# From Complexity to Clarity: Evaluating Explainability of Biomedical Machine Learning Models

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Marta González Mallo [1], Alfonso Valencia [1,2], **Davide Cirillo** [1]

[1] Barcelona Supercomputing Center, Barcelona, Spain; [2] Institució Catalana de Recerca i Estudis Avançats ICREA, Barcelona, Spain

EVENFLOW
Horizon Europe
GA 101070430

By providing transparent and interpretable insights, **explainable artificial intelligence (XAI)** methods enable better understanding and trust in the predictions made by complex **machine learning (ML)** models. In the context of the Horizon Europe project **EVENFLOW** (GA 101070430), we conducted an evaluative study to assess the explanations offered by a number of popular XAI methods. We implemented an **evaluation framework** that facilitates a deeper understanding of the factors influencing the decisions made by complex ML models, while also assessing and contrasting various XAI methods based on **recommendations** that we provide.

**USE CASE:** The study focuses on using ML models to analyze **breast cancer RNASeq data from TCGA** [1]. Out of 35 targets, six were selected for **classification tasks** (**Fig.1A-B**) after filtering out specific classes (<75 samples, non-informative labels, etc.). The classification difficulty was quantified using an **aggregation score** (**Fig.1C**), i.e. the number of closest neighbors from the same class for each instance, weighted inversely to the class percentage and the number of neighbors, and averaged across all instances. Four ML methods of varying complexity (logistic regression, decision tree, random forest, XGBoost) were applied. **XGBoost**, the least interpretable ML method, was selected to evaluate XAI methods on **histological type** classification (low aggregation score) (**Fig.1D**).

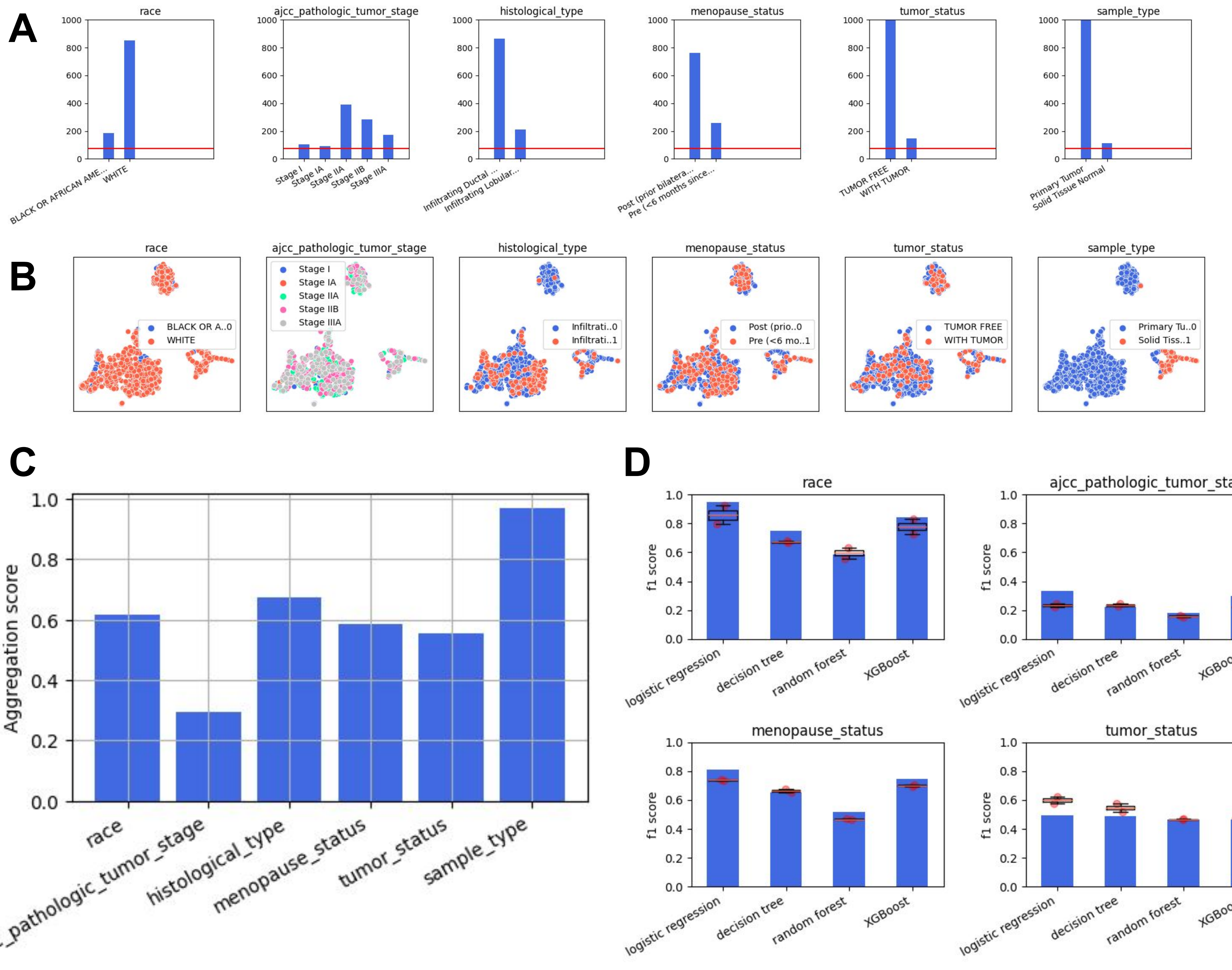


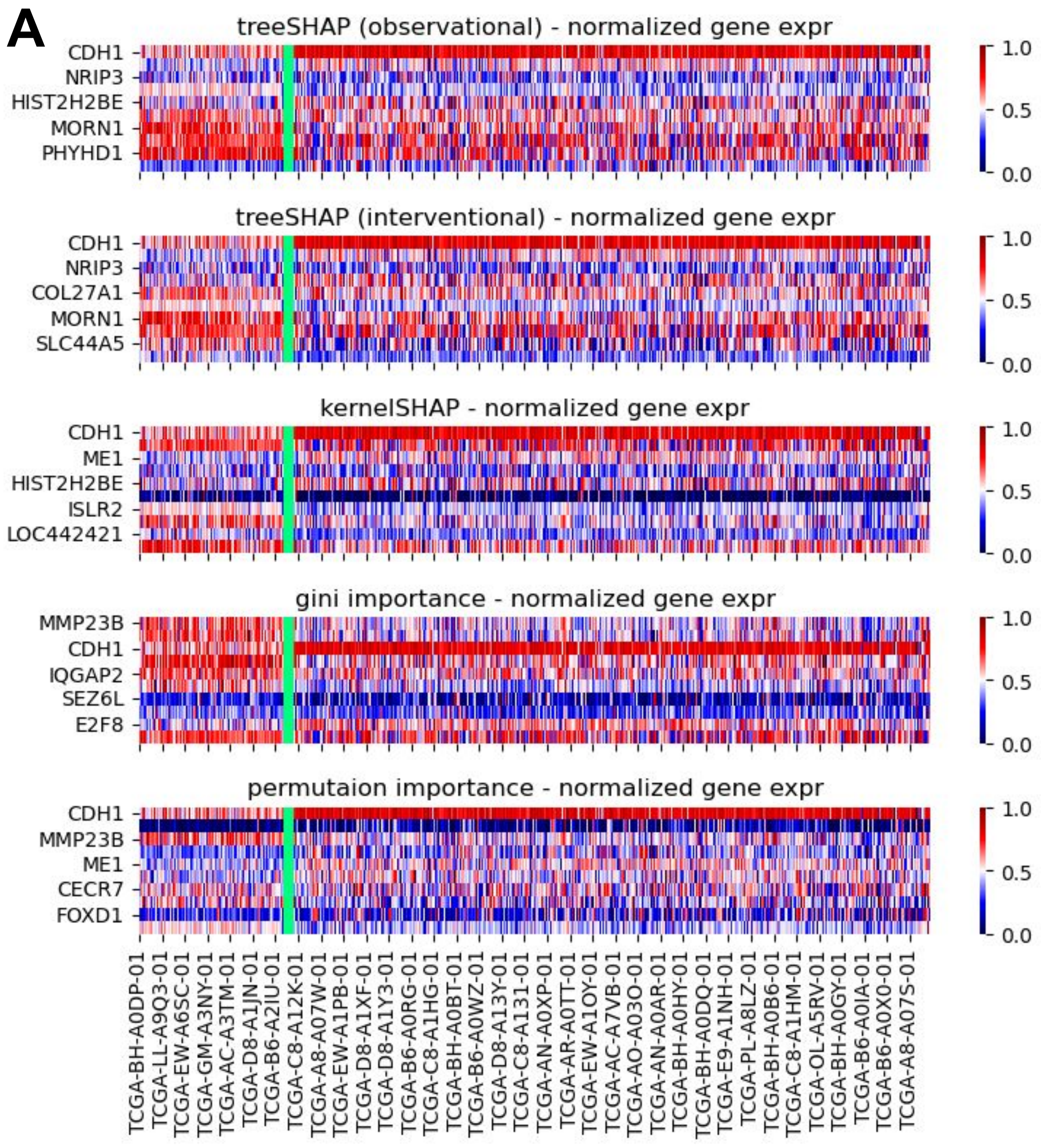
*Figure 1. A) Six selected targets for classification from TCGA RNASeq breast cancer dataset. Red line indicates samples threshold. B) Distribution of labels in each target (PaCMAP plots [2]). C) Aggregation score of the selected targets. While sample type classification has a low level of difficulty, AJCC staging classification is a much harder problem. D) F1 scores of four ML methods with increasing complexity.*

| XAI method | Description | Type of explanations | Typical caveats |
|---|---|---|---|
| Gini Index | It measures the impurity decrease from feature splits, weights it by sample count per node, and averages across all ensemble trees. | Global attribution | Variables with **high cardinality** are prone to inflation as they possess a greater number of potential cutpoints. |
| Permutation Importance | It permutates the features and observes the error in the generated samples (the higher the error, the higher the importance) | Global attribution | **Correlated features** can lead to unrealistic instances and result in smaller assigned importance. |
| KernelSHAP [3] | It leverages a kernel-weighted linear model with perturbations of the input sample to estimate Shapley values for each feature. | Local attribution | Computationally expensive and requires **large samples** for accurate Shap value estimates. |
| TreeSHAP [4] (interventional) | It leverages the decision path to efficiently compute exact Shapley values for each feature. In the interventional variant, missing features in the combinations are estimated disregarding the joint data distribution. | Local attribution | Limited to tree-based models. It has the potential to generate **unrealistic instances**. |
| TreeSHAP [4] (observational) | As above. In the observational variant, missing features in the combinations are estimated based on the joint data distribution | Local attribution | Limited to tree-based models. It has the potential to distribute importance across **correlated feature** and attribute importance to neglected ones. |

*Table 1. Evaluated XAI methods. Global attribution determines feature contributions across the dataset, while local attributions focus on individual predictions.*

**XAI METHODS ASSESSMENT:** The evaluated XAI methods are reported in **Table 1**. The proposed evaluation framework revolves around analyzing sets of genes attribute the highest importance by the assessed XAI methods. This includes visually inspecting their expression values (**Fig.2A**), assessing their predictive power (**Fig.2B**), analyzing their overlap in pairwise comparisons (**Fig.2C**), and exploring their functional relationships (**Fig.2C**). These analyses provide insights into XAI models and differences in importance attribution.

[1] https://www.cancer.gov/tcga
[2] Wang et al. (2021) JMLR:v22:20-1061
[3] Lundberg & Lee (2017) NIPS:30
[4] Lundberg et al. (2020) Nat Mach Intell.
[5] https://reactome.org/
[6] Menche et al. (2015) Science



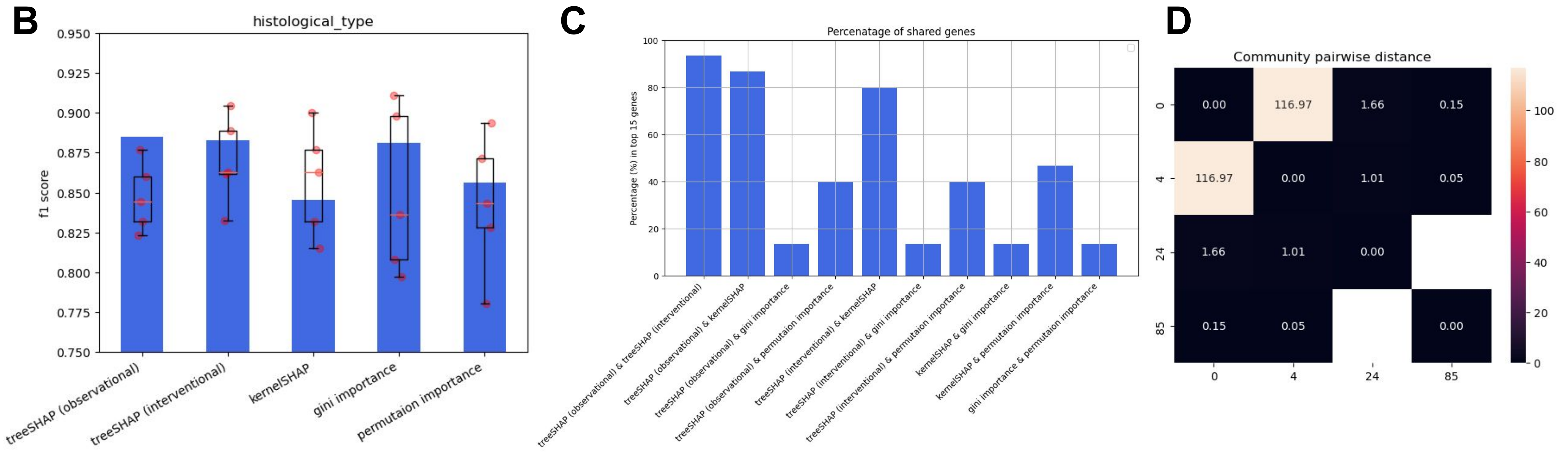


*Figure 2. XAI evaluations based on gene sets with highest importance. A) Expression heatmap of important genes; B) Assessment of important gene predictive power; C) Pairwise overlap analysis of important genes; D) Association of important genes by separation (distance) of label-propagated Reactome communities [5, 6].*

## RECOMMENDATIONS FOR THE USE AND EVALUATION OF XAI METHODS

- ***DO USE A SLEDGEHAMMER TO CRACK A NUT*** - Choose a model appropriate for the task's complexity.
- ***ALL THAT GLITTERS IS NOT GOLD*** - Avoid hastily selecting the first XAI method you come across.
- ***THE DEVIL IS IN THE DETAILS*** - Scrutinize outcomes and algorithmic intricacies of different XAI methods.
- ***LOOK BEFORE YOU LEAP*** - Employ comprehensive quality metrics to thoroughly assess the explanations.
- ***MEASURE TWICE, CUT ONCE*** - Rely on expert guidance and accumulated knowledge for validation.