Montesi *et al. BioData Mining* (2026) 19:64
https://doi.org/10.1186/s13040-026-00568-8

BioData Mining

SOFTWARE

Open Access

TENTACLES: a consensus machine learning tool for robust biomarker discovery in heterogeneous data

Giorgio Montesi[1], Gabriel Dos Santos Mouta[1], Maria Novedrati[1], André F. Cunha[2,3], Alessandro Fuschi[4], Simone Lucchesi[1], Chiara Sonnati[1], Annalisa Ciabattini[1], Francesco Santoro[1], Donata Medaglini[1*] and Helder I. Nakaya[2,3,5*]

*Correspondence:
Donata Medaglini
donata.medaglini@unisi.it
Helder I. Nakaya
helder.nakaya@einstein.br
[1]Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, Siena, Italy
[2]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil
[3]Institut Pasteur de São Paulo, São Paulo, Brazil
[4]Department of Physics and Astronomy, University of Bologna, Bologna, Italy
[5]Hospital Israelita Albert Einstein, São Paulo, Brazil

Abstract

Background Transcriptomic biomarker discovery often fails to produce reproducible gene signatures across independent cohorts due to model-specific biases and dataset heterogeneity. While single-algorithm approaches may perform well on training data, they frequently fail to generalize effectively. Ensemble methods have proven effective in general machine learning applications, yet their systematic integration for consensus-based feature prioritization remains underexplored in transcriptomics.

Results We developed TENTACLES (Transcriptomic Exploration Tool through Aggregation of Classifiers), an open-source modular framework for robust biomarker discovery through multi-algorithm consensus. The tool is an open-source R package that integrates up to 15 supervised learning algorithms and 6 unsupervised clustering methods. The tool utilizes a modular architecture to automate data preprocessing, multi-algorithm feature prioritization, and cross-cohort validation. By aggregating variable importance across multiple models, TENTACLES identifies gene signatures resilient to algorithm-specific biases. We validated the framework using Crohn's disease as a high-heterogeneity case study across 689 samples from four independent publicly available RNA-seq cohorts. TENTACLES identified a 28-gene consensus panel that achieved superior cross-cohort generalizability compared to single-algorithm-derived signatures and conventional differential expression methods while using, compared to the latter, 95% fewer features. This signature was further refined to a minimal 5-gene core that maintained robust discriminatory power in completely unsupervised validation. These results confirm the tool's ability to extract stable biological signals from complex, noisy datasets.

Conclusions TENTACLES provides a scalable, disease-agnostic solution for identifying minimal reproducible gene signatures from heterogeneous transcriptomic data. By bridging the gap between complex ensemble modeling and practical biomarker discovery, the software could serve as a versatile resource for researchers aiming to derive reproducible biomarkers across diverse disease contexts.

Keywords Machine learning, Biomarker discovery, Complex diseases, Crohn's disease, Ensemble learning

Background

High-throughput RNA sequencing has emerged as a powerful approach for elucidating the molecular mechanisms underlying complex diseases. Transcriptomic analyses have successfully identified disease-associated gene expression patterns capable of distinguishing affected individuals from healthy controls. When combined with machine learning (ML) approaches, these datasets have yielded promising multigene signatures with diagnostic and prognostic potential [1–4]. However, current computational pipelines predominantly employ a conventional two-step approach: initial identification of differentially expressed genes (DEGs) followed by training of a single classification model [3–7]. This methodology presents several critical limitations. First, it ignores interactions among genes and subtle but consistent multivariate patterns that may be biologically meaningful. Second, it relies on single-algorithm classifiers that are vulnerable to model-specific biases and may yield signatures that perform well on training data but generalize poorly to independent cohorts with different experimental conditions, patient populations, or technical platforms. Third, the arbitrary statistical thresholds used in DEG filtering may inadvertently exclude genes with modest but consistent discriminatory power across multiple analytical approaches. Fourth, using class-informed DEGs as input features for classifiers trained on the same dataset could introduce data leakage, whereby apparent model performance partly reflects label rediscovery rather than genuine predictive signal. These limitations underscore the need for more robust, generalizable approaches to biomarker discovery in transcriptomic datasets.

Ensemble methods, which aggregate predictions or feature importance across multiple algorithms, have demonstrated superior performance and generalizability in numerous ML applications but remain underutilized in transcriptomic biomarker discovery. By leveraging the complementary strengths of diverse algorithms—from linear models sensitive to additive effects to tree-based methods capable of capturing complex interactions—ensemble approaches can identify more stable and reproducible feature sets.

To address these methodological gaps, we developed TENTACLES (Transcriptomic exploration tool through aggregation of classifiers), a supervised ensemble framework specifically designed for robust feature prioritization in heterogeneous transcriptomic datasets. Unlike conventional approaches that rely on single algorithms, TENTACLES systematically aggregates feature importance rankings across multiple diverse ML classifiers to identify consensus gene signatures that are both predictive and reproducible across independent cohorts.

To validate TENTACLES, we selected Crohn's disease (CD) as a challenging heterogeneous case study. CD is a chronic and relapsing form of inflammatory bowel disease (IBD) characterized by transmural segmental inflammation that can affect any portion of the gastrointestinal tract [8]. The disease presents with distinctive discontinuous lesions ("skip lesions") that may involve the entire digestive tract from mouth to anus. Clinical manifestations are highly variable, ranging from abdominal pain and diarrhea to severe extra-intestinal complications, often resulting in diagnostic delays [9] and the need for invasive diagnostic procedures [1, 10]. These delays carry significant clinical

consequences, as progressive tissue damage frequently leads to strictures, penetrating disease, and the eventual need for surgical intervention in up to 70% of patients [11]. This clinical and molecular heterogeneity poses a substantial challenge for transcriptomic biomarker discovery, making CD an ideal test case for evaluating the robustness of our consensus-based approach.

In this study, we applied TENTACLES to four independent CD cohorts encompassing 689 samples from four bulk RNA-Seq datasets of intestinal biopsies. Our analytical strategy employed a rigorous four-phase validation approach (1): consensus gene identification through multi-algorithm feature importance aggregation in a discovery cohort (2), comprehensive evaluation of the consensus signature in an independent evaluation cohort using complementary analytical strategies (3), refinement and identification of the minimal gene combination through unsupervised clustering in a third independent cohort, and (4) unsupervised validation of the minimal gene signature in a fourth independent cohort.

Implementation

Acquisition and preprocessing of public RNA-Seq data from CD intestinal samples

RNA sequencing datasets related to CD were obtained from the Sequence Read Archive (SRA). Specifically, intestinal biopsy samples from the following BioProjects were downloaded: *PRJNA248469* [12, 13] (Dataset 1, discovery cohort), *PRJNA565216* [14] and *PRJNA985602* [15] (Dataset 2, external evaluation cohort), *PRJNA702434* [16] (Dataset 3, signature refinement cohort), *PRJNA728646* [17] and *PRJNA961704* [18] (Dataset 4, independent validation cohort). Inclusion criteria were restricted to datasets originating from healthy controls (HC) and patients diagnosed with CD. All 689 resulting samples were uniformly processed using a standardized custom pipeline to ensure consistency across datasets. Quality control was performed with **fastp** [19] (v1.0.1), followed by alignment to the human reference genome GRCh38.p14 (NCBI RefSeq assembly GCF_000001405.40) using **STAR** [20] (v2.7.10) with default parameters. Read strandedness was assessed with **RSeQC** [21] (v5.0.1), and transcript quantification was carried out using **Rsubread** [22] (v2.22.1).

TENTACLES implementation

TENTACLES is organized around five sequential, modular functions that together implement a complete pipeline for consensus-based transcriptomic biomarker discovery (Fig. 1). *preProcess()* performs library-size normalization and optional batch-effect correction. *runClassifiers()* trains and tunes up to fifteen supervised classifiers simultaneously and computes variable importance from each final fitted model. *getConsensus()* aggregates these importance scores across classifiers and retains only genes identified as important by at least a user-specified minimum number of algorithms. *testConsensus()* evaluates a gene signature on any expression dataset through both univariate and multivariate methodologies – PCA, per-gene AUROC and fold-change analysis, hierarchical clustering, and a within-dataset MLP classifier – to characterize the panel discriminatory capacities. Finally, *valConsensus()* computes all possible combinations of genes from a gene panel to identify the minimal gene-combination that maximizes class separability through six different unsupervised clustering algorithms.

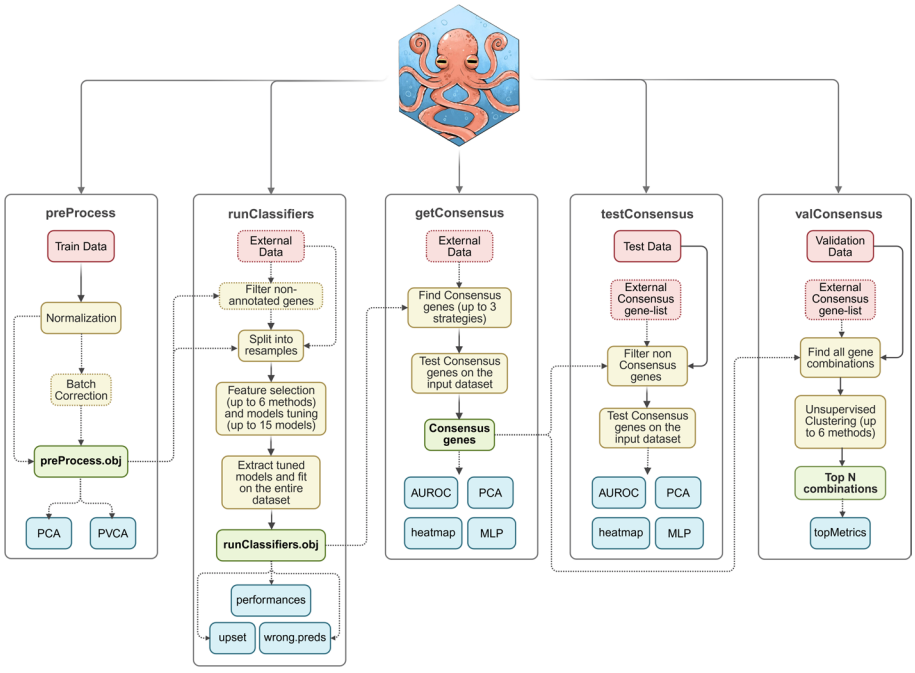

Fig. 1 Overview of TENTACLES main functionalities. TENTACLES comprises five core modules – *preProcess*, *runClassifiers*, *getConsensus*, *testConsensus*, and *valConsensus* – that together implement an end-to-end pipeline for discovering, testing, and validating minimal consensus gene signatures from high-dimensional expression data. Starting from a training dataset, *preProcess* performs normalization, optional batch-effect correction, and exploratory dimensionality reduction (PCA, PVCA), returning a *preProcess.obj* containing the transformed dataset. In *runClassifiers*, the pre-processed data (either the output of *preProcess* or an externally pre-processed dataset) can optionally be filtered to retain only genes annotated in KEGG or GO databases; up to six feature-selection methods and up to 15 supervised learning algorithms are then tuned, refitted on the full training set, and stored in a *runClassifiers.obj*, together with performance summaries, variable-importance measures, and diagnostic plots (including model performance across four metrics, an UpSet plot of important variables across models, and visualizations of misclassified samples). *getConsensus* aggregates feature-importance information across models to derive consensus gene panels using one of three alternative strategies. *testConsensus* evaluates either internally derived or user-defined consensus gene lists on the input data (eventually an independent test or external cohort) by restricting to the selected genes and generating PCA and loading plots, heatmaps, AUROC and fold-change summaries, and multilayer perceptron (MLP) classifiers. Finally, *valConsensus* systematically explores all possible combinations of either internally derived or user-defined consensus genes, applying up to six unsupervised clustering algorithms, and ranks gene sets according to internal clustering metrics, returning the top N subsets (topMetrics). Dotted arrows indicate optional steps in the workflow. Red boxes denote input data, yellow boxes represent internal processing steps, green boxes indicate function outputs and blue boxes indicate optional plots generated by the functions

TENTACLES modular architecture allowed its implementation through a four-phase strategy: ensemble-based signature discovery, cross-cohort signature evaluation, minimal signature identification, and independent unsupervised validation. Each phase was applied to a separate, independently pre-processed RNA-seq cohort following the same workflow (Fig. 2). In all phases, prior to any analysis, data preprocessing was performed using the *preProcess()* function normalizing gene counts with the **edgeR** [23] (v4.4.2) package and batch-effect correction when applicable with the **sva** [24] (v3.54.0) package. The quality of the batch-effect correction, when applied, was examined using Principal Variance Component Analysis (PVCA; **pvca** [25] v1.46.0) plots. Genes without annotation in GO or KEGG databases were optionally filtered out prior to classification (limiting the analysis to approximately 10,000 genes) using the filter argument of the *runClassifiers()* function, in order to focus the analysis on biologically interpretable features and reduce computational load. Genes with CPM > 1 in at least 10% of the samples were retained.

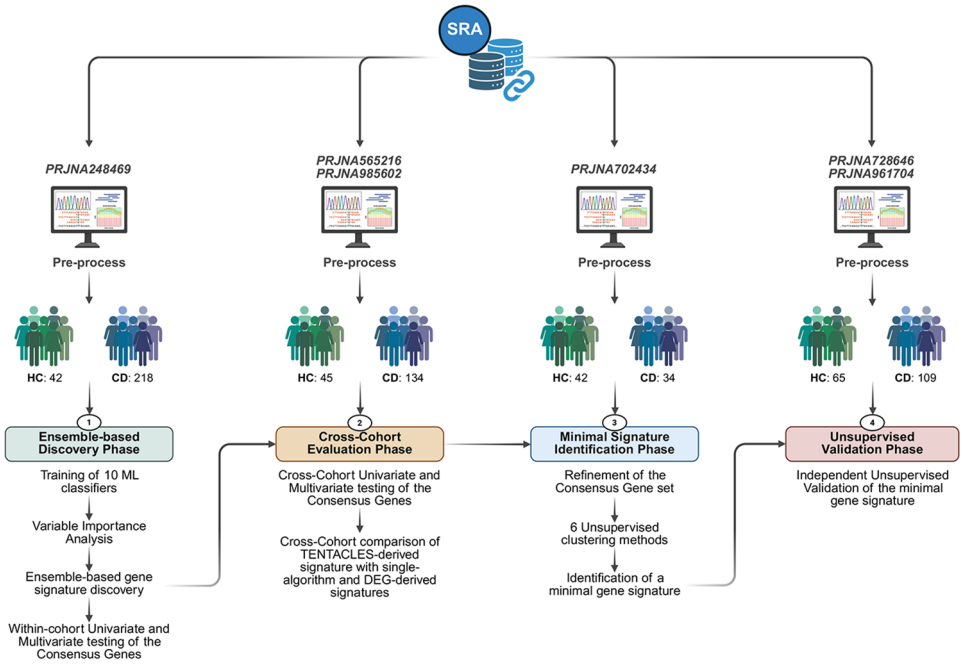

Fig. 2 TENTACLES four-phase workflow for gene-signature discovery and validation in Crohn's disease. Four independent RNA-seq cohorts of intestinal biopsies were retrieved from the NCBI sequence read archive (SRA) and processed through a uniform preprocessing pipeline. (1) ensemble-based discovery phase: dataset 1 (*PRJNA248469*; HC = 42, CD = 218) was used for training 10 ML classifiers, variable importance analysis, ensemble-based consensus gene signature discovery, and within-cohort univariate and multivariate testing of the consensus genes. (2) cross-cohort evaluation phase: dataset 2 (merged *PRJNA565216* and *PRJNA985602*; HC = 45, CD = 134) was used for cross-cohort univariate and multivariate testing of the consensus genes, refinement of the consensus gene set, and benchmarking of the TENTACLES-derived signature against single-algorithm and DEG-derived signatures. (3) minimal signature identification phase: dataset 3 (*PRJNA702434*; HC = 42, CD = 34) was used for exhaustive evaluation of all gene combinations through six unsupervised clustering methods and identification of the minimal gene signature. (4) unsupervised validation phase: dataset 4 (merged *PRJNA728646* and *PRJNA961704*; HC = 65, CD = 109) was used for independent unsupervised validation of the minimal gene signature. PRJNA identifiers denote NCBI SRA BioProjects. HC, healthy controls; CD, Crohn's disease

Ensemble-based signature discovery phase

To identify a robust consensus gene signature from transcriptomic data, the discovery phase was performed using the preprocessed study *PRJNA248469* (Dataset 1), applying the *runClassifiers()* and *getConsensus()* functions from TENTACLES. The workflow comprised sequential steps that included within-fold multi-algorithm training paired with dedicated feature selection strategies, within-fold hyperparameter tuning, final model fitting, variable importance computation and consensus-based gene selection.

Ten supervised classifiers spanning complementary methodological families were trained and tuned using a 5-fold cross-validation strategy: Random Forest (**ranger** [26] v0.17.0), Multi-Layer Perceptron (MLP) and Bagged-MLP (**nnet** [27] v7.3–20), Support Vector Machine with a linear kernel (SVM; **kernlab** [28] v0.9–33), C5-rules (**C50** [29] v0.2.0), LightGBM (**lightgbm** [30] v4.6.0), Decision Tree (**rpart** [31] v4.1.24), Multivariate Adaptive Regression Splines (MARS; **earth** [32] v5.3.4), Extreme Gradient Boosting (XGBoost; **xgboost** [33] v1.7.11.1), and Partial Least Squares (PLS; **mixOmics** [34] v6.30.0).

For all classifiers, data within each cross-validation fold were processed by applying near-zero variance removal, unit-variance-scaling (*step_nzv()* and *step_normalize()*; **recipes** [35] v1.3.1) and majority-class down sampling (*step_downsample()*; **themis** [36]

v.1.0.3). Moreover, each classifier was paired with a specific feature selection strategy: correlation-based filtering with a threshold of 0.95 was applied to Random Forest to prevent highly redundant features from inflating impurity-based importance score; ROC-based ranking (*step_select_roc()*; **colino** [37] v0.0.1) with a threshold of 0.95 was used for MLP, Bagged-MLP, and LightGBM to remove genes with low individual discriminatory power; Boruta (**Boruta** [38] v8.0.0, accessed via **colino** function *step_select_boruta()*) with a maximum of 100 iterations was applied to SVM, C5-rules, Decision Tree, MARS, and PLS to identify features informative in a multivariate, non-linear context; XGBoost was run without additional feature selection to serve as a full-feature-space reference.

All feature selection steps were implemented as recipe steps within the **tidymodels** framework (**workflowsets** [39] v1.1.1) and executed independently on the training portion of each cross-validation fold, so that no information from the held-out data influenced feature selection or scaling parameters. The same procedure was then reapplied on the full training dataset when fitting the final models.

Model hyperparameters were optimized independently for each classifier through a grid-search across five candidate configurations within each fold of the cross-validation. The following hyperparameters were tuned: Random Forest (number of variables randomly sampled at each split, *mtry*; minimum node size, *min_n*), MLP and Bagged-MLP (number of hidden units, weight decay penalty, number of training epochs), linear SVM (regularization cost, margin parameter), C5-rules (number of boosting iterations, *min_n*), LightGBM (*mtry*, number of trees, *min_n*, maximum tree depth, learning rate, minimum loss reduction required for further partitioning), Decision Tree (maximum tree depth, *min_n*), MARS (number of terms, maximum interaction degree), XGBoost (*mtry*, number of trees, *min_n*, maximum tree depth, learning rate, minimum loss reduction, subsample size, early stopping rounds), and PLS (number of components, proportion of predictors used). The optimal configuration for each model was selected based on the highest F1-score across folds, which, computed on the held-out data, served as the primary performance estimate. Accuracy, Precision, and Recall were additionally recorded for each configuration. Resubstitution metrics, are reported alongside cross-validated estimates as apparent performance.

After selecting the optimal hyperparameter configuration, each classifier was refitted on the complete training dataset and variable importance was computed from this final model using model-native methods via the **vip** [40] (v0.4.1) package. For MLP-based models, the Olden algorithm (**NeuralNetTools** [41] v1.5.3) was used. For Bagged-MLP, Olden importances were averaged across bootstrap replicate networks. For SVM-linear, a permutation-based fallback (metric: ROC-AUC) was used since the model architecture does not expose coefficients directly. All remaining classifiers used model-native, non-negative importance metrics: impurity decrease for Random Forest and Decision Tree, gain for LightGBM and XGBoost, rule usage counts for C5-rules, number of model subsets for MARS, and PLS-VIP scores for PLS (Supplementary Table 1). A gene was considered important for a given classifier if its variable importance score was non-zero in case of signed importances, or greater than zero for non-signed importances.

To identify features with consistent discriminatory power across diverse algorithmic approaches, a consensus panel was derived from the variable importance outputs of the fitted classifiers using TENTACLES *getConsensus()* function. Classifiers with markedly poor cross-validated performance were excluded from consensus formation to prevent

unstable, algorithm-specific gene selections from influencing the shared feature set. Among the remaining classifiers, a gene was included in the consensus panel if it was identified as important by at least a minimum number of algorithms (*n.min*), defined a priori as a majority threshold across the contributing classifiers. As part of the *getConsensus()* output, the consensus gene set was immediately evaluated on Dataset 1 itself to quantify its discriminatory capacities through an internal call to *testConsensus()*, applying three complementary analyses: (i) PCA with loading analysis to visualise multivariate class separation and identify the genes driving it; (ii) per-gene AUROC and log2 fold-change (**pROC** [42] v1.18.5) to quantify univariate discriminatory power and expression directionality; and (iii) a naive MLP classifier trained and evaluated within Dataset 1 to provide a multivariate performance estimate of the consensus gene set. For the within-dataset MLP, data were partitioned into a 70% training and 30% test set using stratified sampling; hyperparameters (number of hidden units, penalty, and number of epochs) were tuned via 3-fold cross-validation with a grid of ten candidate configurations, selecting the best configuration by highest ROC-AUC (**tune** [43] v1.3.0). Final performance was evaluated on the held-out test set in terms of AUROC, Accuracy, and Brier score, and variable importance was computed using the Olden algorithm (**NeuralNetTools**). This internal evaluation provides a numerical characterisation of the consensus genes' discriminatory behaviour on the discovery dataset, establishing a reference point for comparison with subsequent independent cohorts.

Cross-cohort signature evaluation phase

To assess whether the consensus gene signature retained discriminatory capacity in an external independent cohort and to further refine it to a minimal set of features with consistent cross-cohort behavior, the external evaluation phase was applied to Dataset 2, a merged cohort derived from two independent studies (*PRJNA565216* and *PRJNA985602*). Batch effects between the two studies were accounted for during preprocessing by including a batch variable accounting for both the study identifier and the internal batch in the *preProcess()* function. The variance associated to the biological class variable was preserved. The same three complementary analyses performed internally on Dataset 1 (PCA with loading analysis, per-gene AUROC and log2 fold-change, and within-cohort MLP evaluation) were then applied to Dataset 2 via the *testConsensus()* function. The within-cohort MLP was tuned and evaluated following the same procedure described above, with performance assessed on the held-out test set in terms of AUROC, Accuracy, and Brier score, and variable importance computed using the Olden algorithm. Running the same set of analyses on both Dataset 1 and Dataset 2 allowed direct comparison of each gene discriminatory behavior across independent cohorts, informing the signature refinement step.

Specifically, genes were carried forward to the subsequent minimal signature identification phase based on the concordance of their signed Olden importance scores between the two within-dataset MLP models: a gene was retained if its importance carried the same sign in both the Dataset 1 and Dataset 2 evaluations—positive (associated with case prediction) or negative (associated with control prediction)—indicating a consistent and reproducible directional contribution across independent datasets.

Minimal signature identification phase

To identify the minimal gene combination capable of stratifying samples in a completely unsupervised manner, the signature refinement phase was applied to Dataset 3 (*PRJNA702434*) using the refined gene set derived from the preceding evaluation phase. All possible non-empty subsets of the candidate genes were systematically evaluated using the *valConsensus()* function, which applies six different unsupervised clustering algorithms—each with the number of clusters fixed at two—to the expression matrix restricted to each gene subset. Gaussian Mixture Models (GMM; **mclust** [44] v6.1.1) and Hierarchical Clustering (HiCl) were performed on normalized gene counts directly, while K-Means clustering was applied both to the raw expression space and to data projected onto dimensionality reduction embeddings, including PCA, Uniform Manifold Approximation and Projection (UMAP; **umap** [45] v0.2.10.0), and t-distributed Stochastic Neighbor Embedding (t-SNE; **Rtsne** [46, 47] v0.17). For each gene combination and clustering method, the resulting unsupervised partition was compared against the reference class-labels using Accuracy, Precision, Recall and F1-score. All subsets were ranked by mean F1-score across the six clustering methods, and the top-performing combinations were reported. The gene combination achieving the best overall ranking was selected as the minimal core signature and carried forward to the final validation phase.

Independent unsupervised validation phase

The minimal core signature was validated on Dataset 4, an independent cohort assembled by merging two BioProjects (*PRJNA728646* and *PRJNA961704*), preprocessed jointly using *preProcess()* with batch-effect correction applied using the study identifier as the batch variable. The variance associated to the biological class variable was preserved. The expression matrix was restricted to the signature genes and subjected to t-SNE dimensionality reduction, followed by K-Means clustering on the resulting two-dimensional embedding with the number of clusters set to two. Class labels were used exclusively a posteriori to evaluate clustering performance in terms of Accuracy, Precision, Recall, and F1-score, providing an unbiased estimate of the signature's ability to stratify samples by disease status in a fully unsupervised setting.

Differential gene expression analysis

As a complementary approach, two differential gene expression analysis were performed independently on Dataset 1 (*PRJNA248469*) and Dataset 2 (obtained by merging *PRJNA565216* and *PRJNA985602)*. For both datasets, samples with low total read counts (<500,000) were excluded to ensure sufficient coverage, and only genes with CPM > 1 in at least 10% of the samples were retained to reduce noise from lowly expressed transcripts.

Differential expression analysis was performed using the **DESeq2** [48] R package (version 1.46.0). **DESeq2** was used to normalize the count data and to compute gene-wise statistics using a negative binomial generalized linear model. Normalized counts were obtained using the variance-stabilizing transformation (VST) for downstream visualization and comparative analysis. For Dataset 1, a **DESeq2** object was created using the class variable (CD vs HC) as the design formula. In contrast, Dataset 2 included samples from multiple studies, and the design formula was adjusted to include the study id and the internal batch as batch correction. The resulting differential expression results were

filtered for multiple testing using the Benjamini–Hochberg method, and significantly differentially expressed genes were defined based on an adjusted p-value < 0.05.

Benchmarking against single-algorithm and differential expression-based methods

To contextualize the performance of the TENTACLES consensus signature relative to conventional biomarker discovery approaches, a benchmarking analysis was performed on Dataset 2 using the same MLP-based *testConsensus()* framework described above. Two classes of comparators were evaluated. First, differential expression-based gene sets of increasing size were derived from DEG analysis performed on Dataset 1 by ranking them by ascending adjusted p-value and selecting the top k genes at several thresholds. Second, for each individual classifier trained during the Discovery Phase, all genes identified as important by that classifier (i.e., all genes passing the per-model importance filter) were used as a separate comparator gene set. Each comparator was evaluated on Dataset 2 using the same MLP-based *testConsensus()* procedure, recording Accuracy, Brier score, and ROC-AUC, enabling a direct comparison of the consensus-based signature against signatures derived by both threshold-driven differential expression and single-algorithm feature selection strategies under identical evaluation conditions.

Reproducibility of differential expression signatures across cohorts

To independently assess the cross-cohort reproducibility of differentially expressed genes and to contextualize the overlap between conventional differential expression and TENTACLES consensus panel, the set of differentially expressed genes derived from DEG analysis performed on Dataset 2 was intersected with the DEG list derived from Dataset 1. This allowed the quantification of the proportion of differentially expressed features reproducibly detected across both independent cohorts. In addition, the overlap between the TENTACLES consensus panel and the union of DEGs identified across both datasets was evaluated to determine what fraction of consensus-selected genes were also supported by classical differential expression evidence, and conversely, to what extent the consensus approach could capture discriminatory features that do not meet conventional differential expression thresholds.

Computational cost analysis

To provide a practical assessment of the computational requirements of TENTACLES, a computational cost analysis was performed for the main functions used throughout the workflow. Benchmarking was carried out on a local Windows 11 workstation equipped with an Intel Core i5-13420H processor with 32 GB RAM. Runtime and peak memory usage were monitored using the **peakRAM** [49] v1.0.2, which was used to record elapsed time and maximum RAM allocation for each function call. When applicable, parallel execution was enabled using the base R **parallel** v4.4.3. Representative runs of *preProcess()*, *getConsensus()*, *testConsensus()*, *runClassifiers()*, and *valConsensus()* were profiled under the execution settings adopted in this study, including serial or parallel configurations and, when relevant, runs with or without plot generation.

Results

Dataset composition and preprocessing

TENTACLES framework (Fig. 1) was applied to four independent RNA-seq cohorts of CD patients and healthy controls (Fig. 2 and Table 1). Dataset 1 (*PRJNA248469*; 260 samples: 218 CD, 42 HC) served as the discovery cohort. Dataset 2 combined two independent studies (*PRJNA565216* and *PRJNA985602*; 179 samples total: 134 CD, 45 HC) and was used as the external evaluation cohort. Dataset 3 (*PRJNA702434*; 76 samples: 34 CD, 42 HC) served as the signature refinement cohort, and Dataset 4 combining two independent studies (*PRJNA728646* and *PRJNA961704*; 174 samples total: 109 CD, 65 HC) served as the independent unsupervised validation cohort. All cohorts were processed through the same standardized pipeline. For Dataset 2 and Dataset 4, batch correction was applied to account for technical variability between the two merged BioProjects, and PVCA plots confirmed effective reduction of study-related variance with preservation of the biological class signal (Supplementary Figure 1).

Ensemble-based signature discovery

Multi-algorithm classification performance on the discovery cohort

Ten supervised machine learning classifiers, encompassing complementary methodological approaches including tree-based methods (Random Forest, C5-rules, Decision Tree), neural networks (MLP, Bagged-MLP), kernel methods (SVM-linear), gradient boosting (LightGBM, XGBoost), regression splines (MARS), and dimensionality reduction (PLS) were trained on Dataset 1. Performance assessment during cross-validated hyperparameter optimization and subsequent full-dataset refitting revealed distinct algorithmic behaviors and convergent classification patterns (Fig. 3A and Supplementary Table 2). Nine of ten algorithms demonstrated robust and consistent performance overall, with median [IQR] cross-validated accuracy of 0.80 [0.79–0.82] and F1-score of 0.60 [0.57–0.64] (corresponding apparent performance after full-dataset refitting was 0.87 [0.85–0.88] and 0.70 [0.68–0.73]). SVM-linear achieved the highest cross-validation performance (accuracy = 0.85 ± 0.02; F1 = 0.67 ± 0.03; precision = 0.54 ± 0.04; recall = 0.90 ± 0.05), with corresponding apparent values of 0.91, 0.78, 0.64, and 1.00 after

Table 1 Datasets used across the analytical workflow. Dataset 1 (*PRJNA248469*) was used during the ensemble-based signature discovery phase for classifier training, variable importance analysis and consensus gene selection. Dataset 2, composed by two independent studies (*PRJNA565216* and *PRJNA985602*), served as the external evaluation cohort for cross-cohort signature testing and benchmarking. Dataset 3 (*PRJNA702434*) was used for signature refinement through exhaustive unsupervised gene combination ranking. Dataset 4, composed by two independent studies (*PRJNA728646* and *PRJNA961704*), provided an independent cohort for unsupervised validation of the refined gene signature. For each dataset, the table reports the count and proportion of healthy controls (HC) and Crohn's disease (CD) samples

Stage	Composing Dataset	Study ID	Health Control (% of total)	Crohn's Disease (% of total)
Ensemble-based Signature Discovery Phase	*Dataset 1*	*PRJNA248469*	42/260 (16%)	218/260 (84%)
Cross-cohort Signature Evaluation Phase	*Dataset 2*	*PRJNA565216*	37/149 (25%)	112/149 (75%)
		PRJNA985602	8/30 (27%)	22/30 (73%)
Minimal Signature Identification Phase	*Dataset 3*	*PRJNA702434*	42/76 (55%)	34/76 (45%)
Unsupervised Validation Phase	*Dataset 4*	*PRJNA728646*	5/22 (23%)	17/22 (77%)
		PRJNA961704	60/152 (39%)	92/152 (61%)

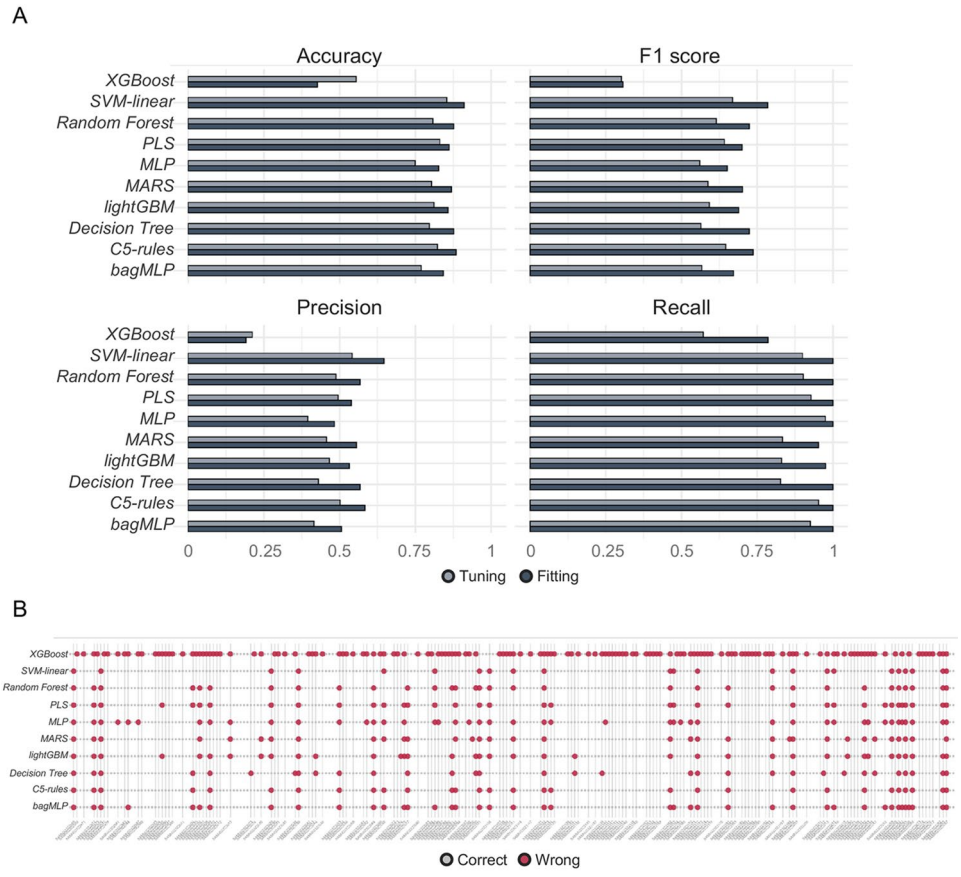

Fig. 3 Classification performances of the ten machine learning classifiers on the discovery cohort. (**A**) barplots reporting cross-validated tuning and full-model fitting performance for ten supervised classifiers – Random Forest, multi-Layer Perceptron (MLP), bagMLP (Bagged-MLP), linear Support Vector machine (SVM-linear), C5-rules, LightGBM, Decision Tree, multivariate Adaptive Regression Splines (MARS), XGBoost, and Partial least Squares (PLS) – evaluated on the training dataset (CD = 218, HC = 42). Metrics include Accuracy, F1-score, precision, and recall. All models were trained with optimized hyperparameters and the feature selection strategy specified in Supplementary table 1. Light bars ("tuning") summarize cross-validated performance during hyperparameter search; dark bars ("fitting") show performance after refitting the selected configuration on the full dataset. (**B**) Sample-level error concordance across models. Each column represents a sample and each row a classifier. Red and grey dots indicate misclassified and correctly classified samples, respectively

full-dataset refitting. Tree-based methods, including C5-rules, Random Forest, and Decision Tree, all exceeded a cross-validated accuracy of 0.79 (>0.87 after refitting). Notably, XGBoost exhibited markedly poor cross-validated performances (accuracy = 0.55 ± 0.07; F1 = 0.30 ± 0.03; precision = 0.21 ± 0.02; recall = 0.57 ± 0.06; apparent values after refitting: 0.43, 0.30, 0.19, and 0.78, respectively). A sensitivity analysis exploring alternative feature selection strategies at the standard tuning grid size confirmed consistently poor performance across configurations (Supplementary Figure 2A–B). Competitive results were approached only when the tuning grid was iteratively expanded (Supplementary Figure 2C–D), indicating that XGBoost requires substantially more extensive hyperparameter search than the remaining classifiers. To avoid introducing configurational asymmetries within the ensemble, XGBoost was excluded from consensus formation. Sample-level error analysis (Fig. 3B) confirmed that the majority of misclassifications originated from XGBoost, while 85% of samples were consistently and correctly classified by all high-performing algorithms. The small subset of samples misclassified by multiple algorithms

likely represented cases with intermediate transcriptomic profiles, possibly reflecting disease heterogeneity, early-stage disease, or technical variability inherent to biological samples.

Consensus feature selection identifies core transcriptomic signatures

Feature importance analysis across the nine well-performing classifiers revealed substantial but structured overlap in gene prioritization (Fig. 4A). Random Forest yielded the largest individual feature set (415 genes), consistent with the tendency of impurity-based importance scores to retain also weakly contributing features. Neural network methods (MLP, Bagged-MLP) contributed the largest feature set intersection with 265 genes exclusively shared between them, while tree-based methods showed strong convergence with neural networks through multiple intermediate intersections ranging from 8 to 42 genes. The intersection architecture demonstrated clear algorithmic families, with the largest single intersection comprising 265 genes exclusively shared by the two MLP-based models, followed by a secondary intersection of 42 genes common to MLPs and Random Forest. Intermediate-size intersections ranging between 8 and 21 genes connected these core methods with PLS, SVM-linear, and C5-rules in various combinations, while smaller intersections involving fewer than four genes displayed high model connectivity, often bridging classifiers from distinct methodological families. Notably, MARS contributed only marginally to the shared feature space, and XGBoost selections appeared largely unique with weak intersections with other classifiers, consistent with its unstable behavior during training and further supporting its exclusion from consensus formation.

Based on these convergence patterns, a 28-gene consensus panel was derived by selecting features deemed important by at least six of the nine high-performing classifiers ($n.min = 6$), corresponding to a strict majority across the ensemble. The sensitivity of the resulting panel size and downstream performance to the choice of *n.min* was assessed across thresholds ranging from 1 to 8, confirming the robustness of the selected threshold (Supplementary Figure 3). The resulting signature encompassed genes with established roles across multiple biological domains critical to CD pathogenesis. The panel included key innate and adaptive immune regulators such as *CTSS*, *CXCL8*, *OSM*,

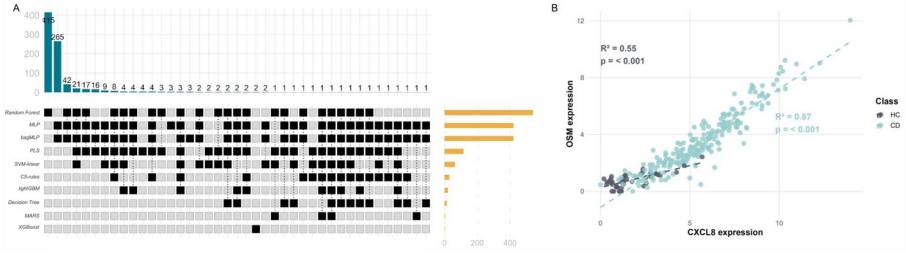

Fig. 4 UpSet plot of important variables across the ten classifiers and co-expression of the two most recurrently selected consensus genes. (**A**) UpSet plot summarizing overlap among genes prioritized by each of the ten classifiers – Random Forest, multi-Layer Perceptron (MLP), bagMLP (Bagged-MLP), linear Support Vector machine (SVM-linear), C5-rules, LightGBM, Decision Tree, multivariate Adaptive Regression Splines (MARS), XGBoost, and Partial least Squares (PLS) – based on variable-importance analyses; bars indicate the size of intersecting gene sets; rows correspond to individual models, with black cells denoting model inclusion in the intersection. Bar plots on the right indicate the total number of important genes per classifier. (**B**) pairwise scatterplot of *CXCL8* versus *OSM* expression (log-normalized counts). Points are colored by class (HC, healthy controls; CD, Crohn's disease). R^2 and *p*-values are reported separately for each class, showing co-expression in CD samples ($R^2 = 0.87$) and healthy controls ($R^2 = 0.55$)

CSF3R, *FCGR1A*, *FCGR3A*, and *FCER1A*, which mediate inflammatory cell recruitment, activation, and antibody-dependent cellular responses. Metabolic reprogramming and cellular stress response were represented by *ACSL1*, *HK2*, *HK3*, *CYP2E1*, *CHAC1*, *XBP1*, and *TXNDC5*, reflecting the altered energy metabolism and endoplasmic reticulum stress characteristic of inflamed intestinal tissue.

Signal transduction and cellular communication pathways were captured through *IL15RA*, *SEMA3E*, *WNT5A*, *HCAR2*, and *ADRA1B*, which regulate immune cell survival, tissue remodeling, and neurotransmitter signaling. Transcriptional and structural cellular regulators including *ELL2*, *RPA1*, and *LIMK2* represented the altered gene expression programs and cytoskeletal dynamics associated with tissue inflammation and repair. The consensus panel was completed by additional genes with established or emerging roles in immune and metabolic contexts, including *BACE2*, *C6*, *BATF2*, *TNFAIP6*, *ADCYAP1*, and *AQP9*, many of which have been implicated in recent IBD-focused studies but had not previously been identified through traditional differential expression approaches.

Among the 28-gene consensus panel, *CXCL8* and *OSM* were the most recurrently selected genes across the nine classifiers, appearing in the important feature sets of eight and nine models respectively. Consistent with this broad algorithmic consensus, *CXCL8* and *OSM* also displayed the strongest coordinated expression in CD samples ($R^2 = 0.87$, $p < 0.001$), compared to moderate co-expression in controls ($R^2 = 0.55$, $p < 0.001$), suggesting a disease-specific amplification of inflammatory signaling (Fig. 4B).

Internal evaluation of the consensus gene panel on the discovery cohort

Following consensus gene selection, the discriminatory capacity of the 28-gene signature was characterized on Dataset 1 through complementary analyses targeting both individual gene behavior and collective classification performance. PCA with loading analysis was applied to examine multivariate class separation and identify the principal drivers of variance within the panel (Fig. 5A–B); per-gene AUROC and fold-change were computed to assess the univariate discriminatory contribution of each gene (Fig. 5C); and an MLP classifier was trained and evaluated within Dataset 1 to quantify the combined predictive power of the signature (Fig. 5D).

Principal component analysis revealed clear class separation for both gene sets, with distinct structural characteristics. The consensus panel showed tighter clustering of healthy controls with greater dispersion among CD patients, consistent with disease heterogeneity. Loading analysis identified key metabolic genes (*HK2*, *BACE2*, *TXNDC5*) and inflammatory mediators (*AQP9*, *TNFAIP6*, *CSF3R*, *CXCL8*, *OSM*) as the primary drivers of PC1, driving the discriminatory pattern.

Univariate AUROC analysis identified eight consensus genes (*CXCL8*, *OSM*, *AQP9*, *FCGR3A*, *FCGR1A*, *TNFAIP6*, *HK2*, *ELL2*) achieving near-perfect discrimination (AUROC > 0.9), while five genes (*C6*, *SEMA3E*, *RPA1*, *ADRA1B*, *FCER1A*) showed protective associations (AUROC < 0.1). Notably, discriminatory power was not strictly correlated with fold-change magnitude, emphasizing the value of multivariate selection approaches.

The MLP classifier trained on the 28-gene panel and evaluated on the held-out 30% test partition achieved excellent performances (Accuracy = 0.962, AUROC = 0.990, and Brier score = 0.111), with *CXCL8*, *ELL2*, *FCGR3A*, *CYP2E1*, and *TXNDC5* emerging as

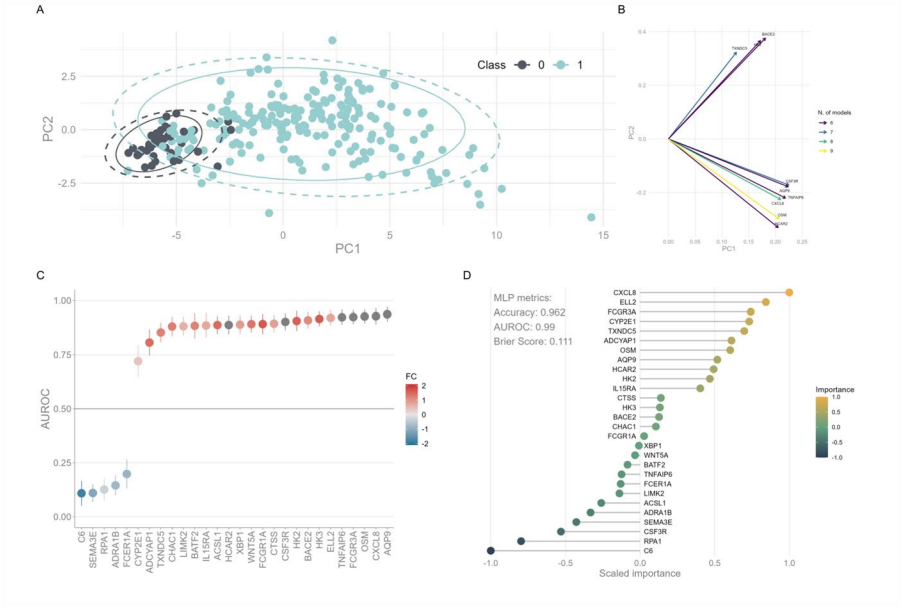

Fig. 5 Univariate and multivariate characterization of the 28-consensus gene panel on the discovery cohort. (**A**) Principal Component analysis (PCA) of the Dataset 1 restricted to the 28 consensus genes. Points are colored by class (0 = HC, 1 = CD); axes report variance explained). (**B**) PCA loading plot for the 28 genes; arrow length reflects loading magnitude, and color encodes how many classifiers selected each gene in the training stage (higher values indicate recurrent selection across models). (**C**) Per-gene univariate discriminative performance reported as AUROC (y-axis, genes ordered by ascending AUROC), with point color indicating log2 Fold change (CD vs HC). Horizontal line marks AUROC = 0.5. (D) Variable importance of a MLP classifier trained on the 28-gene panel and evaluated on a held-out 30% test partition (Accuracy = 0.962, AUROC = 0.99, Brier score = 0.111). Scaled Olden importance scores are shown for all 28 genes; color encodes direction and magnitude of contribution (positive values associated with CD prediction; negative values associated with HC prediction). MLP, multi-Layer Perceptron; HC, healthy Controls; CD, Crohn's disease

the top contributors to the classification decision. Taken together, these results provided a comprehensive characterization of the 28-gene signature on the discovery dataset, serving as a reference baseline against which its performance in independent cohorts could be evaluated.

Cross-cohort signature evaluation

Preservation of discriminatory capacity in an independent cohort

To assess whether the 28-gene consensus panel retained its discriminatory structure in an independent cohort, the same analytical approaches applied internally on Dataset 1 were applied to Dataset 2 (*PRJNA565216* and *PRJNA985602*). PCA revealed partial but directionally consistent class separation between CD and healthy controls along PC1, consistent with the discovery cohort (Supplementary Figure 4A).

Univariate AUROC analysis confirmed that most genes retained their classification trends, with *BACE2*, *HK2*, *WNT5A*, and *ELL2* remaining top performers and *ADRA1B*, *C6*, and *RPA1* continuing to associate with healthy controls (Supplementary Figure 4B). AUROC values were moderately reduced compared to the discovery cohort but remained consistent for an independent dataset.

Multivariate evaluation through a within-cohort MLP classifier trained using the 28-gene panel confirmed these trends, achieving an AUROC of 0.821, an accuracy of 0.8, and a Brier score of 0.18 (Fig. 6). This result confirmed that the consensus gene set supported supervised classification in a new and independent cohort even when a local

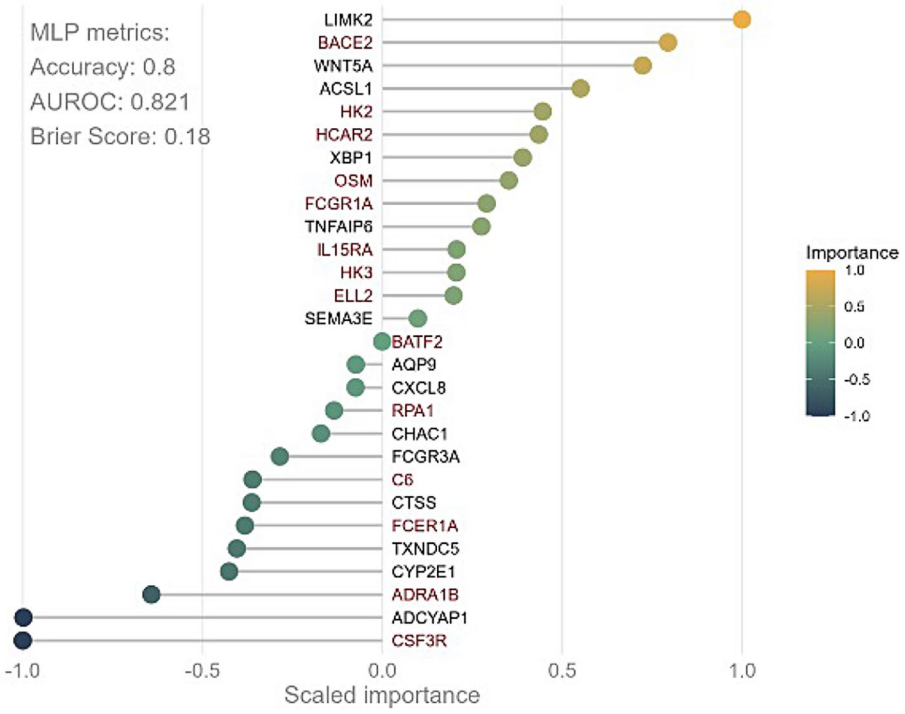

Fig. 6 External evaluation of the 28-gene consensus panel on the independent evaluation cohort. Performance and variable importance of a within-cohort MLP classifier trained on the 28 genes and tested on the external evaluation cohort (Dataset 2). Reported metrics include accuracy, AUROC, and Brier Score. Scaled Olden importance scores are shown for all 28 genes; color encodes direction and magnitude of contribution (positive values associated with CD prediction; negative values associated with HC prediction). Gene names highlighted in red indicate the 14 genes showing concordant importance direction across both Dataset 1 and Dataset 2 evaluations, which were selected as the refined signature and carried forward to the minimal signature identification phase

model was trained from scratch. Notably, MLP variable importance analysis identified 14 genes with consistent directional contributions when compared to Dataset 1 (Figs. 5D and 6).

Benchmarking against single-algorithm and differential expression-based methods

Performance evaluation on the external evaluation cohort

To contextualize the performance of the TENTACLES 28-gene consensus signature relative to conventional biomarker discovery approaches, a systematic benchmarking analysis was performed on Dataset 2. Two classes of comparators were evaluated under identical conditions: DESeq2-derived gene sets assembled from Dataset 1 at multiple size thresholds, and the sets of genes identified as important by each individual classifier during the Discovery Phase. Each gene set was assessed on Dataset 2 using a within-Dataset 2 MLP classifier trained under cross-validation, yielding Accuracy, ROC-AUC, and Brier score as performance metrics (Table 2).

Differential expression analysis of Dataset 1 yielded 582 statistically significant DEGs (padj < 0.05 and $|\log2FC| \geq 1$, Supplementary Material 1 and Fig. 7). These were used to assemble gene sets of increasing size (top28, top50, top100, top200, top300, top400 and all 582 DEGs), ranked by adjusted p-value, for benchmarking against the TENTACLES signature.

Table 2 Benchmarking of gene signatures on the external evaluation cohort (Dataset 2). Classification performance of the TENTACLES 28-gene consensus signature compared to signatures derived from individual classifiers and from differential expression-based gene sets of increasing size, all evaluated on Dataset 2 under identical conditions using a within-cohort MLP classifier trained under cross-validation. For single-algorithm comparators, the signature corresponds to all genes passing each model's variable importance filter during the discovery phase. DEG-based comparators were assembled by ranking the 582 statistically significant DEGs identified in Dataset 1 (padj < 0.05, |log2FC| ≥1) by adjusted *p*-value and selecting the top k genes at predefined thresholds. Performance is reported as Accuracy, ROC-AUC, and Brier score (with 1 – Brier score shown in parentheses). The aggregated score is computed as the sum of Accuracy, ROC-AUC, and (1 – Brier score)

Model producing signature	Number of important genes	Accuracy	ROC-AUC	Brier Score (1 – Brier Score)	Aggregated Score
C5 Rules	29	0.72	0.79	0.18 (0.82)	2.33
PLS	116	0.61	0.62	0.25 (0.75)	1.98
Random Forest	540	0.65	0.72	0.19 (0.81)	2.18
Bag-MLP	421	0.63	0.72	0.20 (0.80)	2.15
Decision Tree	14	0.74	0.87	0.18 (0.82)	2.43
MLP	421	0.63	0.72	0.20 (0.80)	2.15
SVM-linear	64	0.76	0.82	0.18 (0.82)	2.40
XGBoost	2	0.72	0.55	0.21 (0.79)	2.06
lightGBM	20	0.70	0.67	0.19 (0.81)	2.18
MARS	4	0.72	0.66	0.19 (0.81)	2.19
TENTACLES	**28**	**0.8**	**0.82**	**0.18 (0.82)**	**2.44**
DEG – top28	28	0.74	0.82	0.19 (0.81)	2.37
DEG – top50	50	0.78	0.76	0.17 (0.83)	2.37
DEG – top100	100	0.76	0.76	0.18 (0.82)	2.34
DEG – top200	200	0.65	0.62	0.23 (0.77)	2.04
DEG – top300	300	0.63	0.70	0.21 (0.79)	2.12
DEG – top400	400	0.58	0.61	0.24 (0.76)	1.95
DEG all set	582	0.72	0.60	0.19 (0.81)	2.13

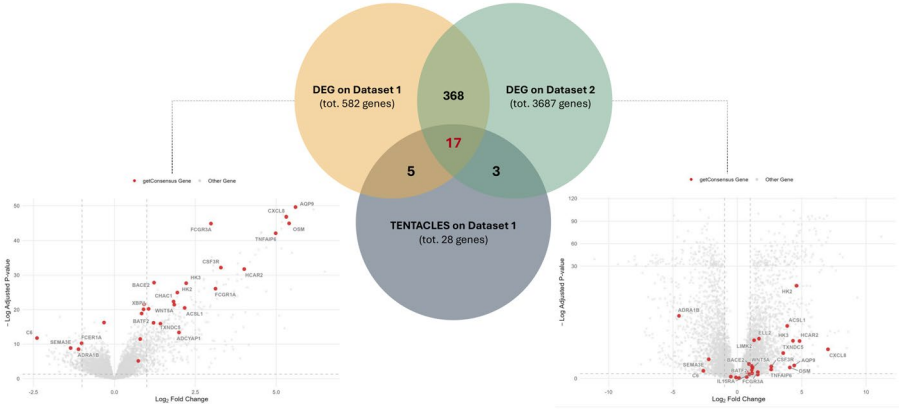

Fig. 7 Cross-cohort reproducibility of differential expression signatures and overlap with the TENTACLES consensus panel. Three-way venn diagram illustrating the overlap among the DEG set from Dataset 1 (582 genes; padj < 0.05, |log2FC| ≥1), the DEG set from Dataset 2 (3687 genes; same thresholds), and the TENTACLES consensus panel derived from Dataset 1 (28 genes). Of the 582 Dataset 1 DEGs, only 363 were also detected in Dataset 2, indicating that approximately 90% of Dataset 2 DEGs were not reproduced in Dataset 1. In contrast, 17 of the 28 TENTACLES consensus genes were independently identified as differentially expressed in both cohorts. Flanking volcano plots display the differential expression landscape of Dataset 1 (left) and Dataset 2 (right); red points highlight the 28 consensus genes in each cohort

The TENTACLES 28-gene consensus signature achieved the highest Accuracy across all comparators (0.8), a ROC-AUC of 0.82, and a Brier score of 0.18. Among single-classifier-derived signatures, that of the Decision Tree achieved the highest ROC-AUC of the entire benchmarking (0.87), but at the cost of lower Accuracy (0.74) relative to the TENTACLES signature. The SVM-linear-derived signature matched the TENTACLES signature on both ROC-AUC (0.82) and Brier score (0.18) but achieved lower Accuracy (0.76). Overall, no individual classifier-derived signature, however, ranked consistently among the top performers across all three metrics simultaneously. Among DEG-derived gene sets, performance did not increase monotonically with gene set size: the top50 DEG-derived signature yielded the lowest Brier score of the entire comparison (0.17) but showed lower ROC-AUC (0.76) and Accuracy (0.78) relative to the TENTACLES signature. At larger thresholds, performance of DEG-derived signatures deteriorated markedly, with gene sets of 200 genes or more showing Accuracy below 0.65 and Brier scores above 0.21.

To assess overall discriminatory balance, an aggregate score was computed as the sum of Accuracy, ROC-AUC, and (1 – Brier score) for each comparator. The TENTACLES signature obtained the highest aggregate score (2.44), followed by the Decision Tree-derived signature (2.43) and the SVM-linear-derived signature (2.40), confirming that its advantage lies not in any single metric but in sustained performance across all three metrics simultaneously.

Reproducibility of differentially expressed genes across cohorts

To further contextualize the robustness of the TENTACLES consensus approach, an independent DEG analysis was performed on Dataset 2 and its results were compared with those from Dataset 1. Dataset 2 analysis identified 3,687 differentially expressed genes (padj < 0.05, |log2FC| ≥1; Supplementary Material 2 and Fig. 7). When intersected with the 582 DEGs from Dataset 1, only 368 genes were found in common, meaning that about 90% of Dataset 2 DEGs were not detected as differentially expressed in Dataset 1 (Fig. 7). This marked instability reflected the sensitivity of differential expression analysis to cohort-specific factors such as patient demographics, study design, and underscores the difficulty of deriving reproducible gene signatures through this approach alone.

In stark contrast, the TENTACLES consensus panel demonstrated markedly greater cross-cohort stability: 17 of the 28 consensus genes (~61%) were independently identified as differentially expressed in both Dataset 1 and Dataset 2, including *FCGR3A*, *CSF3R*, *HK2*, *TNFAIP6*, *WNT5A*, *CXCL8*, *ACSL1*, *HK3*, *ADRA1B*, *C6*, *TXNDC5*, *SEMA3E*, *BATF2*, *HCAR2*, *AQP9*, *OSM*, and *BACE2*. Achieving this level of cross-cohort reproducibility with only 28 features, approximately 95% fewer than the Dataset 1 DEG set and 99% fewer than the Dataset 2 DEG set, highlighted the efficiency of ensemble-based feature selection in discovering a compact yet stable discriminatory signature. Equally important, the remaining 11 consensus genes, which did not meet conventional DEG thresholds in at least one cohort, represented features with consistent multivariate discriminatory power that a threshold-based approach would have missed entirely.

Taken together, these findings indicated that the TENTACLES consensus approach captured a more reproducible and informative subset of disease-relevant genes than conventional differential expression strategies.

Minimal signature identification phase

Signature refinement based on cross-cohort importance concordance

The signed importance profiles of the Dataset 1 and Dataset 2 within-cohort MLP models were compared to identify genes whose directional contribution to classification was consistent across both independent cohorts. Of the 28 consensus genes, 14 showed concordant importance direction, positive (associated with CD) or negative (associated with HC), in both evaluations: *BACE2*, *HK2*, *HCAR2*, *OSM*, *FCGR1A*, *IL15RA*, *HK3*, *ELL2*, *BATF2*, *RPA1*, *C6*, *FCER1A*, *ADRA1B*, and *CSF3R* (Figs. 5D and 6). The remaining 14 genes showed discordant signs between the two cohorts, a pattern likely reflecting the influence of dataset-specific covariates, such as differences in patient age, tissue composition, or residual technical variability, rather than genuinely opposing biological roles. These 14 concordant genes were therefore selected as the refined signature and carried forward to the Minimal Signature Identification Phase.

Minimal signature identification

The 14-gene refined set was carried into Dataset 3 with the dual objective of identifying the minimal gene combination sufficient for accurate disease stratification and of confirming that the discriminatory signal identified during the supervised phases was robust enough to emerge in a fully unsupervised setting. All possible combinations of the 14 candidate genes were systematically evaluated on Dataset 3 (*PRJNA702434*), applying six unsupervised clustering methods – GMM, hierarchical clustering, k-means, and k-means applied to PCA, t-SNE, and UMAP embeddings – to each gene subset and ranking all combinations by mean F1-score across the six different methods.

The top ten gene combinations (C1–C10) demonstrating highest classification performance are reported in Fig. 8. Most combinations achieved accuracy, F1-score, and precision above 75%, confirming strong discriminatory capacity under unsupervised conditions. Hierarchical clustering showed lower precision on average (mean = 0.55, IQR = [0.44–0.78]), while k-means on t-SNE and UMAP embeddings achieved the most consistent results across metrics and combinations.

The best-performing combination (C1) comprised five genes: *CSF3R*, *OSM*, *BACE2*, *HK2*, and *IL15RA*, achieving an accuracy of 0.77, F1-score of 0.72, precision of 0.70, and a recall of 0.79. The second-ranked combination (C2) expanded to nine genes by adding *ELL2*, *ADRA1B*, *RPA1*, and *FCER1A*, maintaining comparably high performance. Additional high-ranking combinations varied in size from four to nine genes, with the five-gene core recurring across the majority of top-ranked combinations. The consistent co-occurrence of these five genes across high-ranking combinations confirmed its centrality as the minimal sufficient signature. This set was therefore carried forward as the pre-specified gene list for a final independent unsupervised validation.

Independent unsupervised validation of the 5-gene core signature

To provide a fully pre-specified assessment of the minimal core signature, the five-gene set (*CSF3R*, *OSM*, *BACE2*, *HK2* and *IL15RA*) was evaluated on the independent cohort Dataset 4 (*PRJNA728646* and *PRJNA961704*), without any additional gene selection or model fitting. Among the six unsupervised clustering methods evaluated during the minimal signature identification phase, k-means on t-SNE embeddings had yielded among the most consistent performance across gene combinations and metrics; it was

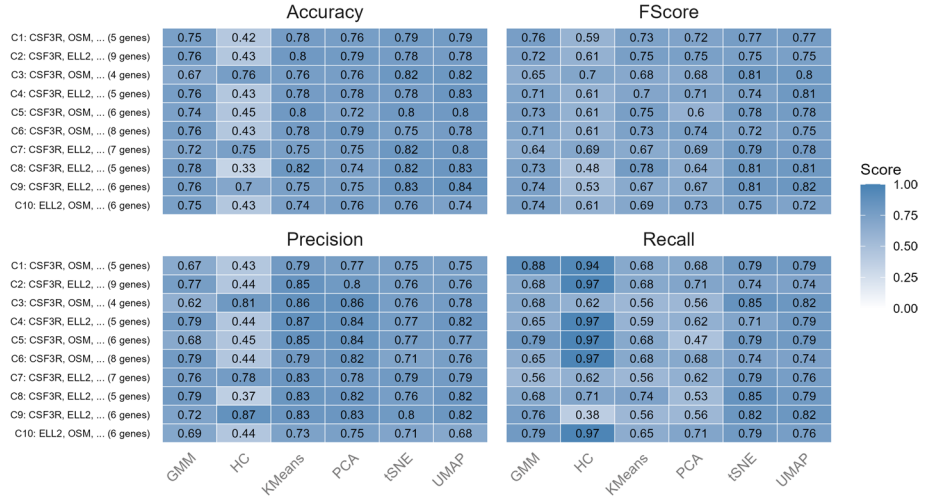

Fig. 8 Performance metrics of the top 10 gene combinations derived from the 14-gene refined signature across six unsupervised clustering methods on the independent validation cohort. Each heatmap panel reports a different evaluation metric: accuracy, F1-score, precision, and recall. Rows correspond to the top 10 gene combinations (C1–C10), selected based on the highest average F1-score across all clustering methods. Columns represent the clustering algorithms applied: gaussian mixture model (GMM), hierarchical clustering (HC), k-means, and k-means on the low-dimensional embeddings from PCA, t-SNE, and UMAP. The color scale reflects the magnitude of each metric, with darker shades indicating better performance. C1: *CSF3R, OSM, BACE2, HK2, IL15RA;* C2: *CSF3R, ELL2, ADRA1B, OSM, RPA1, BACE2, HK2, IL15RA, FCER1A*; C3: *CSF3R, OSM, BACE2, BATF2*; C4: *CSF3R, ELL2, BACE2, HK2, IL15RA;* C5: *CSF3R, OSM, BACE2, HK2, IL15RA, BATF2*; C6: *CSF3R, OSM, RPA1, HCAR2, BACE2, HK2, IL15RA, FCER1A*; C7: *CSF3R, ELL2, OSM, HCAR2, BACE2, HK2, FCER1A*; C8: *CSF3R, ELL2, OSM, BACE2, HK2;* C9: *CSF3R, ELL2, OSM, BACE2, HK2, BATF2*; C10: *ELL2, OSM, RPA1, BACE2, HK2, IL15RA*

therefore applied as the sole method on Dataset 4. The expression matrix was restricted to the five signature genes, subjected to t-SNE dimensionality reduction, and partitioned into two clusters via k-means. Class labels were used exclusively a posteriori to evaluate clustering performance. The 5-gene core signature achieved an accuracy of 0.72, precision of 0.82, recall of 0.71, and F1-score of 0.76 (Supplementary Figure 5), confirming that the discriminatory signal captured by the consensus approach could also generalize to a completely held-out independent cohort under fully unsupervised conditions.

Computational cost analysis

To provide a practical estimate of the computational burden associated with the TENTACLES workflow, runtime and memory usage were profiled for the main functions under the hardware and execution settings adopted in this study (Supplementary Table 3). Overall, *preProcess()*, *getConsensus()*, and *testConsensus()* showed limited computational requirements, particularly when plot generation was disabled, whereas *runClassifiers()* and *valConsensus()* represented the most computationally demanding steps of the pipeline. As expected, the cost of *runClassifiers()* depended on the specific algorithm and configuration used, while the cost of *valConsensus()* increased with the number of genes included in the combinatorial search space. Parallel execution reduced runtime for computationally intensive models when applicable, while peak memory usage remained overall manageable on a standard workstation. These results indicate that, although the most demanding steps of the workflow benefit from adequate computational resources, the TENTACLES pipeline can be executed in practice on a conventional desktop environment.

Discussion

This study presents TENTACLES, a consensus-based machine learning framework designed to prioritize robust and reproducible transcriptional features across heterogeneous RNA-seq datasets. By aggregating outputs from multiple Machine Learning classifiers and applying rigorous validation across four independent cohorts, TENTACLES enabled extraction of compact, generalizable gene signatures that are resilient to dataset-specific biases and modeling assumptions.

Unlike traditional univariate differential expression analyses that evaluate genes in isolation, TENTACLES integrates multiple supervised learning algorithms to identify features carrying predictive power across models and datasets. This multivariate consensus strategy effectively captures non-linear, weakly expressed, or combinatorially active features that conventional statistical thresholds often miss [50]. The framework also incorporates mechanisms to handle imbalanced class distributions, a frequent challenge in clinical datasets. Notably, none of the 28 consensus-derived genes appeared in the KEGG IBD pathway annotation [51], and only 17 overlapped with genes consistently differentially expressed across cohorts. These findings highlight the complementary value of data-driven prioritization frameworks, particularly for diseases with complex, multi-axis pathophysiology such as CD. The marked instability of differential expression results across cohorts, with about 90% of DEGs identified in Dataset 2 absent from those of Dataset 1, further underscores this complementarity, and highlights the fragility of threshold-based approaches when applied to heterogeneous transcriptomic data.

The benchmarking analysis further contextualized the advantage of the ensemble-based approach over conventional strategies. While individual classifier-derived signatures occasionally outperformed TENTACLES on isolated metrics, no single comparator ranked consistently among the top performers across all performance metrics simultaneously. DEG-based gene sets showed markedly degraded performance at larger thresholds, confirming that indiscriminate expansion of a differential expression panel does not translate into generalization gains. The TENTACLES 28-gene consensus signature achieved the highest aggregate score across all benchmarked approaches, reflecting a balanced discriminatory capacity across multiple performance metrics. While certain single-algorithm-derived signatures, approached comparable results, no individual approach matched the consistency of TENTACLES across all metric performance simultaneously.

A central strength of TENTACLES lies in its ability to extract stable transcriptional signals that generalize across fully independent cohorts, encompassing diverse experimental conditions and patient populations. The 28-gene signature prioritized during the ensemble-based signature discovery phase maintained substantial discriminative power in external cohorts, with 14 genes preserving concordant variable importance and directionality. This refined subset, subsequently evaluated in a third independent cohort through unsupervised clustering, confirmed good overall stratification capabilities and allowed to identify a 5-gene core signature achieving highest discriminatory performance across six different clustering methodologies. The 5-gene core signature was then subjected to unsupervised validation in a fourth fully independent cohort, providing an unbiased confirmation of its generalizability. Among the most stable features, a minimal 5-gene panel comprising *OSM*, *HK2*, *IL15RA*, *BACE2*, and *CSF3R* emerged as sufficient for robust CD classification across all analytical frameworks.

Each gene in the minimal signature has established roles in immune regulation and metabolic dysfunction relevant to CD pathogenesis. *OSM*, a cytokine regulator, drives epithelial barrier disruption and anti-TNF therapy resistance [52–54]. *HK2* gene, encoding a key glycolytic enzyme, correlates with inflammatory severity in IBD patients [55]. *IL15RA* promotes immune cell survival and proliferation, with documented upregulation in IBD [56, 57]. *BACE2*, a membrane protease, represents an emerging IBD biomarker [2, 58]. Finally, *CSF3R*, which regulates neutrophil homeostasis, shows altered expression patterns suggesting disrupted immune cell trafficking in CD [59, 60]. This biological coherence, combined with robust statistical validation, strengthens confidence in the clinical relevance of these signatures.

The identified gene panels offer multiple translational applications. The 5-gene minimal signature could serve as a diagnostic classifier requiring fewer resources than comprehensive transcriptomic profiling. The 14-gene refined panel provides additional granularity for molecular subtyping and therapeutic stratification. Both signatures demonstrated consistent performance across supervised and unsupervised analytical frameworks, suggesting potential utility for diverse clinical applications including disease monitoring, treatment response prediction, and patient stratification for clinical trials.

Some limitations merit consideration. Although TENTACLES was validated here exclusively on Crohn's disease transcriptomic data, its disease-agnostic architecture makes it applicable to other inflammatory, autoimmune, and complex disease contexts. Future studies should extend its application to independent immunological and multi-disease datasets to further establish its generalizability beyond Crohn's disease. Furthermore, the analysis was restricted to genes with annotations in GO or KEGG databases, which may have excluded potentially informative transcripts, including poorly characterized genes and long intergenic non-coding RNAs, from the biomarker discovery space. Future analyses applying TENTACLES without this filter may identify additional candidate signatures beyond the annotated gene space. The analysis relied on publicly available bulk RNA-seq datasets with inherent experimental and clinical heterogeneity. Despite rigorous batch correction and quality control measures, residual confounding cannot be entirely excluded. Future studies should validate these signatures in prospectively collected cohorts with standardized protocols. Additionally, single-cell RNA sequencing could provide insights into cell-type-specific expression patterns underlying the bulk tissue signatures. Integration with clinical metadata, including disease phenotypes, treatment responses, and longitudinal outcomes, would further enhance the clinical utility of these biomarkers.

Beyond Crohn's disease, TENTACLES could represent a broadly applicable approach for robust biomarker discovery in complex diseases. The framework's modular design is intended to facilitate adaptation to different data types, disease contexts, and clinical objectives. Its emphasis on consensus-based feature selection and multi-cohort validation addresses persistent reproducibility challenges in computational biology and precision medicine. As transcriptomic datasets continue expanding, such ensemble approaches will become increasingly valuable for extracting reliable, clinically actionable insights from heterogeneous biological data.

Conclusions

This study demonstrates that TENTACLES effectively identifies minimal, reproducible gene signatures for Crohn's disease classification through consensus-based machine learning. The resulting biomarker panels offer biologically interpretable features with potential applications in diagnosis, molecular stratification, and therapeutic monitoring. This framework is intended to represent a scalable, generalizable approach for robust feature discovery that could advance precision medicine efforts across immune-mediated diseases and beyond.

Abbreviation

CD	Crohn's Disease
HC	Healthy Controls
IBD	Inflammatory Bowel Disease
ML	Machine Learning
DEGs	Differentially Expressed Genes
PCA	Principal Component Analysis
PVCA	Principal Variance Component Analysis
MLP	Multi-Layer Perceptron
SVM	Support Vector Machines
MARS	Multivariate Adaptive Regression Splines
XGBoost	Extreme Gradient Boosting
PLS	Partial Least Squares

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-026-00568-8.

Supplementary material 1 - Differential expression analysis of Dataset 1

Supplementary material 2 - Differential expression analysis of Dataset 2

Supplementary material 3 - Supplementary figures and tables

Supplementary material 4 - Code pipeline

Acknowledgements
The authors thank Mattia Granchi for designing the logo of the TENTACLES package.

Author contributions
Conceptualization: GM, MN, HIN. Data curation: AFC, GM, Formal analysis: GM, GDSM, AFC, AF, SL, CS. Methodology: GM, HIN, GDSM, AFC, MN. Project Administration: HIN, DM, AC, FS. Resources: HIN, AC, DM. Software: GM, GDSM, AFC, MN, AF. Supervision: DM, HIN. Validation: GM, GDSM, AFC, MN, SL. Writing – original draft: GM, MN, HIN, GDSM, AFC, AF, SL, CS, AC, FS, DM.

Funding
This work was supported by: Department of Medical Biotechnologies of the University of Siena NextGenerationEU project PNRR MUR Extended Partnership Initiative on Emerging Infectious Diseases project (PNRR PE13 INF_ACT—CUP:B63C22001400007) Calmette & Yersin PhD Grant from the Pasteur Network (AFC).

Data availability
The datasets supporting the conclusions of this article are available in the Zenodo repository at https://zenodo.org/records/19632205. The TENTACLES R package, including all core functions and vignettes, is available on GitHub at https://github.com/Giomu/TENTACLES. The full code used to perform the analyses is provided in Supplementary Material 4 and within the GitHub and Zenodo repositories.

Declarations

Ethical approval
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 January 2026 / Accepted: 14 May 2026
Published online: 10 June 2026

References

1. Nazari MHD, Ghorbaninejad M, Shahrokh S, Meyfour A. Biomarker discovery for non-invasive diagnosis of inflammatory bowel disease using blood transcriptomics. Front Immunol. 2025, Jun, 20;16. https://doi.org/10.3389/fimmu.2025.1570374 .
2. Syed AH, Abujabal HAS, Ahmad S, Malebary SJ, Alromema N. Advances in inflammatory bowel disease diagnostics: machine learning and genomic profiling reveal key biomarkers for early detection. Diagn (Basel). 2024, Jun, 4;14(11):1182. https://doi.org/10.3390/diagnostics14111182 PubMed PMID: 38893707; PubMed Central PMCID: PMC11172026.
3. Bao W, Wang L, Liu X, Li M. Predicting diagnostic biomarkers associated with immune infiltration in Crohn's disease based on machine learning and bioinformatics. Eur J Med Res. 2023, Jul, 26;28(1):255. https://doi.org/10.1186/s40001-023-01200-9.
4. Chen KA, Nishiyama NC, Kennedy Ng MM, Shumway A, Joisa CU, Schaner MR, et al. Linking gene expression to clinical outcomes in pediatric Crohn's disease using machine learning. Sci Rep. 2024, Feb, 1;14(1):2667. https://doi.org/10.1038/s41598-024-52678-0.
5. Shah SNA, Parveen R. Differential gene expression analysis and machine learning identified structural, TFs, cytokine and glycoproteins, including SOX2, TOP2A, SPP1, COL1A1, and TIMP1 as potential drivers of lung cancer. Biomarkers. 2025, Feb, 17;30(2):200–15. https://doi.org/10.1080/1354750X.2025.2461698 PubMed PMID: 39888730.
6. Abbas M, El-Manzalawy Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. BMC Med Genomics. 2020, Aug, 28;13(1):122. https://doi.org/10.1186/s12920-020-00771-4.
7. Ghazal H, El-Absawy ESA, Ead W, Hasan ME. Machine learning-guided differential gene expression analysis identifies a highly-connected seven-gene cluster in triple-negative breast cancer. Biomed (Taipei). 2024;14(4):15–35. https://doi.org/10.37796/2211-8039.1467 PubMed PMID: 39777114; PubMed Central PMCID: PMC11703398.
8. McDowell C, Farooq U, Haseeb M. Inflammatory bowel disease, StatPearls [Internet]. Treasure Island (FL): StatPearls publishing. 2025 [cited 2025 Aug 29] http://www.ncbi.nlm.nih.gov/books/NBK470312/PubMed. PMID: 29262182.
9. Seyedian SS, Nokhostin F, Malamir MD. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. JMedLife. 2019;12(2):113–22. https://doi.org/10.25122/jml-2018-0075 PubMed PMID: 31406511; PubMed Central PMCID: PMC6685307.
10. Nikolaus S, Schreiber S. Diagnostics of inflammatory bowel disease. Gastroenterology. 2007, Nov;133(5):1670–89. https://doi.org/10.1053/j.gastro.2007.09.001 PubMed PMID: 17983810.
11. Scheurlen KM, Parks MA, Macleod A, Galandiuk S. Unmet challenges in patients with Crohn's disease. JCM. 2023, Jan;12(17):5595. https://doi.org/10.3390/jcm12175595.
12. Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J. Clin. Invest. 2014, Aug;124(8):3617–33. https://doi.org/10.1172/JCI75436 PubMed PMID: 25003194; PubMed Central PMCID: PMC4109533.
13. Loberman-Nachum N, Sosnovski K, Di Segni A, Efroni G, Braun T, BenShoshan M, et al. Defining the celiac disease Transcriptome using clinical pathology specimens reveals biologic pathways and supports diagnosis. Sci Rep. 2019, Nov, 7;9(1):16163. https://doi.org/10.1038/s41598-019-52733-1 PubMed PMID: 31700112; PubMed Central PMCID: PMC6838157.
14. Mo A, Krishnakumar C, Arafat D, Dhere T, Iskandar H, Dodd A, et al. African ancestry proportion influences ileal gene expression in inflammatory bowel disease. Cellular Mol Gastroenterol Hepatol. 2020;10(1):203–05. https://doi.org/10.1016/j.jcmgh.2020.02.001 PubMed PMID: 32058087; PubMed Central PMCID: PMC7296223.
15. Garrido-Trigo A, Corraliza AM, Veny M, Dotti I, Melón-Ardanaz E, Rill A, et al. Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. Nat Commun. 2023, Jul, 26;14(1):4506. https://doi.org/10.1038/s41467-023-40156-6 PubMed PMID: 37495570; PubMed Central PMCID: PMC10372067.
16. Friedrich M, Pohin M, Jackson MA, Korsunsky I, Bullers SJ, Rue-Albrecht K, et al. IL-1-driven stromal–neutrophil interactions define a subset of patients with inflammatory bowel disease that does not respond to therapies. Nat Med. 2021, Nov;27(11):1970–81. https://doi.org/10.1038/s41591-021-01520-5 PubMed PMID: 34675383; PubMed Central PMCID: PMC8604730.
17. ENA browser [Internet]. [https://www.ebi.ac.uk/ena/browser/view/PRJNA728646. 2026 Apr 16].
18. Shaw DG, Aguirre-Gamboa R, Vieira MC, Gona S, DiNardi N, Wang A, et al. Antigen-driven colonic inflammation is associated with development of dysplasia in primary sclerosing cholangitis. Nat Med. 2023, Jun;29(6):1520–29. https://doi.org/10.1038/s41591-023-02372-x PubMed PMID: 37322120; PubMed Central PMCID: PMC10287559.
19. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018, Sep, 1;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013, Jan, 1;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
21. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012, Aug, 15;28(16):2184–85. https://doi.org/10.1093/bioinformatics/bts356 PubMed PMID: 22743226.
22. Liao Y, Smyth GK, Shi W. The R package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019, May, 7;47(8):e47. https://doi.org/10.1093/nar/gkz114.
23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010, Jan, 1;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616 PubMed PMID: 19910308; PubMed Central PMCID: PMC2796818.
24. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012, Mar, 15;28(6):882–83. https://doi.org/10.1093/bioinformatics/bts034 PubMed PMID: 22257669; PubMed Central PMCID: PMC3307112.
25. Bioconductor - pvca [Internet]. [cited 2025 Aug 29]. Available from: https://bioconductor.org/packages/release/bioc/html/pvca.html.
26. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Soft. 2017, Mar;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
27. Ripley B, Venables W. Nnet: feed-forward neural networks and multinomial log-linear models [Internet]. 2025. https://cran.r-project.org/web/packages/nnet/index.html. [cited 2025 Aug 29].

28. Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab - an S4 package for kernel methods in R. J Stat Soft. 2004, Nov, 2;11(9):1–20. https://doi.org/10.18637/jss.v011.i09.
29. Kuhn M, Weston S, Culp M, Coulter N. Code) RQ (Author of imported C, code) RR (copyright holder of imported C, et al. C50: C5.0 Decision trees and rule-based models [Internet]. 2025. https://cran.r-project.org/web/packages/C50/index.html. [cited 2025 Aug 29].
30. Shi Y, Ke G, Soukhavong D, Lamb J, Meng Q, Finley T, et al. Lightgbm: light gradient boosting machine [Internet]. 2025. https://cran.r-project.org/web/packages/lightgbm/index.html. [cited 2025 Aug 29].
31. Therneau T, Atkinson B. Port BR (producer of the initial R, maintainer 1999-2017). rpart: recursive partitioning and regression trees [Internet]. 2025. https://cran.r-project.org/web/packages/rpart/index.html. [cited 2025 Aug 29].
32. Milborrow S, Hastie T, Tibshirani R, Miller A, Lumley T. Earth: multivariate adaptive regression splines [Internet]. 2024. https://cran.r-project.org/web/packages/earth/index.html. [cited 2025 Aug 29].
33. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting [Internet]. 2025. https://cran.r-project.org/web/packages/xgboost/index.html. [cited 2025 Aug 29].
34. Rohart F, Gautier B, Singh A, Cao KAL. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017, Nov, 3;13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
35. Kuhn M, Wickham H, Hvitfeldt E, Software P. PBC [Cph, fnd. recipes: preprocessing and feature engineering steps for modeling [Internet]. 2025. https://cran.r-project.org/web/packages/recipes/index.html. [cited 2025 Aug 29].
36. Hvitfeldt E, Software P, PBC. Themis: extra recipes steps for dealing with unbalanced data [Internet]. 2025. https://cran.r-project.org/web/packages/themis/index.html. [cited 2025 Aug 29].
37. Pawley S. stevenpawley/colino [R] [Internet]. 2025. https://github.com/stevenpawley/colino. [cited 2025 Aug 29].
38. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Soft. 2010, Sep;36(11):1–13. https://doi.org/10.18637/jss.v036.i11.
39. Frick H, Kuhn M, Couch S, Software P. PBC [Cph, fnd. workflowsets: create a collection of "tidymodels". Workflows [Internet]. 2025. https://cran.r-project.org/web/packages/workflowsets/index.html. [cited 2025 Aug 29].
40. Greenwell BM, Boehmke B B. Variable importance plots—an introduction to the vip package. R J. 2020;12(1):343. https://doi.org/10.32614/RJ-2020-013.
41. Beck MW. NeuralNetTools: visualization and analysis tools for neural networks. J Stat Soft. 2018, Jul;85(11):1–20. https://doi.org/10.18637/jss.v085.i11.
42. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. 2011, Mar;12(1):77. https://doi.org/10.1186/1471-2105-12-77 PubMed PMID: 21414208; PubMed Central PMCID: PMC3068975.
43. Kuhn [aut M, cre, Software, P, PBC. Tune: tidy tuning tools [Internet]. 2025. https://cran.r-project.org/web/packages/tune/index.html. [cited 2025 Aug 29].
44. Scrucca L, Fraley C, Murphy TB, Raftery AE. Model-based clustering, classification, and density estimation using mclust in R. First. Boca Raton, FL: CRC Press; 2023. p. 1 (Chapman & Hall/CRC the R series).
45. Konopka T. Umap: uniform manifold approximation and projection [Internet]. 2023. https://cran.r-project.org/web/packages/umap/index.html. [cited 2025 Apr 11].
46. van der Maaten Ljp, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008;9(nov):2579–605.
47. van der ML. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014;15(93):3221–45.
48. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014, Dec, 5;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
49. Quinn T. peakRAM: monitor the total and peak RAM used by an expression or function [Internet]. 2017. https://cran.r-project.org/web/packages/peakRAM/index.html. [cited 2026 Apr 18].
50. Liang P, Yang W, Chen X, Long C, Zheng L, Li H, et al. Machine learning of single-cell transcriptome highly identifies mRNA signature by comparing F-score selection with DGE analysis. Mol Ther Nucleic Acids. 2020, Jun, 5;20:155–63. https://doi.org/10.1016/j.omtn.2020.02.004 PubMed PMID: 32169803.
51. Pathway KEGG. Inflammatory bowel disease - reference pathway [Internet]. https://www.kegg.jp/pathway/map05321. [cited 2025 Aug 29].
52. Yang Y, Fu KZ, Pan G. Role of oncostatin M in the prognosis of inflammatory bowel disease: a meta-analysis. World J Gastrointest Surg. 2024, Jan, 27;16(1):228–38. https://doi.org/10.4240/wjgs.v16.i1.228 PubMed PMID: 38328320; PubMed Central PMCID: PMC10845284.
53. Verstockt S, Verstockt B, Machiels K, Vancamelbeke M, Ferrante M, Cleynen I, et al. Oncostatin M is a biomarker of diagnosis, worse disease prognosis, and therapeutic nonresponse in inflammatory bowel disease. Inflamm Bowel Dis. 2021, Oct, 1;27(10):1564–75. https://doi.org/10.1093/ibd/izab032.
54. West NR, Hegazy AN, Owens BMJ, Bullers SJ, Linggi B, Buonocore S, et al. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. Nat Med. 2017, May;23(5):579–89. https://doi.org/10.1038/nm.4307.
55. Weber-Stiehl S, Taubenheim J, Järke L, Röcken C, Schreiber S, Aden K, et al. Hexokinase 2 expression in apical enterocytes correlates with inflammation severity in patients with inflammatory bowel disease. BMC Med. 2024, Oct, 23;22(1):490. https://doi.org/10.1186/s12916-024-03710-7.
56. Silva MA, Menezes J, Deslandres C, Seidman EG. Anti-inflammatory role of Interleukin-15 in Crohn's disease. Inflamm Bowel Dis. 2005, Mar, 1;11(3):219–30. https://doi.org/10.1097/01.MIB.0000160804.52072.6a.
57. Perrier C, Arijs I, Staelens D, Breynaert C, Cleynen I, Covens K, et al. Interleukin-15 receptor α expression in inflammatory bowel disease patients before and after normalization of inflammation with infliximab. Immunology. 2013;138(1):47–56. https://doi.org/10.1111/imm.12014.
58. Hu F, Xiong L, Li Z, Li L, Wang L, Wang X, et al. Deciphering the shared mechanisms of gegen qinlian decoction in treating type 2 diabetes and ulcerative colitis via bioinformatics and machine learning. Front. Med. 2024, Jun, 19;11. https://doi.org/10.3389/fmed.2024.1406149.
59. Ashton JJ, Boukas K, Davies J, Stafford IS, Vallejo AF, Haggarty R, et al. Ileal transcriptomic analysis in paediatric Crohn's disease reveals IL17- and NOD-signalling expression signatures in treatment-naïve patients and identifies epithelial Cells driving differentially expressed genes. J Crohns Colitis. 2021, May, 4;15(5):774–86. https://doi.org/10.1093/ecco-jcc/jjaa236.

60. Carnevale S, Ponzetta A, Rigatelli A, Carriero R, Puccio S, Supino D, et al. Neutrophils mediate protection against colitis and carcinogenesis by controlling bacterial invasion and IL22 production by γδ T Cells. Cancer Immunol Res. 2024, Apr, 2;12(4):413–26. https://doi.org/10.1158/2326-6066.CIR-23-0295.

Publisher's Note