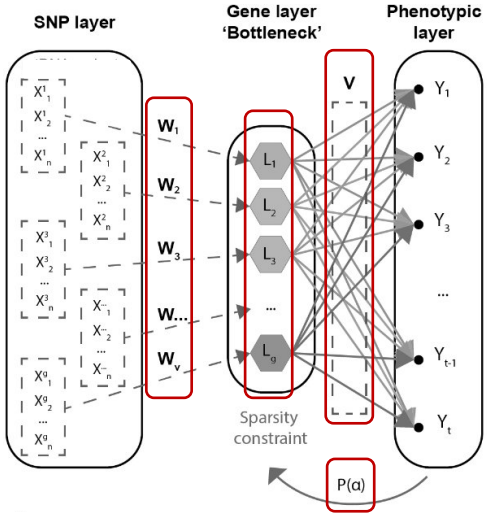
# G2PSR Optimization

**Challenge of over-parametrized neural networks**



# G2PSR Optimization

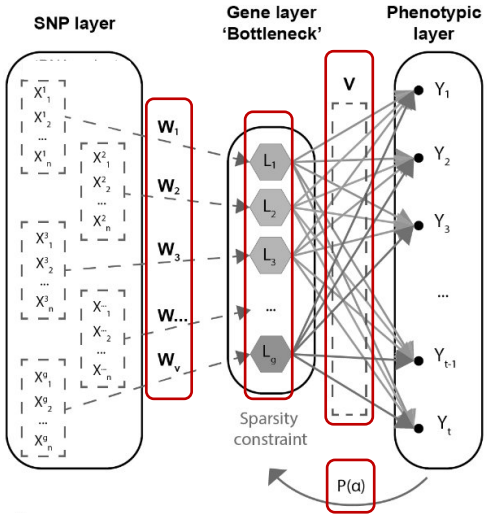
**Challenge of over-parametrized neural networks**




- Regression coefficients per SNPs
- Hidden layer
- Regression coefficients per genes
- Sparsity constraint per genes

# G2PSR Optimization

## Challenge of over-parametrized neural networks



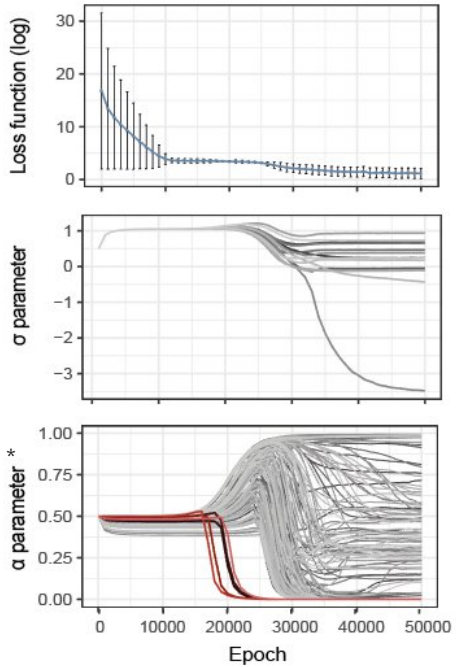

- Regression coefficients per SNPs
- Hidden layer
- Regression coefficients per genes
- Sparsity constraint per genes

## Synthetic experiment

| Attributes | Value |
|---|---|
| Number of genes | 200 (~3500 SNPs) |
| Number of samples | 500 |
| Number of phenotypic features | 15 |
| Noise level | 20% |

# G2PSR Optimization

**G2PSR optimization parameters**




* Red lines correspond to relevant genes

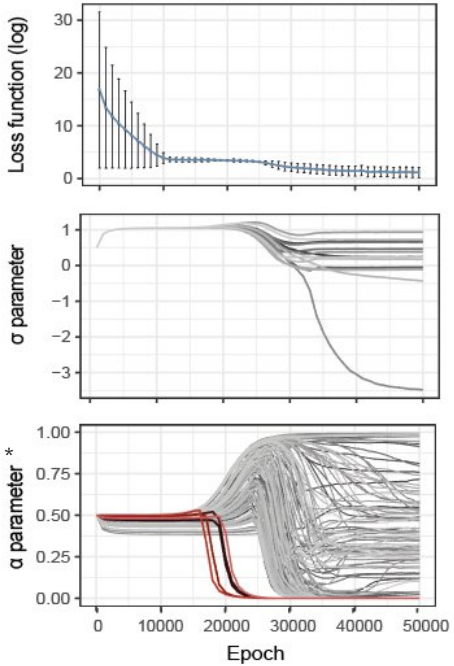
**Overfitting heuristics**

- Over-optimization of the loss
- Decrease of the observational noise variance

→ Identify an early-stopping strategy

# G2PSR Optimization

**G2PSR optimization parameters**



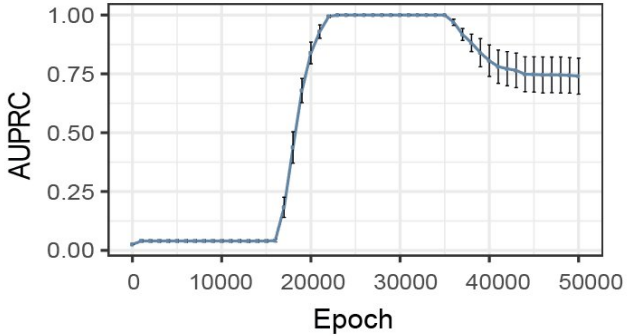

* Red lines correspond to relevant genes

**Overfitting heuristics**

- Over-optimization of the loss
- Decrease of the observational noise variance

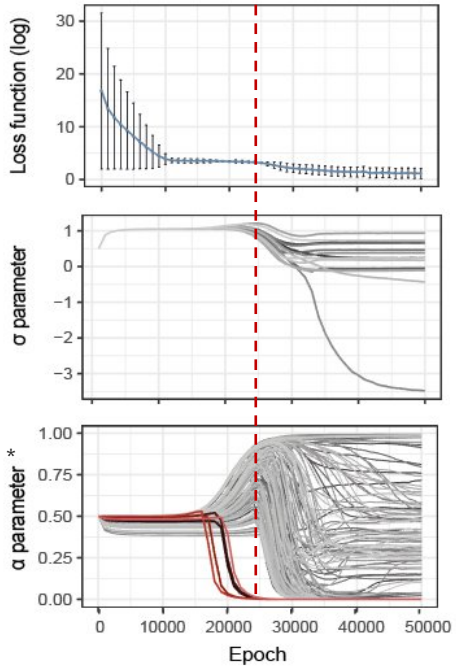
→ Identify an early-stopping strategy

**Area Under the Precision-Recall Curve during G2PSR optimization**



# G2PSR Optimization

**G2PSR optimization parameters**



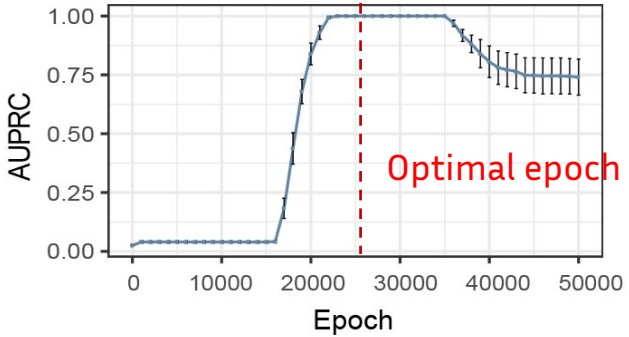

* Red lines correspond to relevant genes

**Overfitting heuristics**

- Over-optimization of the loss
- Decrease of the observational noise variance

→ Identify an early-stopping strategy

**Area Under the Precision-Recall Curve during G2PSR optimization**



# G2PSR applied to Alzheimer's Disease


- Alzheimer's Disease is a progressive disease and the most common type of dementia
- Induces modification in the brain structure (volume variation of different brain regions)
- Mild to serious cognitive impairment (memory loss; disorientation; lack of coherent thinking pattern ...)
- Through the Alzheimer's Disease Neuroimaging Initiative database, we can access numerous biological modalities, such as **genetic**, proteomic, metabolomic and **imaging** data for more than 500 patients.

ADNI
Alzheimer's Disease Neuroimaging Initiative

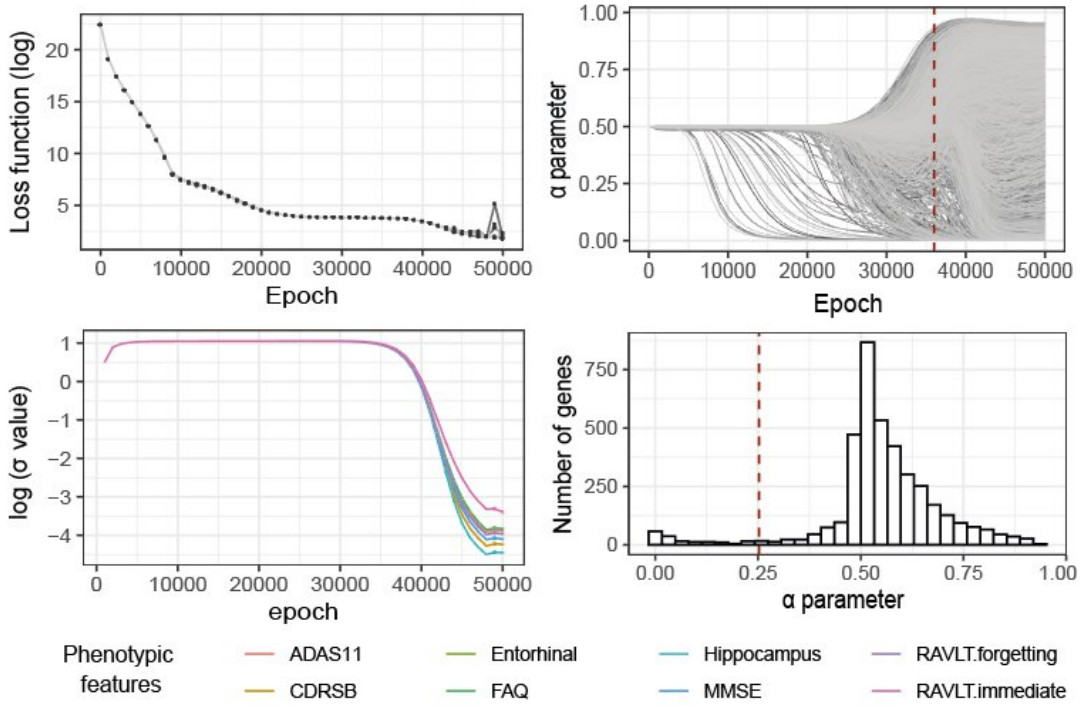
| ADNI test dataset | Value |
|---|---|
| Number of genes | 3953 (104854 SNPs) |
| Number of samples | 491 |
| Number of phenotypic features | 8 |

→ Genes from KEGG pathways (186 pathways)

→ 150 CN, 200 MCI, 139 Dementia

→ Brain volumes (Hyppcampus, Entorhinal), Cognitive tests (ADAS, CDRSB, FAQ, MMSE, RAVLT)

# Primary analysis




- Similarly to applications on synthetic experiments, we reproduce the early stopping strategy.

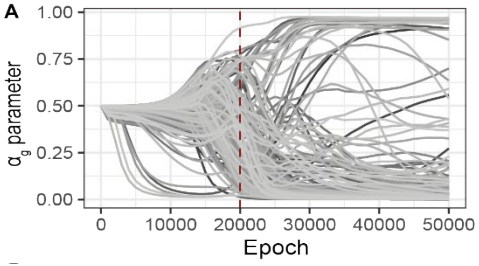
- Analysis of the $\alpha g$ distribution at the optimal epoch:

Reference 'p-value' like distribution

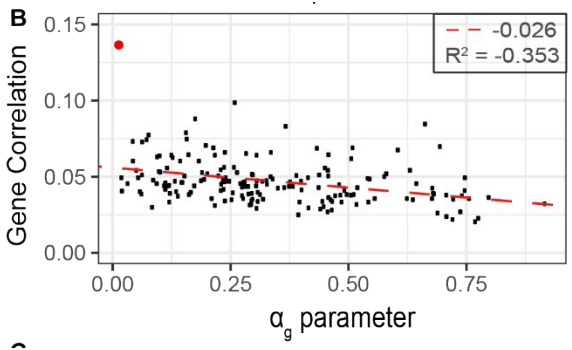
- Selection of 177 'significative' genes with $\alpha g$ < 0,25

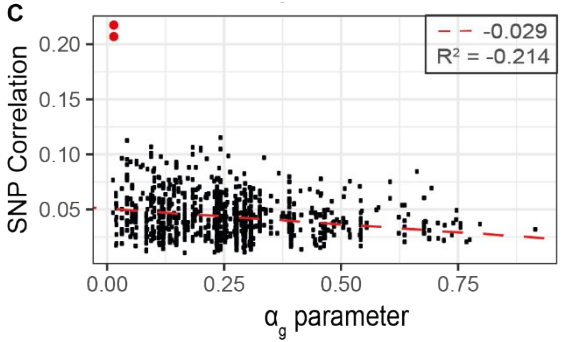
| Gene | $\alpha$ parameter |
|---|---|
| THOC3 | 8,65$^{e}$-5 |
| MAT2B | 9,18$^{e}$-5 |
| CCL28 | 1,54$^{e}$-4 |
| NKD2 | 2,50$^{e}$-4 |
| RICTOR | 2,85$^{e}$-4 |
| AGXT2 | 4,37$^{e}$-4 |
| APOE | 5,59$^{e}$-4 |

# ‘Post-hoc’ analysis




- Refined analysis on the selected genes to improve the experiment settings.
- Analysis of the mean correlation of each Genes and SNPs with all phenotypic features studied
- Negative correlation between SNPs and $\alpha g$ parameter.




| Gene | $\alpha g$ parameter |
|---|---|
| APOE | 0,014 |
| MAT2B | 0,020 |
| THOC3 | 0,021 |
| NKD2 | 0,033 |
| PTPN11 | 0,043 |
| PI3KC2B | 0,044 |
| MCM2 | 0,051 |

# Conclusions

- We proposed a novel Genome-to-Phenome association method, using Bayesian Sparse Regression with biologically inspired constraint, **G2PSR**.
- G2PSR performs better than state-of-the-art methods in synthetic experiments. Yet, each method gives a specific set of information on the Genome-to-Phenome association that cannot be match by our proposed model.
- G2PSR performs well in a real case dataset despite the genomic features/samples ratio. The 2-steps workflow presented here help reduce the number of genes to be analysed, without *a priori* selection.

# Perspectives

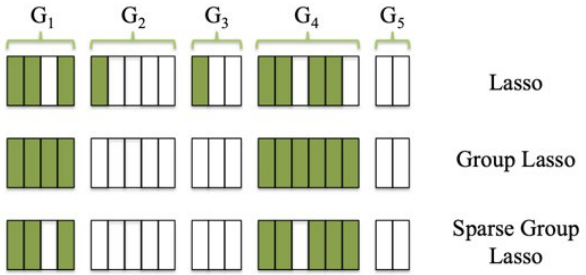
- Implement variant pathogenicity scores as prior in the grouping constraint.
- Benchmark the group constraint to integrate SNPs associated to multiple genes (regulatory regions)
- Integrate other multi-omics data (RNA-seq, functional pathways constraint ...)

Deprez et al, 2022, Frontiers in Molecular Medicine

# Thank you for your attention !

# Comparison with state-of-the-art methods

➢ **Sparse Group Lasso**



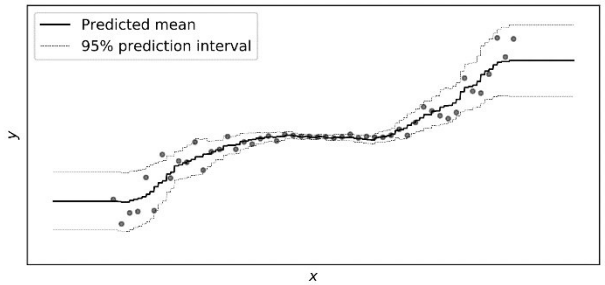

$l_1$–$norm$ SNPs regularization

$l_2$–$norm$ Genes / group regularization

Associated with a single phenotypic feature

[Simon N et al. 2013 *Journal of Computational and graphical Statistics*]

➢ **Bayesian Group Sparse Multi-Task Regression**

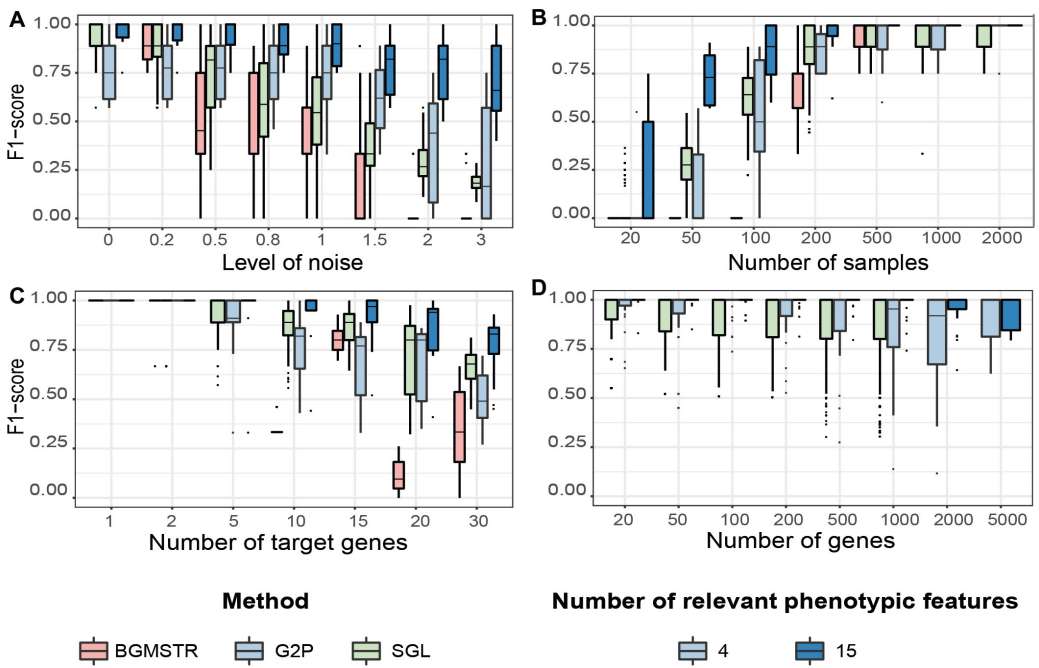



Inspired from Sparse Group Lasso but extended to :

- accommodate multivariate phenotypes,
- produce confidence intervals on regression coefficients.

[Greenlaw K et al. 2017 *Bioinformatics*]

# Comparison with state-of-the-art methods




| Reference dataset attributes | Value |
|---|---|
| Number of genes | 200 |
| Number of samples | 500 |
| Number of phenotypic features | 15 |
| Noise level | 20% |

With all phenotypic features associated to the genome, **G2PSR performs better** than state-of-the-art methods:

- Noise robustness
- Fewer number of samples required (statistical power)
- Improved sensibility to relevant genes
- Improve capacity to analyse large number of genes