Briefings in Bioinformatics, 2026, 27, bbag211
https://doi.org/10.1093/bib/bbag211
Published: 8 May 2026
Review

OXFORD

When complexity does not pay: benchmarking deep learning and ensemble methods for biomarker discovery

Cyrille Mesue Njume[1], Irene Petracci[2], Sonia Bellini[2], Katarzyna Goljanek-Whysall[3], Leo R. Quinlan[3], Agnieszka Fiszer[4], Barbara Borroni[2], Roberta Ghidoni[2], Asli Kumbasar[1], Ali Cakmak[5,*]

[1]Department of Molecular Biology and Genetics, Ayazaga Campus, Istanbul Technical University, Reşitpaşa, Sarıyer, 34467 Istanbul, Turkey
[2]Molecular Markers Laboratory, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, 25125 Brescia, Italy
[3]Discipline of Physiology, School of Medicine, University of Galway, H91 TH33 Galway, Ireland
[4]Department of Medical Biotechnology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
[5]Department of Computer Engineering, Ayazaga Campus, Istanbul Technical University, Reşitpaşa, Sarıyer, 34467 Istanbul, Turkey

*Corresponding author. Department of Computer Engineering, Ayazaga Campus, Istanbul Technical University, Reşitpaşa, Sarıyer, 34467 Istanbul, Turkey. E-mail: ali.cakmak@itu.edu.tr

Abstract

The integration of multi-omics data holds great promise for identifying robust and clinically relevant biomarkers, yet the increasing complexity of computational methods raises questions about their practical utility. In this study, we present a comprehensive benchmarking framework that evaluates 27 feature selection strategies and 11 predictive models across three real-world disease cohorts: Alzheimer's disease, progressive supranuclear palsy, and breast cancer. We compare traditional machine learning, ensemble-based methods, and state-of-the-art deep learning models in terms of predictive performance, stability, and biological interpretability. Our results reveal that ensemble feature selection consistently improves robustness and accuracy, particularly for compact biomarker panels. Surprisingly, deep learning models did not outperform simpler classifiers such as logistic regression (L.Regression), support vector machines, or multilayer perceptrons, which often achieved comparable or superior results with lower computational cost and greater interpretability. Triple-omics yielded the highest validation, followed by dual-omics and then single-omics (Triple > Dual > Single). Biological validation against five independent databases confirmed the clinical relevance of the identified biomarkers, including both well-established and novel candidates. To support reproducibility and community adoption, we provide a web-based tool for applying our benchmarking pipeline. Our findings advocate for a pragmatic approach to biomarker discovery—prioritizing methodological transparency, reproducibility, and biological insight over algorithmic complexity.

Keywords **biomarker discovery, multi-omics integration, feature selection benchmarking, ensemble rank aggregation, integrative bioinformatics**

Introduction

Biomarker discovery has become a cornerstone of precision medicine, enabling earlier disease detection, a more accurate prognosis, and personalized therapeutic strategies. Early studies primarily relied on single-omics approaches, such as transcriptomics or proteomics, which provided valuable insights but often captured only one layer of biological regulation and yielded limited predictive power. With the rapid expansion of high-throughput technologies, researchers can now generate multi-omics datasets that integrate diverse modalities—including, but not limited to, transcriptomics, epigenomics, proteomics, and metabolomics—thereby capturing complementary regulatory layers and offering a more holistic view of disease processes [1]. Integrating these heterogeneous data modalities holds the promise of identifying more robust biomarkers that not only improve clinical decision-making but also provide mechanistic insight into disease biology [2–5]. However, despite the increasing availability of multi-omics data, extracting clinically meaningful and reproducible biomarker panels remains a major challenge.

The core difficulty arises from the intrinsic characteristics of multi-omics data. First, most datasets are high-dimensional yet low-sample ($n \ll p$), where thousands of molecular features are profiled across only a few hundred patients. This imbalance makes analyses highly susceptible to overfitting and spurious associations—false correlations that arise from random noise, confounding, or chance [6–10]. Second, the heterogeneity of omics layers poses additional challenges: gene expression values are continuous, DNA methylation

Received: November 23, 2025. **Revised:** March 11, 2026. **Accepted:** April 2, 2026

profiles are sparse, and mutation calls are discrete, each reflecting distinct biological processes. Integrating such modalities requires harmonization across measurement scales and distributions, which is nontrivial and risks introducing bias [11, 12]. Third, the limited availability of clinically annotated cohorts exacerbates instability and reduces reproducibility across independent datasets. These challenges make biomarker discovery highly sensitive to small perturbations in the data, often yielding panels that fail to generalize. Single-method feature selection approaches, such as univariate statistical tests or machine learning-based rankers, are particularly vulnerable in this setting, frequently producing unstable biomarker lists with inconsistent biological interpretation [12–14]. Ensemble ranking has been proposed to mitigate such variability and improve feature stability and reproducibility across high-dimensional omics datasets [15–18]. However, comprehensive benchmarking across multiple diseases, omics modalities, and model types remains unexplored. Most existing applications are largely limited to single-omics analyses such as transcriptomics [19, 20], proteomics [21], or general omics feature spaces [22–24], while only a few have benchmarked individual feature-selection methods on multi-omics datasets [25, 26]. More broadly, there remains a lack of comprehensive benchmarking pipelines that jointly evaluate feature selection, predictive modeling, and multi-omics integration strategies. Moreover, very few studies systematically perform external bioinformatic validation of discovered biomarker panels by comparing predicted genes or pathways against independent biological databases—the Comparative Toxicogenomics Database (CTD), which curates chemical–gene–disease relationships and pathway annotations [27]; the Human microRNA Disease Database (HMDD), a manually curated resource of microRNA (miRNA)–disease associations [28]; GeneCards, an integrative compendium of human genes and their disease/phenotype links [29]; and the EWAS-ATLAS, a catalog of Cytosine–Phosphate–Guanine site (CpG)–trait associations from epigenome-wide association studies [30]. This step is critical for assessing the biological interpretability and clinical relevance of computational findings. While some studies include *post hoc* validation through overlap or enrichment analyses against known databases or literature [31–33], these are typically limited in scope and do not constitute broad benchmarking across diseases, modalities, and for each method.

Moreover, more recently, several deep learning-based multi-omics integration frameworks, such as Multi-Omics hypeRgraph integration nEtwork (MORE) [34] and Multi-Omics Graph cOnvolutional NETworks (MOGONET) [35] have been proposed to capture cross-modality relationships and exploit complex nonlinear patterns. However, while such frameworks have demonstrated strong performance across several datasets and integration settings, their comparative stability, reproducibility, and consistency against simpler or ensemble-based approaches remain insufficiently explored. In addition, external validation of biomarker panels—essential for establishing biological and clinical relevance—is still rarely performed in a structured and multi-level manner. As a result, it remains unclear whether increasing methodological complexity consistently yields more reliable and interpretable biomarker panels, or whether simpler and ensemble-based approaches may offer equal or greater robustness.

In this study, we present a comprehensive benchmarking framework for multi-omics biomarker discovery across three real-world disease cohorts. We analyze Alzheimer's disease (AD) using the Religious Orders Study and Memory and Aging Project (ROSMAP) cohort [36], PSP using the Mayo RNAseq Study (MayoRNASeq) data [37], and breast cancer (BC) using The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) cohort [38]. We systematically evaluate 27 biomarker selection strategies—spanning single rankers and ensemble ranking methods—and 11 predictive models, ranging from traditional machine learning algorithms to advanced integration approaches such as MORE and MOGONET. Beyond computational benchmarking, we validate biomarker panels against four independent biological reference databases (CTD [27], HMDD [28], GeneCards [29], or EWAS-ATLAS [30], and Gene Functional Enrichment Profiling Tool (g:Profiler) [39] for pathway enrichment cross-checked against CTD-pathways), ensuring both interpretability and clinical relevance. Together, this work establishes a robust, reproducible, and generalizable pipeline for multi-omics biomarker discovery and provides new insights into when methodological complexity offers genuine advantages over simpler, more interpretable strategies.

Materials and methods

This section outlines the experimental design, datasets, and computational framework used in this study. We describe the selection and preprocessing of multi-omics datasets, followed by the feature selection, ensemble ranking, and benchmarking procedures applied to evaluate model performance across integration levels.

Study cohorts

Experiments were conducted using three real-world, publicly available multi-omics datasets: the ROSMAP (AD) [36], the MayoRNASeq cohort (PSP) [37], and the TCGA-BRCA cohort (BC) accessed via the UCSC Xena Browser [38]. Key study characteristics and data types are summarized in Table 1.

Religious Orders Study and Memory and Aging Project

ROSMAP is a longitudinal cohort study combining the ROSEMAP Project to investigate aging and AD [40]. Participants undergo annual cognitive assessments and consent to brain donation, allowing linkage of clinical trajectories with postmortem molecular data. For this study, we used dorsolateral prefrontal cortex profiles of microRNA expression, messenger RNA (mRNA) expression, and DNA methylation, together with diagnostic information on AD. Only individuals with complete molecular and clinical data after quality control were included [41].

The Mayo RNAseq Study

The MayoRNASeq study profiles neurodegenerative diseases using postmortem brain tissue from the Mayo Clinic Brain Bank [37]. We focused on temporal cortex samples from participants diagnosed with PSP and matched controls, with available RNA-Seq gene expression, proteomics, metabolomics, and clinical annotations. Samples with incomplete molecular or phenotypic information were excluded after standard quality control.

Breast Cancer Cohort (TCGA-BRCA)

TCGA-BRCA is a large-scale, multi-omics resource characterizing the molecular and clinical features of BC [42]. We used bulk RNA-seq transcriptomics, microRNA expression, and DNA methylation (Illumina HumanMethylation450) together with clinical annotations obtained via UCSC Xena. Only samples with complete profiles for these modalities and clinical data were retained.

Table 1 Overview of study characteristics

Cohort	Disease/Condition	Data types	Sample size	Data source
ROSMAP	AD	Transcriptomics, microRNA, methylation, clinical	378	Synapse[a]
TCGA-BRCA	BC	Transcriptomics, methylation, microRNA, clinical	1229	UCSC Xena[b]
MayoRNASeq	PSP	Transcriptomics, metabolomics, proteomics, clinical	437	Synapse[c]

[a]ROSMAPDataset. [b]UCSCXenaTCGA-BRCADataset. [c]MayoRNASeqDataset. Note: The sample sizes reported here are based on the available clinical datasets for the specific disease and corresponding controls. Different sample counts may apply to specific omics subsets, as detailed in Supplementary Methods S1.

Table 2 Summary of prepared datasets for single-, dual-, and multi-omics analyses

Cohort	Omics combination	Samples	Features	Classes	Class ratio
AD	miRNA only	378	200	AD/Control	209/169
	mRNA only	378	200	AD/Control	209/169
	Meth only	375	200	AD/Control	207/168
	mRNA + miRNA	375	200 + 200	AD/Control	209/169
	mRNA + Meth	375	200 + 200	AD/Control	207/168
	miRNA + Meth	375	200 + 200	AD/Control	207/168
	mRNA + miRNA + Meth	375	200 + 200 + 200	AD/Control	207/168
PSP	mRNA only	162	200	PSP/Control	84/78
	Prot only	111	200	PSP/Control	83/29
	Metab only	98	200	PSP/Control	78/20
	mRNA + Prot	111	200 + 200	PSP/Control	83/29
	mRNA + Metab	98	200 + 200	PSP/Control	78/20
	Metab + Prot	97	200 + 200	PSP/Control	77/20
	mRNA + Prot + Metab	97	200 + 200 + 200	PSP/Control	77/20
BC	miRNA only	158	200	BC/Control	83/75
	mRNA only	236	200	BC/Control	124/112
	Meth only	375	200	BC/Control	106/96
	mRNA + miRNA	158	200 + 200	BC/Control	83/75
	mRNA + Meth	108	200 + 200	BC/Control	57/51
	miRNA + Meth	175	200 + 200	BC/Control	92/83
	mRNA + miRNA + Meth	108	200 + 200 + 200	BC/Control	57/51

Abbreviations: miRNA, microRNA expression data; mRNA, gene expression data; Meth, DNA methylation data; Prot, proteomics data; Metab, metabolomics data.

Data preprocessing and dataset construction

All data cleaning, preprocessing, and preparation steps were performed within a unified Python 3.11.9 pipeline to ensure consistency across cohorts and omics modalities. Clinical metadata were harmonized with omics expression matrices using biospecimen mapping files, duplicate samples and features were removed, and samples with incomplete multi-omics profiles or missing clinical annotations were excluded. Features with excessive missingness were removed, and the remaining missing entries were imputed using a K-nearest neighbors imputer with distance-weighted averaging. Molecular identifiers (microRNA names and Ensembl gene IDs) were standardized against current reference databases to improve cross-study comparability. Full details of imputation thresholds, identifier mapping, and software versions are provided in Supplementary Methods S1.

To reduce noise and improve interpretability, we applied modality-specific filtering and normalization steps inspired by the MOGONET framework [35]. Low-variance features were removed using modality-specific thresholds, and preselection was performed by Analysis of Variance (ANOVA) F-tests on the training data with false discovery rate (FDR) control. To limit redundancy, we enforced a correlation constraint by pruning features until the first principal component explained $<50\%$ of the variance. Each omics matrix was then scaled to the range [0, 1]; in PSP proteomics and metabolomics, a $\log_{10}$ transformation was applied prior to normalization. A detailed description of these procedures, including variance thresholds by modality and alternative preprocessing choices, is provided in Supplementary Methods S1.

For each cohort, we constructed datasets for three integration settings: (i) single-omics, (ii) all available paired dual-omics combinations, and (iii) triple-omics integration (mRNA + miRNA + methylation in AD and BC; mRNA + proteomics + metabolomics in PSP). This design resulted in different sample sizes and class distributions across disease cohorts. In BC, where substantial class imbalance was observed between tumor and control samples, the majority class was downsampled to $\sim$1.2 times the minority class in all experimental settings. A summary of the resulting datasets is shown in Table 2.

Feature ranking and ensemble aggregation

To identify predictive and biologically meaningful features across heterogeneous, high-dimensional multi-omics datasets, we applied 15 single-feature ranking approaches spanning statistical tests, regularized regression, tree-based importance measures, explainable AI methods, and deep learning-based multi-omics rankers. These rankers capture complementary linear and nonlinear dependencies and are summarized in Supplementary Table S2.

To improve robustness and reduce method-specific biases, we further applied 13 ensemble rank aggregation strategies that combine the outputs of single rankers into consensus feature lists. We considered both rank-based aggregation (which integrates the relative ordering of features across methods) and weight-based aggregation (which

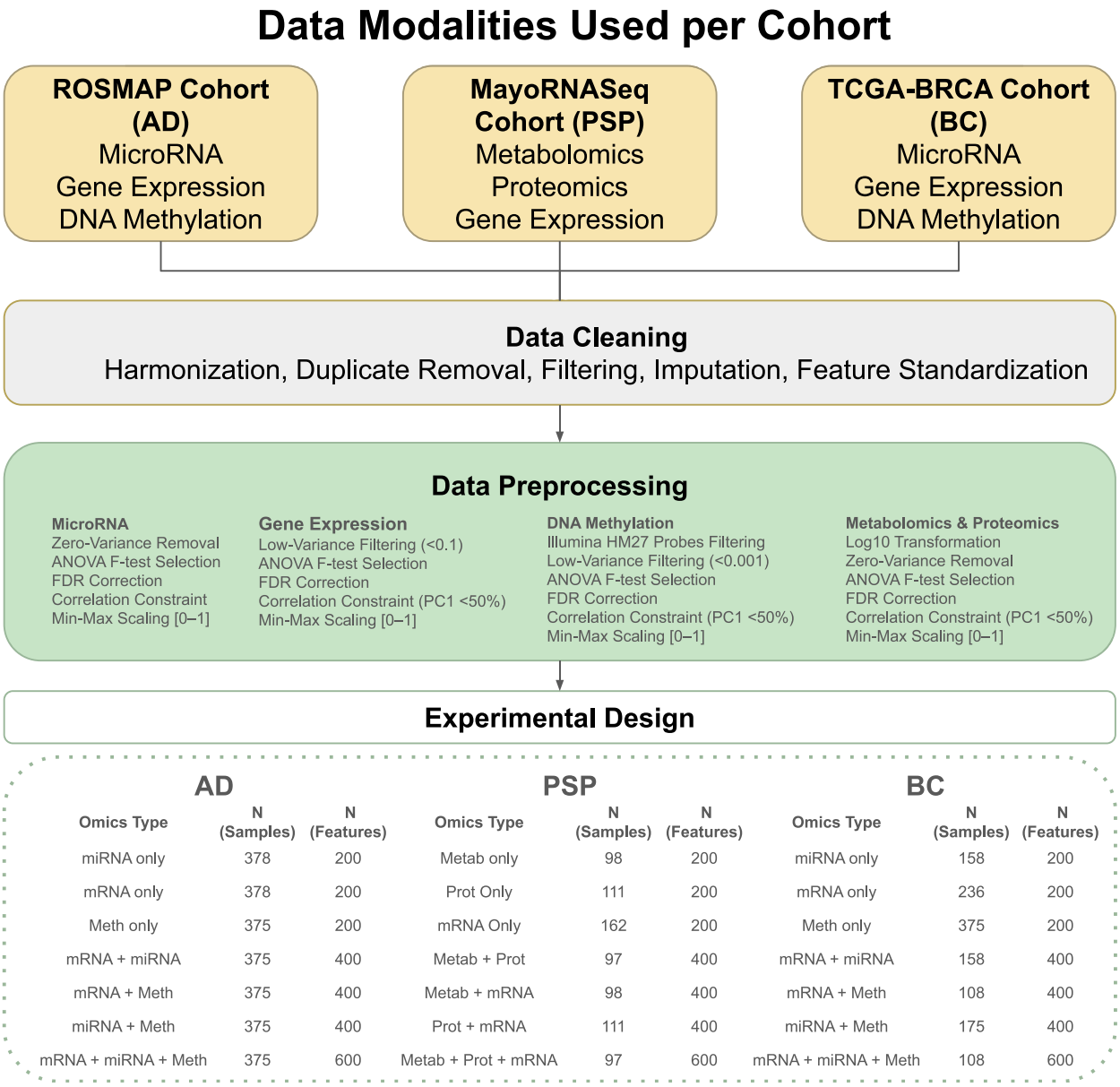

Figure 1 Experimental design and data modalities across cohorts. Overview of datasets and preprocessing steps used in the benchmarking study. Data from AD, PSP, and BC were harmonized through cleaning procedures including duplicate removal, filtering, imputation, and feature standardization, along with modality-specific preprocessing such as variance filtering, ANOVA F-tests, FDR correction, correlation constraints, min–max scaling, and log transformation. The experimental design integrates multiple single-omics, dual-omics, and triple-omics configurations. The lower panel summarizes the number of samples and features available for each omics modality and their combinations per cohort, highlighting the balanced yet heterogeneous structure of the benchmarking framework.

combines normalized importance scores). The ensemble methods are summarized in Supplementary Table S3. We excluded MORE-Ranker and MOGONET-Ranker from the ensembles to preserve a clear comparison between (i) conventional single rankers, (ii) ensemble-based aggregations, and (iii) deep learning-based integration models.

Benchmark setup

Starting from the cleaned and prepared datasets, we designed a benchmarking framework to evaluate feature selection strategies, predictive models, and multi-omics integration levels. Experiments were performed independently for AD, PSP, and BC across three integration settings: single-omics, dual-omics, and triple-omics. In the single-omics setting, each modality (mRNA, miRNA, methylation, proteomics, or metabolomics) was analyzed separately. Dual-omics settings included all pairwise modality combinations available within each cohort, and triple-omics settings integrated three modalities (mRNA + miRNA + methylation for AD and BC; mRNA + proteomics + metabolomics for PSP).

To ensure fair comparison with MORE and MOGONET, which are primarily designed for three-layer integration, we restricted analyses to a common three-modality subset in each cohort. For every omics subset, we derived biomarker panels of increasing size: {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}, plus a full-feature panel (200 features for single-omics, 400 for dual-omics, 600 for triple-omics). This allowed us to study performance as a function of panel sparsity and interpretability.

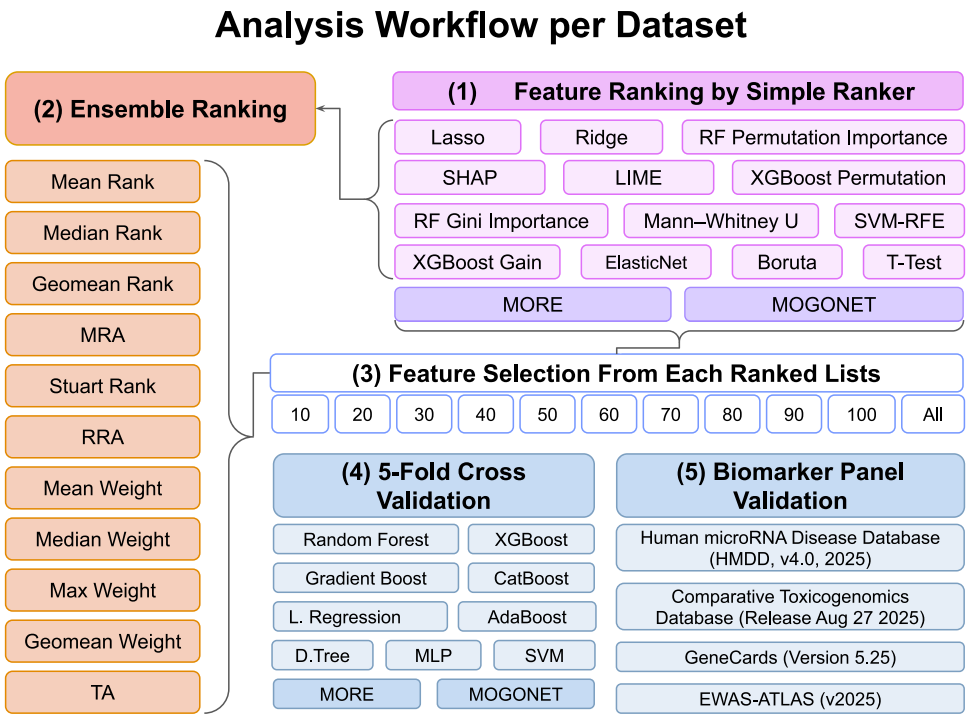

Figure 2 Analysis workflow per dataset. Schematic overview of the benchmarking pipeline applied to each cohort. (1) Features were ranked using a diverse set of methods, including statistical tests (t-test, Mann–Whitney U), regularized models (LASSO, Ridge, Elastic Net), model-based importance methods (Random Forest, XGBoost, Boruta, SVM Recursive Feature Elimination (SVM-RFE)), explainability-based approaches (SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME)), and deep learning-embedded rankers (MORE-Ranker, MOGONET-Ranker). (2) Ranked feature lists were aggregated using ensemble strategies such as mean, median, geometric mean, weighted schemes, robust rank aggregation (RRA), and Stuart rank. (3) From each ranking, biomarker panels of varying sizes (10–100 features and full lists) were selected. (4) These panels were evaluated using five-fold cross-validation across multiple classifiers, including D.Tree, L.Regression, Random Forest, Gradient Boosting, XGBoost, CatBoost, AdaBoost, SVM, and MLP, as well as deep learning frameworks (MORE, MOGONET). (5) Finally, selected biomarker panels were validated against external databases, including HMDD, CTD, GeneCards, and EWAS-ATLAS, to assess biological relevance.

In total, we evaluated *27* feature selection strategies (15 single rankers and 12 ensemble rankers) and *11* predictive models: logistic regression (L.Regression), Random Forest, Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), Adaptive Boosting (AdaBoost), Gradient Boosting (Gradient Boost), Support Vector Machine (SVM), multilayer perceptron (MLP), Decision Tree (D.Tree), and the multi-omics deep learning models MORE and MOGONET. Model training and evaluation were performed using five-fold stratified cross-validation to maintain balanced class distributions across folds.

All analyses were implemented in Python 3.11.9 using `scikit-learn` for traditional machine learning, `xgboost`, `catboost`, and `lightgbm` for gradient boosting methods, and author-provided implementations for MORE and MOGONET. A schematic overview of the workflow is presented in Figs 1 and 2 .

Biomarker panel validation

To assess biological and clinical relevance, each biomarker panel generated during benchmarking was validated against external reference databases at multiple molecular levels. For integrated panels containing features from different omics types, we first split features into modality-specific subsets (genes, microRNAs, and CpG sites) and then queried the most appropriate resource for each type.

MicroRNA–disease associations were evaluated using HMDD (v4.0, 2025) [28]. Gene-level associations were retrieved from the CTD (Release 27 August 2025) [27] and GeneCards (Version 5.25) [29]. DNA methylation sites were cross-referenced against the EWAS-ATLAS (v2025) [30]. For pathway-level validation, we performed functional enrichment using g:Profiler (v0.2, 2025) [39] and compared enriched pathways to curated pathway–disease associations in CTD. Construction of validation datasets and scoring procedures are described in detail in Supplementary Methods S3.

Evaluation metrics

Predictive performance was assessed within the five-fold stratified cross-validation framework using standard classification metrics. We report accuracy, precision (positive predictive value), recall (sensitivity), specificity (true negative rate), negative predictive value (NPV), F_1-score, and the area under the receiver operating characteristic curve (AUC-ROC or AUC). F_1 was used as the primary summary metric under class imbalance, and AUC-ROC was used to compare global discriminative performance independently of decision threshold.

To quantify robustness, we defined performance stability as the width of the observed performance window for each model–selector combination, computed as the difference between the maximum and minimum F_1 (or AUC) across experimental replicates. Narrower

windows indicate more stable and reproducible performance. Formal definitions of all metrics are provided in Supplementary Methods S4.

All benchmarking experiments were implemented in a Python 3.11.9 environment on the National Center for High Performance Computing (UHeM) clusters (hostname: sariyer) running CentOS Linux 7 (Core) with Linux kernel version 3.10.0. The server is equipped with dual Intel Xeon E5-2680 v4 processors (2.40 GHz), providing 28 CPU cores and approximately 128 GB of system memory. All models were trained using central processing unit (CPU) resources only, and graphics processing unit (GPU) acceleration was not used. Experiments were executed using batch jobs through the SLURM workload manager, typically allocating up to 40 parallel tasks on a single compute node to accelerate cross-validation and benchmarking runs. To ensure reproducibility, the complete benchmarking framework, including data preprocessing pipelines, feature selection modules, model implementations, and instructions for reproducing the computational environment, is publicly available at the following GitHub repository: https://github.com/itu-bioinformatics-database-lab/biomarker-benchmark

Results

Baseline model performance without feature selection

Across the three cohorts (AD, PSP, and BC), we first evaluated predictive performance using all available features without applying any feature selection (Table 3). In AD, MLP achieved the highest AUC (0.85 ± 0.05) and a correspondingly strong F_1 score (0.73 ± 0.05), closely followed by L.Regression (AUC = 0.84 ± 0.05, F_1 = 0.71 ± 0.05). Although MOGONET and MORE reached moderate AUCs (0.81 ± 0.03–0.04), they maintained competitive F_1 scores around 0.74 ± 0.03–0.05 due to balanced precision–recall behavior. In contrast, D.Tree performed poorest overall (AUC = 0.61 ± 0.08, F_1 = 0.56 ± 0.13).

In PSP, nearly all models achieved excellent discrimination (AUC > 0.95), with MLP and SVM delivering consistently high F_1 scores (0.98 ± 0.02) and balanced accuracy. MORE achieved the top AUC (0.99 ± 0.01) and perfect recall (1.00 ± 0.00) but exhibited low precision (0.57 ± 0.23), which reduced its F_1 to 0.69 ± 0.19. MOGONET, by contrast, showed a sharp drop in both precision and recall, yielding an anomalously low F_1 (0.29 ± 0.00) despite a relatively high AUC (0.89 ± 0.00), suggesting overfitting or miscalibration in class probability outputs.

In BC, ceiling effects were observed across nearly all classifiers. L.Regression, SVM, and MORE achieved near-perfect performance with AUCs between 0.98 and 1.00 and F_1 scores ranging from 0.96 to 1.00. Gradient-based ensemble models (e.g. XGBoost, CatBoost) also performed at parity, with both AUC and $F_1 > 0.97$, indicating minimal performance differentiation. MOGONET slightly trailed with AUC 0.92 ± 0.00 and F_1 0.96 ± 0.00 despite perfect recall, again hinting at potential imbalance handling issues.

Taken together, these results demonstrate that while deep integration models such as MORE and MOGONET can achieve high AUCs, they do not always translate these into equally strong F_1 scores—particularly in smaller or less balanced cohorts. Conversely, traditional single-model approaches such as L.Regression, SVM, and MLP often deliver more stable and interpretable performance, matching or surpassing complex models when no feature selection is applied.

Impact of feature selection

Applying feature selection led to consistent performance gains across all cohorts (Fig. 3a). Mean F_1-scores increased notably when models were trained on subsets of top-ranked biomarkers rather than the full feature sets, reflecting the benefit of reducing noise and redundancy. The most pronounced improvements were observed in AD and BC, where cross-validation stability was enhanced and several traditional methods matched or outperformed more complex architectures once dimensionality was reduced. PSP models, already near ceiling without selection, showed modest but still measurable gains in precision and F_1, suggesting that targeted feature reduction can fine-tune performance even in highly separable cohorts. Overall, feature selection improved predictive accuracy while also keeping simpler models competitive, making the resulting biomarker panels easier to interpret and more suitable for clinical use.

Effect of omics integration strategies

We next assessed the effect of single-, dual-, and triple-omics integration strategies on predictive performance (Figs 3b and c; and 4). Figure 3b and c summarizes the median of average cross-validation F_1-scores across selector–classifier–panel size combinations, while Fig. 4 shows the peak performance achieved across omics configurations. Together, these complementary perspectives capture both central trends and best-case outcomes. Across all cohorts, predictive performance improved progressively from single-omics to dual-omics and reached its highest levels with triple-omics integration. In BC, all integration levels achieved near-ceiling medians and maxima (0.95–1.0 across metrics), confirming that the dataset is highly separable regardless of modality configuration, though triple-omics still achieved the most consistent top scores. In PSP, single-omics models already performed strongly, but integration—particularly triple-omics—yielded small yet measurable gains in precision and F_1 scores. AD showed the most pronounced benefit from integration: performance increased substantially when moving from single- to dual-omics, and further improved under triple-omics configurations, with the mRNA + methylation + miRNA combination consistently outperforming all alternatives. Overall, these results demonstrate that predictive accuracy and stability improve as the level of omics integration increases, with triple-omics providing the most comprehensive and robust representation of the underlying biological signal, followed by dual-omics and single-omics analyses.

Benchmarking across classifiers

Classifier performance varied across cohorts and integration strategies, but a consistent trend emerged: traditional machine learning approaches often matched or surpassed deep learning frameworks (Figs 3d, 5b and 6). In the median F_1 comparison (Fig. 3d), L.Regression, SVM, ensemble tree models (AdaBoost and CatBoost), and MLP consistently occupied the top tier across cohorts, whereas D.Tree, Gradient Boost, MORE, and MOGONET generally ranked lower with reduced medians. This indicates that well-regularized linear/nonlinear baselines and MLP yield the most reliable central performance, while the more specialized multiview methods do not confer a universal advantage.

Further insights come from the precision–recall analysis in Fig. 5b. Points closest to the top-right corner correspond to classifiers

Table 3 Baseline classification results across cohorts without feature selection, reporting performance for 11 models using triple-omics data.

Cohort	Model	AUC	Accuracy	Precision	Recall	F_1	Specificity	NPV
AD	L.Regression	0.84 ± 0.05	0.73 ± 0.06	0.71 ± 0.08	0.71 ± 0.06	0.71 ± 0.05	0.75 ± 0.10	0.76 ± 0.05
	Random Forest	0.78 ± 0.04	0.70 ± 0.04	0.67 ± 0.04	0.65 ± 0.09	0.66 ± 0.06	0.74 ± 0.05	0.73 ± 0.05
	XGBoost	0.80 ± 0.04	0.71 ± 0.04	0.68 ± 0.05	0.68 ± 0.08	0.68 ± 0.05	0.73 ± 0.06	0.74 ± 0.05
	D.Tree	0.61 ± 0.08	0.61 ± 0.07	0.55 ± 0.09	0.59 ± 0.16	0.56 ± 0.13	0.63 ± 0.03	0.66 ± 0.07
	Gradient Boost	0.79 ± 0.05	0.71 ± 0.05	0.67 ± 0.07	0.69 ± 0.06	0.68 ± 0.04	0.72 ± 0.09	0.74 ± 0.04
	CatBoost	0.81 ± 0.05	0.73 ± 0.05	0.71 ± 0.06	0.68 ± 0.08	0.69 ± 0.05	0.76 ± 0.09	0.75 ± 0.05
	AdaBoost	0.79 ± 0.06	0.73 ± 0.06	0.71 ± 0.06	0.67 ± 0.08	0.69 ± 0.07	**0.77 ± 0.05**	0.75 ± 0.06
	MLP	**0.85 ± 0.05**	0.74 ± 0.06	0.70 ± 0.07	0.74 ± 0.05	0.72 ± 0.06	0.73 ± 0.08	0.78 ± 0.05
	SVM	0.81 ± 0.04	0.72 ± 0.04	0.68 ± 0.02	0.71 ± 0.13	0.69 ± 0.06	0.72 ± 0.06	0.77 ± 0.07
	MORE	0.81 ± 0.03	0.73 ± 0.05	0.70 ± 0.08	**0.75 ± 0.11**	0.72 ± 0.05	0.72 ± 0.13	**0.79 ± 0.06**
	MOGONET	0.81 ± 0.04	**0.76 ± 0.03**	**0.73 ± 0.05**	**0.75 ± 0.06**	**0.74 ± 0.03**	**0.77 ± 0.06**	**0.79 ± 0.03**
PSP	L.Regression	0.98 ± 0.03	0.94 ± 0.07	0.94 ± 0.06	0.99 ± 0.02	0.96 ± 0.04	0.75 ± 0.27	0.90 ± 0.20
	Random Forest	0.92 ± 0.05	0.90 ± 0.07	0.91 ± 0.06	0.97 ± 0.03	0.94 ± 0.04	0.60 ± 0.30	0.85 ± 0.20
	XGBoost	0.96 ± 0.03	0.91 ± 0.06	0.92 ± 0.06	0.97 ± 0.03	0.94 ± 0.04	0.65 ± 0.25	0.88 ± 0.15
	D.Tree	0.64 ± 0.12	0.82 ± 0.07	0.85 ± 0.05	0.94 ± 0.08	0.89 ± 0.04	0.35 ± 0.25	0.57 ± 0.40
	Gradient Boost	0.83 ± 0.15	0.88 ± 0.12	0.90 ± 0.09	0.96 ± 0.05	0.93 ± 0.07	0.55 ± 0.40	0.70 ± 0.40
	CatBoost	0.95 ± 0.05	0.90 ± 0.05	0.89 ± 0.05	**1.00 ± 0.00**	0.94 ± 0.03	0.50 ± 0.22	**1.00 ± 0.00**
	AdaBoost	0.95 ± 0.04	0.93 ± 0.08	0.95 ± 0.05	0.96 ± 0.05	0.96 ± 0.05	0.80 ± 0.19	0.85 ± 0.20
	MLP	0.97 ± 0.03	**0.97 ± 0.03**	**0.97 ± 0.03**	0.99 ± 0.02	**0.98 ± 0.02**	**0.90 ± 0.12**	0.96 ± 0.08
	SVM	0.98 ± 0.02	0.93 ± 0.05	0.93 ± 0.04	0.99 ± 0.02	0.96 ± 0.03	0.70 ± 0.19	0.93 ± 0.13
	MORE	**0.99 ± 0.01**	0.76 ± 0.19	0.57 ± 0.23	**1.00 ± 0.00**	0.69 ± 0.19	0.70 ± 0.24	**1.00 ± 0.00**
	MOGONET	0.89 ± 0.00	0.75 ± 0.00	0.33 ± 0.00	0.25 ± 0.00	0.29 ± 0.00	0.88 ± 0.00	0.82 ± 0.00
BC	L.Regression	**1.00 ± 0.01**	0.98 ± 0.02	0.98 ± 0.03	0.98 ± 0.04	0.98 ± 0.02	0.98 ± 0.04	0.98 ± 0.03
	Random Forest	**1.00 ± 0.01**	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.04	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.03
	XGBoost	0.97 ± 0.04	0.97 ± 0.04	0.98 ± 0.04	0.97 ± 0.04	0.97 ± 0.04	0.98 ± 0.04	0.96 ± 0.04
	D.Tree	0.93 ± 0.08	0.93 ± 0.08	0.94 ± 0.05	0.91 ± 0.14	0.92 ± 0.09	0.94 ± 0.05	0.92 ± 0.11
	Gradient Boost	0.98 ± 0.03	0.96 ± 0.03	0.98 ± 0.03	0.95 ± 0.07	0.96 ± 0.04	0.98 ± 0.04	0.95 ± 0.06
	CatBoost	**1.00 ± 0.00**	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.04	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.03
	AdaBoost	0.97 ± 0.04	0.96 ± 0.03	0.98 ± 0.03	0.95 ± 0.07	0.96 ± 0.04	0.98 ± 0.04	0.95 ± 0.06
	MLP	**1.00 ± 0.01**	0.98 ± 0.02	0.98 ± 0.03	0.98 ± 0.04	0.98 ± 0.02	0.98 ± 0.04	0.98 ± 0.03
	SVM	**1.00 ± 0.01**	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.04	**0.99 ± 0.02**	**1.00 ± 0.00**	0.98 ± 0.03
	MORE	**1.00 ± 0.00**	0.95 ± 0.00	0.91 ± 0.00	**1.00 ± 0.00**	0.95 ± 0.00	0.91 ± 0.00	**1.00 ± 0.00**
	MOGONET	0.92 ± 0.00	0.95 ± 0.00	0.92 ± 0.00	**1.00 ± 0.00**	0.96 ± 0.00	0.91 ± 0.00	**1.00 ± 0.00**

Bold values indicate the best-performing result for each evaluation metric (column).

achieving the best balance between precision and recall. MLP and L.Regression clustered near this frontier across all three cohorts, with CatBoost and Gradient Boost also performing strongly. AdaBoost trailed slightly beneath this group, while D.Tree formed a mid-tier cluster with moderate precision and recall. Random Forest, SVM, and XGBoost generally occupied the upper-mid to upper-right regions, reflecting high recall with somewhat lower precision depending on the dataset. By contrast, MORE and MOGONET consistently populated the lower parts of the panels, indicating simultaneously weaker precision and recall across cohorts.

A more granular perspective is provided by Fig. 6, which tracks classifier performance across feature set sizes and multiple evaluation metrics (AUC, accuracy, F_1, NPV, precision, recall, and specificity) in BC, PSP, and AD. In BC, nearly all classifiers achieved ceiling-level performance across all metrics regardless of feature number. Linear (L.Regression) and nonlinear models (CatBoost, SVM, XGBoost, and MLP) maintained mean values near 1.0 with minimal sensitivity to panel size, underscoring the strength of the dataset signal and suggesting that even simple models with modest panels can achieve near-perfect discrimination. By contrast, PSP revealed greater variability. While AUC, accuracy, and F_1 scores generally stabilized >0.9 across classifiers, specificity declined with larger feature sets, especially in Gradient Boost and XGBoost. L.Regression, SVM, and MLP displayed the most consistent profiles, maintaining balanced sensitivity and precision with narrow fluctuations. D.Tree, Gradient Boost, Random Forest, MORE, and MOGONET lagged, with reduced precision and specificity despite achieving moderate recall. In AD, the most challenging cohort, feature set size had a pronounced effect on model performance. L.Regression, MOGONET, and MLP exhibited steady improvements across all evaluation metrics as biomarker panels expanded, with performance plateauing ∼50–60 features across nearly all metrics. Tree-based ensembles delivered competitive recall but showed greater metric-specific variability: Random Forest and XGBoost achieved strong recall yet more modest precision. These results highlight the influence of panel size on predictive stability and emphasize the need for balanced feature selection in noisy, multi-omics Alzheimer's data.

Collectively, these results demonstrate that metric-specific behaviors are as important as overall accuracy. Precision and specificity tended to degrade more rapidly with larger feature sets, particularly in PSP and AD, while recall remained relatively stable. Traditional methods and MLP consistently balanced these tradeoffs, whereas MORE and MOGONET struggled to maintain parity across metrics. These findings highlight the dual importance of parsimonious feature selection and classifier choice in multi-omics biomarker discovery, and caution against assuming that increased model complexity alone yields superior results.

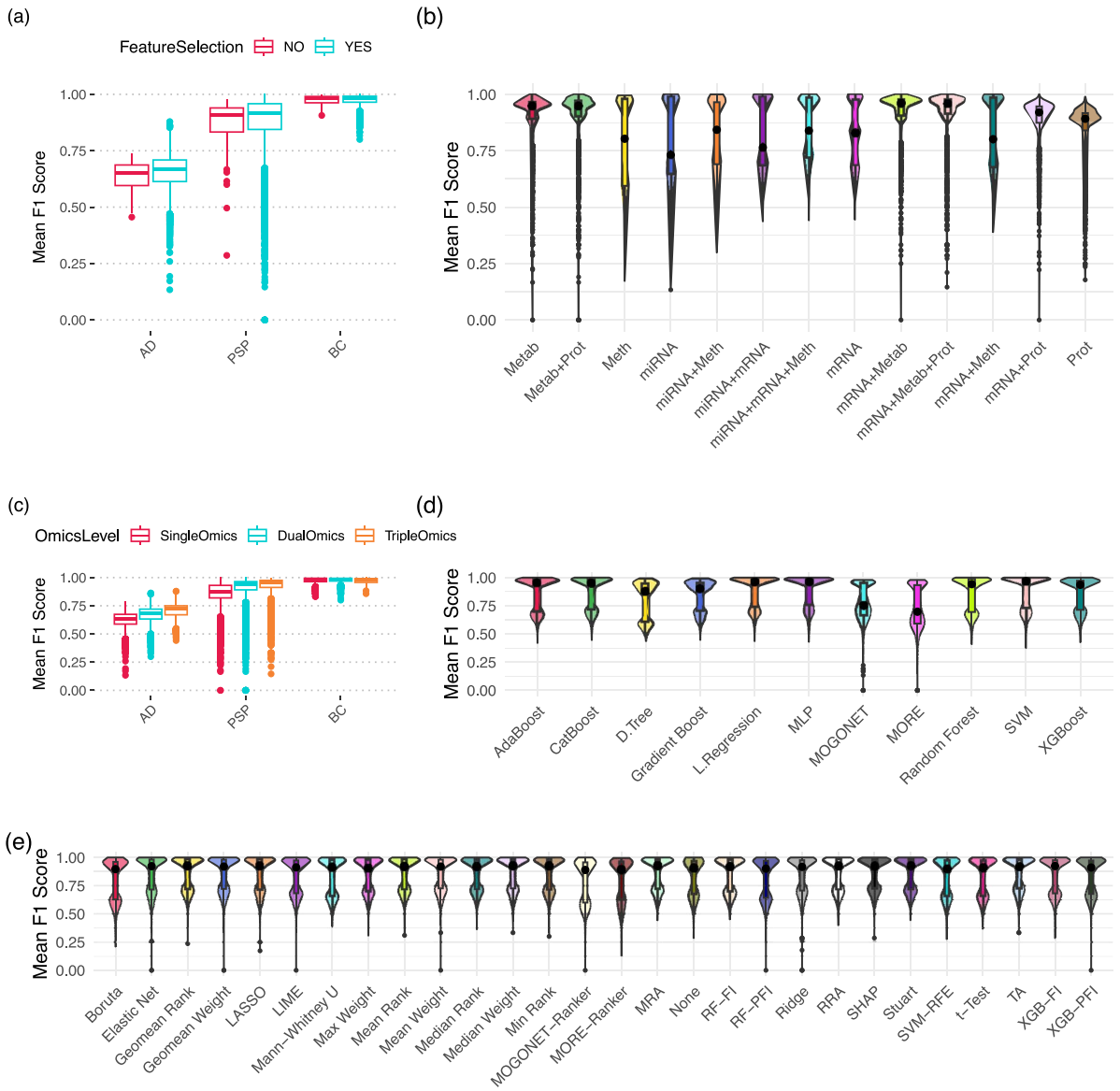

Figure 3 Benchmarking model performance across datasets, omics levels, models, and feature selection strategies. (a) Comparison of mean F1-scores with and without feature selection across AD, PSP, and BC, showing consistent performance improvements with feature selection. (b) Performance by data modality and integration strategy, where triple-omics combinations frequently outperform single- and dual-omics configurations. (c) Comparison of single-omics, dual-omics, and triple-omics integration across cohorts, highlighting the trade-off between information gain and potential noise. (d) Performance of individual machine learning models, showing that traditional linear and tree-based methods remain competitive with deep learning approaches such as MORE and MOGONET. (e) Impact of different feature selection rankers, indicating that ensemble aggregation strategies provide more stable and accurate results compared with individual rankers. Collectively, these panels illustrate how feature selection and multi-omics integration influence predictive performance across benchmarking settings.

Performance comparison of feature selection strategies

We benchmarked 27 feature ranking methods, including MORE and MOGONET, single rankers, and ensemble aggregation approaches. Figure 3e summarizes their mean F_1 score distributions across cohorts. Figure 7 provides a direct comparison between single feature selectors, ensemble aggregation methods, and deep learning-embedded rankers across cohorts . Ensemble-based selectors such as geometric mean rank, mean rank, median rank, median rank algorithm, Stuart rank, max weight, median weight, and RRA rank achieved consistently strong performance with high medians and relatively narrow spreads. Several traditional selectors, including t-Test, Least Absolute Shrinkage and Selection Operator (LASSO), Mann–Whitney U, random forest feature importance (RF-FI), and XGBoost feature importance (XGB-FI), performed comparably well, highlighting their continued competitiveness. By contrast, LIME and the more recent multiview-specific deep learning rankers (MOGONET-Ranker and MORE-Ranker) exhibited lower medians and broader variability, indicating weaker and less stable performance. SHAP fell into an intermediate category, outperforming the weakest selectors but not matching the stability of ensembles or top traditional methods.

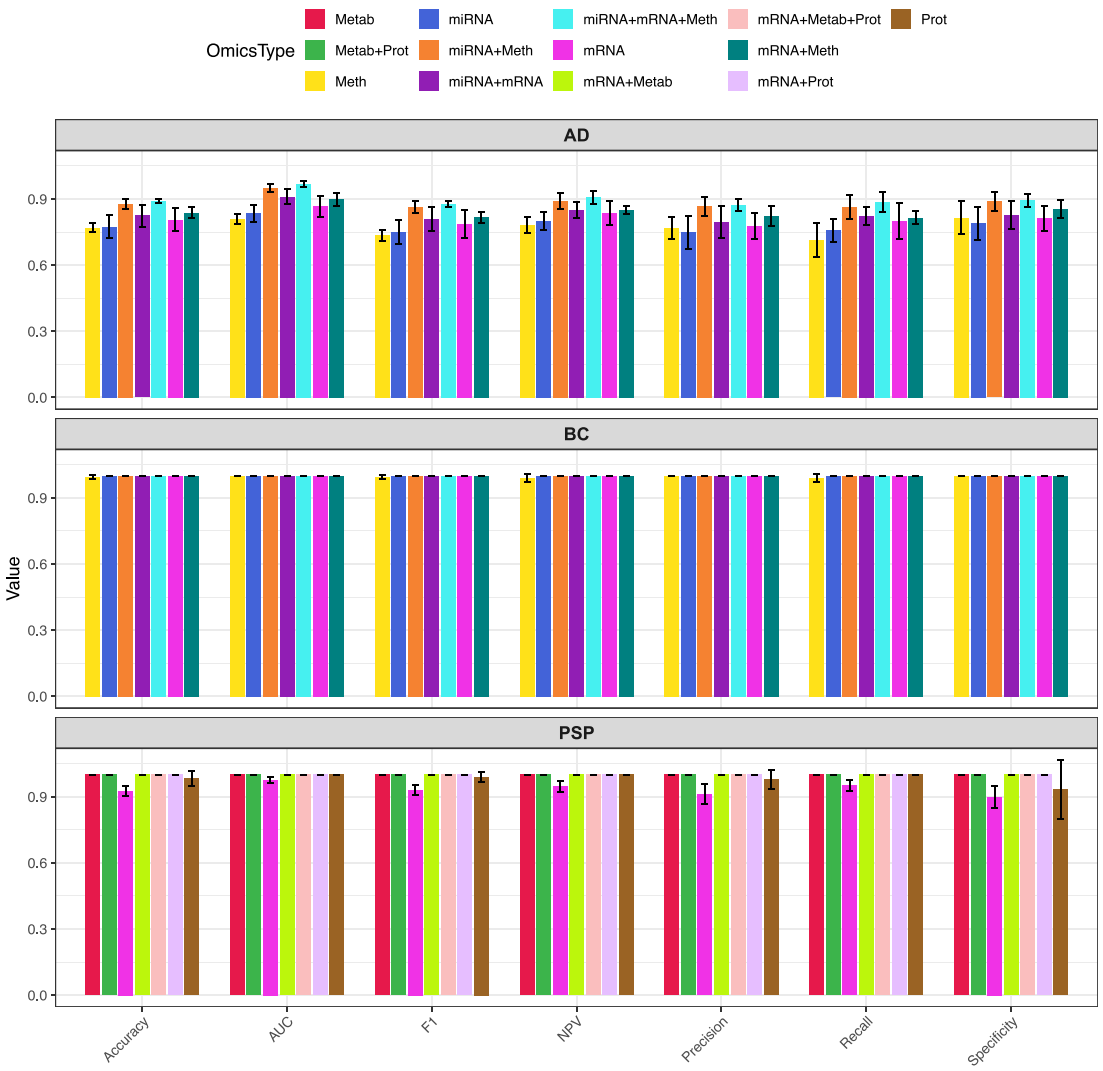

Figure 4 Model performance across omics types and integration strategies in AD, PSP, and BC, where AUC, accuracy, F1-score, NPV, precision, recall, and specificity are reported for the best-performing combinations of feature selector, classifier, and panel size within each omics configuration, with bars representing individual omics inputs (mRNA, miRNA, methylation, proteomics, metabolomics) and their pairwise or three-way integrations, showing that triple-omics configurations achieve the highest performance across most metrics while dual-omics provide intermediate performance between single- and triple-omics settings, illustrating how predictive performance varies across integration strategies.

Figure 5a presents precision–recall scatter plots per selector, with points colored by cohort. The apparent groupings are primarily *cohort effects*: BC concentrates near the upper-right (high precision and recall), PSP forms an intermediate band, and AD lies lower—especially on recall. Against this backdrop, selectors differ mainly in (i) the overall location of this three-cohort triad and (ii) the spread between cohorts. Ensemble aggregators (e.g. mean/median/geometric-mean rank/weight, RRA, Stuart) and strong traditional methods (t-Test, LASSO, Elastic Net, RF-FI, XGB-FI, and Boruta) place the triad higher with tighter dispersion, whereas SHAP, LIME, SVM-RFE, Ridge, Min Rank, Max Weight, and Median Rank Aggregation show mid-level placement and wider spread. The multiview rankers (MORE-Ranker and MOGONET-Ranker) and Threshold Algorithm (TA) tend to sit lower overall and exhibit larger cohort gaps. Thus, the separation visible in the figure reflects differences in cohort separability rather than intrinsic clustering of selectors; methods that both elevate the triad and compress the BC–PSP–AD gap are the most robust. Figure 8 further summarizes how classifier performance varies across feature rankers and how biomarker panel size influences F_1-scores across AD, PSP, and BC.

Biological validation of discovered biomarkers

We quantified biological plausibility as true positives (TPs)—the overlap between the top-k selected biomarkers and the top-1000 entries from external databases (HMDD, CTD-pathways, CTD, GeneCards, and EWAS-ATLAS)—for $k \in \{10, 30, 50, 100\}$ (Figs 9– 12). Across feature selectors, ensemble methods (Geom. Mean (rank/weight), Mean/Median (rank), Stuart, RRA) consistently achieved the highest TP counts, with strong traditional rankers [t-test, LASSO, RF-FI, XGBoost Feature Importance (XGB-FI), SVM-RFE] often close behind; methods such as LIME, TA, MOGONET-Ranker, and MORE-Ranker generally validated less. Across omics strategies, single-omics (top panels) showed the lowest recoveries, while dual-omics (middle) and especially triple-omics (bottom) yielded higher overlap across databases. Cohort differences were evident but stable across cutoffs: BC typically attained

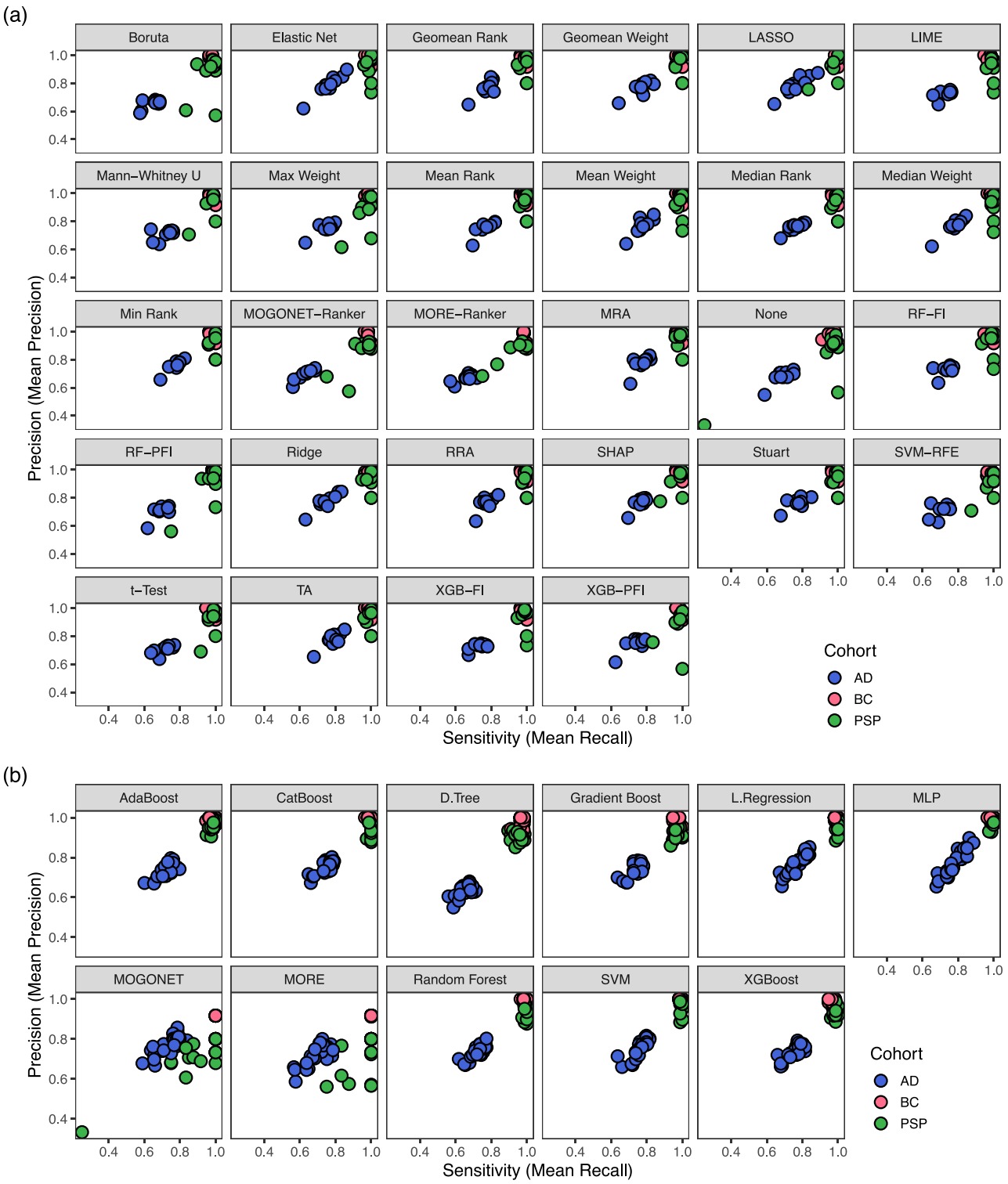

Figure 5 Comparative performance of feature selection strategies and classifiers across cohorts. (a) Precision–recall plots summarizing the trade-off between recall and precision for different feature selection rankers across AD, PSP, and BC, where ensemble rankers demonstrate improved balance compared with individual methods. (b) Precision–recall analysis across classifiers, showing that tree-based ensemble methods (AdaBoost, XGBoost, and CatBoost) and linear models (L.Regression, SVM, and MLP) achieve performance comparable to deep learning models (MORE and MOGONET).

the highest absolute TPs (including CpG validations), AD benefited most in miRNA-based validation (HMDD), and PSP showed moderate gains . Increasing *k* raised absolute TP counts and slightly compressed between-method gaps, but the qualitative ordering—and the advantage of multi-omics integration—remained consistent across all four figures.

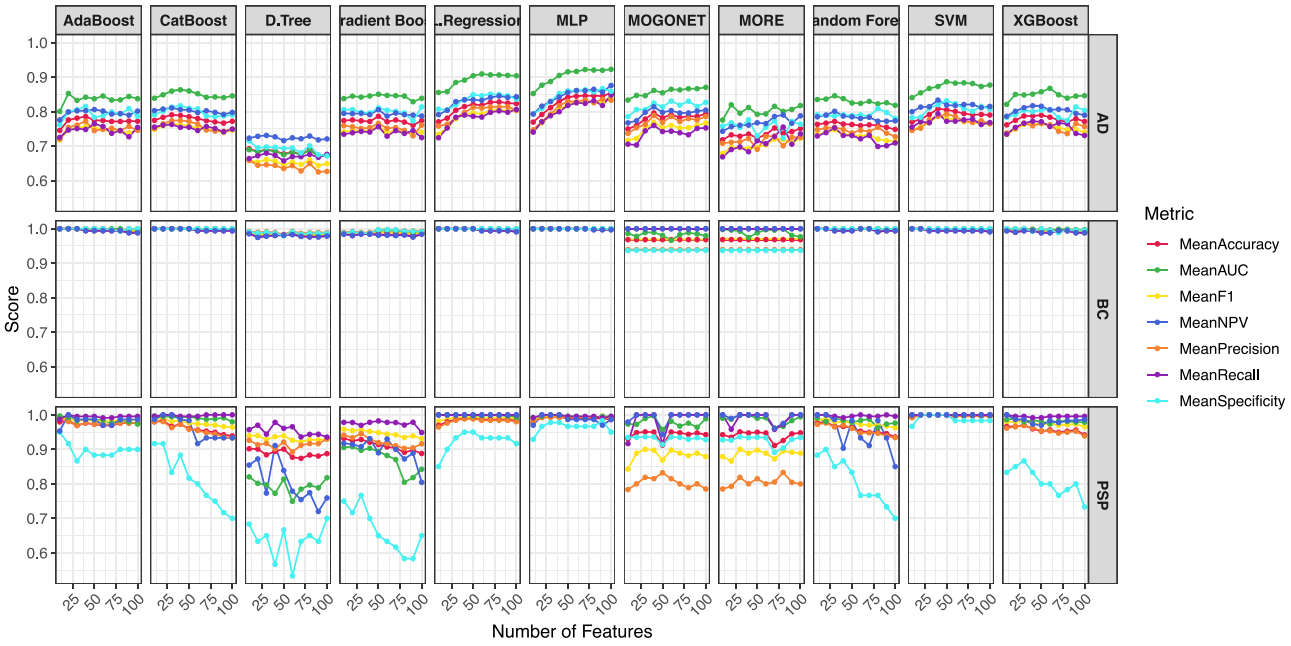

Figure 6 Model performance across biomarker panel sizes in AD, PSP, and BC cohorts, where AUC, accuracy, F1-score, NPV, precision, recall, and specificity are plotted against the number of selected features for multiple classifiers—including traditional linear models (L.Regression and support vector classifier, SVC), tree-based ensemble methods (Random Forest, XGBoost, CatBoost, AdaBoost, and Gradient Boosting), MLP, and deep learning approaches (MORE and MOGONET)—showing rapid performance gains at small panel sizes (10–30 features), stabilization at intermediate sizes (50–75 features), and minimal changes beyond 100 features, illustrating how predictive performance varies with biomarker panel size across modeling approaches.

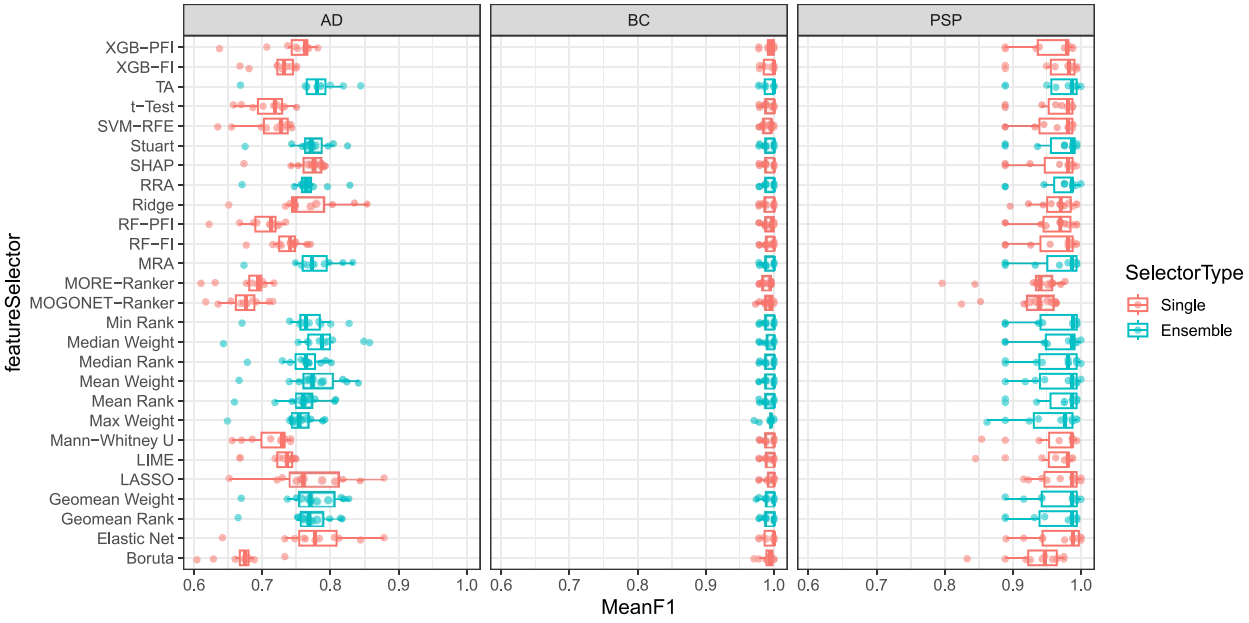

Figure 7 Comparison of single and ensemble feature selection methods across cohorts. Mean F1-scores for individual (single) rankers, ensemble aggregation methods, and embedded rankers from deep learning frameworks (MORE and MOGONET) across BC, PSP, and AD datasets. Bars represent average performance across classifiers and biomarker panel sizes. Ensemble methods such as Mean Weight, Geomean Weight, TA, RRA, and Stuart show higher and more stable performance across cohorts, while regularized single methods (Elastic Net, LASSO, and Ridge) achieve comparable results in some settings, particularly in BC. Deep learning-embedded rankers (MORE and MOGONET) exhibit variable performance relative to these approaches. These results illustrate differences in performance across feature selection strategies.

In this study, we adopt a data-driven definition of biomarker discovery, referring to the identification of molecular features that robustly discriminate between disease and control samples within each omics layer. This definition aligns with the analytical phase of biomarker research, where computational models prioritize disease-associated molecules based on reproducibility and predictive value, rather than immediate clinical applicability. Following the classical framework of biomarker classification [43], our findings should be interpreted

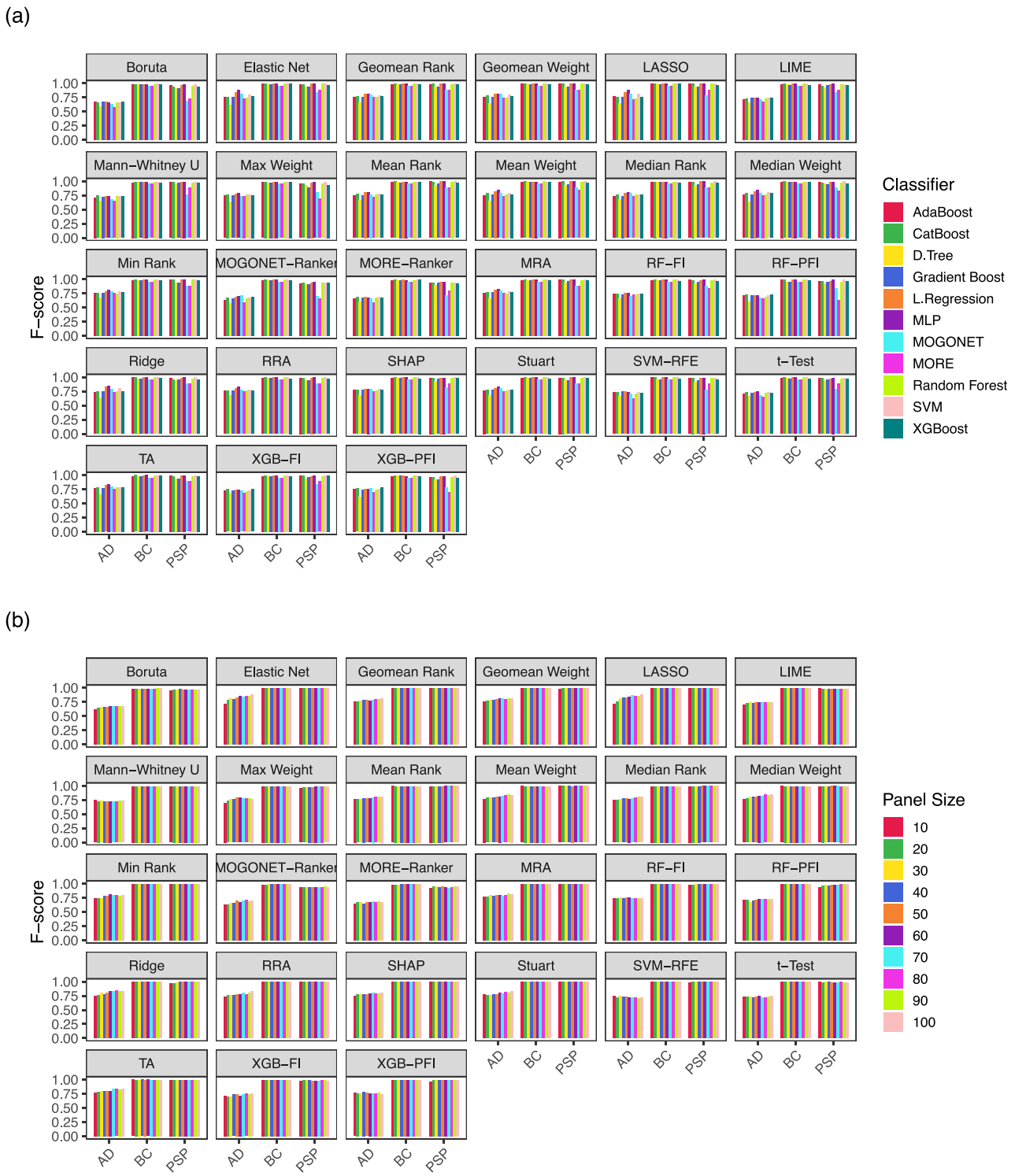

Figure 8 Model performance across rankers, classifiers, and biomarker panel sizes in three cohorts. (a) Mean F1-scores of different classifiers (D.Tree, L.Regression, SVM, Random Forest, Gradient Boosting, AdaBoost, CatBoost, XGBoost, MLP, MORE, MOGONET) when paired with feature subsets generated by multiple rankers across AD, PSP, and BC. Ensemble rankers are associated with higher classifier performance compared to individual rankers, while tree-based ensemble methods and linear models show performance comparable to deep learning frameworks. (b) Effect of biomarker panel size on F1-score across rankers and cohorts. Performance increases with small panel sizes (10–30 features) and stabilizes at larger panel sizes. Together, these panels illustrate how feature selection strategy, classifier choice, and panel size influence predictive performance across cohorts.

Table 4 AD summary table of reproducible cross-omics biomarkers

Feature	FeatureType	OmicsLevel	Selection frequency	N (Selectors)
cg06690548	Meth	Single+Dual+Triple	37	20
cg12981137	Meth	Single+Dual+Triple	35	17
cg12845808	Meth	Single+Dual	34	21
cg19832721	Meth	Single+Dual+Triple	33	18
cg16003238	Meth	Single+Dual+Triple	32	18
cg22442730	Meth	Single+Dual+Triple	29	17
cg12864235	Meth	Single+Dual	24	16
cg24765079	Meth	Single+Dual+Triple	20	16
cg02489552	Meth	Single+Dual+Triple	20	13
APLN	mRNA	Single+Dual+Triple	51	18
CCDC69	mRNA	Single+Dual+Triple	47	17
PPDPF	mRNA	Single+Dual+Triple	43	17
SLC6A12	mRNA	Single+Dual+Triple	42	16
PTPRF	mRNA	Single+Dual+Triple	36	18
MEIS3	mRNA	Single+Dual+Triple	34	14
CNN3-DT	mRNA	Single+Dual	33	18
PLEKHM2	mRNA	Single+Dual+Triple	32	17
RPL29	mRNA	Single+Dual+Triple	32	14
QDPR	mRNA	Single+Dual+Triple	24	15
SPACA6	mRNA	Single+Dual+Triple	21	14
hsa-miR-129-5p	miRNA	Single+Dual+Triple	83	22
hsa-miR-132	miRNA	Single+Dual+Triple	80	22
hsa-miR-885-5p	miRNA	Single+Dual+Triple	55	20
hsa-miR-29a	miRNA	Single+Dual+Triple	37	22
hsa-miR-146b-5p	miRNA	Single+Dual+Triple	27	15
hsa-miR-99a	miRNA	Single+Dual+Triple	23	12
hsa-miR-129-3p	miRNA	Single+Dual+Triple	22	19

as molecular signatures of disease derived from postmortem tissue (AD and PSP) and tumor tissue (BC). While these signatures provide valuable insight into disease mechanisms and may guide future development of accessible biomarkers in biofluids (e.g. blood or cerebrospinal fluid), the current analyses focus on computational prioritization within affected tissue contexts. Recent large-scale efforts, such as the blood-based molecular atlas of AD [44], further emphasize the importance of translating tissue-derived molecular signatures into accessible biofluid biomarkers. Such complementary studies support the view that multi-omics profiling in postmortem tissue can inform the search for peripheral correlates of neurodegenerative processes.

Cross-omics biomarkers identified per cohort

Across cohorts, our framework revealed both validated and novel biomarker candidates reproducibly prioritized across multiple omics levels and feature selection strategies. Selection frequencies and rank heatmaps (Tables 4– 6; Figs 13– 15) summarize the consistency and stability of identified features. To focus on reproducible signals, we restricted the lists to markers that appeared among the top 10 ranked features in at least 20 distinct rankers. For each candidate, the tables report the omics levels in which it was selected, the number of rankers supporting its inclusion (N), and the total frequency of top-10 appearances across omics settings. The corresponding heatmaps display the same set of features, showing their minimum rank (achieved at each omics level) across all rankers, thereby complementing the frequency-based summaries with a visual representation of ranking stability.

In AD, miRNAs emerged as particularly stable cross-omics markers. hsa-miR-129-5p and hsa-miR-132 were selected by 22 rankers with the highest frequencies (83 and 80, respectively), consistent with their strong prior evidence in AD-related synaptic regulation and neuroinflammation. Other validated miRNAs such as hsa-miR-29a and hsa-miR-146b-5p also ranked highly, alongside novel signals like hsa-miR-885-5p, which showed robust reproducibility across rankers. Several mRNAs (APLN, PPDPF, PLEKHM2, MEIS3, CCDC69, and SPACA6) were consistently identified, some with previous links to AD pathology, while others remain poorly characterized. Importantly, CpG methylation sites (cg06690548, cg12981137, cg19832721, cg16003238, and cg22442730) formed a reproducible block of epigenetic candidates, underscoring the contribution of DNA methylation to AD-associated regulatory disruption. Together, these findings demonstrate the ability of ensemble feature selection to recover both known AD biomarkers and underexplored regulatory signals across omics layers.

In PSP, metabolite signals were dominant, reflecting strong metabolic dysregulation. Highly ranked metabolites included trimethylamine N-oxide (TMAO), a known neuroinflammatory marker, along with 1-linoleoyl–GPC (18:2), stachydrine, trigonelline, salicylate, and 1-methyl-5-imidazoleacetate, which were selected with high frequency across rankers and omics levels. Protein markers such as TBCK, ANKRD11, and PPT1 were also reproducibly prioritized, with PPT1 aligning with known lysosomal dysfunction in neurodegeneration. Several novel candidates, including ARHGAP19-SLIT1, ARHGAP35, and ASAH1, were consistently ranked but have limited prior evidence in PSP, suggesting new directions for functional validation. The integration of metabolomics and proteomics proved especially effective in uncovering reproducible signals, highlighting the relevance of metabolic-lysosomal interactions in PSP disease pathology.

In BC, both miRNAs and mRNAs were robustly recovered, with strong overlap with established oncogenic drivers. Validated

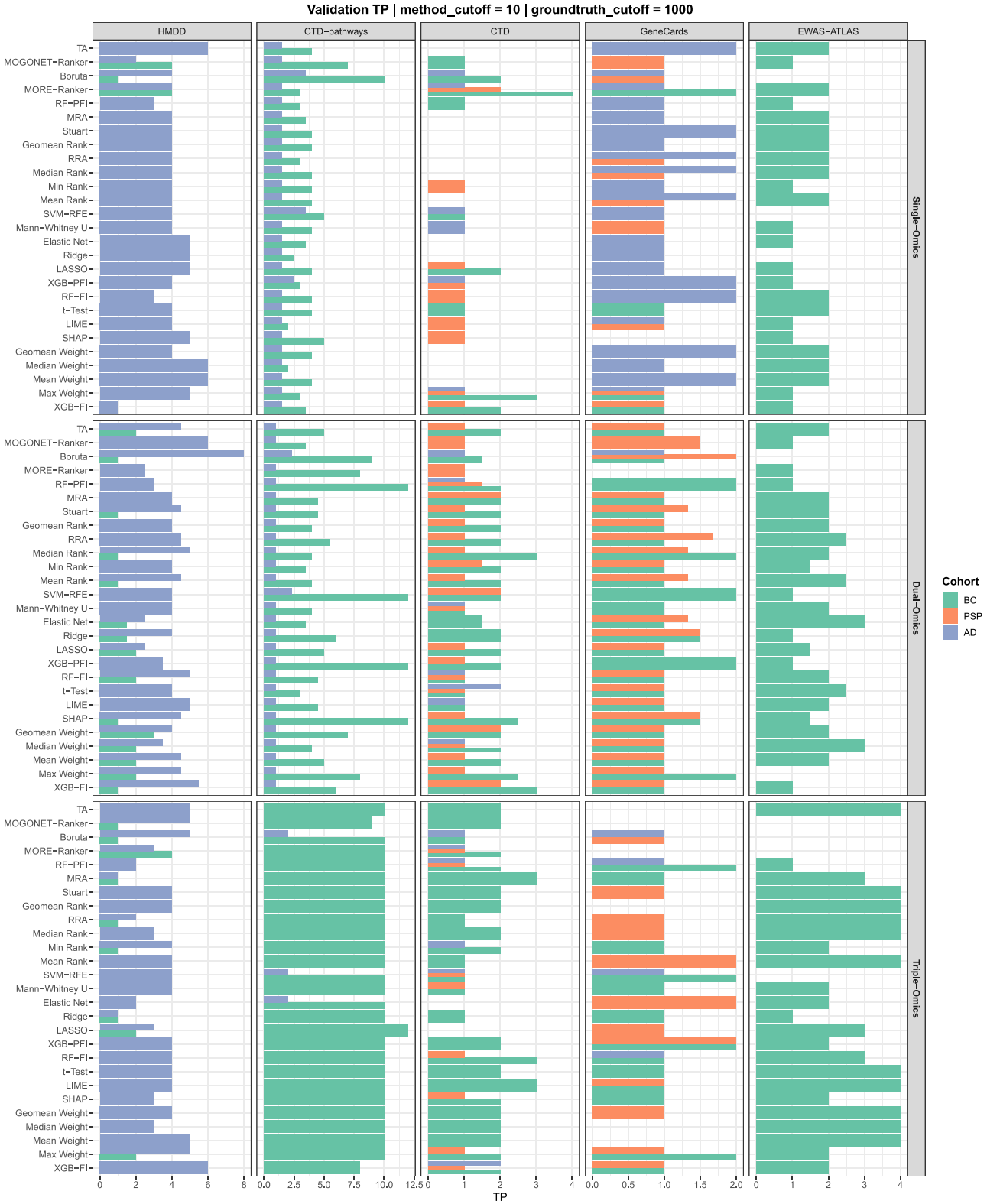

Figure 9 External validation results at the Top-10 cutoff. Bars represent overlaps with curated disease-associated features across multiple databases, including HMDD (miRNAs), CTD-pathways, CTD, GeneCards, and EWAS-ATLAS, for single-, dual-, and triple-omics configurations across AD, PSP, and BC cohorts. Multi-omics integration is associated with higher overlap counts, particularly in CTD-pathways, while ensemble-based methods show consistently strong performance across cohorts.

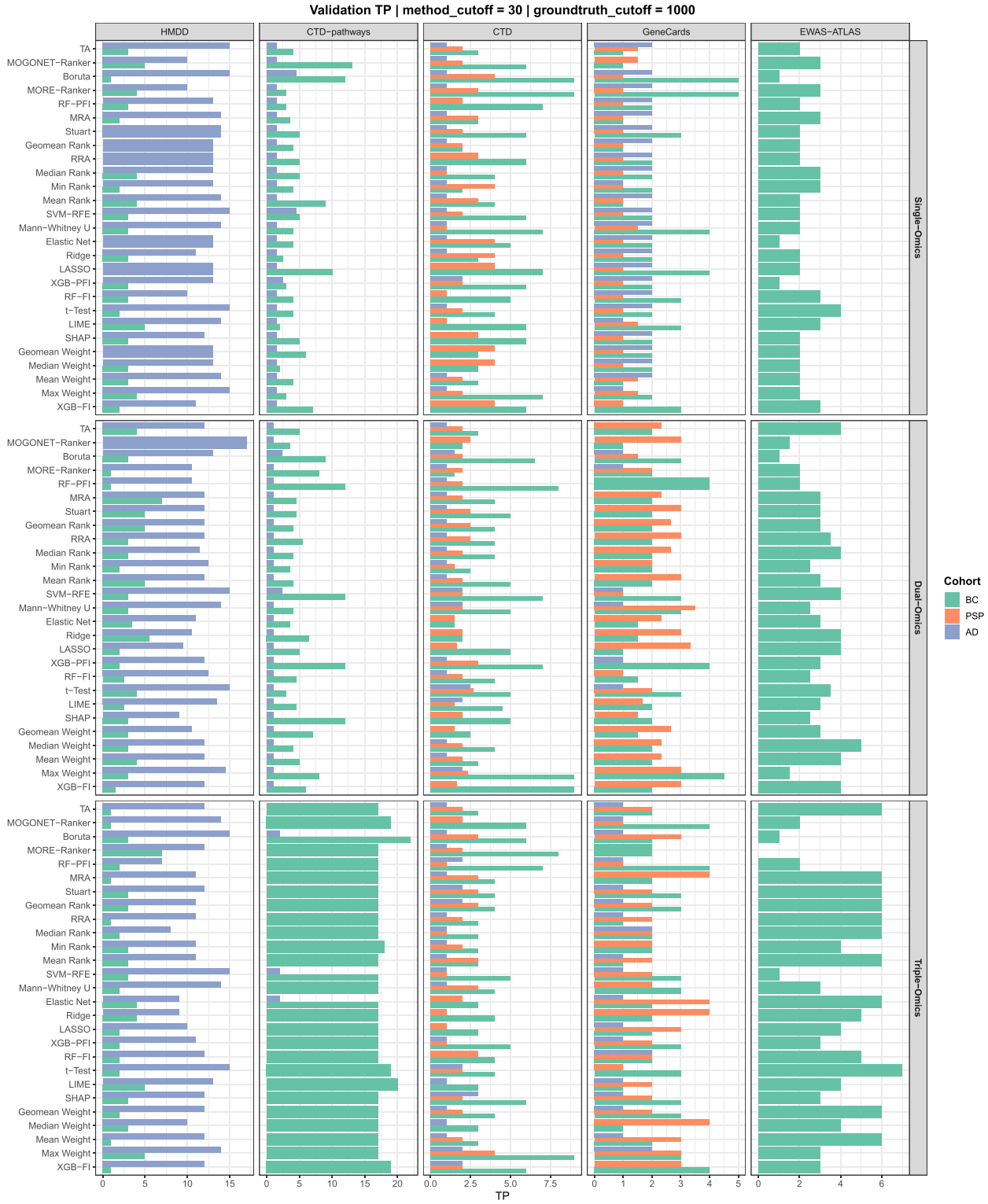

Figure 10 External validation results at the Top-30 cutoff. Bars represent overlaps with curated disease-associated features across HMDD, CTD-pathways, CTD, GeneCards, and EWAS-ATLAS across AD, PSP, and BC cohorts. A wider dynamic range of overlap counts is observed compared with smaller panel sizes, with CTD-pathways and EWAS-ATLAS showing increased validation. The relative ordering of methods remains similar, with ensemble-based approaches generally achieving higher overlap counts.

candidates such as hsa-miR-21-5p, hsa-miR-139-5p/3p, hsa-miR-141-3p, COL10A1, MMP11, and FIGF were consistently prioritized, confirming the framework's ability to detect well-characterized cancer biomarkers. At the same time, several mRNAs lacking strong prior association with BC (PPP1R12B, LRRC3B, TMEM220, PPAPD1C1A, HSD17B6, ADAMTS5, SDPR, and SPRY2) were reproducibly selected

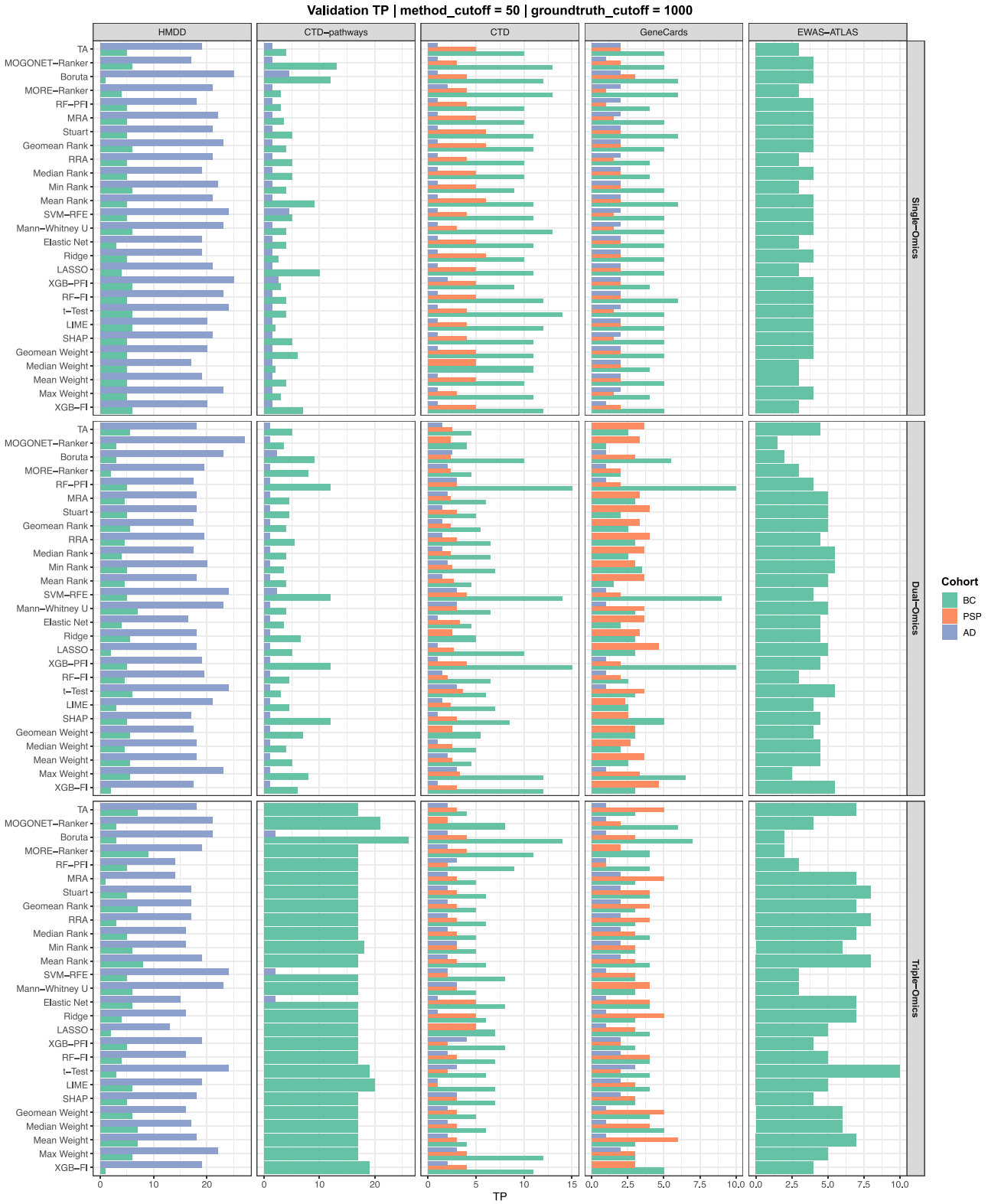

Figure 11 External validation results at the Top-50 cutoff. Bars represent overlaps with curated disease-associated features across multiple validation databases for AD, PSP, and BC cohorts. Overlap counts increase across all cohorts, with dual- and triple-omics configurations generally showing higher values. Differences between feature selection methods become less pronounced at this panel size.

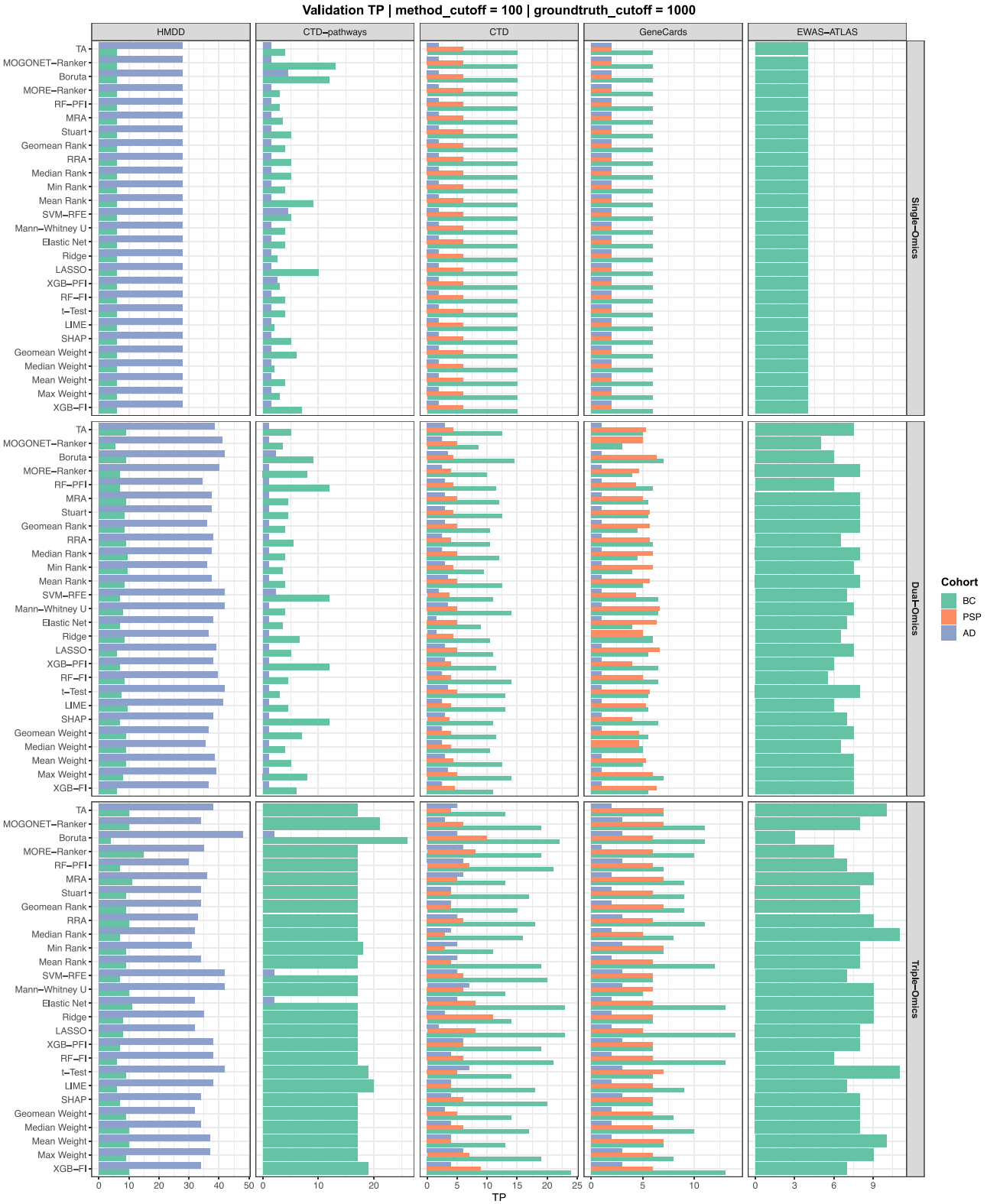

Figure 12 External validation results at the Top-100 cutoff. Bars represent overlaps with curated disease-associated features across validation databases for AD, PSP, and BC cohorts. Overlap counts are highest at this panel size, with signs of saturation in some datasets. Multi-omics configurations continue to show higher overlap values, and the relative performance differences between methods are largely maintained.

Table 5 PSP summary table of reproducible cross-omics biomarkers

Feature	FeatureType	OmicsLevel	Selection frequency	N (Selectors)
N-acetyl-3-methylhistidine*	Metab	Single+Dual+Triple	90	23
stachydrine	Metab	Single+Dual+Triple	88	23
1-linoleoyl-GPC (18:2)	Metab	Single+Dual+Triple	84	24
1-methyl-5-imidazolelactate	Metab	Single+Dual+Triple	80	24
trigonelline (N'-methylnicotinate)	Metab	Single+Dual+Triple	73	20
trimethylamine N-oxide	Metab	Single+Dual+Triple	63	22
1-methyl-5-imidazoleacetate	Metab	Single+Dual+Triple	60	18
1,2-dilinoleoyl-GPC (18:2/18:2)	Metab	Single+Dual+Triple	36	18
12-HHTrE	Metab	Single+Dual+Triple	26	17
salicylate	Metab	Single+Dual+Triple	23	15
2-methylserine	Metab	Single+Dual+Triple	20	15
TBCK	Prot	Single+Dual+Triple	58	22
NTN5	Prot	Single+Dual+Triple	37	18
ANKRD11	Prot	Single+Dual+Triple	35	17
ARHGAP19-SLIT1	Prot	Single+Dual+Triple	30	17
ARHGAP35	Prot	Single+Dual+Triple	27	22
E9PIM6	mRNA	Single+Dual	41	23
PPT1	mRNA	Single+Dual	34	21
LIPA4	mRNA	Single+Dual+Triple	30	15
PCP	mRNA	Single+Dual	24	20
ASAH1	mRNA	Single+Dual	21	20

Table 6 BC summary table of reproducible cross-omics biomarkers

Feature	FeatureType	OmicsLevel	Selection frequency	N (Selectors)
cg15700197	Meth	Single+Dual+Triple	54	19
cg23290344	Meth	Dual+Triple	33	19
cg21504624	Meth	Single+Dual	23	20
cg11052143	Meth	Single+Dual	23	17
cg25691167	Meth	Single+Dual+Triple	22	19
cg26353877	Meth	Dual+Triple	20	15
COL10A1	mRNA	Single+Dual+Triple	72	24
MMP11	mRNA	Single+Dual+Triple	57	24
LRRC3B	mRNA	Single+Dual+Triple	53	22
FIGF	mRNA	Single+Dual	37	17
CD300LG	mRNA	Single+Dual+Triple	36	19
SPRY2	mRNA	Single+Dual	33	19
PPAPDC1A	mRNA	Single+Dual+Triple	28	14
HSD17B6	mRNA	Single+Dual+Triple	27	22
ADAMTS5	mRNA	Single+Dual+Triple	26	19
PPP1R12B	mRNA	Dual+Triple	26	16
TMEM220	mRNA	Single+Dual+Triple	23	12
SDPR	mRNA	Single+Dual	21	15
CPA1	mRNA	Single+Dual+Triple	20	16
CA4	mRNA	Single+Dual+Triple	20	14
hsa-miR-21-5p	miRNA	Single+Dual+Triple	69	24
hsa-miR-139-5p	miRNA	Single+Dual+Triple	35	20
hsa-miR-96-5p	miRNA	Single+Dual+Triple	32	21
hsa-miR-10b-5p	miRNA	Single+Dual+Triple	31	19
hsa-miR-183-5p	miRNA	Single+Dual+Triple	28	20
hsa-miR-139-3p	miRNA	Single+Dual+Triple	28	18
hsa-miR-1307-5p	miRNA	Single+Dual+Triple	24	19

across rankers, highlighting potential novel players in BC biology. Similarly, CpG sites such as cg26353877, cg23290344, and cg11052143 repeatedly achieved top rankings, despite limited prior characterization, suggesting unexplored epigenetic signals. These results emphasize how multi-omics integration not only validates known oncogenic pathways but also systematically identifies reproducible novel features across data modalities.

Together, these analyses show that ensemble ranking consistently recovers disease-relevant biomarkers across AD, PSP, and BC, while also revealing reproducible novel features that warrant further experimental investigation.

Discussion

In this study, we benchmarked 27 feature ranking methods, multiple classifiers, and three levels of omics integration across heterogeneous cohorts (AD, PSP, and BC). By jointly evaluating predictive performance and biological validity, we provide a comprehensive

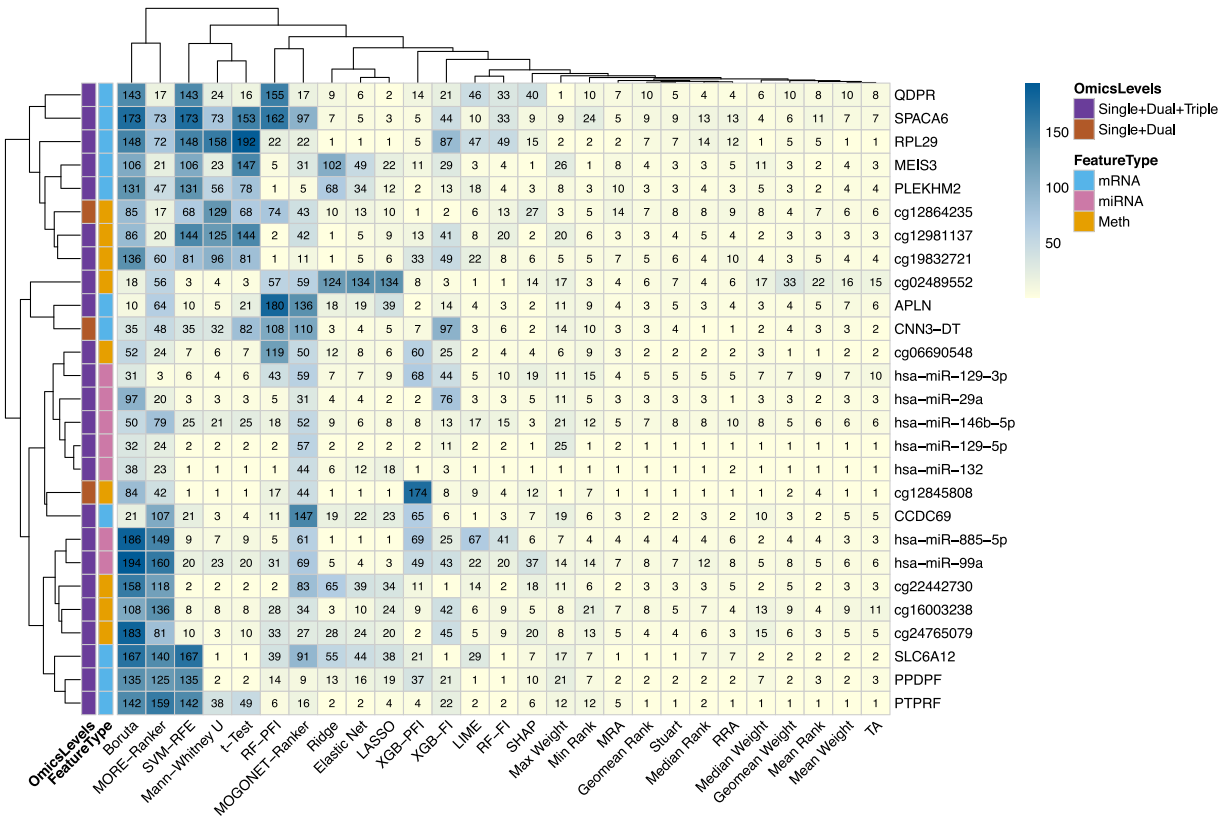

Figure 13 Rank heatmap for the AD cohort showing feature ranking stability across multiple selection methods, where rows represent candidate microRNAs, genes, and CpG methylation sites and columns correspond to single and ensemble feature selection approaches, with each cell indicating the assigned rank (lower values denote higher importance), highlighting consistently top-ranked features such as hsa-miR-129-5p, hsa-miR-132, hsa-miR-146b-5p, ARRDC2, PLEKHM2, and PPDPF, along with additional candidates including hsa-miR-885-5p, APLN, CCDC69, SPACA6, MEIS3, and CNN3-DT, and several CpG sites (e.g., cg12981137, cg22442730, cg16003238, cg19832721, and cg12864235) that also show consistently high rankings across methods.

assessment of how methodological choices influence reproducibility and biomarker reliability in high-dimensional, low-sample omics settings.

Our results show that model complexity does not necessarily translate into higher accuracy. Traditional classifiers—including L.Regression, SVMs, Random Forests, XGBoost, CatBoost, and MLPs—consistently matched or outperformed multiview deep learning models such as MORE and MOGONET. This trend was most pronounced in BC, where nearly all classifiers reached ceiling-level performance, demonstrating that well-regularized shallow models can fully capture strong molecular signals. Deep models showed lower median accuracy and greater variability, reflecting the difficulty of training high-parameter architectures with limited omics samples.

Feature selection emerged as a key determinant of stability and performance. Ensemble aggregation methods—including Mean/Median Rank, Mean/Median Weight, Geometric Means, RRA, Stuart, and TA—consistently provided the most robust panels across cohorts, integration levels, and panel sizes. Several classical selectors (LASSO, Elastic Net, Ridge, and SHAP) also performed strongly, reinforcing their utility in omics pipelines. In contrast, Boruta, MORE-Ranker, and MOGONET-Ranker tended to produce less stable or less predictive panels. These findings highlight the advantages of consensus-based feature aggregation.

Omics integration had a strong and systematic effect on performance. Dual-omics often provided the best trade-off between accuracy and generalization. Triple-omics achieved the highest maximum accuracy across selectors and classifiers, although gains depended on modality balance and cohort characteristics. AD benefitted strongly from miRNA–methylation combinations, PSP from mRNA–proteomics, and BC showed uniformly high accuracy regardless of integration level. These results highlight that multi-omics integration yields synergistic gains when modality quality is balanced.

Biological validation revealed that integration concentrates true disease signal. Across HMDD, CTD, GeneCards, and EWAS-ATLAS, dual- and triple-omics consistently increased true-positive overlaps even though each modality contributed fewer features in fixed-size panels. This pattern suggests that cross-omics prioritization filters modality-specific noise and enriches shared biological mechanisms—an effect observed consistently across cutoffs $k \in \{10, 30, 50, 100\}$.

Cohort differences reflected biological and technical factors. BC achieved the highest TP counts, likely due to sample size and strong transcriptomic signals. AD showed the strongest miRNA recovery (HMDD), consistent with modality strengths. PSP showed moderate but consistent gains across integration levels. Nevertheless, the selector hierarchy and integration benefit remained stable across cohorts.

Taken together, our results recommend multi-omics pipelines that integrate across molecular layers and employ ensemble-based feature aggregation. Mid-sized panels (30–60 features) paired with linear models, SVMs, or MLPs provided the best balance of accuracy,

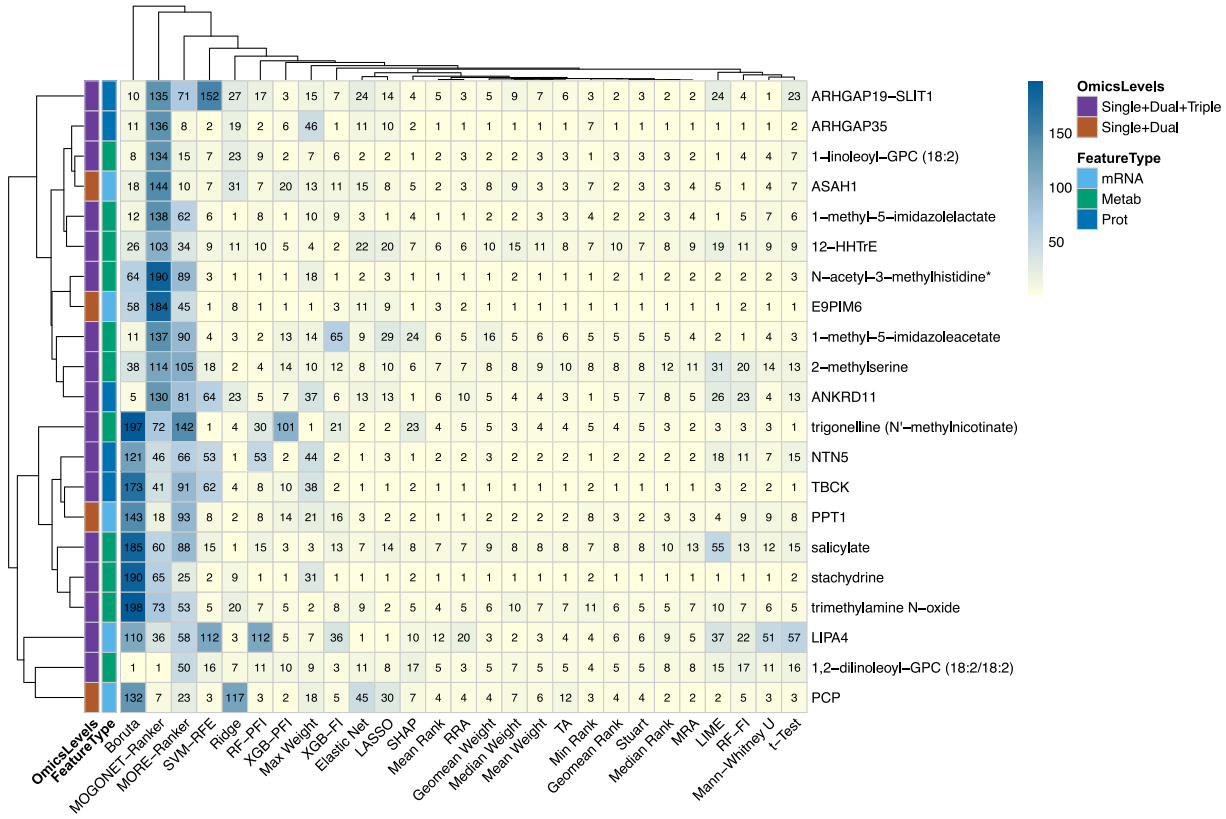

Figure 14 Rank heatmap for BC. Feature ranking stability across multiple selection methods in the BC cohort. Rows represent candidate features, including microRNAs, genes, and CpG methylation sites, while columns correspond to feature selection methods grouped into single and ensemble approaches. Each cell displays the rank assigned to a feature by a given method, with smaller values indicating higher importance. Several features, including hsa-miR-21-5p, hsa-miR-139-5p/3p, hsa-miR-96-5p, COL10A1, MMP11, and FIGF, are consistently ranked highly across methods, along with additional candidates such as PPP1R12B, LRRC3B, TMEM220, PPAPDC1A, HSD17B6, ADAMTS5, SDPR, and SPRY2. Multiple CpG methylation sites (e.g. cg23290344, cg26353877, cg25691167, cg11052143, and cg15700197) also show consistently high rankings across methods.

stability, and interpretability. Large panels ($K \geq 70$) offered only modest improvements, and deep learning models rarely outperformed simpler baselines, suggesting that their use should be reserved for substantially larger datasets with independent validation.

This study has limitations. Cohort sizes—particularly PSP—limit statistical power. Cohort heterogeneity complicates generalization, and technical differences (e.g. CpG mapping, platform variability) may influence rankings. Deep learning models may perform better in larger cohorts or with transfer learning.

Beyond methodological benchmarking, this study also has implications for translational biomarker discovery. By systematically evaluating feature selection strategies across multiple omics modalities and disease cohorts, the proposed framework provides a practical guide for identifying robust candidate biomarkers that warrant further biological and clinical investigation. For example, several top-ranked features identified across cohorts correspond to molecules previously implicated in disease-related pathways, supporting the biological relevance of the selected signatures. While the biomarkers identified in this study require experimental validation and prospective clinical evaluation, the benchmarking framework helps prioritize candidates that demonstrate consistent predictive performance and biological support across datasets and external validation resources. In future work, integrating this framework with additional data modalities such as genomic variants—including gene mutation profiles widely available in resources such as TCGA—may further improve biomarker robustness and facilitate translation into clinically actionable diagnostic or prognostic panels.

Future work should incorporate larger AD cohorts [Alzheimer's Disease Neuroimaging Initiative (ADNI), Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD)], broader multi-omics integration—including richer proteomic, metabolomic, and genomic variant data—and mechanistic validation of novel candidates identified here. Longitudinal prediction and treatment–response modeling represent promising next steps.

To promote transparency, we provide two complementary resources. BioMark is an online platform for user-defined multi-omics analysis that supports dimensionality reduction, feature selection using single- and ensemble-rankers, and predictive modeling [45]. BioMark is available at https://bioinf.itu.edu.tr/biomark/. Furthermore, an interactive benchmark explorer is also provided to inspect results generated using the benchmarking framework presented in this study. The benchmark framework and explorer application are available at https://github.com/itu-bioinformatics-database-lab/biomarker-benchmark. Further details are provided in the Supplementary Information.

Conclusion

This study presents a comprehensive benchmarking framework for multi-omics biomarker discovery, addressing a critical question in bioinformatics and translational research: Does increasing

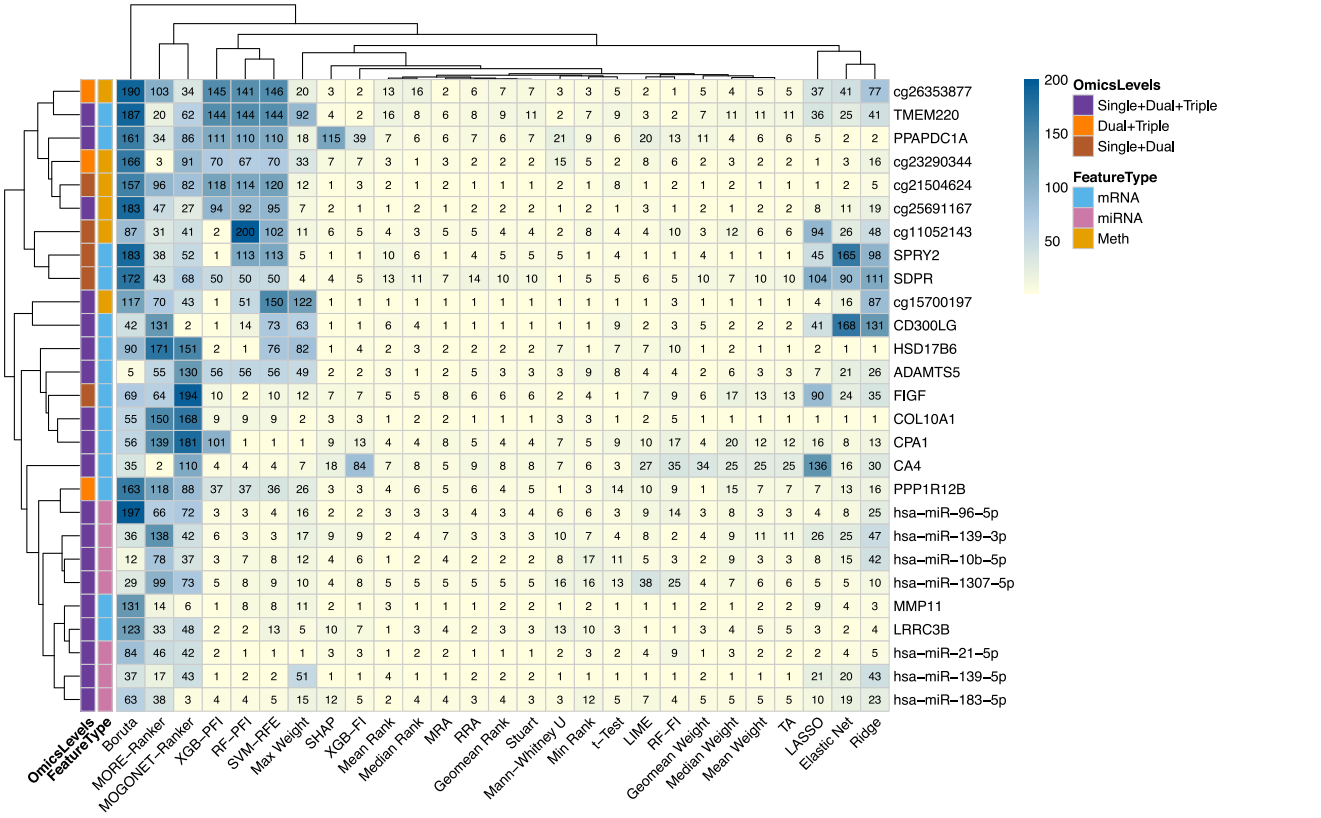

Figure 15 Rank heatmap for the PSP cohort showing feature ranking stability across multiple selection methods, where rows represent candidate mRNA, proteins, and metabolites and columns correspond to single and ensemble feature selection approaches, with each cell indicating the assigned rank (lower values denote higher importance), highlighting consistently top-ranked features such as TMAO, ANKRD11, and PPT1, along with additional candidates including TBCK, ARHGAP19-SLIT1, ARHGAP35, ASAH1, and metabolites such as stachydrine, trigonelline, salicylate, 1-methyl-5-imidazoleacetate, 2-methylserine, 12-HHTrE, and 1-linoleoyl–GPC (18:2).

methodological complexity yield better biomarkers? By systematically evaluating 27 feature selection strategies and 11 predictive models across three diverse disease cohorts—AD, PSP, and BC—we provide robust evidence that challenges the assumption that deep learning and complex integration methods inherently outperform traditional approaches.

Our findings demonstrate that ensemble-based feature selection methods, particularly those leveraging rank and weight aggregation, consistently enhance the stability, reproducibility, and biological relevance of biomarker panels. Notably, traditional machine learning models, such as L.Regression, SVMs, and MLPs often matched or exceeded the performance of advanced deep learning frameworks like MORE and MOGONET, especially in settings with limited sample sizes and high-dimensional data. We show a monotonic trend: dual-omics outperform single-omics, and triple-omics outperform dual-omics—triple-omics generally yields the best overall performance, with gains most evident when data quality and sample size are adequate. Compact biomarker panels (30–60 features) derived from ensemble selectors and interpretable models achieve high predictive accuracy and facilitate biological interpretation.

Biological validation against curated databases reinforces the clinical relevance of our approach. The reproducible identification of known and novel candidates underscores the utility of ensemble strategies. We provide a web-based interactive explorer for visualizing benchmarking results.

In conclusion, our work advocates for a pragmatic and evidence-driven approach to multi-omics biomarker discovery—one that prioritizes robustness, interpretability, and biological validation over algorithmic complexity. We encourage researchers to critically assess the tradeoffs between model sophistication and practical utility, especially in the context of translational applications.

Key Points

- Ensemble feature selection consistently yields more stable and accurate biomarker panels than individual rankers.
- Deep learning models (Multi-Omics hypergraph integration network, Multi-Omics Graph convolutional networks) do not outperform simpler classifiers such as logistic regression, Support Vector Machine, or Multilayer Perceptron.
- Predictive performance improves with integration level (Triple-omics > Dual-omics > Single-omics).
- External validation using Comparative Toxicogenomics Database, Human Gene Database, Human microRNA Disease Database, and the Epigenome-Wide Association Study Atlas confirms the biological and clinical relevance of the discovered biomarkers.
- The benchmarking pipeline promotes transparency, reproducibility, and practical applicability in multi-omics biomarker discovery.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

None declared.

Funding

A. Çakmak and C. Mesue Njume were supported by the Scientific and Technological Research Council of Türkiye (TÜBıTAK) through the EU Joint Programme—Neurodegenerative Disease Research (JPND) (Grant No. 124N069), by the Scientific Research Projects Unit of Istanbul Technical University (ITU BAP) (Grant No. TGA-2025-46998), and by the National Center for High-Performance Computing (UHEM) (Grant No. 1009742021). This research was also funded by the Italian Ministry of Health—EU Joint Programme—Neurodegenerative Disease Research (JPND), "Large scale analysis of OMICS data for drug-target finding in neurodegenerative diseases"—ERP-2023-23684212—ERP-2023-JPND-MyRIAD; by the Polish National Science Centre within the same project [2023/05/Y/NZ3/00160]; and by the Health Research Board, JPND-2023-1: EU Joint Programme—Neurodegenerative Disease Research (JPND), "Large scale analysis of OMICS data for drug-target finding in neurodegenerative diseases."

Data availability

The multi-omics datasets analyzed in this study are subject to data use agreements and cannot be publicly shared within this repository. Access to the raw data can be obtained directly from the respective repositories in accordance with their access policies. AD data, including mRNA, miRNA, and DNA methylation profiles from postmortem brain tissue, are available under controlled access through the Synapse AMP-AD Knowledge Portal (https://www.synapse.org/#!Synapse:syn3219045). PSP data, comprising transcriptomic, proteomic, and metabolomic measurements, are accessible through the AMP-AD Knowledge Portal (https://adknowledgeportal.synapse.org/) under study accession syn5550404. The associated metabolomics dataset, obtained from the same Mayo Clinic PSP samples, is available under the title "The Landscape of Metabolic Brain Alterations" (`syn26401311`). BC data, including gene expression, miRNA expression, and DNA methylation profiles, were obtained from the UCSC Xena Browser (https://xena.ucsc.edu/). Only processed and derived outputs—such as ranked feature lists, performance summaries, and validation statistics—are included in the GitHub repository to comply with institutional and repository data use policies.

Code availability

All code developed for this study is openly available at https://github.com/itu-bioinformatics-database-lab/biomarker-benchmark. The repository contains reproducible pipelines for data preprocessing, feature selection, model benchmarking, and validation analyses. Users can replicate all reported results by obtaining the datasets listed in the *Data Availability* section or apply the framework to their own multi-omics cohorts .

Implementation note. The Python implementations of the Stuart rank aggregation and RRA algorithms were adapted from the `RobustRankAggreg` R package (version 1.2; Kolde & Laur) [46], released under the GPL-2 license. Algorithmic equivalence with the original R implementation was verified using benchmark gene lists.

References

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017; **18**:83.
2. Subramanian I, Verma S, Kumar S *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020; **14**:117793221989905.
3. Huang Z, Zhan X, Xiang S *et al.* Salmon: survival analysis learning with multi-omics neural networks on breast cancer. *Front Genet* 2019; **10**:166.
4. Kim D, Li R, Dudek SM *et al.* Athena: identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Min* 2013; **6**:23.
5. Sun Y, Goodison S, Li J *et al.* Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 2006; **23**:30–7.
6. Clarke R, Ressom HW, Wang A *et al.* The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008; **8**:37–49.
7. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005; **2**:e124.
8. Boulesteix A-L, Slawski M. Stability and aggregation of ranked gene lists. *Brief Bioinform* 2009; **10**:556–68.
9. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA* 2006; **103**:5923–8.
10. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005; **365**:488–92.
11. Lazar C, Meganck S, Taminau J *et al.* Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 2012; **14**:469–90.
12. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017; **35**:498–507.
13. Awada W, Khoshgoftaar TM, Dittman D *et al.* A review of the stability of feature selection techniques for bioinformatics data. In: *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pp. 356–63. Las Vegas, NV, USA: IEEE, 2012.
14. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl Inf Syst* 2006; **12**:95–116.
15. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Comput Applic* 2019; **32**:5951–73.
16. Song X, Waitman LR, Hu Y *et al.* Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inform Assoc* 2019; **26**:242–53.
17. Seijo-Pardo B, Bolón-Canedo V, Alonso-Betanzos A. On developing an automatic threshold applied to feature selection ensembles. *Inf Fusion* 2019; **45**:227–45.
18. Spooner A, Mohammadi G, Sachdev PS *et al.* Ensemble feature selection with data-driven thresholding for Alzheimer's disease biomarker discovery. *BMC Bioinformatics* 2023; **24**:9.

19. Paplomatas P, Krokidis MG, Vlamos P *et al.* An ensemble feature selection approach for analysis and modeling of transcriptome data in Alzheimer's disease. *Appl Sci* 2023; **13**:2353.
20. Claude E, Leclercq M, Thébault P *et al.* Optimizing hybrid ensemble feature selection strategies for transcriptomic biomarker discovery in complex diseases. *NAR Genom Bioinform* 2024; **6**:lqae079.
21. Liang Y, Gharipour A, Kelemen E *et al.* Homogeneous ensemble feature selection for mass spectrometry data prediction in cancer studies. *Mathematics* 2024; **12**:2085.
22. Yao Z, Zhu G, Too J *et al.* Feature selection of omic data by ensemble swarm intelligence based approaches. *Front Genet* 2022; **12**:793629.
23. Le P, Gong X, Ung L *et al.* A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data. *Front Syst Biol* 2024; **4**:1355595.
24. Budhraja S, Doborjeh M, Singh B *et al.* Filter and wrapper stacking ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data. *Brief Bioinform* 2023; **24**:bbad382.
25. Li Y, Mansmann U, Du S *et al.* Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinformatics* 2022; **23**:412.
26. Łukaszuk T, Krawczuk J, Żyła K *et al.* Stability of feature selection in multi-omics data analysis. *Appl Sci* 2024; **14**:11103.
27. Davis A, Wiegers T, Sciaky D *et al.* Comparative toxicogenomics database's 20th anniversary: update 2025. *Nucleic Acids Res* 2024; **53**:D1328–34.
28. Cui C, Zhong B, Fan R *et al.* HMDD v4.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2023; **52**:D1327–32.
29. Rebhan M, Chalifa-Caspi V, Prilusky J *et al.* GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 1998; **14**:656–64.
30. Xiong Z, Li M, Ma Y *et al.* GMQN: a reference-based method for correcting batch effects and probe bias in humanmethylation beadchip. *Front Genet* 2022; **12**:810985.
31. Mihajlović K, Ceddia G, Malod-Dognin N *et al.* Multi-omics integration of scRNA-seq time series data predicts new intervention points for Parkinson's disease. *Sci Rep* 2024; **14**:10983.
32. Tripathy RK, Frohock Z, Wang H *et al.* Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks. *NPJ Syst Biol Appl* 2025; **11**:43.
33. Liang J, Huang X, Li W *et al.* Identification and external validation of the hub genes associated with cardiorenal syndrome through time-series and network analyses. *Aging* 2022; **14**:1351–73.
34. Wang Y, Wang Z, Yu X *et al.* MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification. *Brief Bioinform* 2024; **26**:bbae658.
35. Wang T, Shao W, Huang Z *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* 2021; **12**:3445.
36. Pérez-González AP, García-Kroepfly AL, Pérez-Fuentes KA *et al.* The ROSMAP project: aging and neurodegenerative diseases through omic sciences. *Front Neuroinform* 2024; **18**:1443865.
37. Allen M, Carrasquillo MM, Funk C *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data* 2016; **3**:160089.
38. Goldman MJ, Craft B, Hastie M *et al.* Visualizing and interpreting cancer genomics data via the xena platform. *Nat Biotechnol* 2020; **38**:675–8.
39. Kolberg L, Raudvere U, Kuzmin I *et al.* g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res* 2023; **51**:W207–12.
40. Bennett DA, Schneider JA, Arvanitakis Z *et al.* Overview and findings from the religious orders study. *Curr Alzheimer Res* 2012; **9**:628–45.
41. De Jager PL, Ma Y, McCabe C *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data* 2018; **5**:180142.
42. Weinstein JN, Collisson EA, Mills GB *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013; **45**:1113–20.
43. Mayeux R. Biomarkers: potential uses and limitations. *NeuroRX* 2004; **1**:182–8.
44. Grande G, Valletta M, Rizzuto D *et al.* Blood-based biomarkers of Alzheimer's disease and incident dementia in the community. *Nat Med* 2025; **31**:2027–35.
45. Balikci MA, Njume CM, Cakmak A. Biomark: biomarker analysis tool. *BMC Bioinformatics* 2026; **27**:42.
46. Kolde R, Laur S, Adler P *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012; **28**:573–80.

Briefings in Bioinformatics, 2026, **27**, bbag211
https://doi.org/10.1093/bib/bbag211
Review