Studying Limits of Explainability by Integrated Gradients for Gene Expression Models

Myriam BONTONOU[1], Anaïs HAGET[2], Maria BOULOUGOURI[2], Jean-Michel ARBONA[1], Benjamin AUDIT[3], Pierre BORGNAT[3]

[1]Univ Lyon, ENSL, CNRS, LBMC, Lyon, France [2]LTS2 laboratory, EPFL, Lausanne, Switzerland
[3]Univ Lyon, ENS de Lyon, CNRS, Laboratoire de physique, Lyon, France

ENS DE LYON

EPFL

GraphNEx
Graph Neural Networks
for Explainable Artificial Intelligence

Context

- **Supervised learning problems** are formulated to **decipher** complex **molecular processes** driving cellular life.
  E.g. phenotype prediction from transcriptomic data (gene expression).
- Feature attribution **explainability methods** return the **input features** on which the **individual predictions** are predominantly based.
- These features are often **interpreted as the cause of the phenotype**.

Problematic: What is the relevance of biomarkers identified using explainability methods?

GitHub

Contributions

- Exploration of the relevance of the features identified by explainability.
- Definition of **quantitative metrics**.
- **Simulation of data**, with known discriminative features, mimicking genes.

PyTorch code `https://github.com/mbonto/XAI_for_genomics`.

Definition of quantitative metrics

Sample level [2]

How the prediction of a sample changes when features are set to zero?

- Network f, input x, modified input $\tilde{x}$.

$$\textbf{Prediction gap PG} = \max(f(x) - f(\tilde{x}), 0)$$

- **Area under PG** when an increasing number of features is set to zero with
  - most important removed first ⟶ **PG on Important features (PGI)**.
  - less important removed first ⟶ **PG on Unimportant features (PGU)**.

Model level

How the accuracy of a network changes when genes are set to zero?

- Accuracy obtained with the most important features for the whole dataset.
- Accuracy obtained with random features.

Do known discriminative features stand out among the identified features?

- Number of relevant features $\mathcal{F}$ among the identified features $\mathcal{M}$.

$$\textbf{Feature Attribution FA} = \frac{|\mathcal{F} \cap \mathcal{M}|}{|\mathcal{F}|}$$

Simulation of gene expression data

Generative probabilistic model called **Latent Dirichlet Association** [3].
⟶ Known for document generation.

Individual samples (documents) are generated with a **fixed number** N **of sequencing reads** (words) associated with **metabolic pathways** (subjects).

- Prior $\boldsymbol{\eta}_p$ proportion of genes expressed in pathway p.
- Prior $\boldsymbol{\alpha}_c$ proportion of pathways expressed in class c.
- Proportion of reads appearing in a pathway $\beta_p \sim \text{Dirichlet}(\boldsymbol{\eta}_p)$.

Generation of a sample s with N reads

Step 1 Draw the proportion of pathways $\boldsymbol{\theta}_s \sim \text{Dirichlet}(\boldsymbol{\alpha}_c)$.

Step 2 For each read i,
- pathway assignment $p_i \sim \text{Multinomial}(\boldsymbol{\theta}_s)$,
- drawn gene $g_i \sim \text{Multinomial}(\boldsymbol{\beta}_p)$.

Experimental setting

- **Simulated data** (9900) or Gene expression from **PanCan TCGA** (9680).
- **Classification problem** 33 classes.
- **Algorithm** Logistic Regression (LR), Multilayer Perceptron (MLP), Diffusion layer on a correlation graph (D).
- **Explainability method** Integrated Gradients (IG).

PanCan TCGA [1] - 16335 genes.
SIMU1/2 - 15000 genes. 1500 non-overlapping / 3000 overlapping pathways.

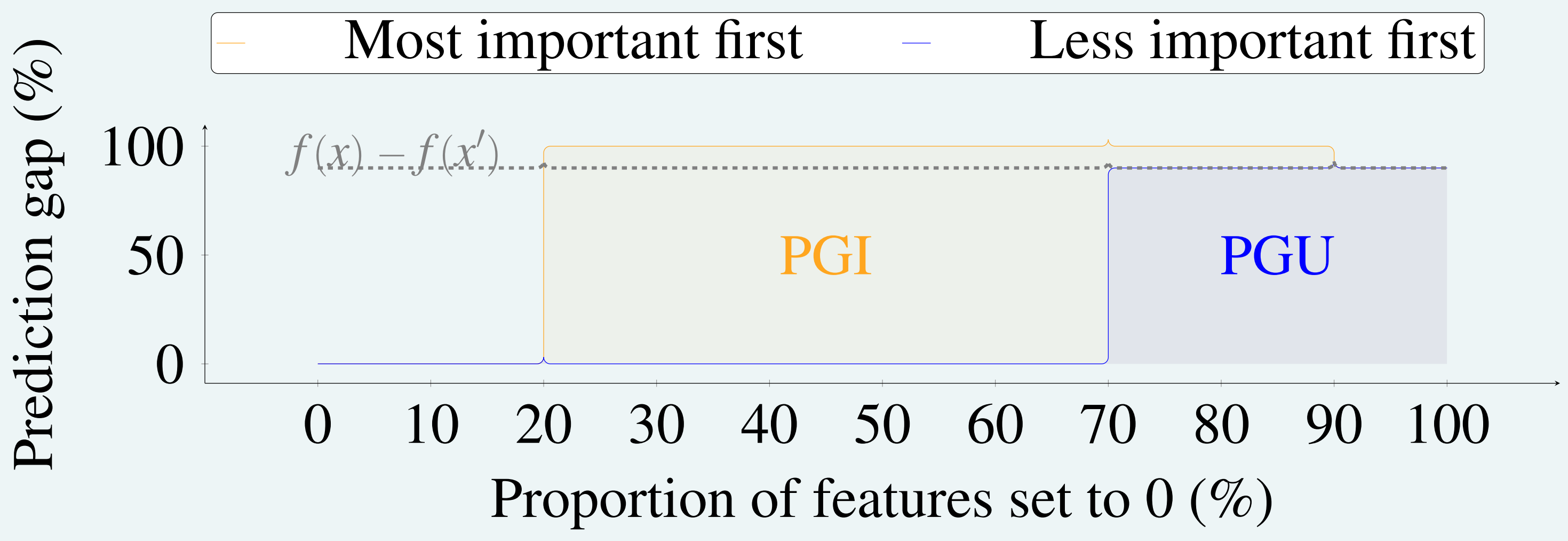

Figure 1 - Scheme describing the Prediction Gaps on Important features (PGI) and on Unimportant features (PGU).

Results

Table 1 - Explainability metrics averaged over test samples.

(a) Pan-Can TCGA

Network	LR	MLP	D + LR	D + MLP
Balanced accuracy (↑)	93.2%	94.7%	92.5%	94.3%
PGI (↑)	0.9570	0.9567	0.9750	0.9652
PGU (↓)	0.0035	0.0197	0.0053	0.0133

(b) Simulations

Dataset	SIMU1		SIMU2	
Network	LR	MLP	LR	MLP
Accuracy (↑)	99.5%	99.5%	99.9%	100%
PGI (↑)	0.9905	0.9714	0.9881	0.9842
PGU (↓)	0.0007	0.0036	0.0007	0.0039
FA (↑)	0.72	0.76	0.43	0.45
D + FA (↑)	1	1	0.96	1

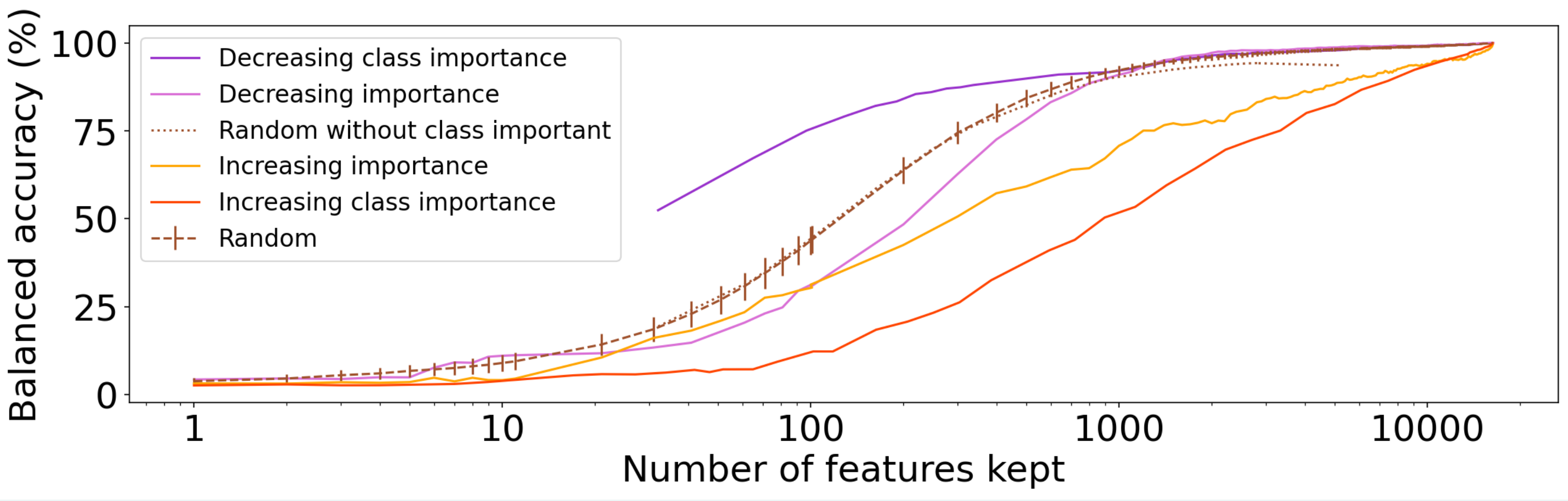

Figure 2 - Explainability metrics on Pan-Can TCGA data with LR.

Conclusion

- Evaluation of the **complexity of the real dataset PanCan TCGA**.
  - Set of 50 genes sufficient to classify each sample (PGU).
  - But not necessary (PGI).
- **Analyse** of the pertinence of the selected features **on simulated data** (FA).
- **Well behaved explanatory features** are **ambiguous**.

[1] `https://portal.gdc.cancer.gov/`.

[2] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.