



HAL
open science

Extensive benchmark of machine learning methods for quantitative microbiome data

Sébastien Fromentin, Florian Plaza Oñate, Nicolas Maziers, Samar Berreira Ibraim, Guillaume Gautreau, Oscar Gitton-Quent, Manolo Laiola, Soufiane Maski, Raphaëlle Momal, Florence Thirion, et al.

► **To cite this version:**

Sébastien Fromentin, Florian Plaza Oñate, Nicolas Maziers, Samar Berreira Ibraim, Guillaume Gautreau, et al.. Extensive benchmark of machine learning methods for quantitative microbiome data. JOBIM, Jul 2021, Paris, France. hal-04163473

HAL Id: hal-04163473

<https://hal.science/hal-04163473v1>

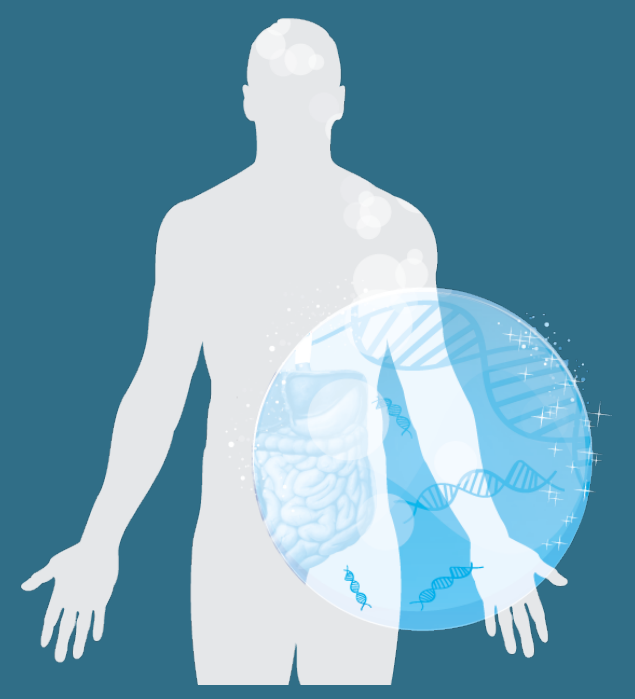
Submitted on 17 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



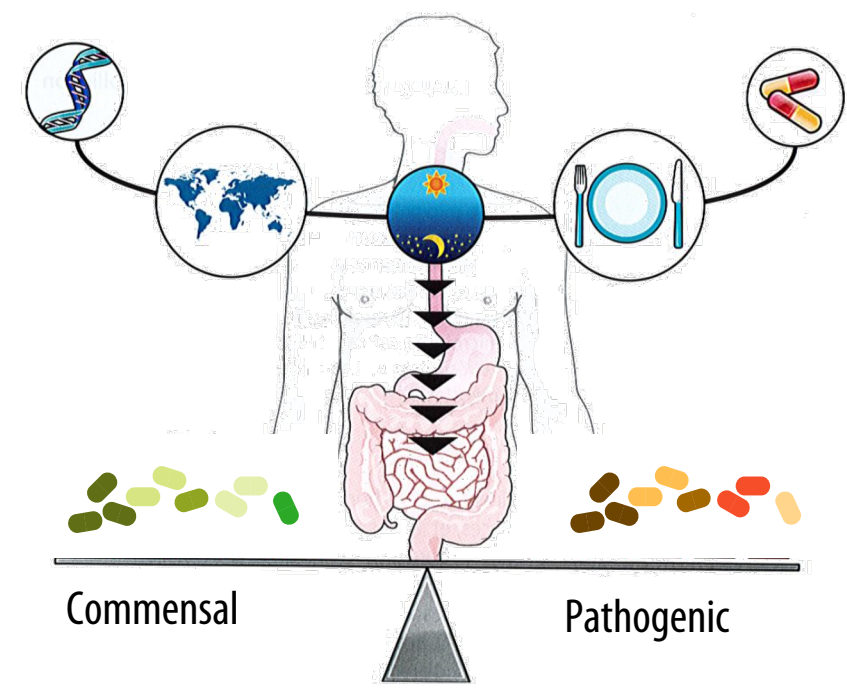
Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License



Sébastien Fromentin¹, Florian Plaza Oñate¹, Nicolas Maziers¹, Samar Berreira Ibraim¹, Guillaume Gautreau¹, Oscar Gitton-Quent¹, Manolo Laiola¹, Soufiane Maski¹, Raphaëlle Momal¹, Florence Thirion¹, Franck Gautier¹, Nicolas Pons¹ and Magali Berland¹

Introduction

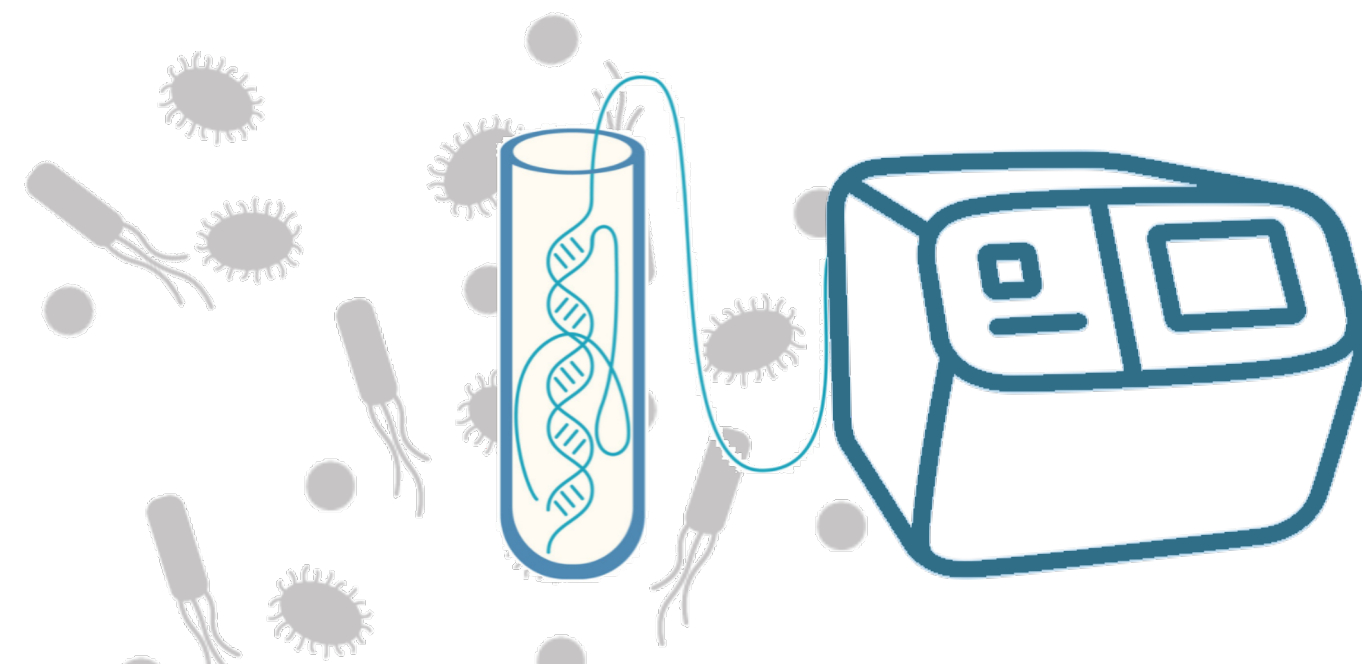
Microbial dysbiosis is associated with many chronic diseases and often correlates with severity.



Prediction of phenotypic features can help to stratify patients:

- ▶ Clinical status (classification)
- ▶ Disease severity (regression)

Which machine learning (ML) method should be used?



Omics technologies – in particular shotgun metagenomics – allows a highly precise microbiome profiling

Miaw

Microbiome artificial intelligence workflow

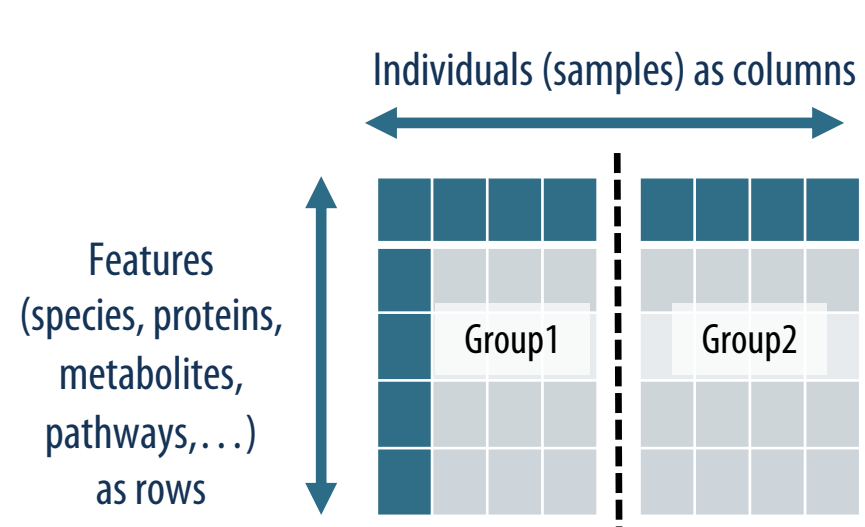
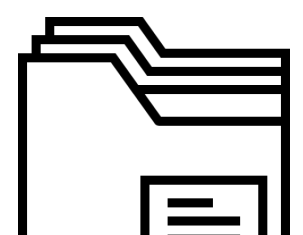
The code is based in the R package Caret (topepo.github.io/caret) to train and tune ML models. The Activeeon Proactive solution was used to efficiently distribute the computing load on multiple servers.



Machine Learning workflow

Input data

Abundance tables and metadata



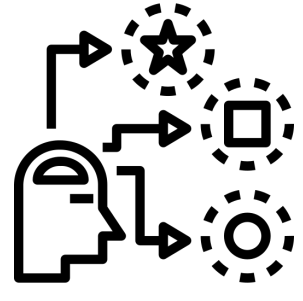
Standard preprocessings

- Downsizing
- Normalization by gene length

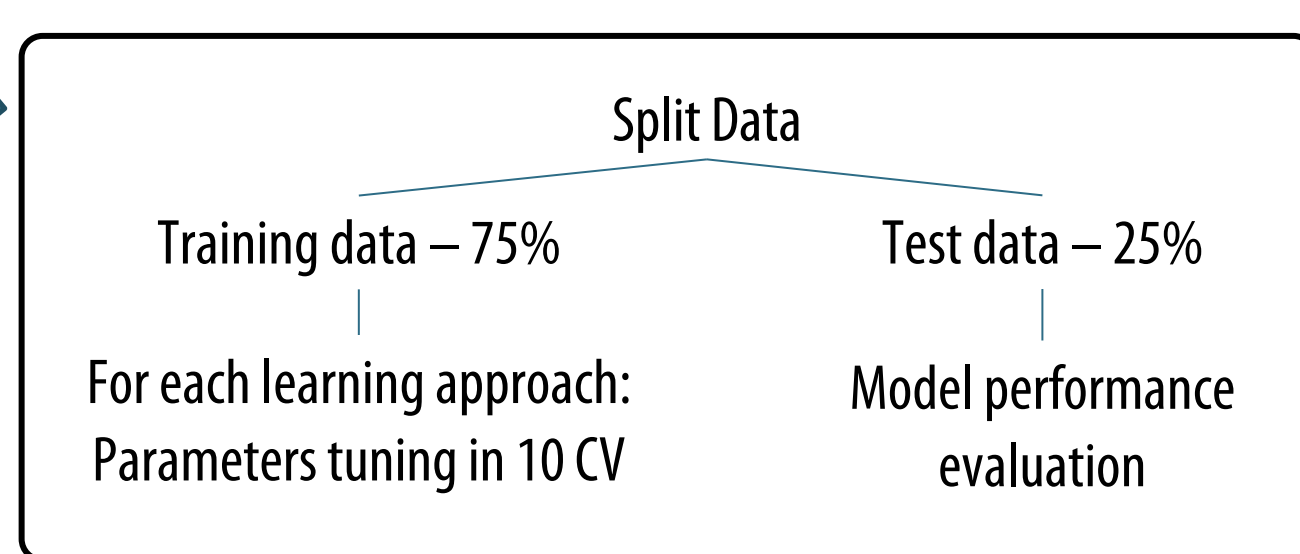
Specific preprocessings

- Near zero variance
- Linear combos / correlated predictors

Validation scheme

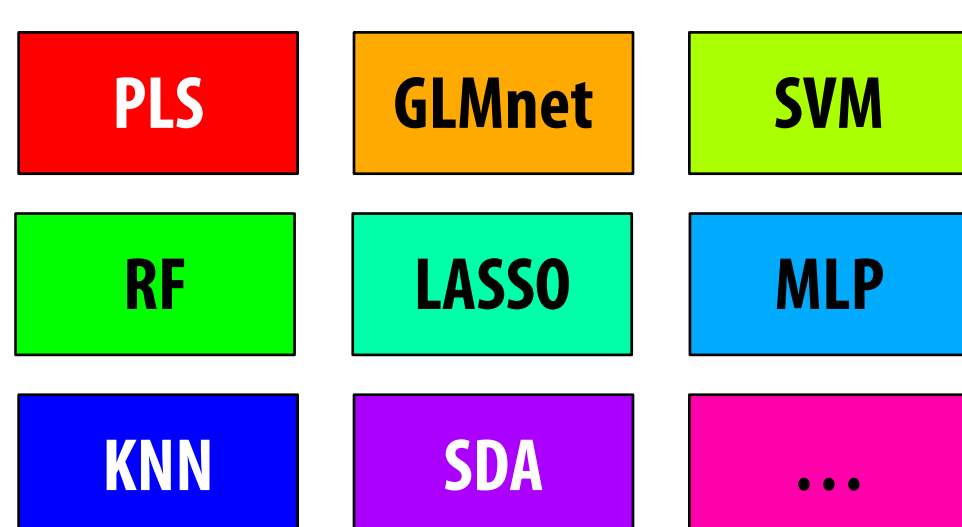


N = 100



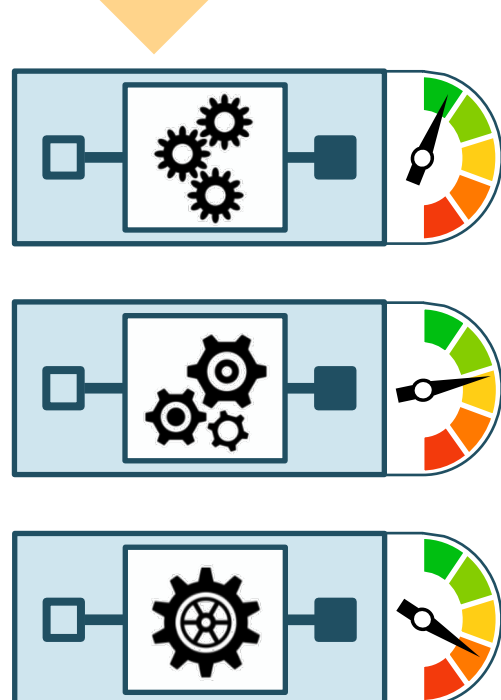
- Tuneable parameters

Benchmarked methods

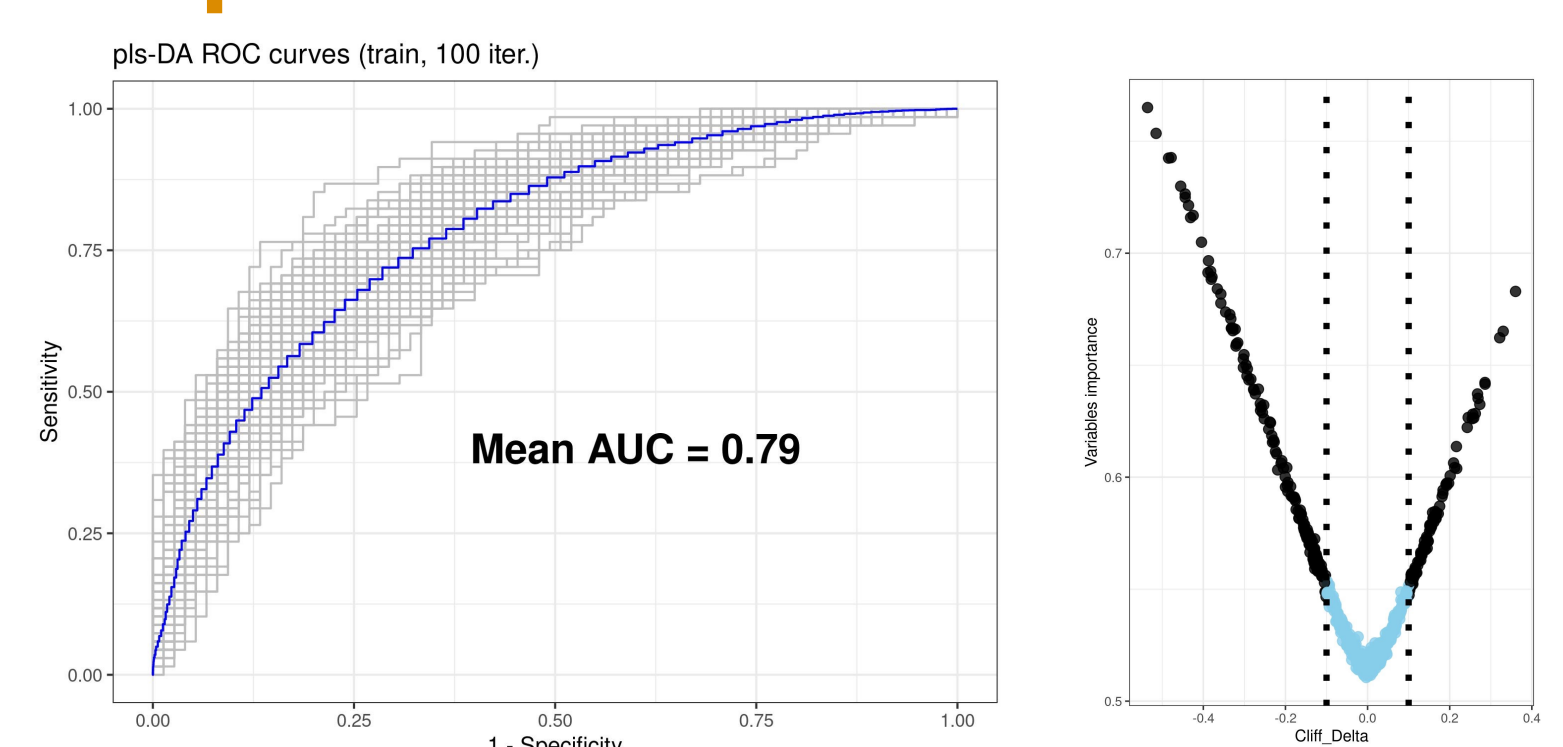


- 69 ML classification methods
- 53 ML regression methods

Outputs



- Indicators of predictive performance
- Variable importance

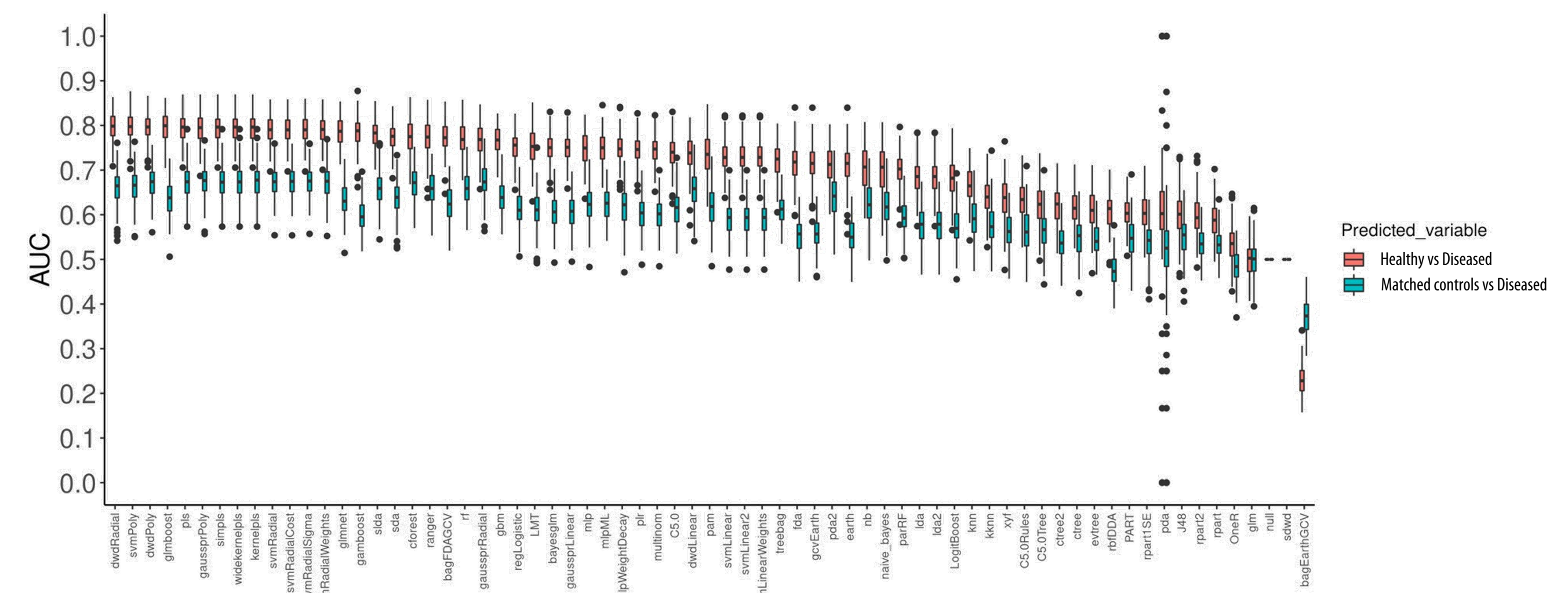


Results

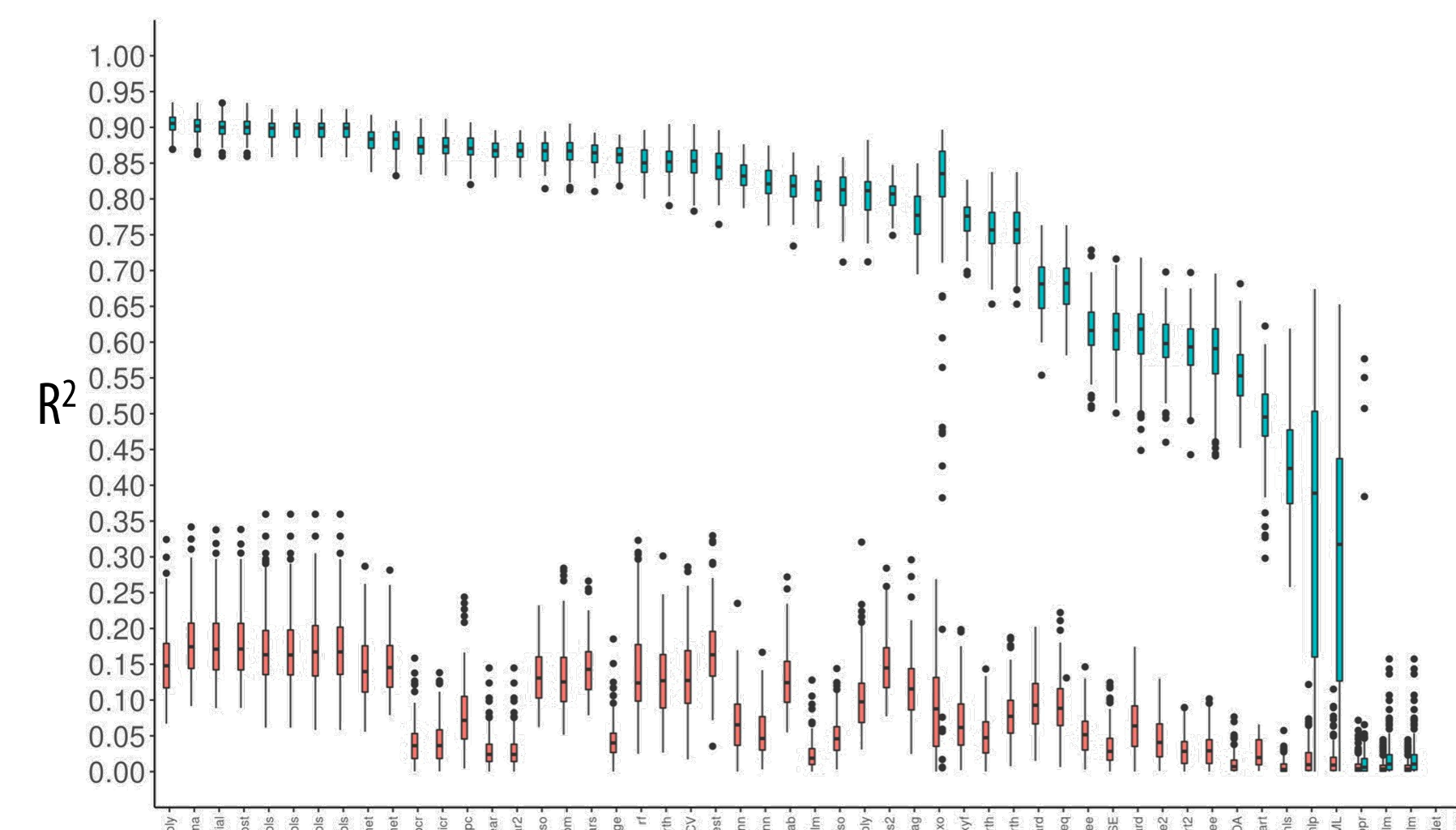
Study Case	Liver cirrhosis	Coronary diseases	Coronary diseases (deconfounded)	Coronary diseases	Vegan	Anorexia	Chronic kidney disease
Sample size	237	578	744	99	150	147	240
Omic type	Metagenomics	Metagenomics	Metagenomics	Metaproteomics	Metagenomics	Metagenomics	Metagenomics
# features	436	729	729	755	619	580	754
Best mean AUC	0.95	0.80	0.68	0.82	0.93	0.89	0.63

Example of results (coronary diseases - metagenomics):

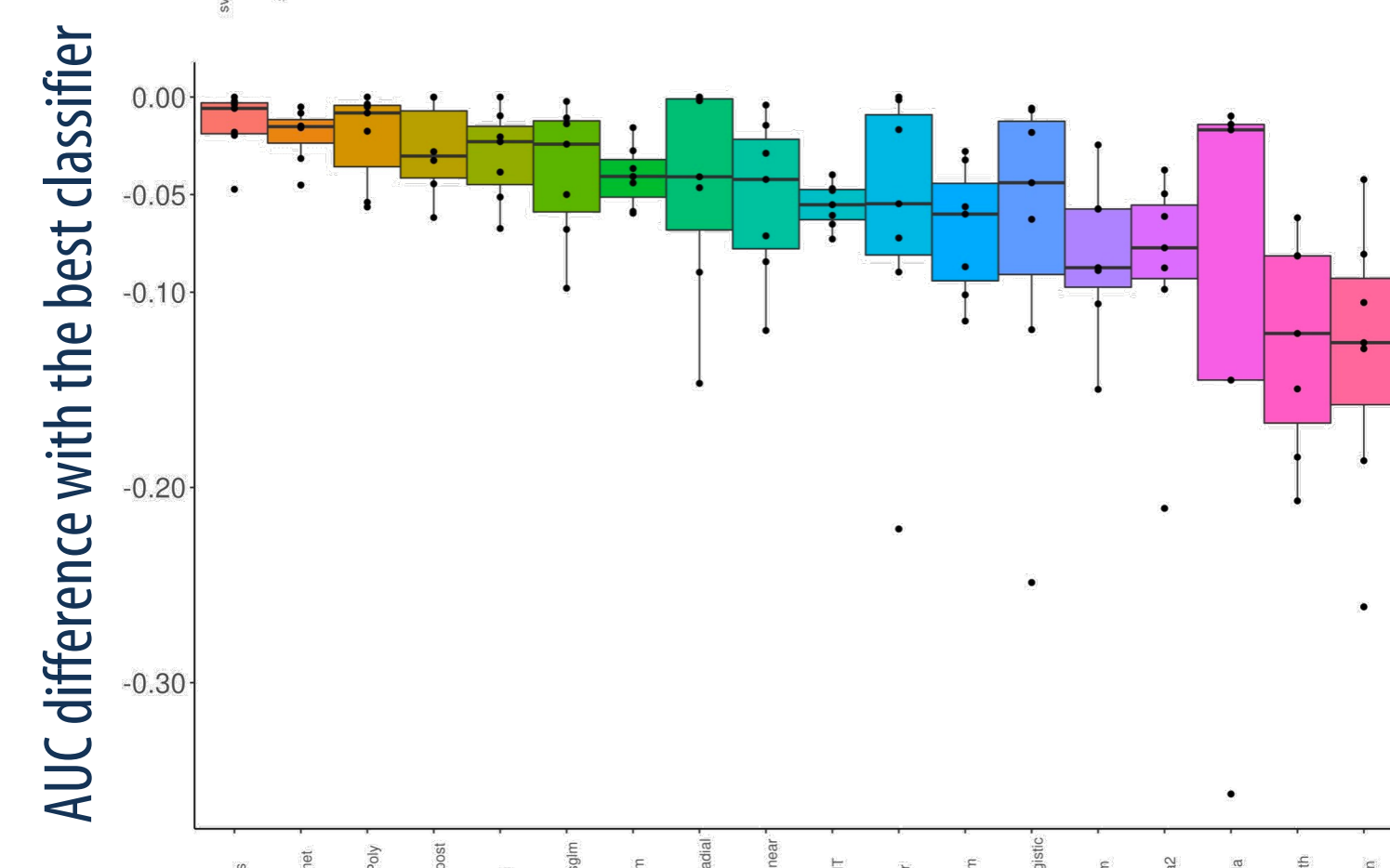
- Many classification models provide similar performances both for an easy prediction task (Healthy vs Diseased) and a difficult prediction task (individuals matched on the principal confounders [Age, BMI, Diabetes, Gender] vs Diseased)



- Regression model performances are more heterogeneous than those of classification models



- Performances on 7 datasets
- PLS, SVM (poly) and GLMnet were the best predictive approaches to classify microbiome data
- Some methods lack of stability



Conclusion

- Miaw allows an easy benchmark across methods and datasets
- New methods can be easily added with the caret wrapping functions

- Depending on the dataset, some methods perform better or worse; so far PLS, GLMnet and SVM performed generally well
- Results will be extended to other public datasets (database under construction)

