UNIVERSITÉ DE LYON

ENS DE LYON

GraphNEx
Graph Neural Networks
for Explainable Artificial Intelligence

Discussing gene selection methods based on gene expression data

GT LEGO, 23 november 2023

Myriam Bontonou, Anaïs Haget, Maria Boulougouri, Jean-Michel Arbona, Benjamin Audit, Pierre Borgnat

chist-era

anr

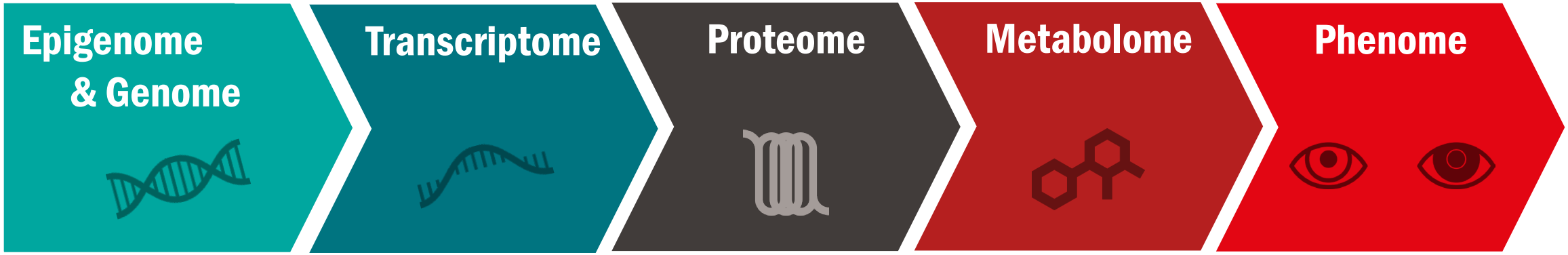
Overview

- Functional genomics: objective and available data
- Selecting relevant genes with statistical and machine learning methods
- Application to cancer classification on bulk RNA-seq data

Overview

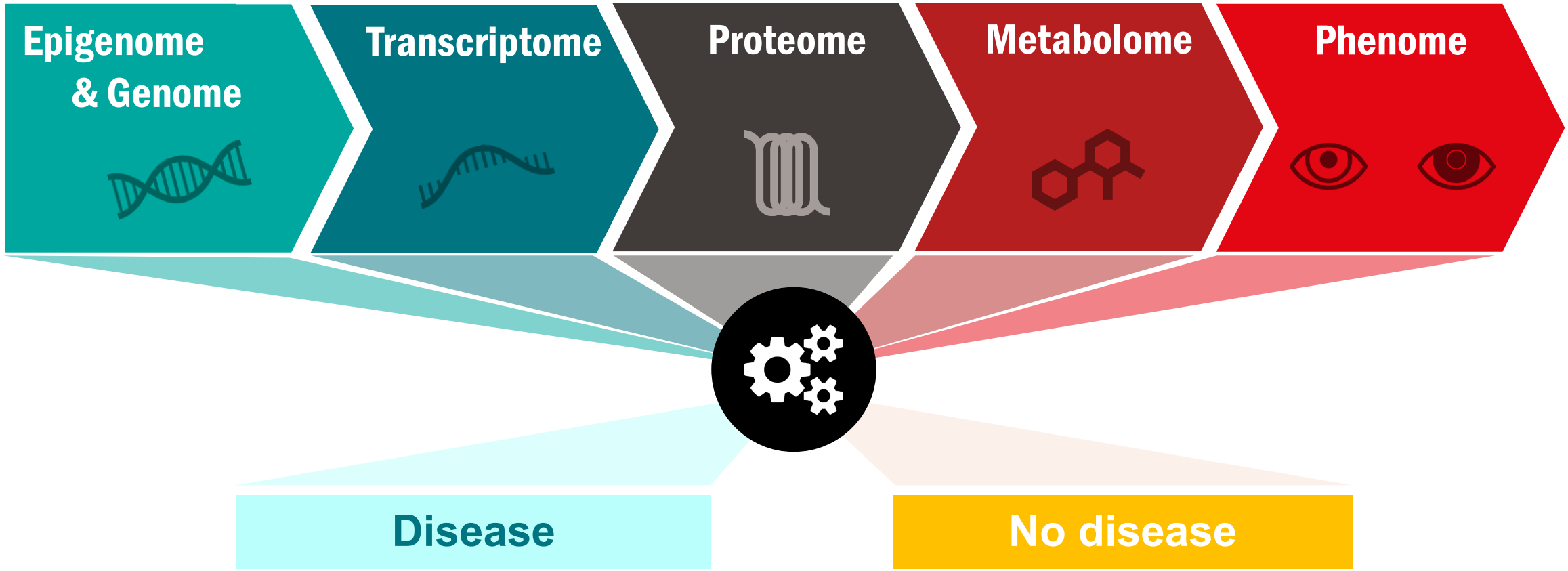
- **Functional genomics: objective and available data**
- Selecting relevant genes with statistical and machine learning methods
- Application to cancer classification on bulk RNA-seq data

Introduction to functional genomics



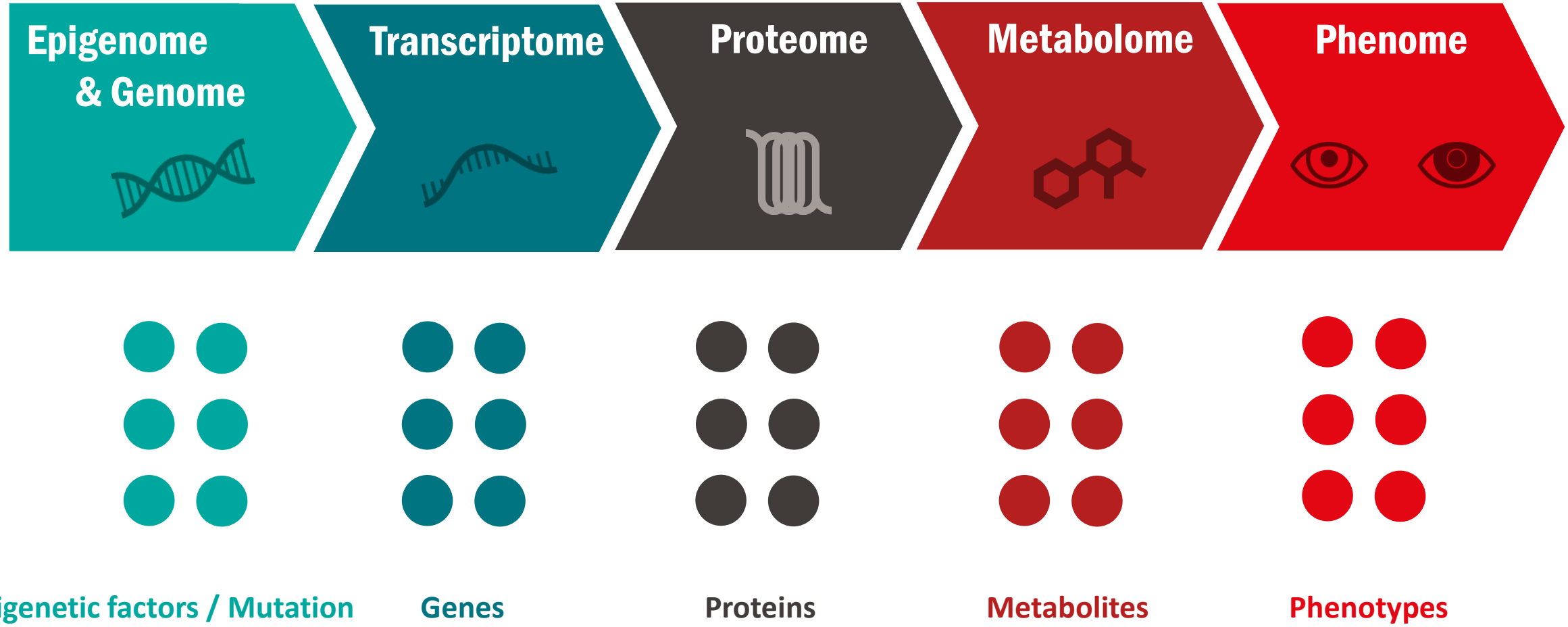
Adapted from Zitnik et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion (2019).

Introduction to functional genomics

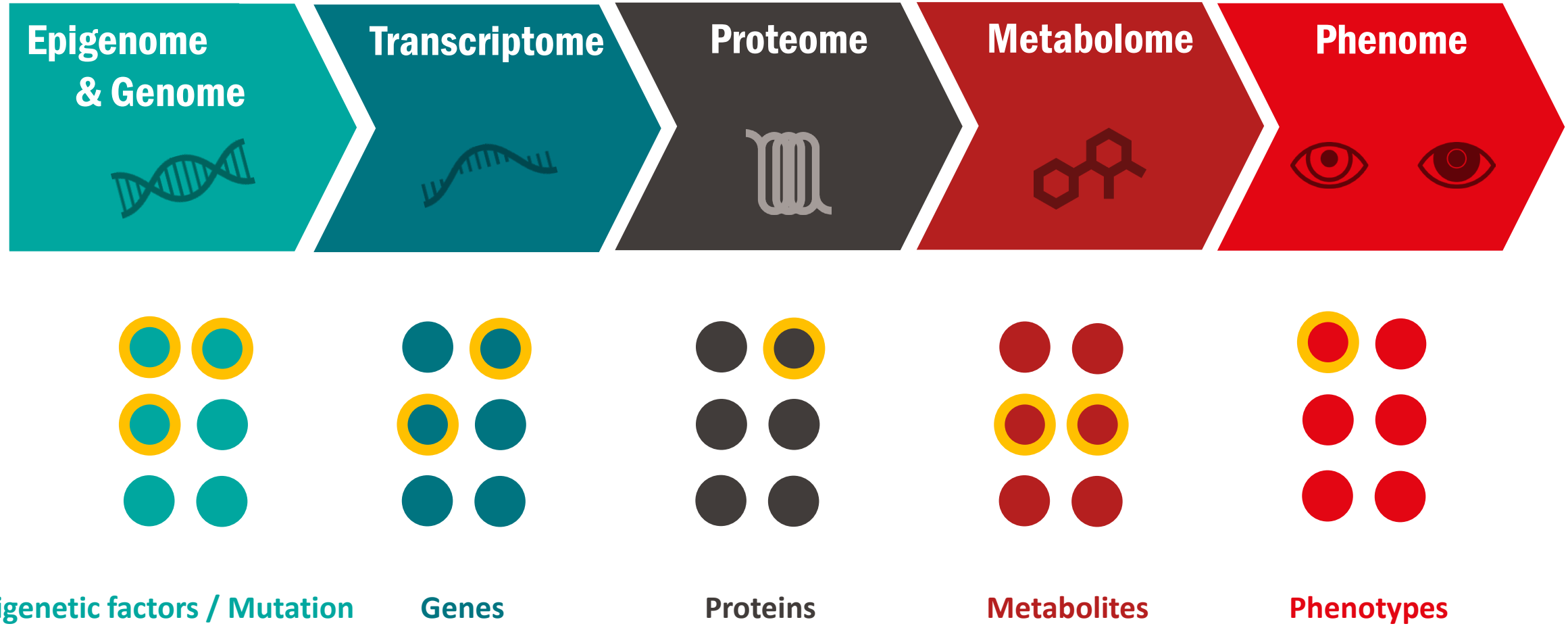


Adapted from Zitnik et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion (2019).

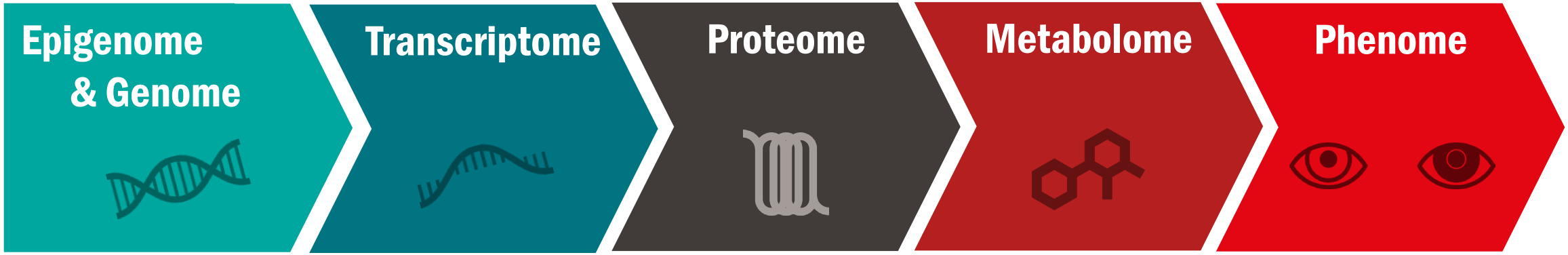
Introduction to functional genomics

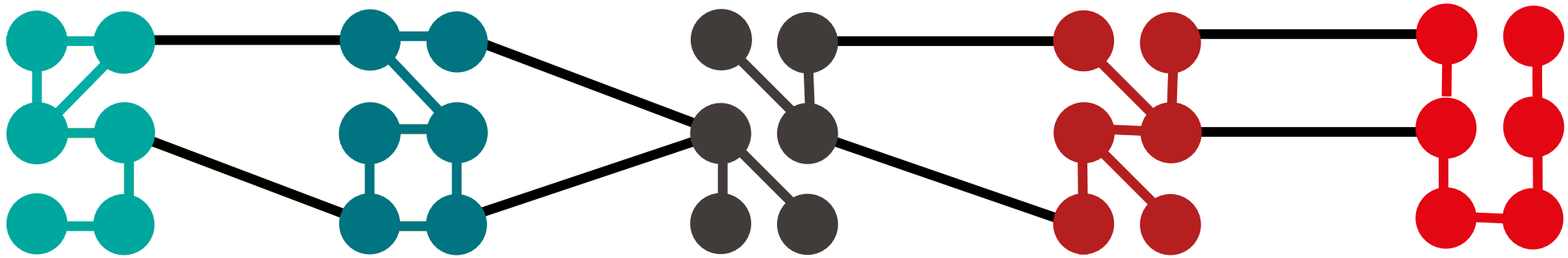


Introduction to functional genomics



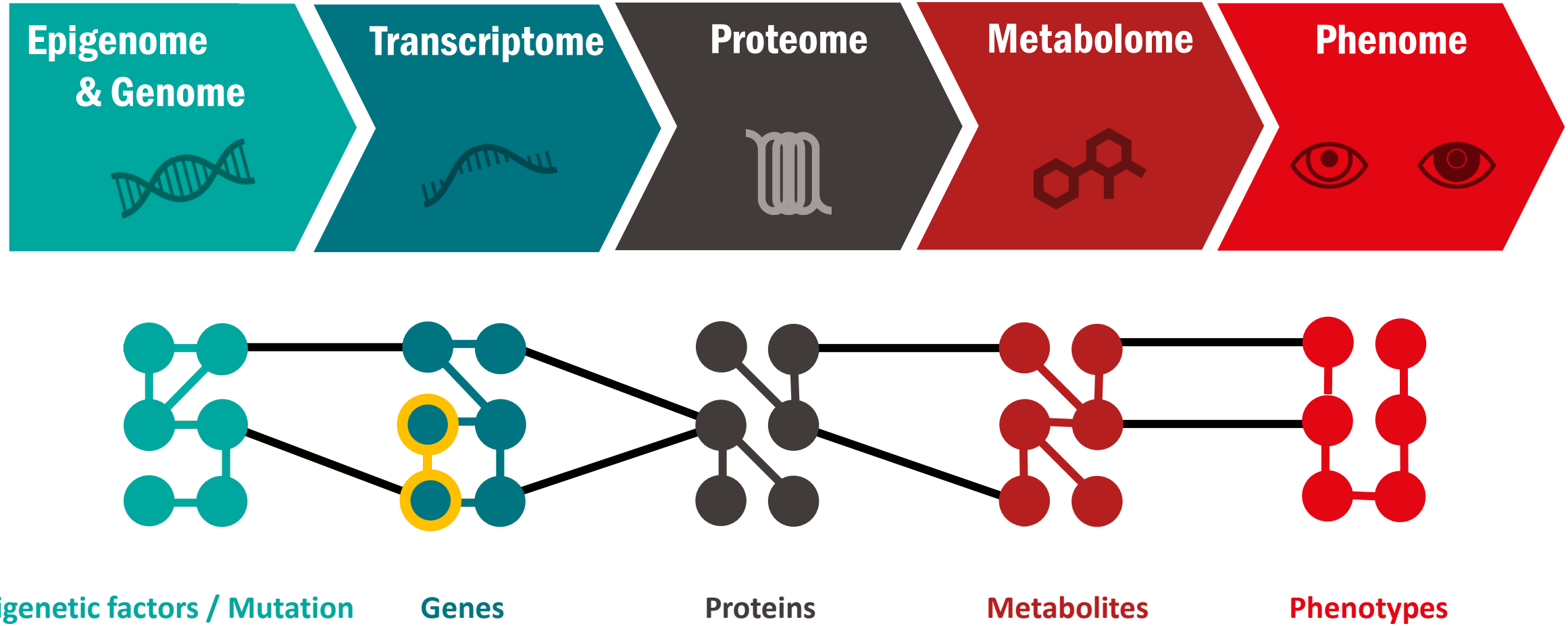
Introduction to functional genomics



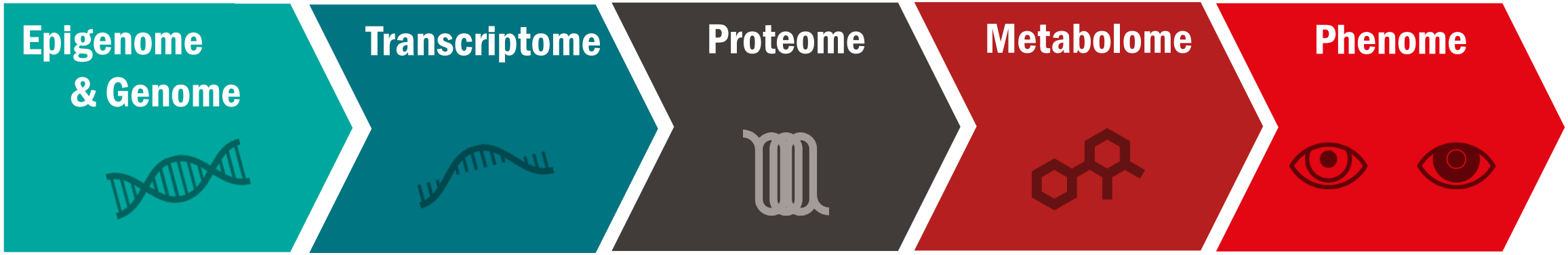


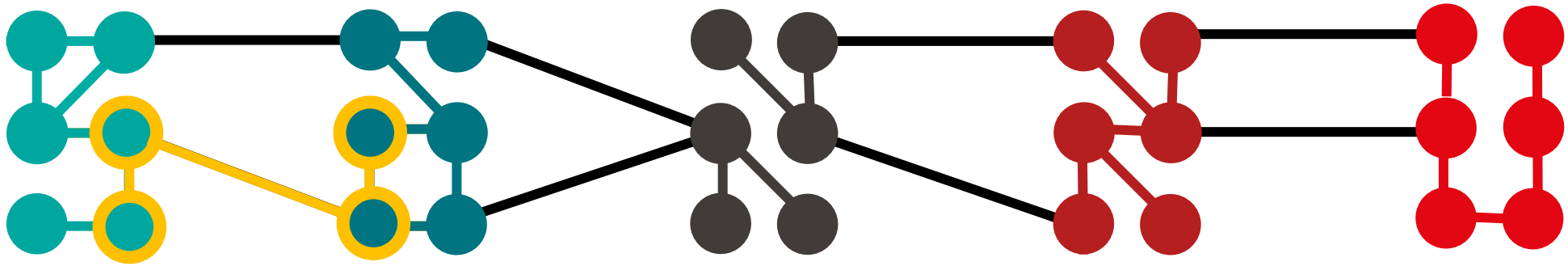
Epigenetic factors / Mutation — Genes — Proteins — Metabolites — Phenotypes

Introduction to functional genomics



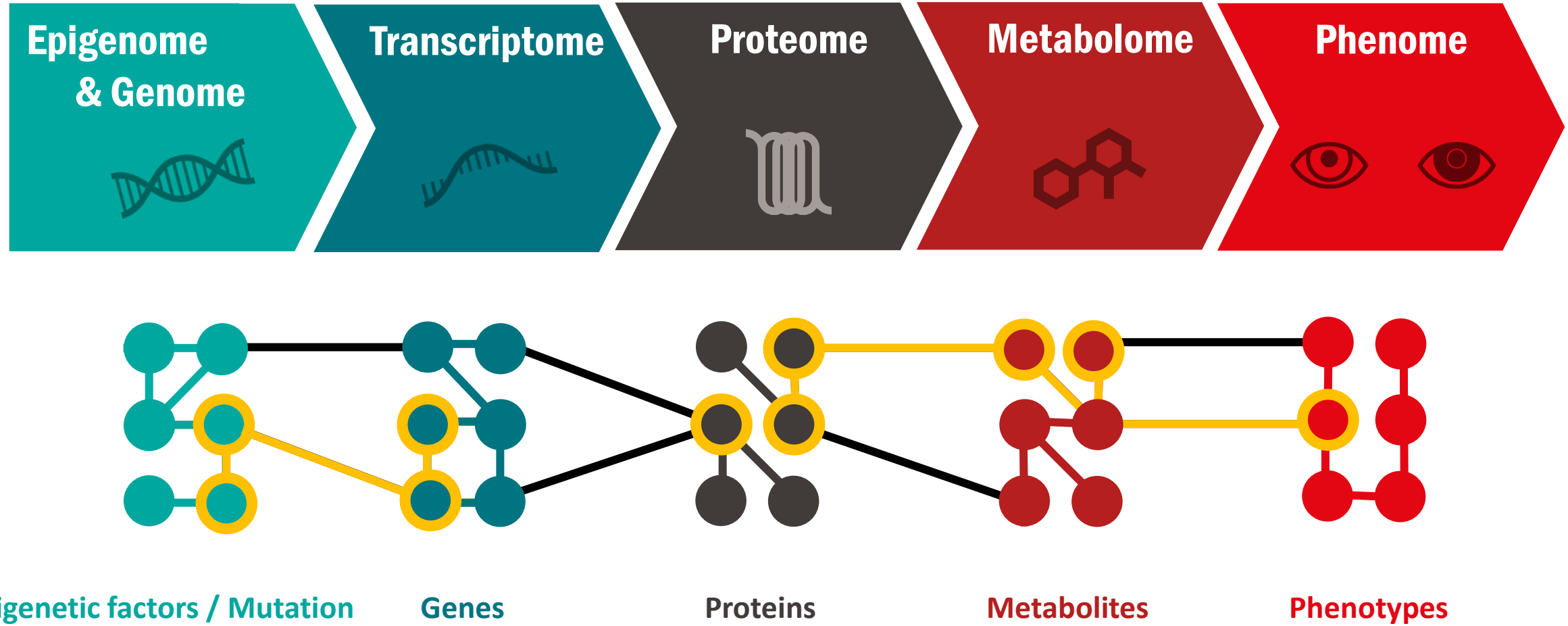
Introduction to functional genomics



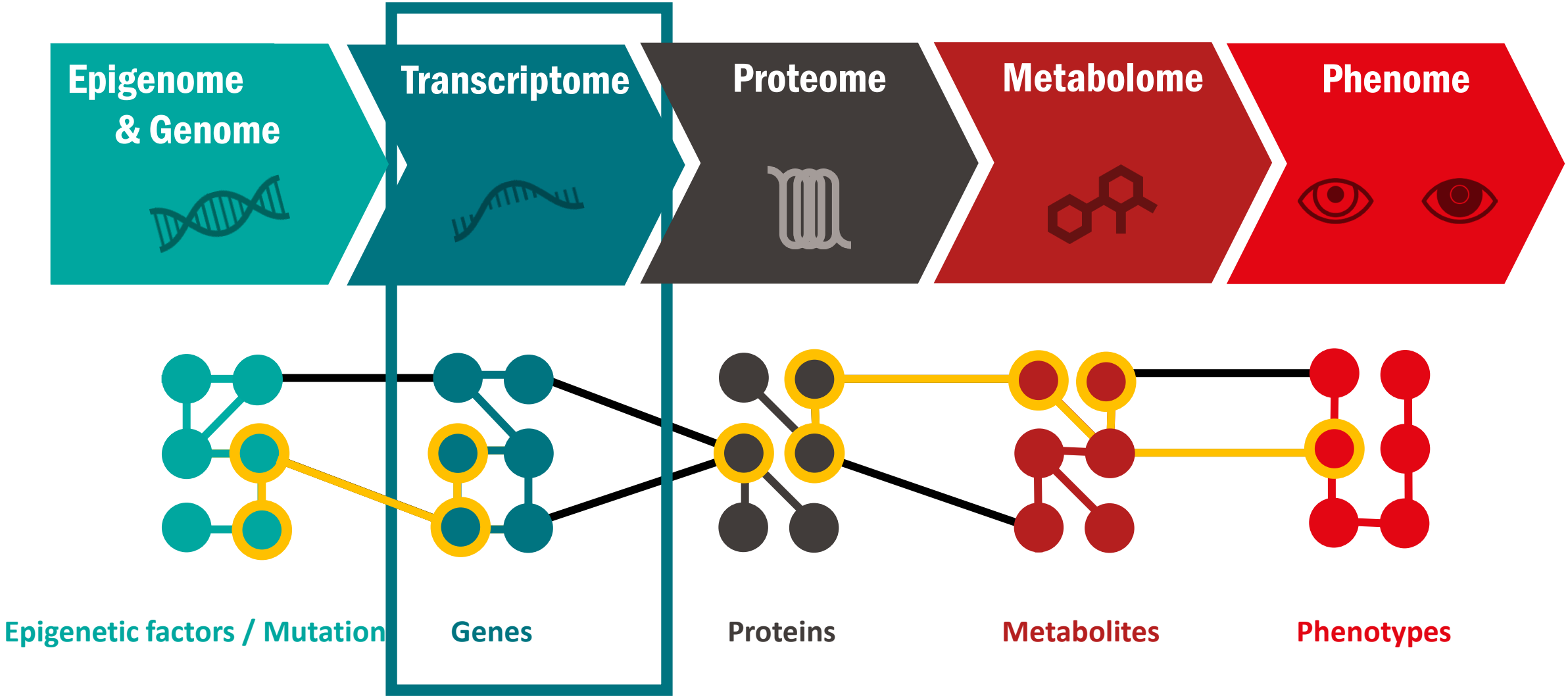


Epigenetic factors / Mutation | Genes | Proteins | Metabolites | Phenotypes

Introduction to functional genomics



Introduction to functional genomics



Overview

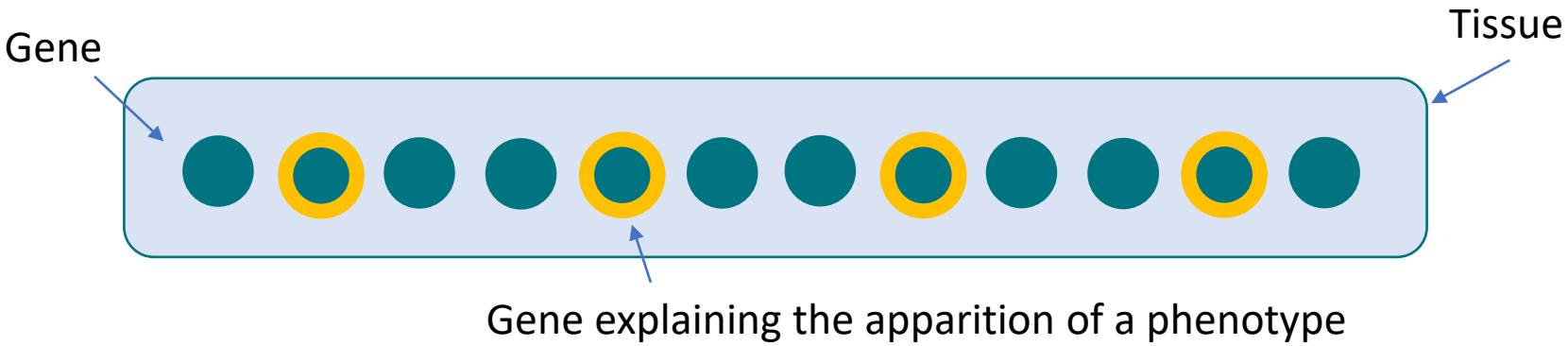
- Functional genomics: objective and available data
- **Selecting relevant genes with statistical and machine learning methods**
- Application to cancer classification on bulk RNA-seq data

Statistics and machine learning on gene expression data

Objective Understand why certain phenotypes appear based on the genes expressed in a tissue.

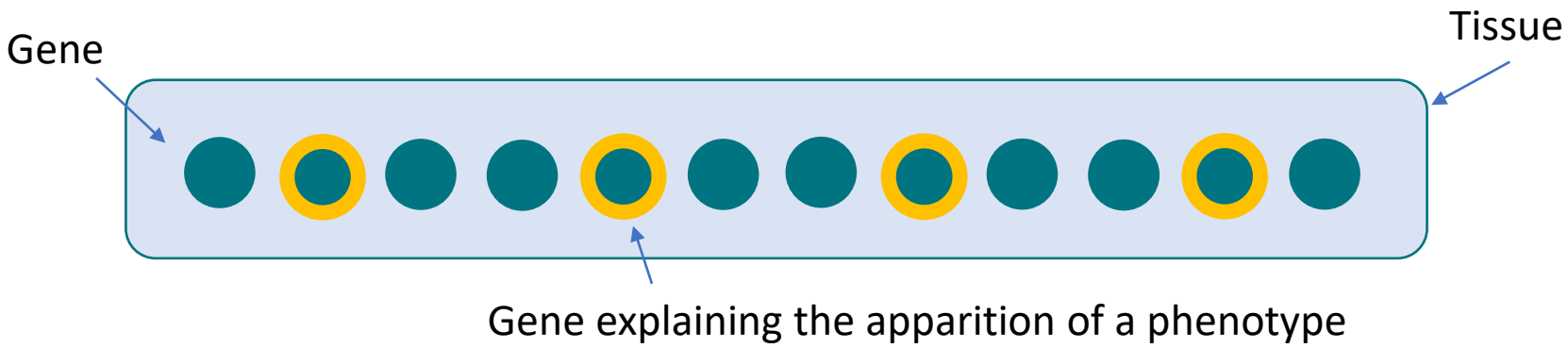
Statistics and machine learning on gene expression data

Objective Understand why certain phenotypes appear based on the genes expressed in a tissue.



Statistics and machine learning on gene expression data

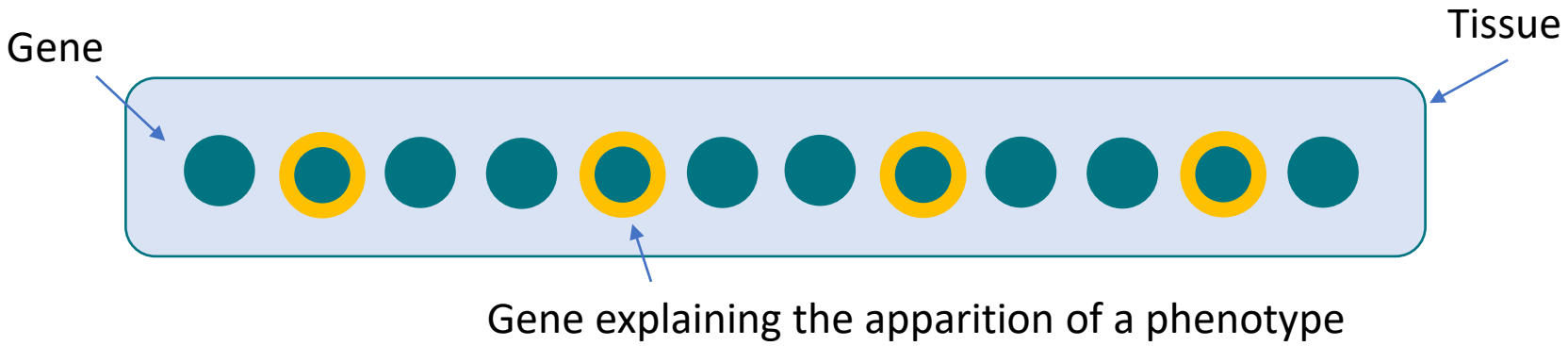
Objective Understand why certain phenotypes appear based on the genes expressed in a tissue.



Datasets The Cancer Genome Atlas (TCGA), TARGET, Genotype-Tissue Expression project (GTEx).

Statistics and machine learning on gene expression data

Objective Understand why certain phenotypes appear based on the genes expressed in a tissue.



Datasets The Cancer Genome Atlas (TCGA), TARGET, Genotype-Tissue Expression project (GTEx).

Dataset	Task	Source	# classes	# samples (max / min per class)	# genes
ttg-breast	Healthy vs Tumor	TCGA TARGET GTEx	2	1384 (1092 / 292)	14373
BRCA-pam	PAM50 classes	Legacy TCGA	5	916 (421 / 67)	13896

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Genes are ranked according to...

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Genes are ranked according to...

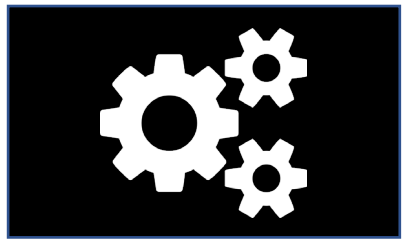
Intrinsic score related to a gene

Statistics and machine learning on gene expression data

Algorithm

Task

Data



Decision

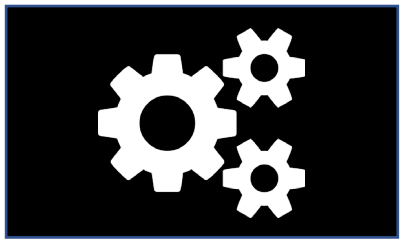
User

Statistics and machine learning on gene expression data

Algorithm

Task

Data



Decision

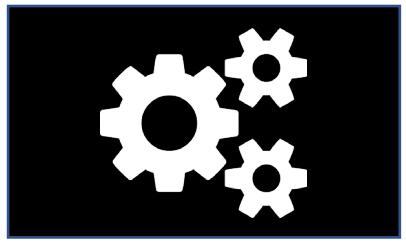
Medical diagnosis

User

Statistics and machine learning on gene expression data

Algorithm

Task

Data →  → Decision

Medical diagnosis

- ✓Genome 
- ✓Transcriptome 
- ✓Proteome 
- ✓Phenome…  

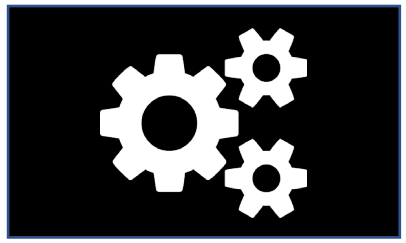
User

Statistics and machine learning on gene expression data

Algorithm

Task

Data



Decision

Medical diagnosis

✓Genome

✓Transcriptome

✓Proteome

✓Phenome…

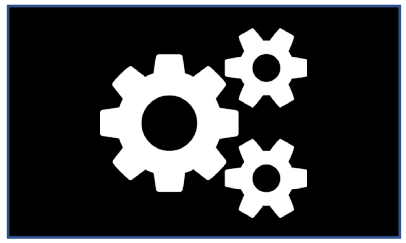
Cancer

User

Statistics and machine learning on gene expression data

Algorithm

Task

Data →  → Decision

Medical diagnosis

✓Genome 

✓Transcriptome 

✓Proteome 

✓Phenome… 

Cancer

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Genes are ranked according to...

Intrinsic score related to a gene

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters → Embedded methods → Post-hoc explainability

Genes are ranked according to...

Intrinsic score related to a gene

Intrinsic classifier score

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters → Embedded methods → Post-hoc explainability

Genes are ranked according to...

Filters	Embedded methods	Post-hoc explainability
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters	Embedded methods	Post-hoc explainability
Genes are ranked according to...		
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier
Examples		

Statistics and machine learning on gene expression data

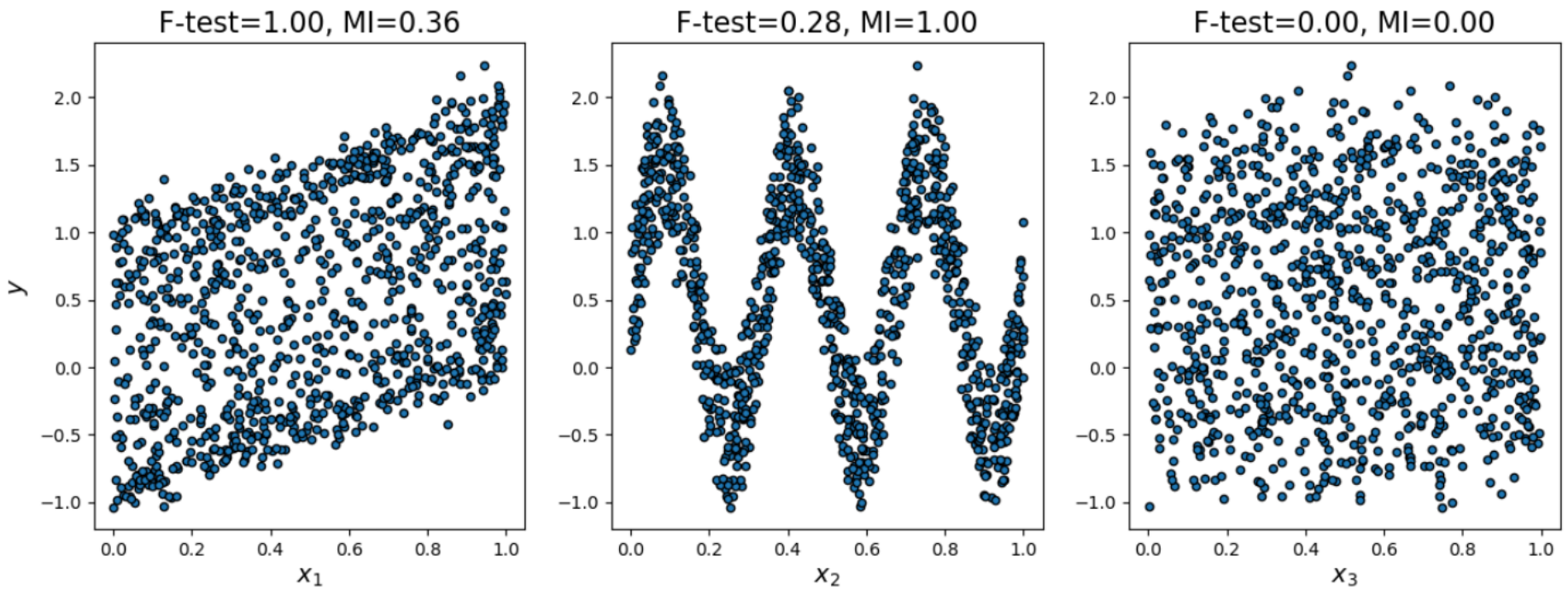
How genes are selected with statistical and machine learning methods?

Feature selection

Filters	Embedded methods	Post-hoc explainability
Genes are ranked according to...		
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier
Examples		
Mutual information		

Statistics and machine learning on gene expression data

Mutual information (MI)



Example on scikit-learn.

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters	Embedded methods	Post-hoc explainability
Genes are ranked according to...		
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier
Examples		
Mutual information DESeq2		

Statistics and machine learning on gene expression data

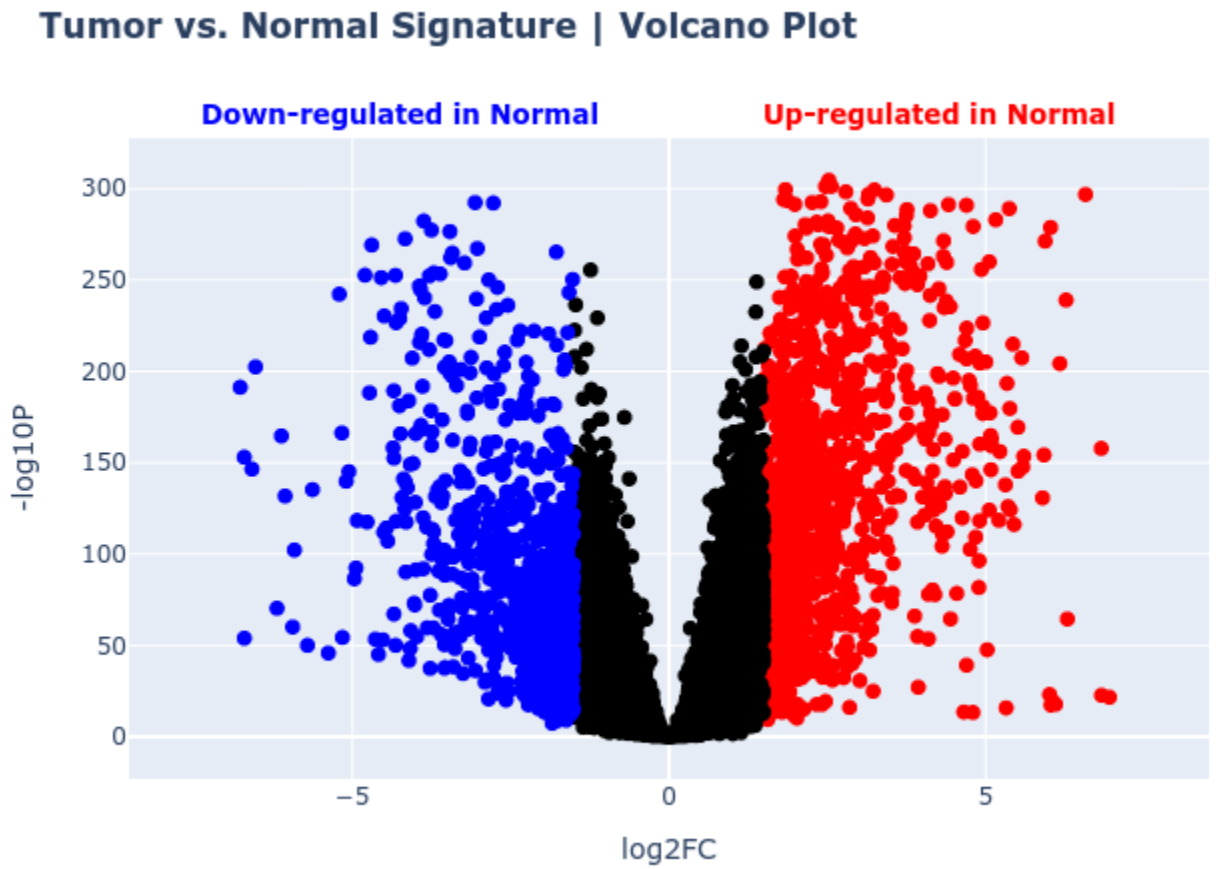
DEseq2 (DE)

Matrix with read counts associated with the genes

Modelisation of the counts

Log2 fold change of each gene

Testing for the significance of the changes



Love et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology (2014).

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters	Embedded methods	Post-hoc explainability
Genes are ranked according to...		
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier
Examples		
Mutual information DESeq2	Logistic regression with L1 penalty	

Statistics and machine learning on gene expression data

Logistic regression trained with a L1 penalty (LR+L1)

$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + \boldsymbol{w}_1\boldsymbol{x}_1 + \boldsymbol{w}_2\boldsymbol{x}_2 + \boldsymbol{w}_3\boldsymbol{x}_3 + \boldsymbol{w}_4\boldsymbol{x}_4 + \ \ldots + \boldsymbol{w}_{10000}\boldsymbol{x}_{10000})$$

Statistics and machine learning on gene expression data

Logistic regression trained with a L1 penalty (LR+L1)

$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + \boldsymbol{w}_1\boldsymbol{x}_1 + \boldsymbol{w}_2\boldsymbol{x}_2 + \boldsymbol{w}_3\boldsymbol{x}_3 + \boldsymbol{w}_4\boldsymbol{x}_4 + \ \ldots + \boldsymbol{w}_{10000}\boldsymbol{x}_{10000})$$

↓ after training

$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + 0 + \boldsymbol{w}_2\boldsymbol{x}_2 + 0 + 0 + \ \ldots + \boldsymbol{w}_{10000}\boldsymbol{x}_{10000})$$

Statistics and machine learning on gene expression data

How genes are selected with statistical and machine learning methods?

Feature selection

Filters	Embedded methods	Post-hoc explainability
Genes are ranked according to...		
Intrinsic score related to a gene	Intrinsic classifier score	Explainability score computed after training a classifier
Examples		
Mutual information DESeq2	Logistic regression with L1 penalty	Logistic regression Multilayer perceptron Graph neural network

Statistics and machine learning on gene expression data

Logistic regression (LR+L2)

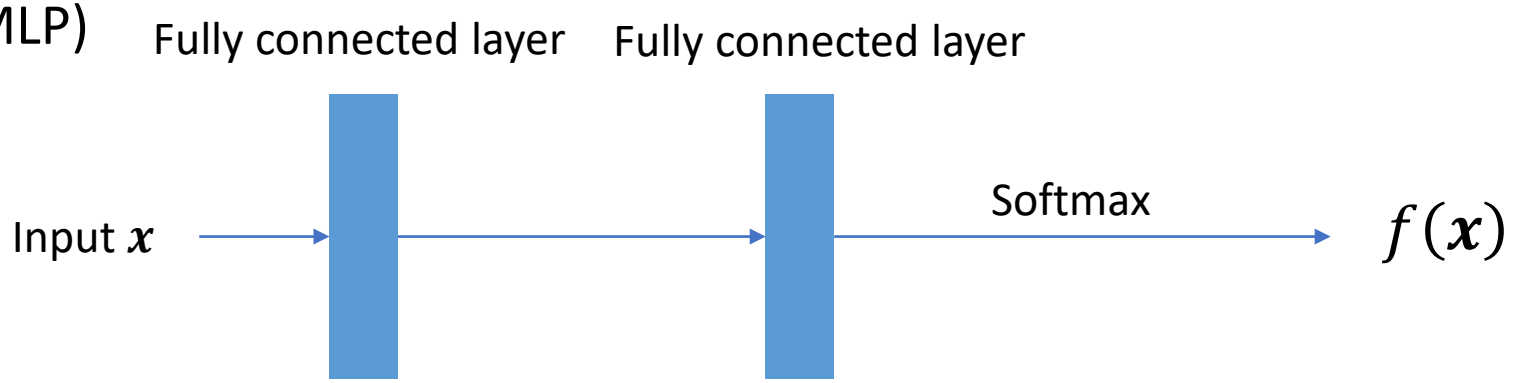
$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + \boldsymbol{w}_1 x_1 + \boldsymbol{w}_2 x_2 + \boldsymbol{w}_3 x_3 + \boldsymbol{w}_4 x_4 + \ \ldots + \boldsymbol{w}_{10000} x_{10000})$$

Statistics and machine learning on gene expression data

Logistic regression (LR+L2)

$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + \boldsymbol{w}_1 x_1 + \boldsymbol{w}_2 x_2 + \boldsymbol{w}_3 x_3 + \boldsymbol{w}_4 x_4 + \ \ldots + \boldsymbol{w}_{10000} x_{10000})$$

Multilayer perceptron (MLP)

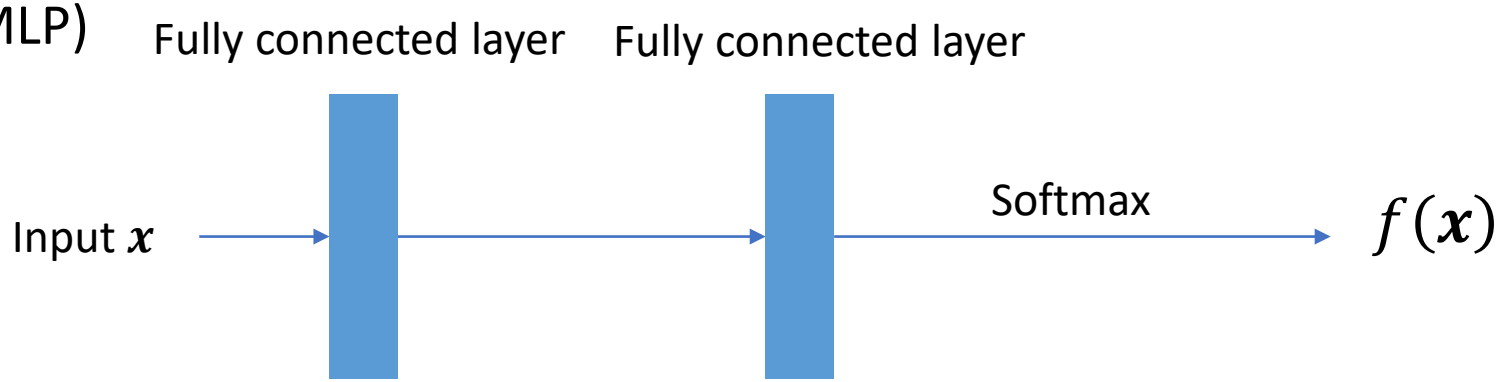


Statistics and machine learning on gene expression data

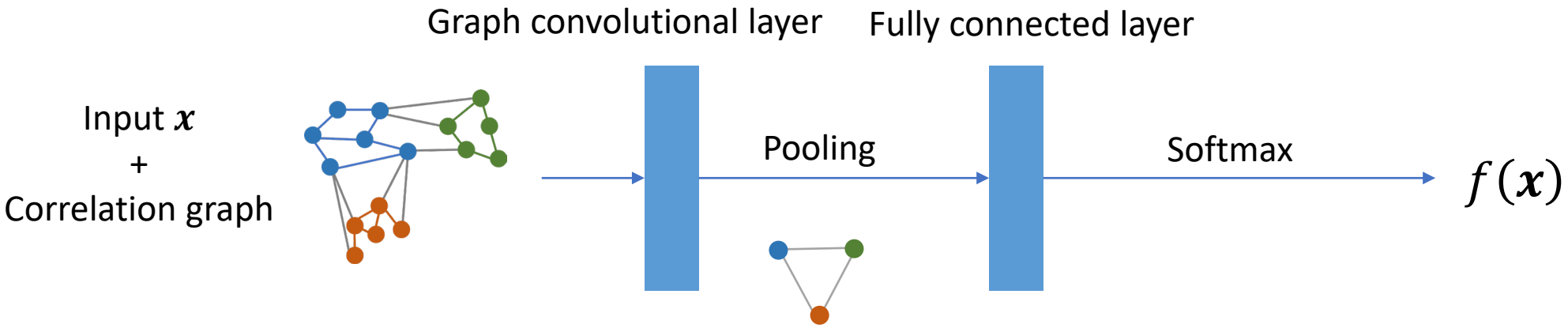
Logistic regression (LR+L2)

$$f(\boldsymbol{x}) = softmax(\boldsymbol{w}_0 + \boldsymbol{w}_1 x_1 + \boldsymbol{w}_2 x_2 + \boldsymbol{w}_3 x_3 + \boldsymbol{w}_4 x_4 + \ \ldots + \boldsymbol{w}_{10000} x_{10000})$$

Multilayer perceptron (MLP)



Graph convolutional network (GCN)



Post-hoc explainability

Input $\boldsymbol{x} \in \mathbb{R}^F$ (tumor tissue) Reference $\boldsymbol{x}'$ (average of healthy tissues)

Model f: $\mathbb{R}^F \rightarrow \mathbb{R}^C$

Sundararajan et al. Axiomatic Attribution for Deep Networks. ICML (2017).

Statistics and machine learning on gene expression data

Post-hoc explainability

Input $\boldsymbol{x} \in \mathbb{R}^F$ (tumor tissue) Reference $\boldsymbol{x}'$ (average of healthy tissues)

Model f: $\mathbb{R}^F \rightarrow \mathbb{R}^C$

Explainability method attributing to each gene i of $\boldsymbol{x}$ a score $\phi_i(\boldsymbol{x})$. *Here, integrated gradients.*

$$\phi_i(\boldsymbol{x}) = (\boldsymbol{x_i} - \boldsymbol{x_i}') \int_{\alpha=0}^{1} \frac{\partial \boldsymbol{f_c}(\boldsymbol{z})}{\partial x_i} d\alpha \text{ with } z = (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{x}' + \boldsymbol{\alpha x}$$

Sundararajan et al. Axiomatic Attribution for Deep Networks. ICML (2017).

Statistics and machine learning on gene expression data

Post-hoc explainability

Input $\boldsymbol{x} \in \mathbb{R}^F$ (tumor tissue) Reference $\boldsymbol{x}'$ (average of healthy tissues)

Model f: $\mathbb{R}^F \rightarrow \mathbb{R}^C$

Explainability method attributing to each gene i of $\boldsymbol{x}$ a score $\phi_i(\boldsymbol{x})$. *Here, integrated gradients.*

$$\phi_i(\boldsymbol{x}) = (\boldsymbol{x_i} - \boldsymbol{x_i}') \int_{\alpha=0}^{1} \frac{\partial \boldsymbol{f_c}(\boldsymbol{z})}{\partial x_i} d\alpha \text{ with } z = (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{x}' + \boldsymbol{\alpha x}$$

Genes ranked in decreasing order of average scores.

Sundararajan et al. Axiomatic Attribution for Deep Networks. ICML (2017).

Overview

- Functional genomics: objective and available data
- Selecting relevant genes with statistical and machine learning methods
- **Application to cancer classification on bulk RNA-seq data**

Application

Dataset	Task	# classes	# samples (max / min per class)	# genes
ttg-breast	Healthy vs Tumor	2	1384 (1092 / 292)	14373
BRCA-pam	PAM50 classes	5	916 (421 / 67)	13896

Application

Dataset	Task	# classes	# samples (max / min per class)	# genes
ttg-breast	Healthy vs Tumor	2	1384 (1092 / 292)	14373
BRCA-pam	PAM50 classes	5	916 (421 / 67)	13896

Model	Hyperparameter	ttg-breast	BRCA-pam
MLP	# layers	1	1
	# hidden feat.	20	20
GNN	# layers	1	1
	# hidden feat.	1	2

Application

Dataset	Task	# classes	# samples (max / min per class)	# genes
ttg-breast	Healthy vs Tumor	2	1384 (1092 / 292)	14373
BRCA-pam	PAM50 classes	5	916 (421 / 67)	13896

Dataset	LR+L1	LR+L2	MLP	GNN
ttg-breast	14374	14374	287541	12497
BRCA-pam	69485	69485	278085	118568

Number of parameters

Dataset	LR+L1	LR+L2	MLP	GNN
ttg-breast	262	10	10	24
BRCA-pam	673	6	7	15

Average training duration

Application

Dataset	LR+L1	LR+L2	MLP	GNN
ttg-breast	99.7 ± 0.0	98.9 ± 0.2	99.4 ± 0.3	99.0 ± 0.3
BRCA-pam	92.3 ± 0.0	88.7 ± 1.0	87.4 ± 1.8	87.9 ± 0.9

Average balanced accuracy (%)

Healthy vs Tumor

PAM50 breast cancer subtypes

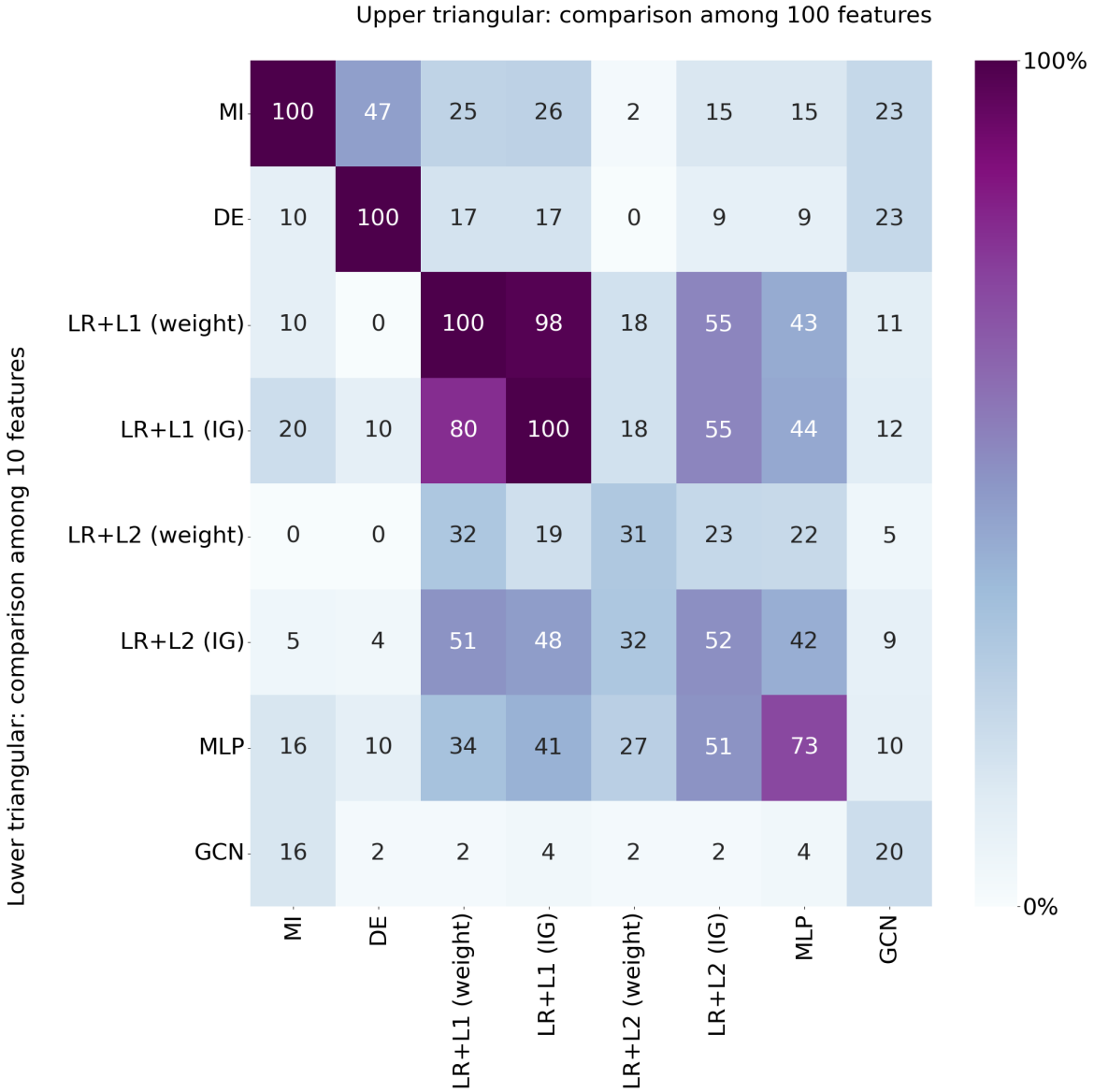
Dataset	LR+L1	LR+L2	MLP	GNN
ttg-breast	99.5 ± 0.0	98.5 ± 0.1	99.3 ± 0.2	99.1 ± 0.2
BRCA-pam	91.8 ± 0.0	87.4 ± 1.0	87.7 ± 1.3	87.8 ± 0.7

Average accuracy (%)

Healthy vs Tumor

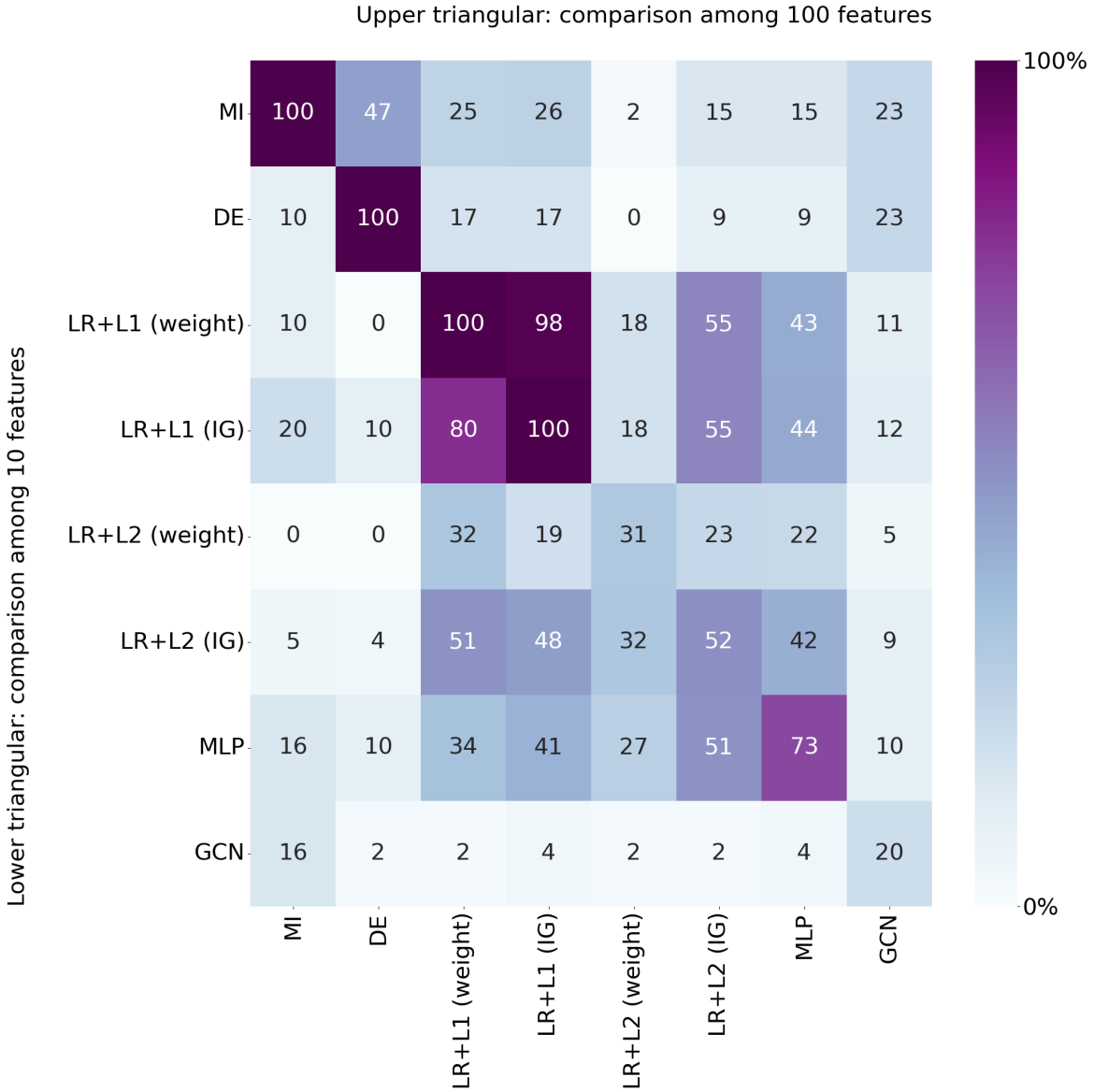
PAM50 breast cancer subtypes

Application

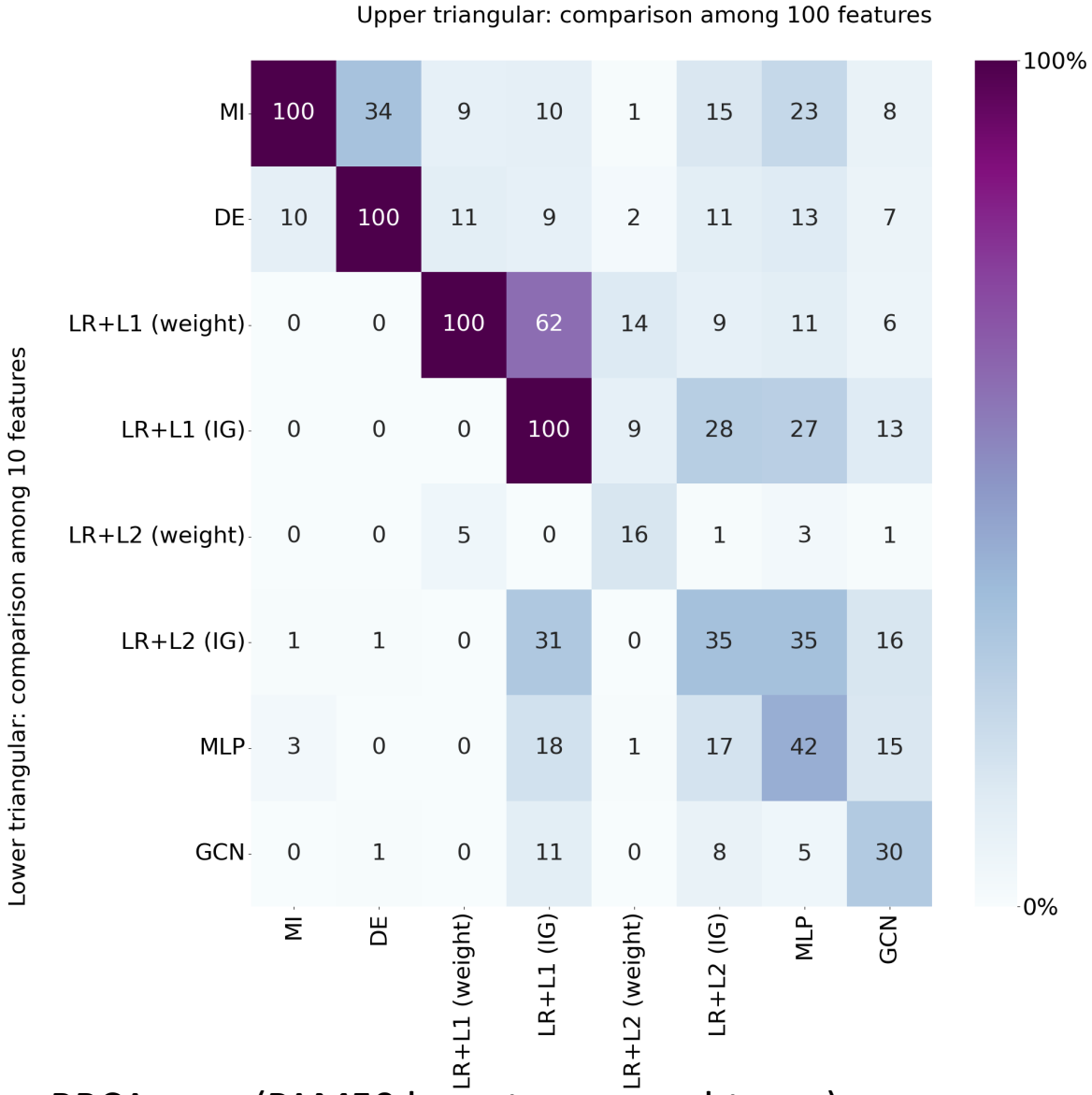


ttg-breast (healthy vs tumor)

Application

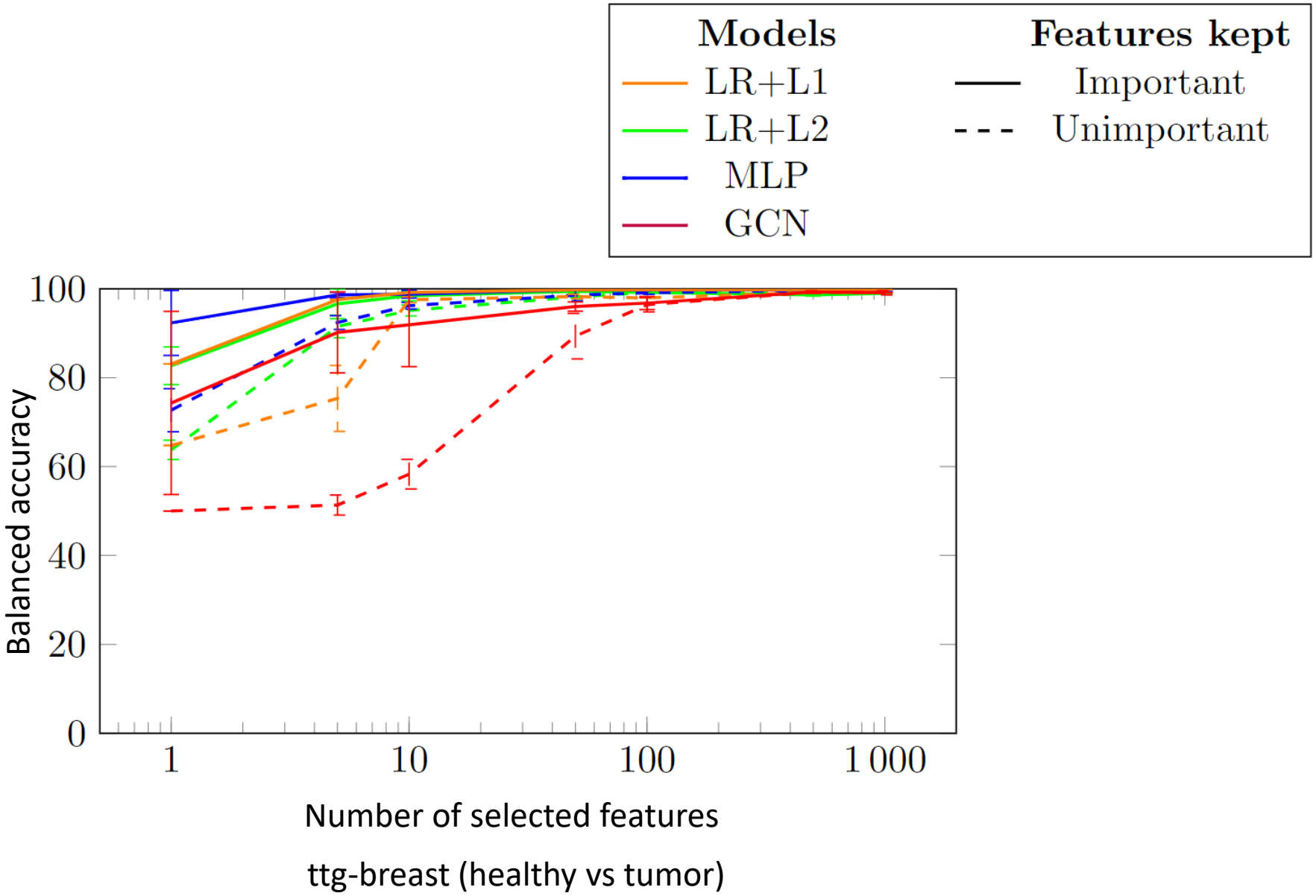


ttg-breast (healthy vs tumor)



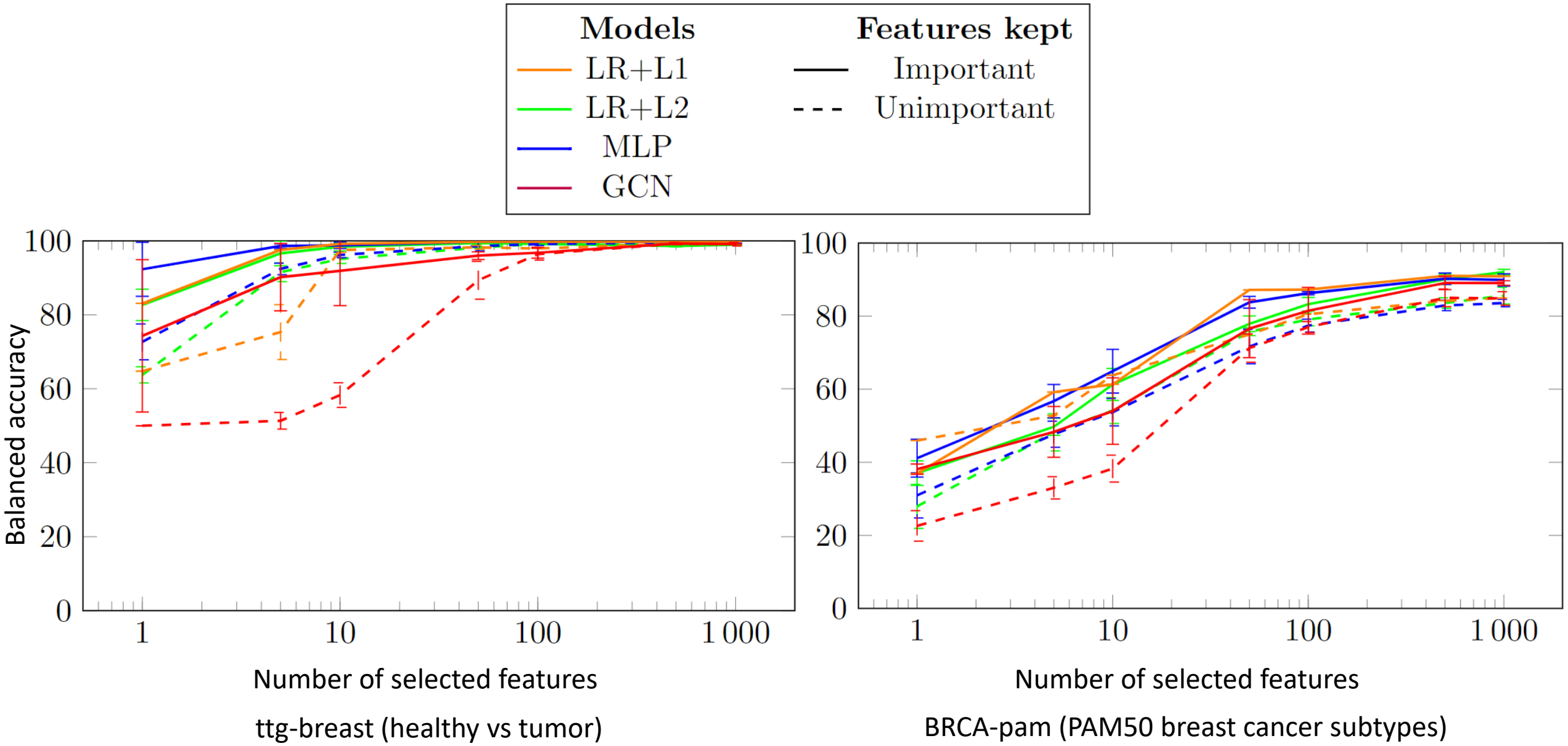
BRCA-pam (PAM50 breast cancer subtypes)

Application

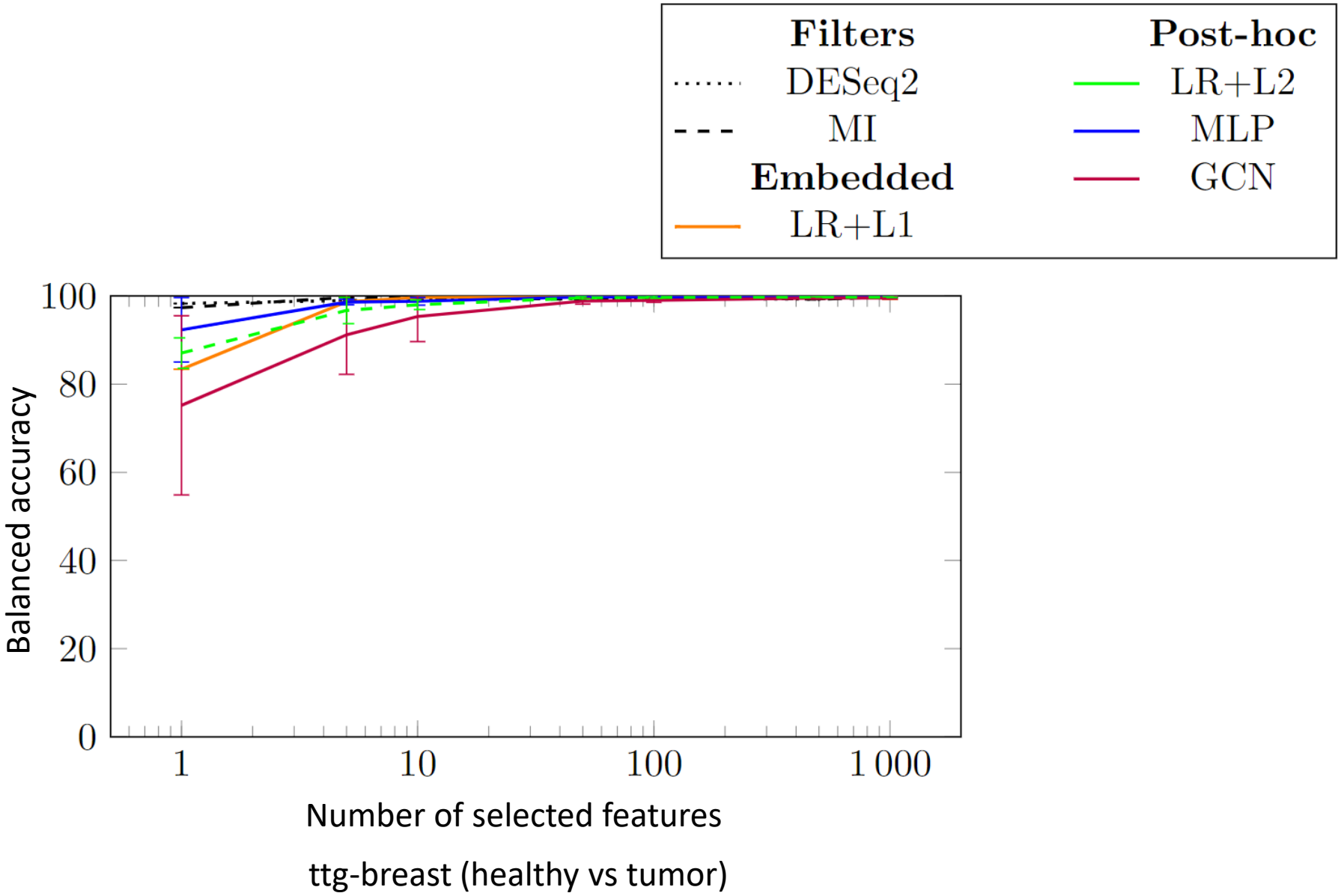


ttg-breast (healthy vs tumor)

Application

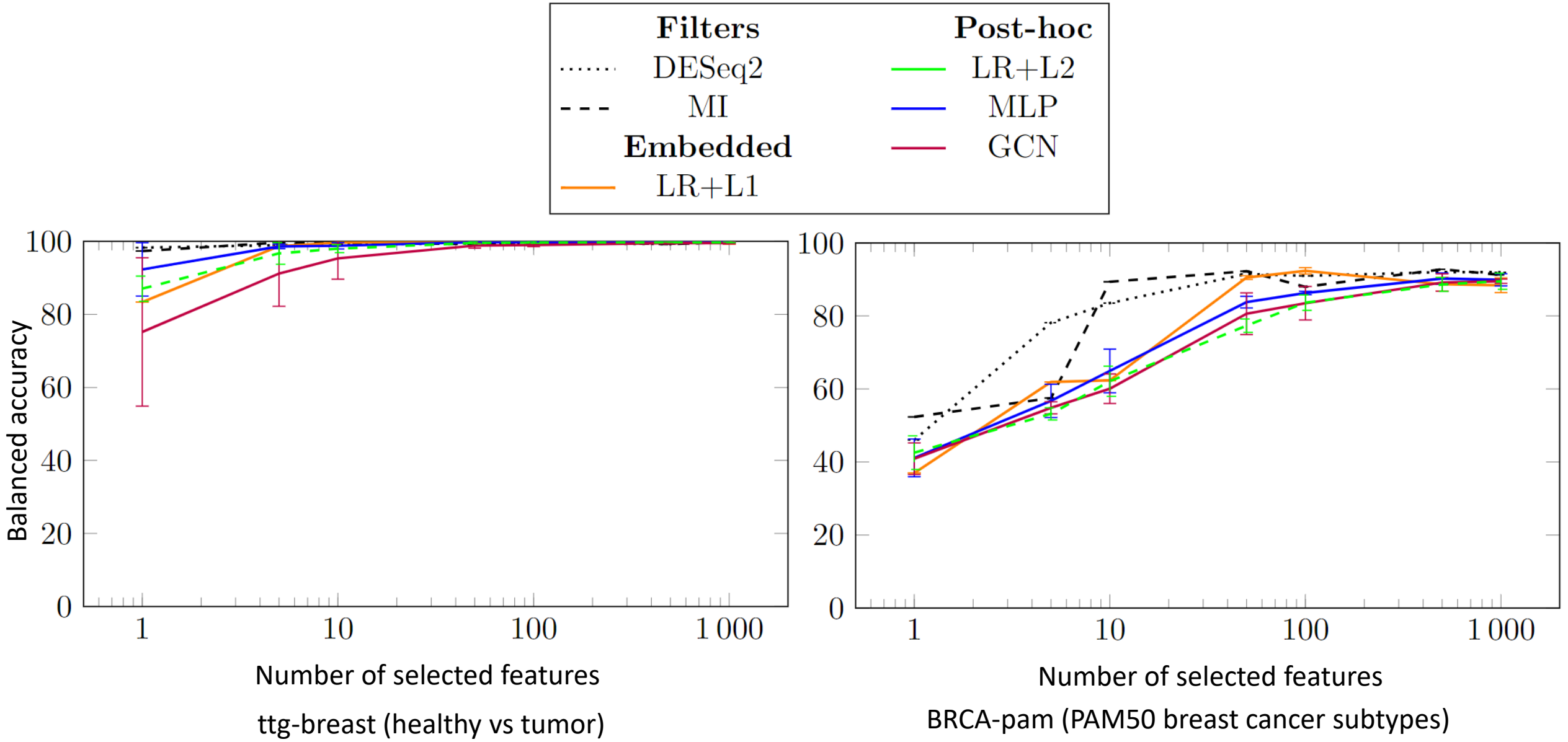


Application



ttg-breast (healthy vs tumor)

Application



Conclusion

Summary

- Challenges of functional genomics
- Machine learning methods selecting relevant genes
- Discussion on 2 datasets
  - ❖ Good classification performance, even with linear models.
  - ❖ Small sets containing the best-ranked genes are sufficient to achieve a good classification.
  - ❖ Sets generated by the different methods differ significantly.
  - ❖ Similar classification performance can be achieved with lower ranked genes.

Conclusion

Summary

- Challenges of functional genomics
- Machine learning methods selecting relevant genes
- Discussion on 2 datasets
  - ❖ Good classification performance, even with linear models.
  - ❖ Small sets containing the best-ranked genes are sufficient to achieve a good classification.
  - ❖ Sets generated by the different methods differ significantly.
  - ❖ Similar classification performance can be achieved with lower ranked genes.

Next steps

- Extension to other classification problems
- Extension to multiomics datasets
- Automatically extract explanations at a relevant scale