Exploring Explainability and its Limits in Machine Learning in Gene Expression Models

Myriam BONTONOU[1,2], Jean-Michel ARBONA[1], Benjamin AUDIT[2], Pierre BORGNAT[2]

Univ Lyon, ENS de Lyon, CNRS, LBMC[1], Laboratoire de physique[2], Lyon, France

GraphNEx Graph Neural Networks for Explainable Artificial Intelligence

CHIST-ERA-19-XAI-006 for GRAPHNEX ANR-21-CHR4-000

Context

- **Supervised learning problems** can be formulated to **decipher** complex **molecular processes** driving cellular life. E.g. phenotype prediction from molecular data.
- **Explainability methods** return the **important features** (e.g. molecules) on which the predictions are predominantly based.
- These features are often interpreted as an **indication of the cause of the predictions**.

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

GitHub

Contributions

- Exploration of the relevance of features from a model's perspective.

⟶ Are the identified features sufficient/necessary for the predictions?

- Emphasis on **quantitative metrics** and **experiments**.
- Experiments on **gene expression data** and **simulated data** with known discriminative features mimicking genes.

Methodological framework

Objective - Assess the relevance of explanatory factors in a classification task.

- **Data sample:** $\mathbf{x} \in \mathbb{R}^F$.
- **Supervised model:** $f : \mathbb{R}^F \mapsto \mathbb{R}^C$.
  E.g. neural network, multilayer perceptron (MLP), logistic regression (LR).
- **Chosen explainability method:** integrated gradients [3]. Given a baseline $\mathbf{x}'$ and f_c the output of f associated with the class c of $\mathbf{x}$, the score attributed to the i^{th} feature of $\mathbf{x}$ is:

$$\phi_i(\mathbf{x}) = (\mathbf{x}_i - \mathbf{x}'_i) \int_{\alpha=0}^{1} \frac{\partial f_c(x)}{\partial \mathbf{x}_i}\bigg|_{x=\mathbf{x}'+\alpha(\mathbf{x}-\mathbf{x}')} d\alpha \,.$$

  Note: contrary to classical feature selection methods, this method assigns a different importance value to each of the features within each example.
- **Feature rankings**
  - **Local (for a data sample x):** features ranked according to the scores $\phi_i(\mathbf{x})$.
  - **Global (for all data samples):** features ranked according to the scores ϕ_i^g. Given M data samples and the Euclidean norm $\|\cdot\|$,

$$\phi_i^g = \frac{1}{M} \sum_{k=1}^{M} \frac{\phi_i(\mathbf{x}^k)}{\|[\phi_1(\mathbf{x}^k), \ldots, \phi_F(\mathbf{x}^k)]\|}$$

Metrics measuring ranking relevance

To what extent do predictions vary when some features are masked?

Given a feature ranking, the values of the features of a sample $\mathbf{x}$ *can be replaced by the values of the reference* $\mathbf{x}'$ *one after the other.*

Let us denote $\tilde{\mathbf{x}}_p$ the data sample $\mathbf{x}$ containing p masked variables and e_p the prediction error with p masked features $e_p = \max(f(\mathbf{x}) - f(\tilde{\mathbf{x}}_p), 0)$. The **prediction gap (PG)** is the area under the curve e_p as a function of p:

$$\text{PG} = \sum_{p=1}^{F} \frac{\max(f(\mathbf{x}) - f(\tilde{\mathbf{x}}_p), 0)}{F} \,.$$

- **PGI** Masking the most important features first (decreasing scores).
- **PGU** Masking the least important features first (increasing scores).
- **PGR** Masking the features in a random order.

Do known discriminative features stand out among the identified features?

In the case of simulated data, the concordance of a ranking with the features that are really important for classifying an sample can be measured.

Consider a set $\mathcal{E}_r$ of truly important features and a set $\mathcal{E}_i$ consisting of the $|\mathcal{E}_r|$ most important features identified by integrated gradients, the **feature agreement (FA)** is:

$$\text{FA} = \frac{|\mathcal{E}_r \cap \mathcal{E}_i|}{|\mathcal{E}_r|}.$$

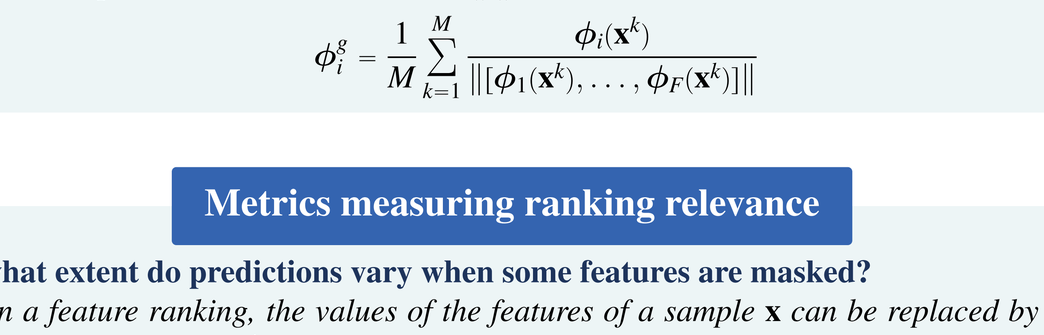


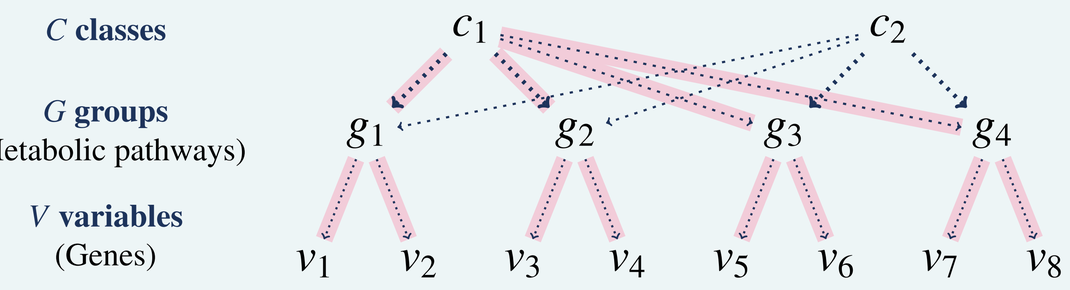
Figure 1 - Scheme describing the metrics PGI and PGU.

Experimental setting

- Gene expression datasets from TCGA [1]
  - PanCan: 33 classes (cancer type), 9680 samples, 16335 features.
    → Baseline: null vector.
  - BRCA: 2 classes (healthy vs tumours), 1210 samples, 58274 features.
    → Baseline: average of the healthy samples.
- Simulation (see Fig.2): 33 classes, 9900 samples, 15000 features, 370 informative per class.
  → Baseline: null vector.

The scores are computed for each sample correctly classified of the test set.
Remark: for BRCA, the scores are computed only on tumour samples.

Simulated dataset



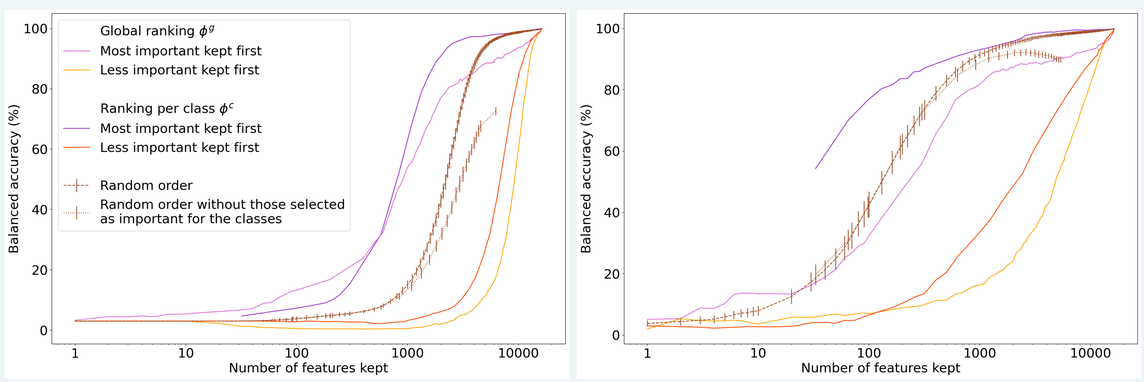
Prior on the proportion of groups within a class c $\boldsymbol{\alpha}_c \in \mathbb{R}_+^G$.
Prior on the proportion of variables within a group g $\boldsymbol{\eta}_g \in \mathbb{R}_+^V$.
Proportion of variables within a group g $\boldsymbol{\beta}_g \sim \text{Dirichlet}(\boldsymbol{\eta}_g)$.
Proportion of groups within a data sample s of class c_1 $\boldsymbol{\theta}_s \sim \text{Dirichlet}(\boldsymbol{\alpha}_{c_1})$.

Figure 2 - LDA model used to simulate a dataset [2]. A data sample s of class c_1 is generated by drawing a large number of variables using the two multinomial distributions associated to the sample: 1/ $g \sim \text{Multinomial}(\boldsymbol{\theta}_s)$, 2/ $v \sim \text{Multinomial}(\boldsymbol{\beta}_g)$.

Results

Table - Average metrics (%) and standard deviations calculated on the test samples of each dataset after training 10 times each model with random initialisations.
Each PGR is averaged over 30 random rankings.

	MLP			LR		
Dataset	PanCan	BRCA	Simulation	PanCan	BRCA	Simulation
Balanced accuracy (↑)	94.5 ± 0.3	99.6 ± 0.1	99.9 ± 0.1	93.7 ± 0.4	96.6 ± 0.3	99.8 ± 0.1
PGR	23.2 ± 0.6	53.5 ± 2.1	25.5 ± 0.2	3.1 ± 0.2	88.8 ± 0.2	3.7 ± 0.1
Local ranking ϕ						
PGI (↑)	96.1 ± 0.2	98.7 ± 0.3	98.5 ± 0.1	96.0 ± 0.2	99.9 ± 0.1	99.1 ± 0.1
PGU (↓)	4.8 ± 1.7	0.9 ± 0.2	0.3 ± 0.1	0.7 ± 0.7	1.1 ± 0.1	0.1 ± 0.1
FA (↑)	-	-	74.2 ± 0.3	-	-	72.8 ± 0.4
Global ranking ϕ^g						
PGI (↑)	59.0 ± 1.8	98.2 ± 0.3	49.9 ± 0.3	33.4 ± 1.2	99.9 ± 0.1	42.4 ± 0.5
PGU (↓)	17.2 ± 1.3	1.6 ± 0.3	20.0 ± 0.2	9.4 ± 0.4	1.9 ± 0.1	4.6 ± 0.3
FA (↑)	-	-	100.0 ± 0.1	-	-	99.5 ± 0.1



(a) MLP (b) LR

Figure 3 - Curves obtained on the test samples of the PanCan dataset.
The ϕ^c are calculated in the same way as the ϕ^g, considering only the samples in class c. To plot the curves, the features appearing first in the rankings ϕ^c of each class c are kept. The random curves are averaged over 30 trials. Error bars represent standard deviations.

Remark: the balanced accuracy is the average of recalls calculated for each class.

Conclusion

- Evaluation of the **complexity of two real datasets from a model's perspective**.
  E.g. MLP on PanCan.
  - *For each data sample, a set of around 784 features is sufficient for prediction.*
    → *Keeping the 784 most important features enables to maintain good predictions. Keeping less does not (local PGU).*
    → *After masking the first 637 features, predictions deteriorate (local PGI).*
  - *Similarly, for the whole dataset, keeping a set of 2810 features is sufficient (global PGU). However, this set is not necessary.*
    → *Predictions are degraded once the 6697 most important features have been masked (global PGI). Masking only the 2810 features do not deteriorate the prediction.*
- **Analyse** of the pertinence of the selected features **on simulated data** (FA).
- **Well behaved explanatory features** are **ambiguous**.

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

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

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

[3] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.