

Gpredomics: rapid, interpretable and accurate prediction models for compositional data

Raynald DE LAHONDÈS¹, Louison LESAGE^{1,2}, Vadim PULLER¹, Fabien KAMBU MBUANGI^{3,4}, Eugeni BELDA^{3,4}, Jean-Daniel ZUCKER^{3,4}, Edi PRIFTI^{3,4}

¹ GMT Science, 75013 Paris, France.
² Univ Rouen Normandie, Normandie Univ, Master Bioinformatique, F-76000 Rouen, France.
³ IRD, Sorbonne University, UMMISCO, 32 Avenue Henri Varagnat, F-93143 Bondy, France.
⁴ Sorbonne Université, INSERM, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière, France.

Artificial intelligence is revolutionizing biological data analysis, expanding our understanding of health and disease. However, metagenomic studies often involve high-dimensional data with limited sample sizes, leading many state-of-the-art methods to overfit and generalise poorly. Moreover, traditional AI models are often too complex for use in medical or regulatory settings. We previously introduced the Predomics approach, proposed as an R package, for building interpretable models that match or outperform standard methods in predicting outcomes from microbiome datasets (Prifti *et al.*)¹. We now present Gpredomics, a Rust-based, GPU-compatible version that is ~1000× faster and incorporates a number of new features (Fig. 1).

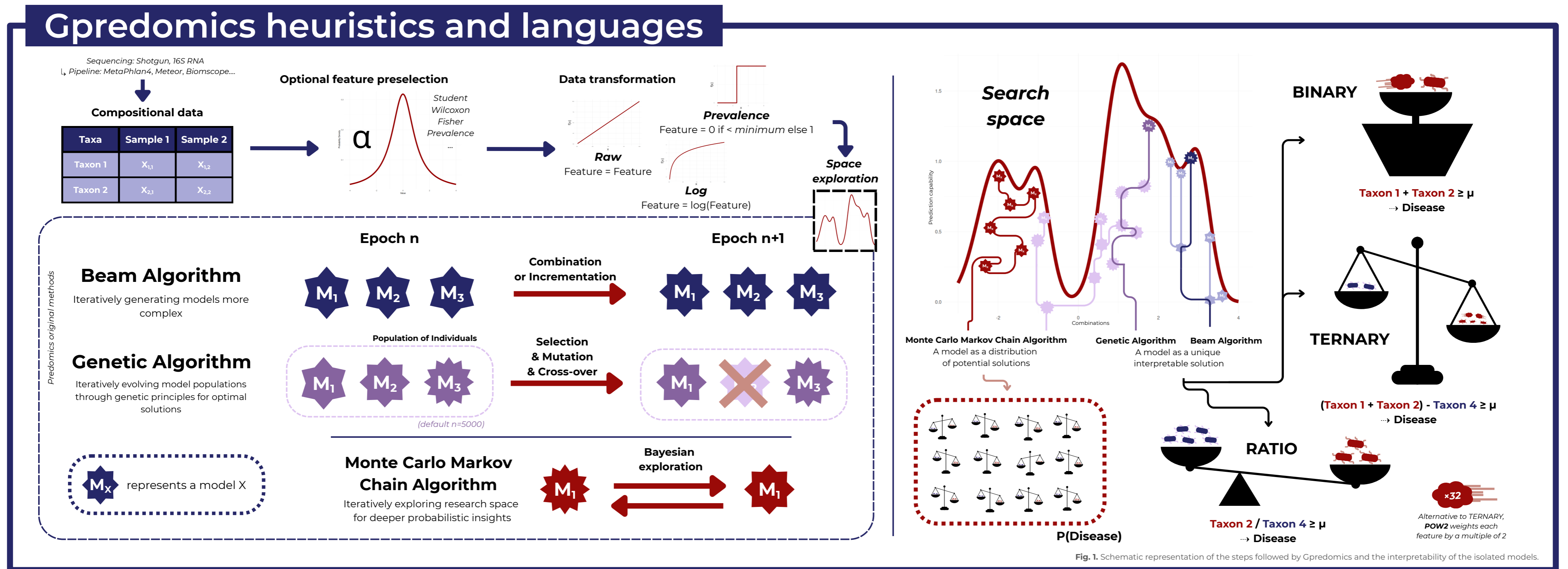
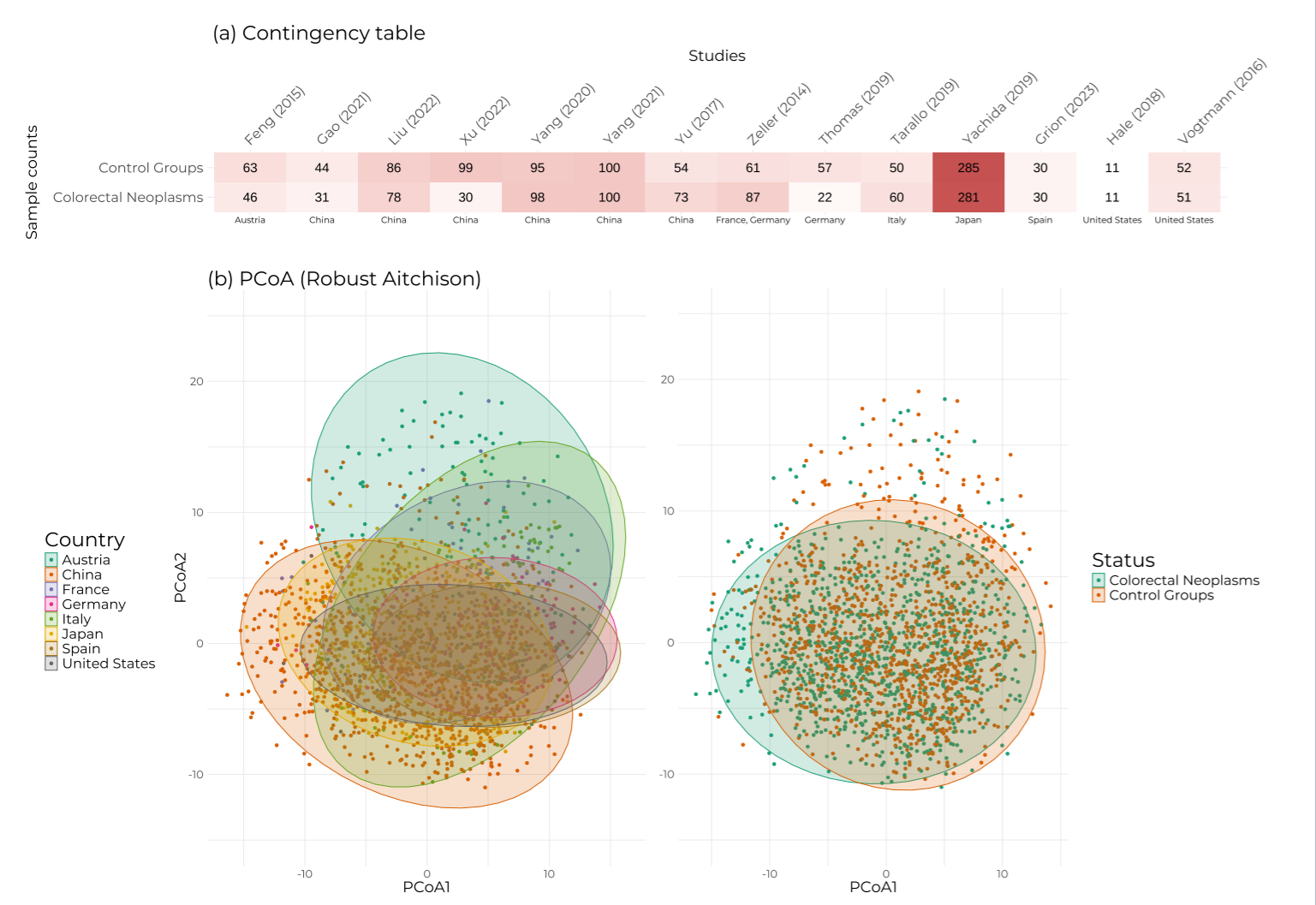


Fig. 1. Schematic representation of the steps followed by Gpredomics and the interpretability of the isolated models.

Performs like traditional methods

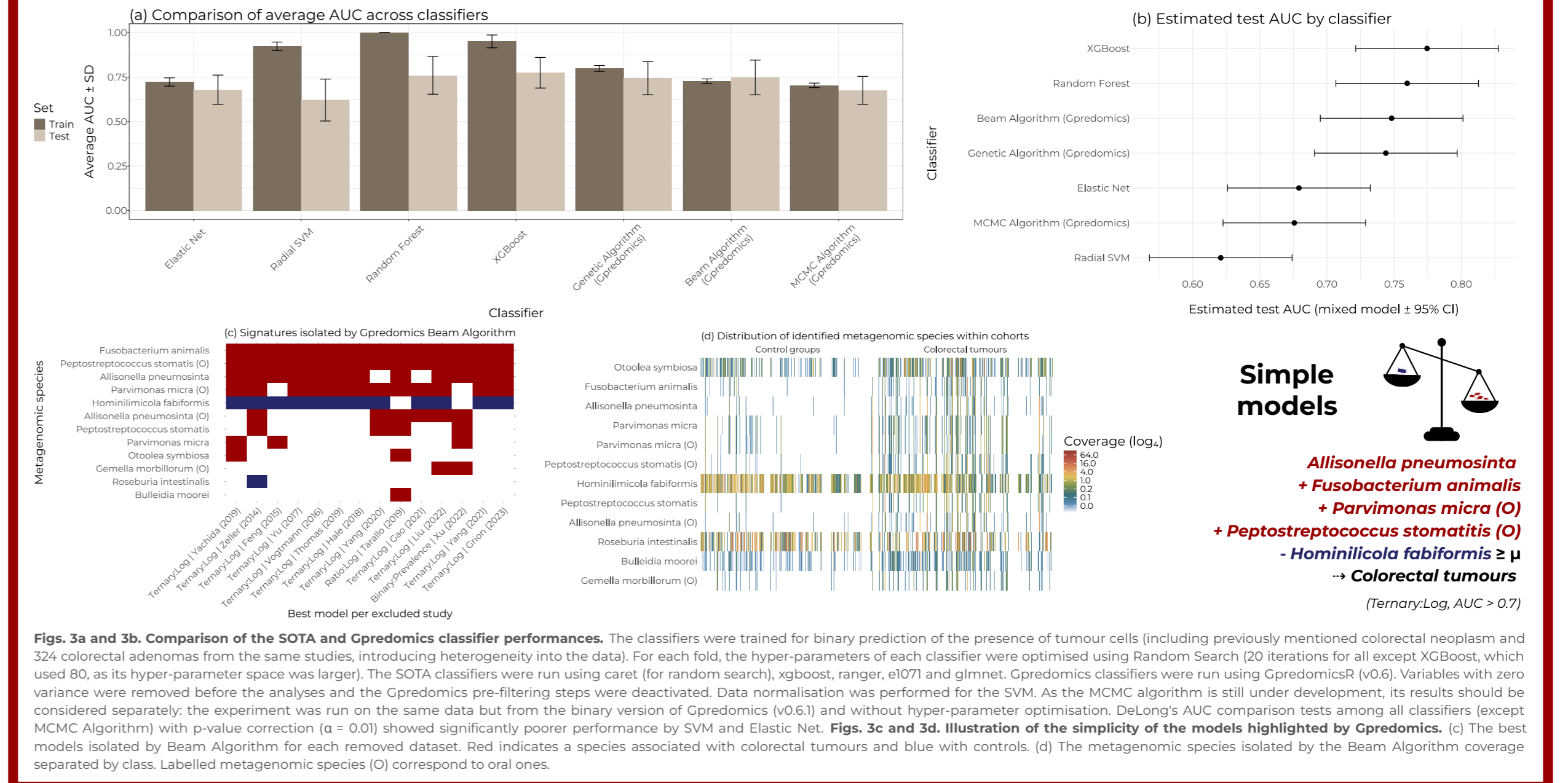
Several experiments were carried out to characterise the performance of Gpredomics to predict colorectal neoplasms in 2085 samples from 14 different studies originating from different countries (Figs. 2). The data were pre-processed using GMT Science's proprietary *Biomscope* pipeline (Puller *et al.*)³, leading to the characterisation of the gene coverage of 2834 intestinal and oral metagenomic species. Models found via Gpredomics were compared with other SOTA classifiers by adopting a Leave-One-Dataset-Out approach (LODO) and by optimising hyper-parameters in 5-CV with random search on each training fold group, in accordance with standard nested cross-validation practices (Fig. 3). Furthermore, Gpredomics performances in the prediction of colorectal neoplasms were compared with the performance published in a meta-study (Piccinno *et al.*)⁴ in a paired-trial approach (Fig. 4). Finally, a performance comparison based on the language, data pre-processing and Gpredomics algorithm was carried out, using a LODO approach (Figs. 5). In all these comparisons, Gpredomics demonstrated its ability to combine performance with biological interpretability.

Data from different regions



Figs. 2a and 2b. (a) Contingency table by status (control or colorectal neoplasm) of data analysed via the *Biomscope* pipeline and used for analyses. (b) PCA projection of robust Aitchison distances between samples, highlighting their origins and associated status.

SOTA-level performance with interpretability



Figs. 3a and 3b. Comparison of the SOTA and Gpredomics classifier performances. The classifiers were trained for binary prediction of the presence of tumour cells (including previously mentioned colorectal neoplasm and 324 colorectal adenomas from the same studies, introducing heterogeneity into the data). For each fold, the hyper-parameters of each classifier were optimised using Random Search (20 iterations for all except XGBoost, which used 80, as its hyper-parameter space was larger). The SOTA classifiers were run using caret (for random search), xgboost, ranger, e1071 and gbm. Gpredomics classifiers were run using Gpredomics (v0.6). Variables with zero variance were removed before the analyses and the Gpredomics pre-filtering steps were deactivated. Data normalisation was performed for the SVM. As the MCMC algorithm is still under development, its results should be considered separately; the experiment was run on the same data but from the binary version of Gpredomics (v0.6.1) and without hyper-parameter optimisation. Delong's AUC comparison tests among all classifiers (except MCMC Algorithm with p-value correction ($\alpha = 0.01$)) showed significantly poorer performance by SVM and Elastic Net. Figs. 3c and 3d. Illustration of the simplicity of the models highlighted by Gpredomics. (c) The best models isolated by Beam Algorithm for each removed dataset. Red indicates a species associated with colorectal tumours and blue with controls. (d) The metagenomic species isolated by the Beam Algorithm covered separately by class. Labelled metagenomic species (O) correspond to oral ones.

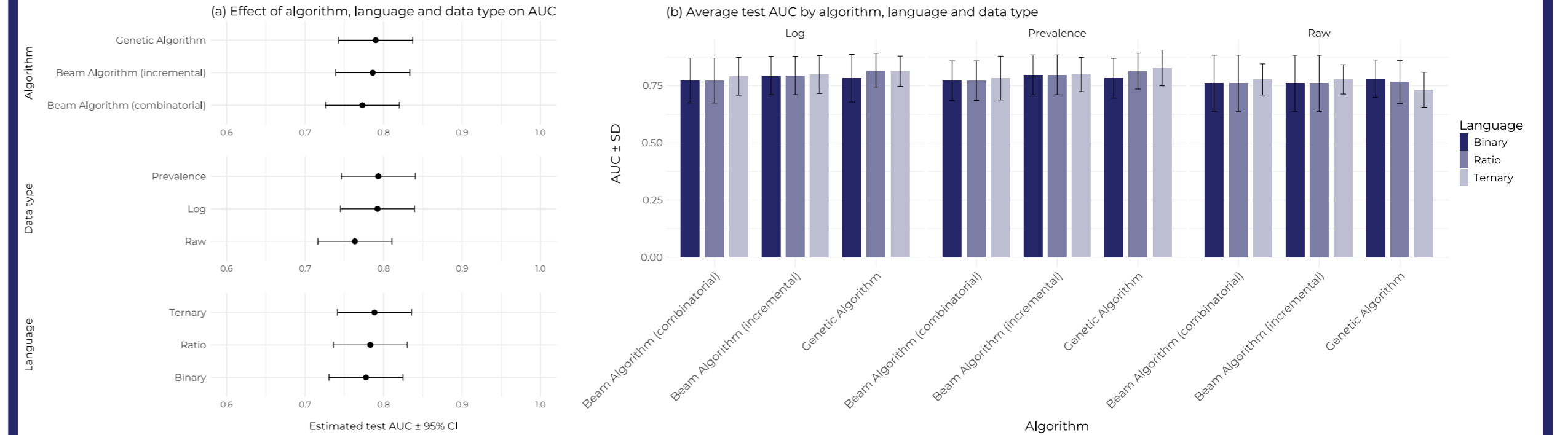
Comparison with Piccinno *et al.*

Comparison of performance against Piccinno *et al.* (2024) meta-study

Training set	Feng (2015)	Liu (2022)	Thomas (2019)	Vogtmann (2016)	Wille (2018)	Yang (2020)	Yang (2021)	Yu (2017)	Zeller (2014)
(P) Random Forest Feng (2015)	0.92	0.70	0.82	0.63	0.79	0.60	0.58	0.71	0.70
(P) Random Forest Liu (2022)	0.73	0.96	0.59	0.71	0.75	0.79	0.74	0.72	0.73
(P) Random Forest Thomas (2019)	0.69	0.57	0.75	0.48	0.66	0.74	0.61	0.68	0.61
(P) Random Forest Vogtmann (2016)	0.65	0.84	0.59	0.68	0.85	0.75	0.75	0.80	0.76
(P) Random Forest Wille (2018)	0.80	0.84	0.71	0.73	0.83	0.88	0.79	0.85	0.77
(P) Random Forest Yang (2020)	0.71	0.80	0.71	0.69	0.83	0.96	0.77	0.87	0.85
(P) Random Forest Yang (2021)	0.75	0.85	0.68	0.73	0.86	0.84	0.87	0.87	0.76
(P) Random Forest Yu (2017)	0.81	0.83	0.75	0.74	0.88	0.92	0.82	0.88	0.81
(P) Random Forest Zeller (2014)	0.80	0.86	0.71	0.69	0.82	0.92	0.78	0.87	0.86
(C) Pow2Log Feng (2015)	0.89	0.69	0.61	0.56	0.45	0.60	0.58	0.59	0.63
(C) Pow2Log Liu (2022)	0.69	0.88	0.60	0.66	0.66	0.70	0.68	0.68	0.65
(C) BinaryLog Thomas (2019)	0.85	0.78	0.79	0.70	0.76	0.77	0.69	0.81	0.72
(C) BinaryLog Vogtmann (2016)	0.77	0.73	0.64	0.70	0.66	0.59	0.66	0.74	0.65
(C) TernaryPrevalence Wille (2018)	0.60	0.74	0.63	0.65	0.80	0.78	0.69	0.81	0.68
(C) TernaryLog Yang (2020)	0.72	0.77	0.67	0.68	0.84	0.92	0.75	0.80	0.78
(C) BinaryPrevalence Yang (2021)	0.82	0.80	0.71	0.70	0.83	0.78	0.79	0.84	0.73
(C) TernaryLog Yu (2017)	0.80	0.84	0.74	0.73	0.88	0.88	0.80	0.85	0.79
(C) BinaryRaw Zeller (2014)	0.76	0.73	0.66	0.60	0.76	0.83	0.69	0.77	0.82

Fig. 4. Comparison of performance between the Piccinno *et al.* (2024) meta-study (P) and Gpredomics models (C) in predicting colorectal neoplasms. For Gpredomics, variables present in at least 20% of the samples were selected. The languages and data types were optimised upstream using grid-search 5-CV. Intra-study performances were obtained using 20/10-CV in order to estimate generalisation as accurately as possible and to reproduce the conditions of the study. Inter-study performances were obtained by applying the 200 isolated intra-study models to the other studies. The analyses were carried out using Gpredomics v0.6 and v0.6.1.

Various algorithms, languages and types



Figs. 5a and 5b. Comparison of the performance of Gpredomics languages, data types and algorithms in distinguishing control patients from patients with colorectal neoplasms. Performances were measured using a LODO approach. The variables were pre-filtered internally using Gpredomics (v0.6.1). Pow2Log language has been excluded from the comparison because it is a language specific to the genetic algorithm for the moment. The same applies to the MCMC Algorithm, which is currently under development and only includes a Generic language adapted from Ternary.

Perspectives

Gpredomics is already a solid alternative to traditional methods, and it is expected to evolve further over the coming weeks and months in order to improve its capabilities. These include multi-class classification, the addition of meta-modelling based on voting, optimisation of hyper-parameters by grid or Bayesian search, the addition of new methods for exploring the search space and improvements to existing heuristics. Future enhancements will also introduce new interface options, such as Python and a Shiny web interface, to complement current technical improvements. With these improvements, Gpredomics aims to provide increasingly effective solutions for the research community, hopefully driving significant advancements.