When Complexity Doesn't Pay: Benchmarking Deep Learning and Ensemble Methods for Biomarker Discovery

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

[1] Department of Molecular Biology and Genetics, Ayazaga Campus, Istanbul Technical University, Reşitpaşa, Sarıyer, 34467 Istanbul, Turkey

[2]Department of Computer Engineering, Ayazaga Campus, Istanbul Technical University, Reşitpaşa, Sarıyer, 34467 Istanbul, Turkey

[3] Molecular Markers Laboratory, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, 25125 Brescia, Italy

[4] Discipline of Physiology, School of Medicine, University of Galway, H91 TH33 Galway, Ireland

[5] Department of Medical Biotechnology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

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*

Key Points

- Ensemble feature selection consistently yields more stable and accurate biomarker panels than individual rankers.
- Deep learning models (MORE, MOGONET) do not outperform simpler classifiers such as logistic regression, SVM, or MLP.
- Predictive performance improves with integration level (Triple-omics ¿ Dual-omics ¿ Single-omics).
- External validation using CTD, GeneCards, HMDD, and EWAS-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.

1 Introduction

Biomarker discovery has become a cornerstone of precision medicine, enabling earlier disease de-

*Corresponding author: ali.cakmak@itu.edu.tr

tection, 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] [3] [4] [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][7][8][9][10]. Second, the heterogeneity of omics layers poses additional challenges: gene expression values are continuous, DNA methylation 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 [13][12][14]. Ensemble ranking has been proposed to mitigate such variability and improve feature stability and reproducibility across high-dimensional omics datasets [15][16][17][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][23][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 miRNA–disease associations [28]; **GeneCards**, an integrative compendium of human genes and their disease/phenotype links [29]; and the **EWAS-ATLAS**, a catalog of 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, 32, 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 MORE [34] and MOGONET [35] have been proposed to capture cross-modality relationships and exploit complex non-linear 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 ROSMAP cohort [36], progressive supranuclear palsy (PSP) using the

MayoRNASeq cohort [37], and breast cancer (BC) using the 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 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.

2 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. Each subsection details a specific component of the workflow, from data preparation to biomarker validation and performance evaluation.

2.1 Study cohort

Experiments in this study were conducted using three real-world, publicly available datasets—the Religious Orders Study and Memory and Aging Project (ROSMAP; Alzheimer's disease, **AD**) [36], the MayoRNASeq cohort (progressive supranuclear palsy, **PSP**) [37], and the TCGA-BRCA cohort (breast cancer, **BC**) accessed via the UCSC Xena Browser [38]. The characteristics of these datasets are summarized in Table 1, and brief descriptions of each dataset are provided in the following sections. Henceforth, we refer to the three cohorts by their disease-based shorthand: **AD** (ROSMAP), **PSP** (MayoRNASeq), and **BC** (TCGA–BRCA). To avoid ambiguity, when we mean the disease entity rather than the cohort, we spell it out in full (e.g., "Alzheimer's disease," "progressive supranuclear palsy," "breast cancer").

Religious Orders Study and Memory and Aging Project (ROSMAP)

ROSMAP is a longitudinal cohort study designed to investigate the neuropathological and molecular mechanisms underlying aging and Alzheimer's disease [40]. ROSMAP combines two harmonized prospective cohorts: the Religious Orders Study, which recruits older Catholic nuns, priests, and brothers across the United States, and the Memory and Aging Project, which enrolls lay older individuals from the Chicago area. Participants are cognitively assessed annually, and all consent to brain donation at the time of death, enabling comprehensive linkage between clinical trajectories and postmortem molecular data. AD encompasses extensive phenotypic and molecular information, including demographic and clinical variables, neuropsychological testing, neuropathological indices, genetic variation, transcriptomics, epigenomics, proteomics, and metabolomics [41]. For the present study, we focused on molecular profiles generated from postmortem dorsolateral prefrontal cortex (DLPFC), including microRNA expression, mRNA expression, and DNA methylation data, along with diagnostic information on Alzheimer's disease. The analytic cohort included participants with available molecular profiles and clinical diagnoses, after standard quality control and preprocessing. Individuals with incomplete molecular or phenotypic information were excluded.

The Mayo RNAseq Study (MayoRNAseq)

The MayoRNASeq study is a large-scale project designed to characterize the multi-omics profiles of neurodegenerative diseases, including progressive supranuclear palsy, Alzheimer's disease, and age-matched controls [37]. The dataset was generated from postmortem brain tissues obtained from the Mayo Clinic Brain Bank, with samples collected from both the temporal cortex and cerebellum. Participants were recruited as part of neuropathological investigations and classified into diagnostic categories based on established consensus criteria integrating clinical assessments with postmortem neuropathological evaluations. PSP provides extensive molecular and phenotypic information, including RNA-Seq expression profiles, proteomics, clinical diagnoses, and metabolomics data. For the present study, we focused exclusively on postmortem temporal cortex samples, analyzing RNA-Seq gene expression, proteomics, and metabolomics patterns from participants diagnosed with progressive supranuclear palsy and matched controls. After applying standard quality

Table 1: Overview of study characteristics

Cohort	Disease / Condition	Data Types	Sample Size	Data Source
ROSMAP	Alzheimer's disease	Transcriptomics, microRNA, methylation, clinical	378	Synapse[a]
TCGA-BRCA	Breast cancer	Transcriptomics, methylation, microRNA, clinical	1,229	UCSC Xena[b]
MayoRNASeq	Progressive Supranuclear Palsy	Transcriptomics, metabolomics, proteomics, clinical	437	Synapse[c]

[a] ROSMAP Dataset
[b] UCSC Xena TCGA-BRCA Dataset
[c] MayoRNASeq Dataset
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 Section 2.2.

control and preprocessing steps, only participants with complete molecular profiles and clinical information were included in the analytic cohort.

Breast Cancer Cohort (TCGA-BRCA)

The TCGA-BRCA dataset was obtained from The Cancer Genome Atlas (TCGA) project via the UCSC Xena data portal [42]. TCGA-BRCA is a large-scale, multi-omics study designed to comprehensively characterize the molecular and clinical features of breast cancer. The dataset includes transcriptomics, microRNA expression, DNA methylation, copy number variation, somatic mutations, and matched clinical annotations. For the present study, we focused on transcriptomic profiles derived from bulk RNA sequencing (Illumina HiSeq 2000), microRNA expression data, and DNA methylation data generated using the Illumina HumanMethylation450 platform. After applying standard quality control and preprocessing steps, only samples with complete molecular profiles and clinical data were included in the analytic cohort.

2.2 Data Preprocessing Pipeline

All data cleaning, preprocessing, and preparation steps were performed as part of a unified data preprocessing pipeline to ensure consistency across cohorts and omics modalities.

Data Cleaning

For each dataset, clinical metadata were harmonized with the corresponding omics expression matrices using biospecimen mapping files to ensure consistent sample identifiers across modalities. Duplicate samples and features were removed, and features with more than 30% missing values were excluded. Remaining missing values were imputed using the KNNImputer from the Scikit-learn library (version 1.7.0), with distance-weighted averaging based on the five nearest neighbors. To ensure consistency of molecular identifiers, microRNA names were standardized using miRBase (release version obtained in 2025), while Ensembl gene IDs were mapped to HGNC gene symbols using the MyGene Python API (version 3.1.0). This mapping process improved cross-study comparability and downstream interpretability. For each dataset, only samples with complete multi-omics profiles and matched clinical annotations were retained for downstream analyses. All data cleaning and preprocessing steps were performed in a dedicated Python 3.11.9 environment.

Data Preprocessing

To reduce noise, remove experimental artifacts, and improve interpretability of downstream analyses, we adopted data preprocessing protocols adapted from the MOGONET framework [35]. For DNA methylation data, only probes corresponding to the Illumina Infinium HumanMethylation27 BeadChip were retained to ensure biological interpretability. Features with no signal (mean value equal to zero) or low variance were removed. We applied variance-based filtering thresholds according to data modality: (i) mRNA expression: features with variance < 0.1 were removed, (ii) DNA methylation: features with variance < 0.001 were removed, and (iii) miRNA expression, proteomics, and metabolomics: only features with zero variance were excluded, given the limited number of available features.

To further reduce redundancy, we performed feature preselection using ANOVA F-tests within the training set to identify features significantly

associated with the target labels. False Discovery Rate (FDR) control was applied to correct for multiple testing. To avoid selecting highly correlated features, we imposed an additional constraint such that the first principal component of the preselected features explained $< 50\%$ of the variance. After feature selection, each omics dataset was individually scaled to the range [0, 1] using min-max normalization before model training. For the PSP metabolomics and proteomics datasets, a $\log_{10}$ transformation was applied prior to preprocessing to stabilize variance and normalize feature distributions.

Data Preparation

For each cohort, datasets were organized to support three experimental settings: (i) single-omics, (ii) paired dual-omics, and (iii) integrated triple-omics analyses. This design resulted in varying dataset sizes and class distributions across the three cohorts (AD, PSP, and BC). In BC, substantial class imbalance was observed between tumor and control samples. To mitigate potential bias in model training, the dominant class (tumor) was downsampled to approximately 1.2 times the size of the minority class (control) in all experimental settings. A detailed summary of dataset characteristics, including the number of samples, feature counts per omics type, and class distributions for each data subset, is provided in Table 2. Table rows are grouped by cohort to clearly distinguish dataset-specific characteristics.

2.3 Single Rankers

To identify predictive and biologically meaningful features across heterogeneous, high-dimensional multi-omics datasets, we employed a comprehensive suite of 15 single-feature ranking approaches, each representing a different paradigm of feature importance estimation, summarized in Table 3. These single rankers were selected to ensure diversity in algorithmic foundations, ranging from model-agnostic explainers to regularized regression, statistical hypothesis testing, and deep learning–based multi-omics rankers. The combination of these complementary strategies increases robustness by capturing both linear and non-linear dependencies while mitigating biases introduced by any single method.

2.4 Ensemble Rankers

To improve robustness and reduce biases from individual feature selectors, we applied a comprehensive ensemble rank aggregation strategy that combined the results of 13 single rankers (rankers in Table 3 except MORE-Ranker and MOGONET-Ranker). Ensemble learning mitigates the instability of individual ranking algorithms by integrating their outputs into a consensus ranking. We implemented two main categories of ensemble methods: **rank-based aggregation**, which integrates the relative ordering of features across rankers, and **weight-based aggregation**, which combines continuous importance scores assigned by different rankers. A summary of the ensemble methods is provided in Table 4.

MORE-Ranker and **MOGONET-Ranker** were excluded from the ensemble because our objective was to systematically benchmark three distinct tiers of methods: (1) conventional single rankers, (2) ensemble-based aggregations of these single rankers, and (3) advanced deep learning–based models such as MORE and MOGONET. Including the latter in the ensemble would have blurred the distinction between traditional aggregation-based strategies and complex representation-learning frameworks. Instead, they were retained as independent benchmarks to evaluate how ensemble rank aggregation compares with state-of-the-art deep learning–based feature selection methods.

These ensemble rankers combine complementary evidence from diverse single-feature selectors. Rank-based methods emphasize consensus ordering, whereas weight-based methods leverage continuous importance scores. Together, they enhance stability, robustness, and reproducibility of biomarker discovery in high-dimensional, multi-omics datasets.

2.5 Benchmark Setup

Starting from the cleaned and prepared datasets described in Section 2.2, we designed a comprehensive benchmarking framework to evaluate feature selection strategies, predictive models, and multi-omics integration methods. The benchmarking experiments were performed independently for each cohort (**AD**, **PSP**, and **BC**) and across three integration levels: single-omics, dual-omics, and triple-omics analyses. In the single-omics setting, each omics type (e.g., mRNA, miRNA, methylation, proteomics, or metabolomics) was analyzed separately to assess its individual predictive power. The dual-omics setting included all pairwise combinations of omics modalities available within each cohort—for example, AD included mRNA–miRNA, mRNA–methylation, and miRNA–methylation pairs; PSP included mRNA–proteomics, mRNA–metabolomics, and metabolomics–proteomics; and BC included

Cohort	Omics Combination	Samples	Features	Classes	Class Ratio
AD	miRNA only	378	200	Alzheimer's disease / Control	209 / 169
	mRNA only	378	200	Alzheimer's disease / Control	209 / 169
	Meth only	375	200	Alzheimer's disease / Control	207 / 168
	mRNA + miRNA	375	200 + 200	Alzheimer's disease / Control	209 / 169
	mRNA + Meth	375	200 + 200	Alzheimer's disease / Control	207 / 168
	miRNA + Meth	375	200 + 200	Alzheimer's disease / Control	207 / 168
	mRNA + miRNA + Meth	375	200 + 200 + 200	Alzheimer's disease / Control	207 / 168
PSP	mRNA only	162	200	Progressive supranuclear palsy / Control	84 / 78
	Prot only	111	200	Progressive supranuclear palsy / Control	83 / 29
	Metab only	98	200	Progressive supranuclear palsy / Control	78 / 20
	mRNA + Prot	111	200 + 200	Progressive supranuclear palsy / Control	83 / 29
	mRNA + Metab	98	200 + 200	Progressive supranuclear palsy / Control	78 / 20
	Metab + Prot	97	200 + 200	Progressive supranuclear palsy / Control	77 / 20
	mRNA + Prot + Metab	97	200 + 200 + 200	Progressive supranuclear palsy / Control	77 / 20
BC	miRNA only	158	200	Breast cancer / Control	83 / 75
	mRNA only	236	200	Breast cancer / Control	124 / 112
	Meth only	375	200	Breast cancer / Control	106 / 96
	mRNA + miRNA	158	200 + 200	Breast cancer / Control	83 / 75
	mRNA + Meth	108	200 + 200	Breast cancer / Control	57 / 51
	miRNA + Meth	175	200 + 200	Breast cancer / Control	92 / 83
	mRNA + miRNA + Meth	108	200 + 200 + 200	Breast cancer / Control	57 / 51

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

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

mRNA–miRNA, miRNA–methylation, and mRNA–methylation combinations. The triple-omics setting involved the joint integration of all available omics layers in each cohort (mRNA + miRNA + methylation for AD and BC, and mRNA + proteomics + metabolomics for PSP) to evaluate the added value of full multi-omics integration.

Although each cohort provides additional omics modalities beyond these three, we restricted our analyses to a common subset to ensure methodological comparability with **MORE** and **MOGONET**, which are primarily designed for three-layer integration. This selection enabled fair benchmarking under consistent experimental conditions.

This hierarchical setup enabled a systematic comparison across three methodological tiers—conventional single rankers, ensemble rank aggregation methods, and deep learning–based models (MORE and MOGONET). A detailed summary of all data subsets, including the number of samples, feature counts, and class distributions, is provided in Table 2.

For every omics subset, we generated biomarker panels of increasing size to systematically evaluate model performance under varying levels of feature sparsity. Candidate panel sizes were set to {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}, in addition to a full-feature panel containing all available features: **200** for single-omics datasets, **400** for dual-omics subsets, and **600** for triple-omics subsets. This enabled a structured exploration of predictive performance while balancing panel interpretability and computational cost.

A total of *27 feature selection strategies* were evaluated, spanning single rankers and ensemble-based rankers. Single rankers included model-agnostic explainers such as SHAP and LIME, statistical tests (two-sample t-Test and Mann–Whitney U-test), regularized regression methods (LASSO, Ridge, and Elastic Net), tree-based selectors (Boruta, RF Gini Importance (RF-FI), RF Permutation Importance (RF-PFI), XGBoost Gain (XGB-FI), and XGBoost Permutation Importance (XGB-PFI)), wrapper-based selection via SVM-RFE, and deep learning-based methods (MORE-Ranker and MOGONET-RANKER). In addition, we incorporated ensemble feature aggregation methods, categorized into two groups: *rank-based* (Mean Rank, Median Rank, Min Rank, Geomean Rank, Median Rank Algorithm (MRA), Stuart Rank Aggregation (Stuart), and Robust Rank Aggregation (RRA)) and *weight-based* (Mean Weight, Median Weight, Max Weight, Geomean Weight, and the Threshold Algorithm (TA)). For rank-based selectors, features were ordered according to their relative importance, while weight-based selectors integrated normalized feature importance scores to produce consensus rankings.

For each cohort, omics subset, feature selector, and biomarker panel size, we trained *11 predictive models* spanning diverse learn-

Table 3: Summary of non-ensemble (single) feature ranking methods used in this study.

Method	Model/Approach	Type	Description	Ref.
Model-Agnostic Explainable AI				
SHAP	XGBoost (post-hoc)	Model-agnostic	Decomposes predictions into additive contributions using Shapley values from cooperative game theory.	[43, 44]
LIME	Random Forest (surrogate)	Model-agnostic	Approximates local decision boundaries with interpretable surrogate models to explain feature influence.	[45]
Regularized Regression-Based Methods				
LASSO	Penalized regression	Multivariate	ℓ_1 penalty enforces sparsity, selecting a small subset of important features.	[46]
Ridge	Penalized regression	Multivariate	ℓ_2 penalty shrinks coefficients smoothly, mitigating multicollinearity.	[47]
Elastic Net	Penalized regression	Multivariate	Combines ℓ_1 and ℓ_2 penalties to balance sparsity and stability.	[48]
Statistical Hypothesis Testing				
t-Test	Statistical test	Univariate	Two-sample t-test comparing group means under approximate normality.	–
Mann–Whitney U	Non-parametric test	Univariate	Rank-based test comparing group distributions without assuming normality.	[49]
Tree-Based Feature Selection and Importance				
Boruta	Random Forest wrapper	Multivariate	Compares real feature importance with permuted "shadow" features; retains only relevant features.	[50]
RF-FI	Random Forest	Multivariate	Gini (impurity) importance aggregated across trees.	[51]
RF-PFI	Random Forest	Model-agnostic	Permutation importance: drop in performance after permuting each feature.	–
XGB-FI	Gradient boosting	Multivariate	XGBoost gain-based importance aggregated across boosted trees.	[52]
XGB-PFI	Gradient boosting	Model-agnostic	Permutation importance computed on trained XGBoost models.	–
Wrapper-Based Recursive Feature Elimination				
SVM-RFE	Linear Support Vector Machine (SVM)	Multivariate	Iteratively removes the least influential features based on SVM weights, yielding a ranking.	[53]
Deep Learning–Based Multi-Omics Feature Ranking				
MORE-Ranker	Deep autoencoder integration	Multivariate	Learns joint latent representations across omics and derives feature importance via gradient-based attribution.	[34]
MOGONET-Ranker	Graph neural network integration	Multivariate	Uses modality-specific graphs and cross-view attention to assign feature importance scores.	[35]

Note: "MORE-Ranker" and "MOGONET-Ranker" denote the feature rankings generated by the respective frameworks' biomarker-ranking scripts via RFE; in this benchmark, these rankings are used as standalone selectors and are not included in the ensemble aggregations.

ing paradigms: L.Regression (**L.Regression**), Random Forest, Extreme Gradient Boosting (**XGBoost**), CatBoost, AdaBoost, Gradient Boosting (**Gradient Boost**), Support Vector Machine (**SVM**), Multilayer Perceptron (**MLP**), Decision Tree (**D.Tree**), MORE, and MOGONET. The two multi-omics–specific deep models—MORE and MOGONET—are designed to exploit cross-modality dependencies for integrative biomarker discovery. Model evaluation used **5-fold stratified cross-validation**, ensuring balanced class distributions within each fold and enabling consistent comparisons of predictive performance. Henceforth, we use the abbrevia-

Table 4: Summary of ensemble rank aggregation methods applied in this study.

Method	Description	Ref.
Rank-Based Aggregation		
Mean Rank	Computes the arithmetic mean of feature ranks across all rankers; features with consistently high ranks receive better scores.	–
Median Rank	Uses the median rank to reduce sensitivity to extreme rankings from outlier rankers.	–
Min Rank	Assigns each feature its best (lowest) rank observed across rankers, emphasizing consistently strong features.	–
Geomean Rank	Aggregates feature ranks using their geometric mean, favoring features that perform well across most rankers.	–
Median Rank Algorithm (MRA)	A consensus aggregation algorithm designed to maximize Kendall's τ correlation with individual rankers by iteratively aligning medians.	[54]
Stuart Rank Aggregation (Stuart)	A probabilistic meta-analysis-based method that models rank distributions to estimate significance levels for aggregated ranks.	[55]
Robust Rank Aggregation (RRA)	Assigns p-values to features by testing whether their observed ranks deviate from a uniform null distribution, robust to noise and missing data.	[55]
Weight-Based Aggregation		
Mean Weight	Computes the arithmetic mean of normalized importance weights across rankers, producing a consensus feature score.	–
Median Weight	Uses the median importance score to minimize the influence of extreme weights from unstable rankers.	–
Max Weight	Assigns each feature its highest observed weight across rankers, prioritizing features that are strongly favored by any ranker.	–
Geomean Weight	Aggregates feature weights using the geometric mean, favoring features consistently scored high across rankers.	–
Threshold Algorithm (TA)	Iteratively identifies features exceeding a dynamic weight threshold across rankers, retaining only those showing consistent importance.	[56]

tions in parentheses.

All scripts and experiments were implemented in **Python 3.11.9**, leveraging `scikit-learn` for traditional machine learning methods, i.e., `xgboost`, `catboost`, and `lightgbm` for gradient boosting methods, and customized Python implementations for **MORE** and **MOGONET**, as provided by the authors of these works. A schematic overview of the entire benchmarking workflow, including data preparation, feature selection, biomarker panel generation, model evaluation, and performance assessment, is presented in Figures 1 and 2. A detailed description of the evaluation metrics applied within the cross-validation loop is provided in Section 2.7.

2.6 Biomarker Panel Validation

To ensure biological relevance and clinical interpretability, each biomarker panel generated during the benchmarking experiments was systematically validated against external reference databases. For every panel size in each experimental setting described in Section 2.5, we performed validation at multiple molecular levels using publicly available, expert-curated resources. In experiments involving dual-omics or triple-omics integration, the selected panel could contain features from multiple data modalities. In such cases, panels were decomposed into omics-specific subsets (e.g., genes, microRNAs, methylation sites) prior to validation, ensuring that each feature type was evaluated using the most appropriate reference resource.

Validation was conducted at four distinct levels corresponding to specific molecular and functional categories: (1) **microRNAs** — disease associations were validated using the *HMDD (v4.0, 2025)* [28]; (2) **genes** — validated using two complementary resources, the *CTD (Release Aug 27 2025)* [27] and *GeneCards (Version 5.25)* [29], both of which provide curated disease–gene associations with evidence scores; (3) **DNA methylation sites** — validated using the *EWAS-ATLAS (v2025)* [30], a comprehensive repository of epigenome-wide associations linking CpG sites to diseases and traits; and (4) **pathways** — as-

sessed using *g:Profiler (v0.2, 2025)* [39] for functional enrichment, with enriched pathways subsequently cross-referenced against pathway–disease associations from CTD for independent confirmation.

Generation of Validation Datasets

Validation datasets were compiled from the aforementioned resources to ensure consistent reference standards across all experiments. From the *CTD* database, we retrieved all disease-associated genes and pathways relevant to Alzheimer's disease, breast cancer, and progressive supranuclear palsy. Each gene–disease association was accompanied by a confidence score provided by CTD. For pathway-level validation, we computed a custom association score for each pathway based on the proportion of disease-associated genes present within the pathway relative to its total gene content. Similarly, gene–disease associations and evidence scores were obtained from the *GeneCards* database, ensuring complementary coverage of disease-relevant targets. For methylation sites, validation datasets were extracted from *EWAS-ATLAS*, which links CpG sites to diseases via large-scale epigenome-wide association studies. MicroRNA–disease associations and supporting evidence were obtained from the *HMDD* database, which compiles experimentally validated miRNA–disease relationships. Protein- and metabolite-level validation was not performed due to the lack of sufficiently reliable and comprehensive reference datasets as of 2025. Because CTD curates pathway–disease associations, we treat this as a separate validator (CTD-pathways). For each gene set derived from gene-expression analyses, we performed pathway enrichment with g:Profiler and validated the predicted pathways against CTD-pathways.

2.7 Evaluation Metrics

To comprehensively assess the predictive performance and reliability of biomarker panels across models and experimental settings, we computed multiple evaluation metrics within the 5-fold stratified cross-validation framework. These metrics capture different aspects of model performance, including classification accuracy, sensitivity, specificity, predictive values, and overall discriminative ability. *Accuracy* measures the proportion of correctly classified samples among all predictions and provides a general measure of model performance but can be misleading when class imbalance is present. To address this, we considered additional metrics that separately evaluate positive and negative classifications. *Precision* (also known as the Positive Predictive Value, PPV) quantifies the proportion of correctly predicted positive samples among all positive predictions, reflecting the reliability of positive classifications. *Recall* (also referred to as Sensitivity or True Positive Rate) measures the proportion of true positive samples correctly identified by the model, capturing its ability to detect relevant cases. *Specificity* (True Negative Rate) evaluates the model's ability to correctly classify negative samples, complementing sensitivity. Together, these measures provide a balanced view of model performance for both classes. To combine precision and recall into a single summary metric, we computed F_1 , which represents their harmonic mean and provides a robust indicator of predictive performance, particularly under class imbalance. In addition, the *Negative Predictive Value* (NPV) was calculated to quantify the proportion of correctly predicted negative samples among all negative predictions, complementing PPV when evaluating models in datasets with skewed class distributions. Finally, we evaluated the *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)*, which measures the overall ability of the model to discriminate between positive and negative classes across varying classification thresholds. AUC-ROC is particularly relevant for comparing models independently of specific decision thresholds and provides an aggregated measure of model discrimination.

Together, these metrics provide a comprehensive framework for evaluating predictive accuracy, sensitivity, specificity, and discriminative power across diverse biomarker panels, feature selectors, and predictive models.

Performance Stability. In addition to standard evaluation metrics, we quantified *stability* as the variability of model performance across experimental replicates. Specifically, stability was defined as the width of the performance window, i.e., the difference between the maximum and minimum observed F1-scores (or AUCs) for a given model–selector combination. A narrow window indicates consistent behavior across experimental conditions, reflecting higher robustness and reproducibility.

3 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 fea-

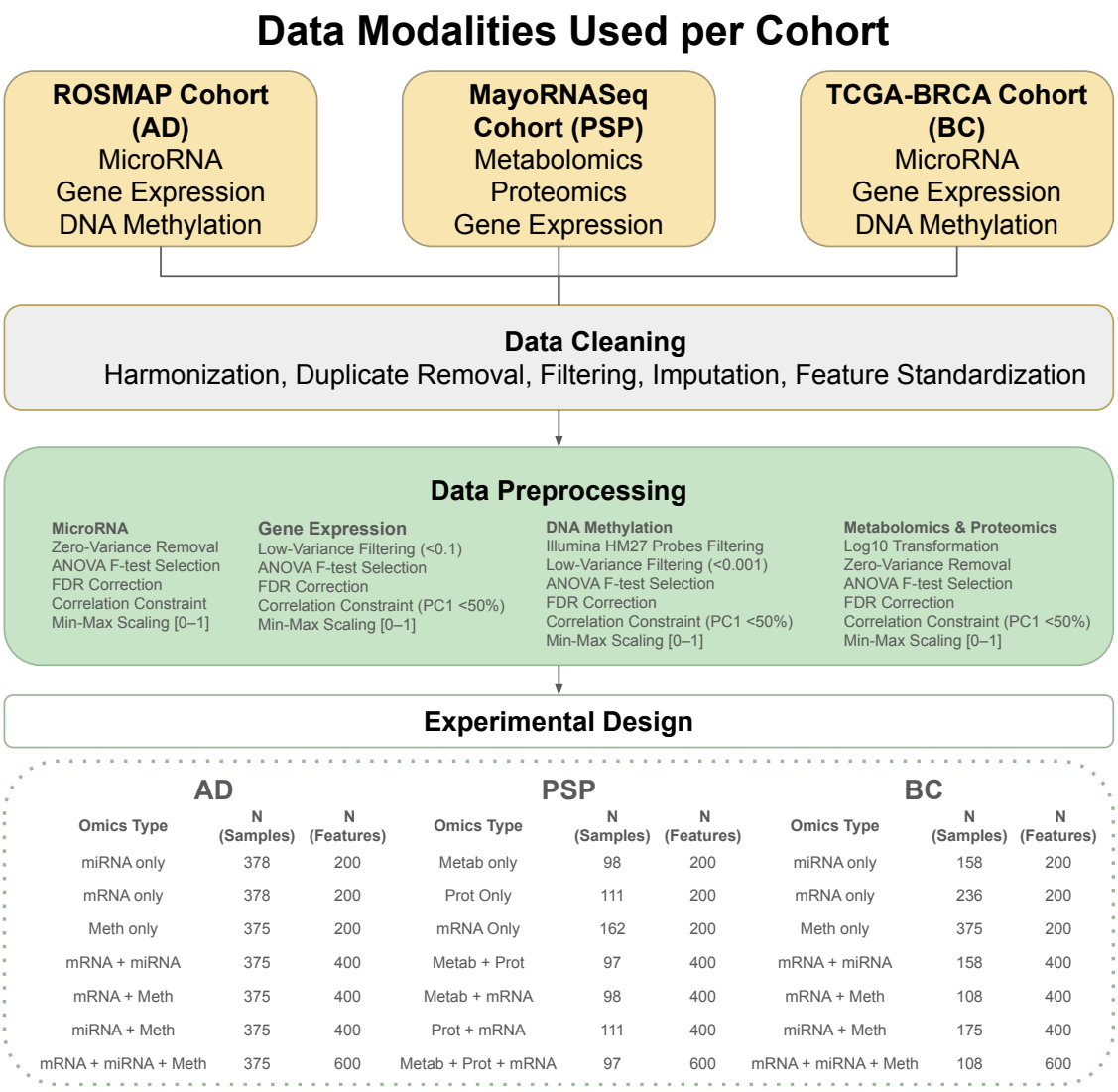


Figure 1: Experimental design and data modalities across cohorts. Overview of datasets and preprocessing steps used in the benchmarking study. Data from AD, PSP, and BC were harmonized through cleaning (duplicate removal, filtering, imputation, feature standardization) and modality-specific preprocessing (e.g., variance filtering, ANOVA F-tests, FDR correction, correlation constraints, min–max scaling, or log transformation). The experimental design combined 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 design of the benchmarking framework.

ture selection (Table 5). 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 $= 0.84 \pm 0.05$, $F_1 = 0.71 \pm 0.05$). 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 $= 0.61 \pm 0.08$, $F_1 = 0.56 \pm 0.13$).

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

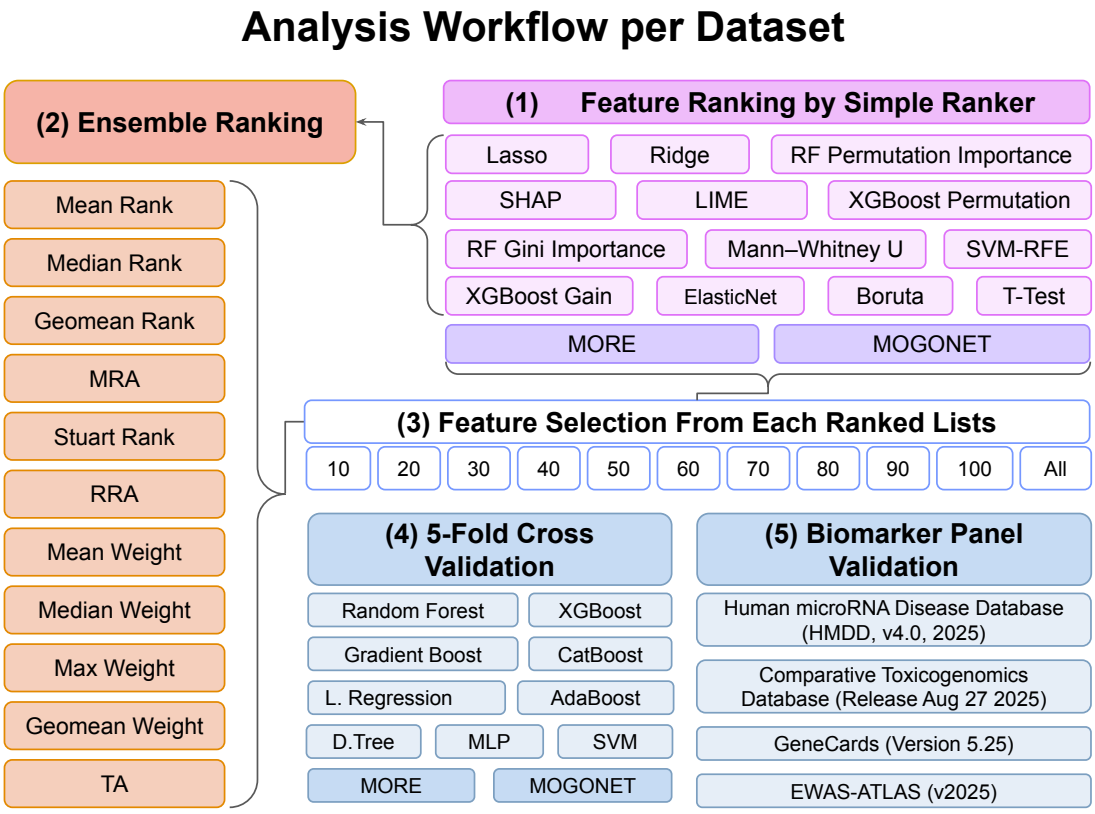


Figure 2: Analysis workflow per dataset. Schematic overview of the benchmarking pipeline applied to each cohort. (1) Features were ranked using a diverse set of single rankers, including statistical tests (t-Test, Mann–Whitney U), regularized models (LASSO, Ridge, Elastic Net), model-based importances (Random Forest, XGBoost, Boruta, SVM-RFE), explainability-based scores (SHAP, LIME), and deep learning–embedded rankers (MORE-Ranker, MOGONET-Ranker). (2) Ranked lists were aggregated using ensemble methods 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 via 5-fold cross-validation using a variety of traditional ML-based classifiers (D.Tree, L.Regression, Random Forest, Gradient Boost, XGBoost, CatBoost, AdaBoost, SVM, MLP) and deep learning frameworks (MORE, MOGONET). (5) Finally, selected panels were validated against external databases, including HMDD, CTD, GeneCards, and EWAS-ATLAS, to assess biological relevance.

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 $\pm$ 0.00 and F_1 0.96 $\pm$ 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 (Figure 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

Table 5: Baseline classification results across cohorts without feature selection. Performance is reported for 11 models in triple-omics.

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

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 (Figure 3b–c; Figure 4). Figures 3b–c summarize the median of average cross-validation F_1-scores across selector–classifier–panel size combinations, while Figure 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.

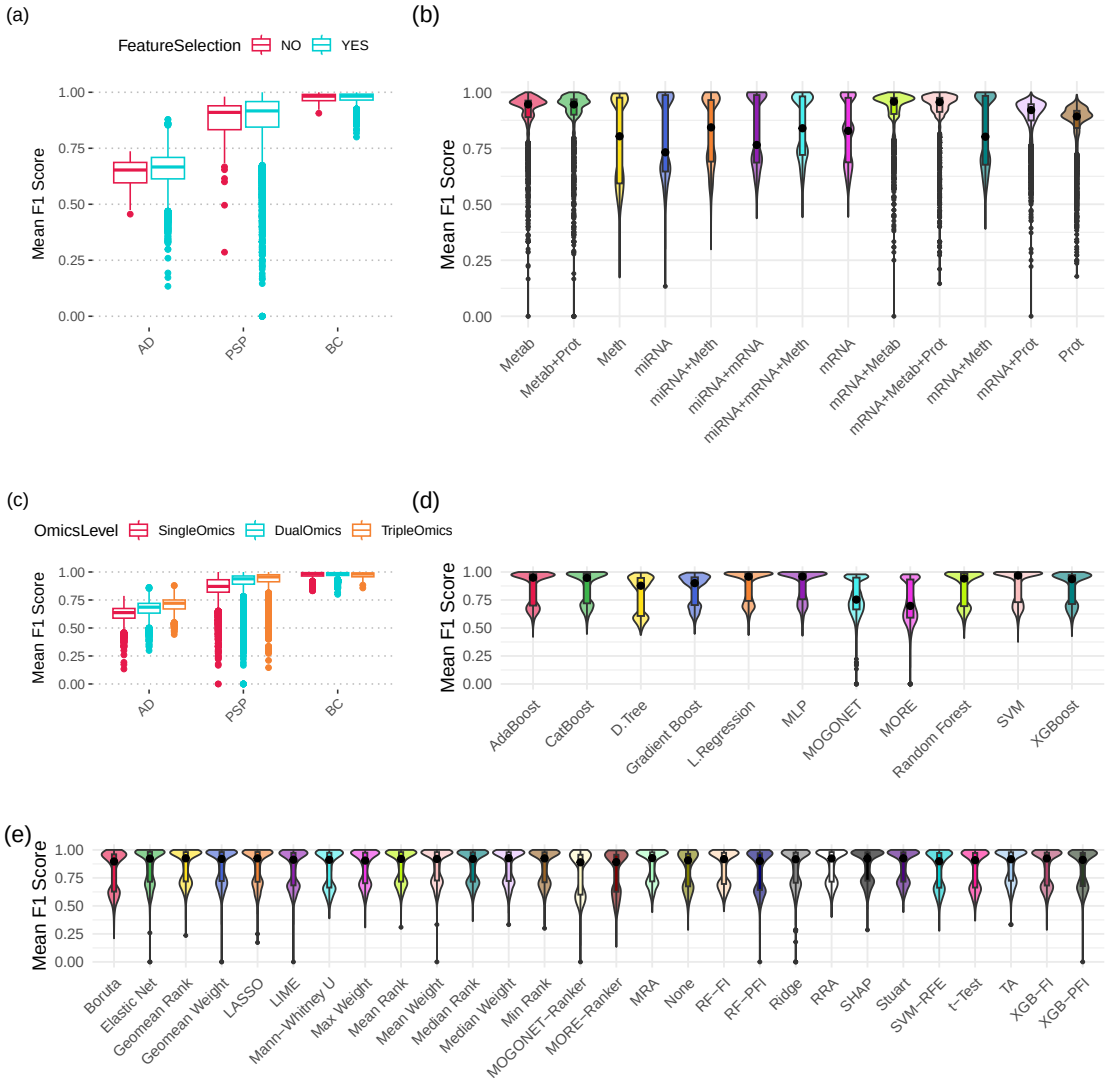


Figure 3: Benchmarking model performance across datasets, omics levels, models, and feature selection strategies. (a) Comparison of mean F1-scores with and without feature selection across AD, PSP, and BC, showing consistent gains with feature selection. (b) Performance by data modality and integration strategy, where triple-omics combinations frequently outperform other integrations. (c) Comparison of single-omics, dual-omics, and triple-omics integration across cohorts, highlighting the balance between information gain and noise. (d) Performance of individual machine learning models, where traditional linear and tree-based methods remain competitive with deep learning approaches such as MORE and MOGONET. (e) Impact of different feature selection rankers, showing that ensemble aggregation strategies yield more stable and accurate results compared to single rankers. Collectively, these panels demonstrate that while feature selection and judicious integration enhance performance, added model complexity does not always translate to better outcomes.

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 (Figures 3d, 5b, 6). In the median F_1 comparison (Figure 3d), L.Regression, SVM, ensemble tree models (AdaBoost, CatBoost), and MLP consistently occupied the top tier across cohorts, whereas D.Tree, Gradient Boost, MORE, and MOGONET generally ranked lower with reduced

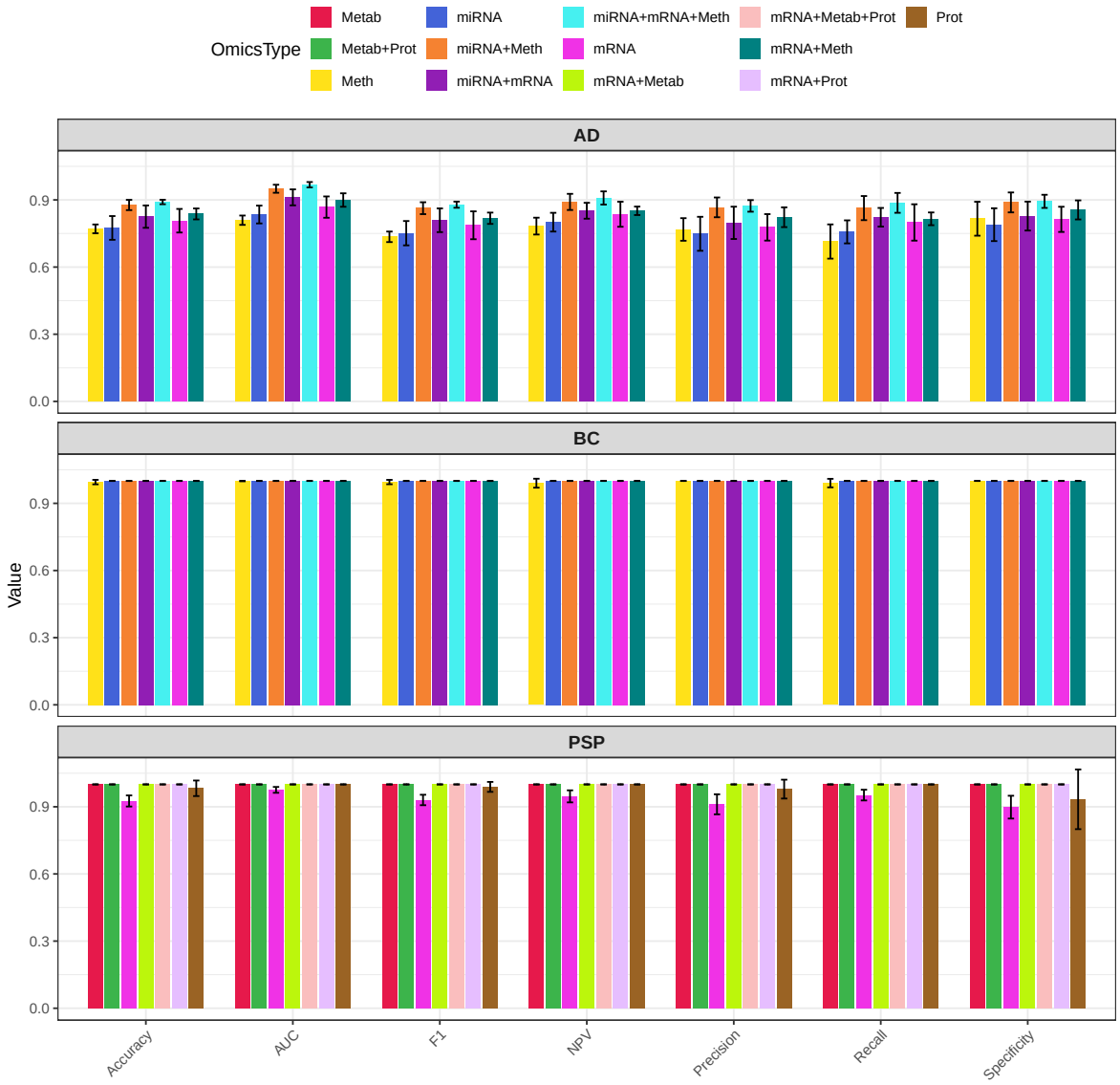


Figure 4: Model performance across omics types and integration strategies. Performance metrics (AUC, accuracy, F1, NPV, precision, recall, and specificity) are shown for the best-performing combinations of feature selector, classifier, and panel size within each omics configuration, evaluated across AD, PSP, and BC. Bars correspond to specific omics inputs (e.g., mRNA, miRNA, methylation, proteomics, metabolomics, and their pairwise or three-way integrations). Results show that triple-omics integration consistently achieved the highest overall performance, surpassing both single- and dual-omics settings. Dual-omics configurations ranked second, offering a strong balance between complementary information gain and model stability. These results emphasize that integrating multiple omics layers enhances predictive robustness, with triple-omics providing the most comprehensive view of underlying biological variation.

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 Figure 5b. Points closest to the top-right corner correspond to classifiers 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 Figure 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, 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 above 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 around 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 trade-offs, 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.

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. 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, LASSO, Mann–Whitney U, random forest importance (RF-FI), and XGBoost importance (XGB-FI), performed comparably well, highlighting their continued competitiveness. By contrast, LIME and the more recent multiview-specific deep learning rankers (MOGONET-Ranker, 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 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, Boruta) place the triad higher with tighter dispersion, whereas SHAP, LIME, SVM-RFE, Ridge, Min Rank, Max Weight, MRA show mid-level placement and wider spread. The multiview rankers (MORE-Ranker, MOGONET-Ranker) and `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.

Figures 7a–b illustrate how feature selectors interact with classifiers and feature panel sizes. The strongest selectors—including traditional statistical and model-based methods (t-Test, LASSO,

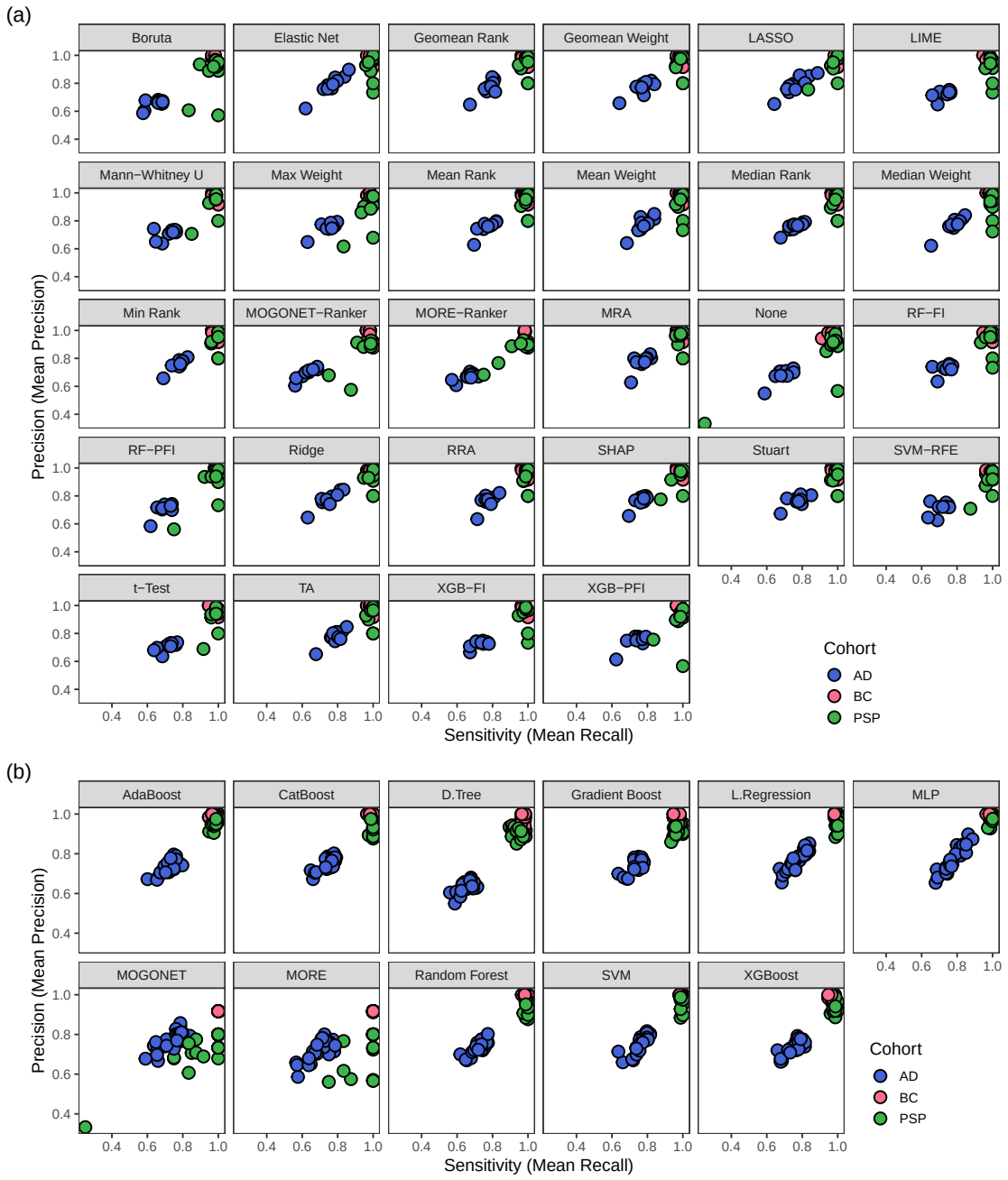


Figure 5: Comparative performance of feature selection strategies and classifiers across cohorts. (a) Precision–recall plots summarizing the trade-off between sensitivity (mean recall) and mean precision for different feature selection rankers across AD, PSP, and BC, showing that ensemble rankers generally achieve better balance than single methods. (b) Similar precision–recall analysis across classifiers, highlighting that tree-based ensembles (AdaBoost, XGBoost, CatBoost) and linear models (L.Regression, SVM, MLP) maintain competitive performance compared to deep learning models (MORE, MOGONET).

Elastic Net, Ridge, Boruta, RF-FI, XGB-FI) and ensemble aggregators (Mean Rank/Weight, Median Rank/Weight, Geomean Rank/Weight, Stuart, RRA)—consistently supported high F_1-scores across models and panel sizes. These selectors enabled linear (L.Regression), kernel (SVM), MLP, and ensemble tree classifiers (XGBoost, Random Forest, CatBoost) to reach stable, near-

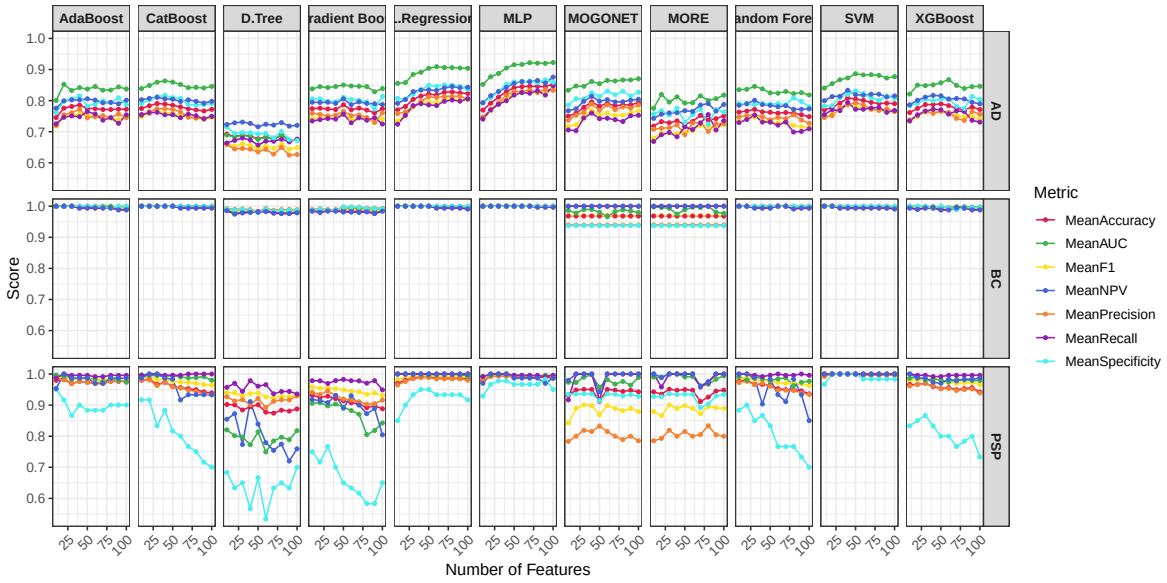


Figure 6: Model performance across biomarker panel sizes in different cohorts. Performance metrics (AUC, accuracy, F1, NPV, precision, recall, and specificity) as a function of the number of selected features for AD, PSP, and BC. Curves represent different classifiers, including traditional linear models (L.Regression, SVC), tree-based ensembles (Random Forest, XGBoost, CatBoost, AdaBoost, Gradient Boost), multilayer perceptron (MLP), and deep learning–based methods (MORE, MOGONET). Across cohorts and models, performance improves rapidly with small panels (10–30 features), reaches near-optimal values by intermediate panel sizes (50–75), and shows limited gains beyond 100 features. These results indicate that compact biomarker panels can capture most of the predictive signal, making them both efficient and clinically feasible.

optimal performance across all cohorts. In contrast, weaker or less consistent selectors—such as MORE-Ranker, MOGONET-Ranker, LIME, SVM-RFE, and TA—produced lower F_1-scores and greater variability, particularly in AD. Even strong classifiers like SVM and Gradient Boost could not compensate for the poor feature sets generated by these methods. Performance gains generally plateaued around 50–60 features across most models and cohorts, indicating that this range balances information capture with redundancy control. Ensemble and well-regularized selectors maintained robustness across small and large panels, while weaker methods exhibited sharper declines and inconsistent stability. Collectively, these results emphasize that selector quality determines classifier potential, and that optimal panel sizes should align with selector reliability to ensure reproducible biomarker discovery.

Figure 8 compares selectors across cohorts, stratified by ensemble versus single methods. In BC, nearly all selectors—ensemble and single alike—achieved ceiling-level performance (F_1 1.0), reflecting the strong discriminative signal in this dataset. In PSP, ensemble selectors generally concentrated at the top with consistently high mean F_1 values and tighter spreads, though several traditional methods (t-Test, LASSO, RF-FI, XGB-FI) performed equally well, indicating redundancy of complex aggregation in this cohort. In AD, however, differences were more pronounced: ensembles (e.g., Geomean Rank/Weight, Median Rank/Weight, Stuart, RRA) clearly outperformed most single methods, which displayed wider variability and lower means. Only a few single selectors (SHAP, Ridge, LASSO, Elastic Net) approached ensemble-level stability in this challenging dataset. Together, these patterns show that while ensembles provide cross-cohort robustness, traditional single rankers remain competitive in high-signal settings, whereas noisy, heterogeneous data like AD reveal the distinct advantage of consensus aggregation.

Together, these results demonstrate that ensemble feature selection methods consistently outperformed most individual rankers across cohorts, panel sizes, and evaluation metrics (Figures 3e, 5a, 7, 8). At the same time, traditional methods such as t-Test, LASSO, and RF-FI remained among the stronger single rankers, underscoring their continued relevance in biomarker discovery pipelines.

The similarity and divergence of feature selec-

tors across cohorts can be understood by combining the PCA clustering (Figures 9a–c) and UpSet overlap plots (Figures 9d–f) with earlier benchmarking results. Across all three cohorts, ensemble rankers formed distinct clusters away from single rankers in PCA space, indicating systematic differences in their selection behavior. In AD and PSP, ensembles (Mean, Median, Geomean, RRA, Stuart) occupied cohesive regions, while single rankers such as SVM-RFE, SHAP, and Ridge grouped more closely together, suggesting internally consistent selection patterns within each class of methods. Deep learning–embedded selectors (MORE-Ranker and MOGONET-Ranker) consistently fell outside both the ensemble and single ranker clusters, highlighting their tendency to produce idiosyncratic feature sets that diverge from consensus patterns.

The UpSet plots further emphasize these relationships by quantifying the number of features uniquely selected by each method. In AD, intersections were dominated by traditional selectors such as XGB-FI, RF-PFI, and t-Test, which exhibited large counts of unique features (up to 19 features selected exclusively by a single method), indicating limited overlap with other selectors. In contrast, ensemble rankers converged on smaller, more coherent subsets, with very few unique features (often fewer than four across cohorts), reflecting stronger agreement with other methods. Similar patterns were observed in PSP, where traditional rankers displayed unique selections of up to 23–24 features, while ensembles showed far fewer distinct features, emphasizing their integrative behavior. In BC, nearly all selectors showed relatively low uniqueness due to the dataset's strong signal and separability, yet ensemble selectors still maintained more balanced intersections compared to the scattered and idiosyncratic profiles of weaker methods (e.g., LIME, TA, SVM-RFE).

Together, these results confirm that while traditional single rankers (t-Test, LASSO, RF-FI, XGB-FI) often behave similarly within their own class but produce more method-specific feature sets, ensemble aggregation enhances cross-method consensus by identifying features shared across diverse selection strategies. In contrast, specialized selectors such as LIME and deep learning–embedded rankers generate more divergent outputs, reinforcing the conclusion that increased methodological complexity does not necessarily yield more convergent or interpretable selection patterns. This convergence of evidence from PCA and UpSet benchmarking highlights that ensembles and top-performing traditional methods are not only competitive but also systematically more similar in their selections, whereas deep learning–based selectors consistently diverge from the rest.

Comprehensive Analysis of Interactions across Feature Selectors, Panel Sizes, Models, and Omics Levels

To systematically evaluate how feature selectors interact with models, panel sizes, and omics integration strategies, we generated heatmaps summarizing maximum F1-scores across selector–model–omics and selector–panel size–omics combinations. For each cohort (AD, PSP, BC), biomarker panels of increasing size ($K = 10$–100) were constructed from ranked feature lists produced by 27 different selection strategies, spanning statistical tests, regularized regressors, tree-based importance scores, wrapper methods, explainability methods, and ensemble aggregation techniques. Each panel was evaluated using 11 classifiers covering linear, tree-based ensemble, kernel, and deep learning paradigms, alongside the multi-omics-specific models MORE and MOGONET. Performance was averaged across cohorts, and the maximum F1-score was retained either per selector–panel size–omics combination or per selector–omics–model combination. The resulting heatmaps (Figures 11 and 10) provide a trivariate view of how selectors align with model families and panel sizes under different levels of omics integration.

Across models and omics levels (Figure 10), **triple-omics achieved the highest F_1-scores overall**, followed closely by dual-omics, while single-omics remained the weakest configuration. The best-performing cells were concentrated where ensemble selectors (Mean, Median, or Geomean ranks; RRA; Stuart) paired with traditional classifiers such as L.Regression, SVM, MLP, and CatBoost. In contrast, embedded selectors (MORE-Ranker, MOGONET-Ranker) formed a low-F_1 band across models, indicating poor alignment with both optimal model families and informative omics combinations. Notably, not only ensemble selectors but also single methods such as LASSO, Ridge, and Elastic Net achieved competitive or even higher performances in several settings, though their outcomes were less consistent across omics levels. Overall, the strongest three-way association was Ensemble or single selector; $\leftrightarrow$; traditional model; $\leftrightarrow$; Multi-omics (Triple > Dual > Single), and deviations from this synergy resulted in marked performance drops. These patterns align with the

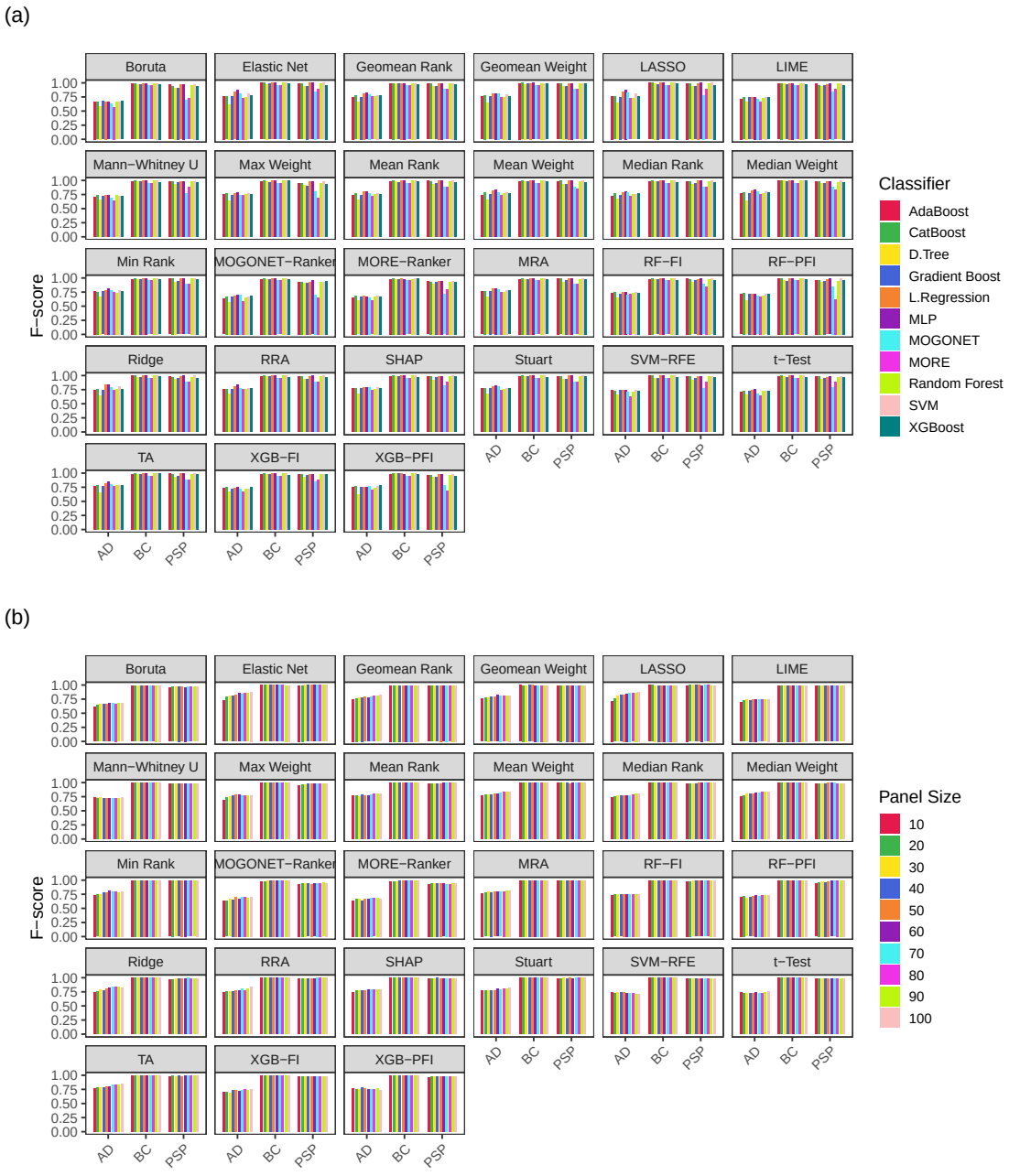


Figure 7: Model performance across rankers, classifiers, and biomarker panel sizes in three cohorts. (a) Mean F1-scores of different classifiers (D.Tree, L.Regression, SVM, Random Forest, Gradient Boost, AdaBoost, CatBoost, XGBoost, MLP, MORE, MOGONET) when paired with feature subsets generated by multiple rankers across AD, PSP, and BC. Ensemble rankers consistently improved classifier performance compared to single rankers, with tree-based ensembles and linear models performing competitively with deep learning frameworks. (b) Effect of biomarker panel size on F1-score across rankers and cohorts. Performance increased rapidly with small panels (10–30 features) and plateaued at larger panel sizes, demonstrating that compact feature sets retain most predictive signal. Together, these panels highlight the benefit of ensemble feature selection and compact biomarker panels across diverse models and cohorts.

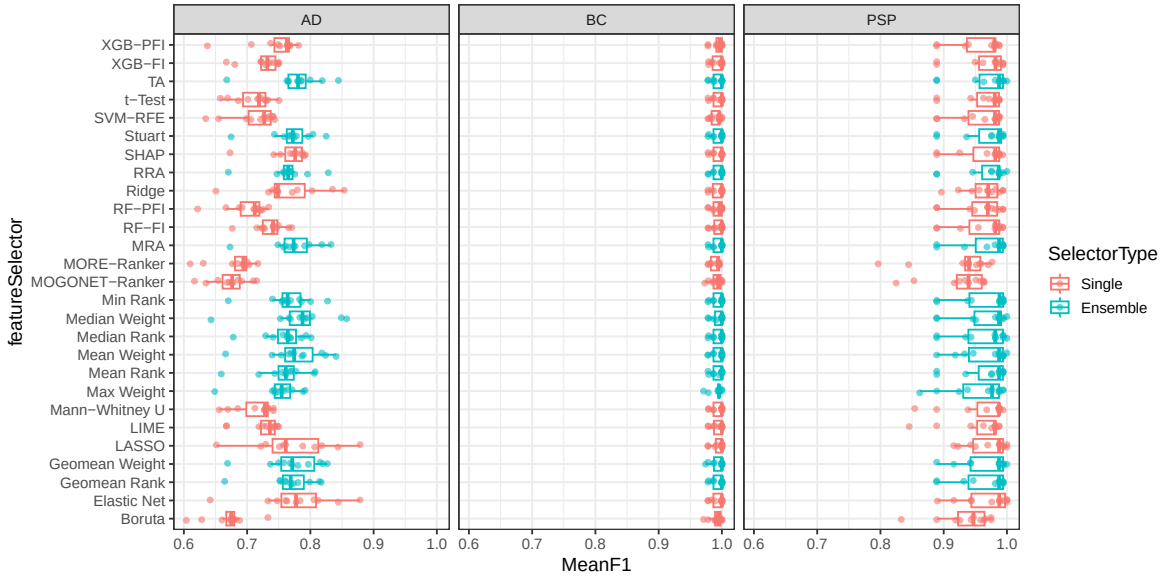


Figure 8: Comparison of single and ensemble feature selection methods across cohorts. Mean F1-scores for individual (single) rankers, ensemble aggregation methods, and embedded rankers from deep learning frameworks (MORE, MOGONET) across BC, PSP, and AD datasets. Bars represent average performance across classifiers and panel sizes. Ensemble methods such as Mean Weight, Geomean Weight, TA, RRA, and Stuart consistently achieved strong and stable performance across cohorts. Regularized single methods like Elastic Net, LASSO, and Ridge also performed competitively, particularly in BC, where the distinction between ensemble and single rankers was most pronounced. In contrast, deep learning–embedded rankers (MORE, MOGONET) did not consistently outperform either group, underscoring the reliability of aggregation-based and regularized feature selection strategies.

broader benchmarking trends, where ensemble- and regression-based selectors demonstrated greater stability and interpretability compared to deep-learning rankers.

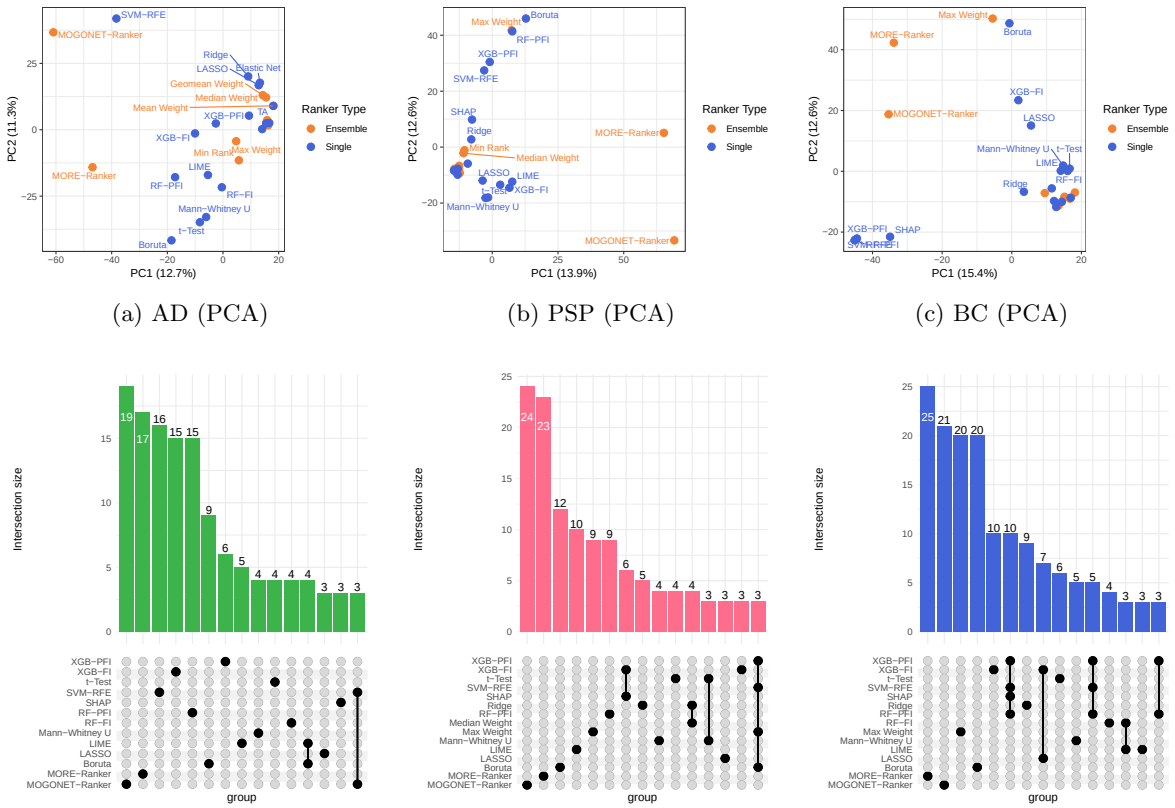
F1 values increased sharply from $K \approx 10$ to $K \approx 30$–60 and then plateaued (Figure 11), showing a positive correlation with panel size that saturated, particularly in dual-omics, which displayed the densest block of near-ceiling F1-scores across selectors. Single-omics exhibited a stronger dependence on K, with low scores at small K and moderate gains at mid K, while triple-omics showed broader variance across K, reflecting sensitivity to redundancy and noise. Ensemble selectors were least sensitive to K, maintaining high F1-scores across the full range, whereas weaker or embedded selectors showed earlier plateaus and reduced slopes. The favorable tri-variate regime emerged as: *ensemble selector* $\leftrightarrow$ *dual-omics* $\leftrightarrow$ *mid-sized panels* ($K \approx 30$–60). Expanding K beyond 60 provided little additional gain and in some cases reduced stability, consistent with the broader analysis that compact panels capture most relevant signal and that dual-omics balances information and noise.

The strongest tri-variate correlations consistently converged on the same corner: ensemble aggregation combined with traditional models, evaluated in triple-omics settings with mid-sized panels. This is where similarity, stability, and peak F1 coincided, whereas embedded selectors, single-omics, and very small or very large K systematically underperformed or exhibited unstable behavior.

Biological validation of discovered biomarkers

Here's a concise rewrite that covers all four figures without going deep into specifics:

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, EWAS-ATLAS)—for $k \in \{10, 30, 50, 100\}$ (Figures 13–16). 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, XGB-FI, SVM-RFE) often close



(a) AD (PCA) (b) PSP (PCA) (c) BC (PCA)

(d) AD (UpSet) (e) PSP (UpSet) (f) BC (UpSet)

Figure 9: **Selector Similarity and Divergence.** We evaluated the relationships between feature selectors using PCA and UpSet analyses, where *similarity* reflects the degree of feature overlap or clustering proximity among rankers, and *divergence* indicates systematic differences in their selection behavior. Panels Panels (d–f) show UpSet plots of feature intersections across different rankers for AD, PSP, and BC, respectively, computed using the top 30 features selected by each method. While individual rankers selected highly variable feature sets, ensemble methods consistently identified overlapping subsets, reflecting greater methodological coherence within each cohort. This is evident from the fact that most ensemble methods do not appear on the UpSet plot's y-axis among top selectors with unique feature sets, indicating that their selections largely overlap with one another rather than being distinct. Panels (a–c) show principal component analysis (PCA) of feature selection rankers across cohorts. Each profile encodes the union of top-k features (k = 10–100) as binary presence/absence vectors. (a) AD: PC1 = 12.7%, PC2 = 11.3%. (b) PSP: PC1 = 13.9%, PC2 = 12.6%. (c) BC: PC1 = 15.4%, PC2 = 12.6%. Across all cohorts, ensemble rankers cluster separately from single rankers, while deep learning–embedded rankers (MORE, MOGONET) occupy distinct positions, highlighting systematic divergence in selection behavior.

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 (green) typically attained the highest absolute TPs (including CpG validations), AD (blue) benefited most in miRNA-based validation (HMDD), and PSP (orange) 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.

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,

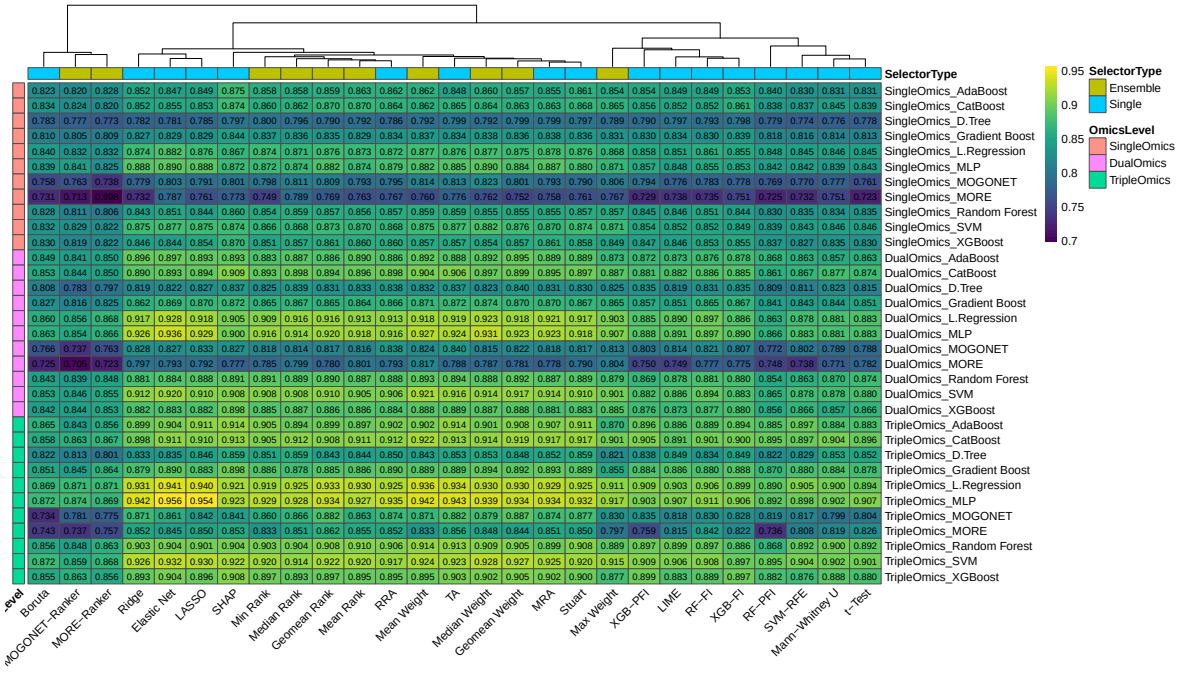


Figure 10: **Heatmap of maximum F_1-scores for selector–omics level–model combinations.** Maximum F_1-scores are reported for each selector–omics level pairing after averaging performance across cohorts and taking the best score across panel sizes. Rows correspond to feature selection strategies (single rankers, ensemble rankers, and embedded rankers from MORE and MOGONET), while columns represent classifiers spanning linear, tree-based ensemble, and deep learning models. Each cell reports the maximum F_1-score achieved for a given selector–omics configuration across models. Triple-omics integration consistently delivers the strongest and most stable performance, surpassing both single- and dual-omics settings across most selectors and classifiers. Ensemble rankers combined with traditional classifiers (e.g., L.Regression, MLP, SVM) achieve results comparable to or even exceeding those of deep learning models, underscoring their efficiency and interpretability. In addition, single regularized methods such as LASSO, Ridge, and Elastic Net also achieved competitive or higher F_1-scores in several settings, though with slightly lower consistency across omics levels. Dual-omics configurations remained competitive—particularly with L.Regression, SVM, and MLP—yet triple-omics consistently outperformed when the model architecture and constraints were held constant. These findings highlight that the synergistic information captured through triple-omics integration markedly enhances predictive robustness, while both ensemble and regularized feature selection methods remain key drivers of high model performance.

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 [57], our findings should be interpreted 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 CSF), the current analyses focus on computational prioritization within affected tissue contexts. Recent large-scale efforts, such as the blood-based molecular atlas of Alzheimer's disease [58], 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 6, 7, 8; Figures 17, 19, and 18) summarize the consistency and stability of identified features. To focus on repro-

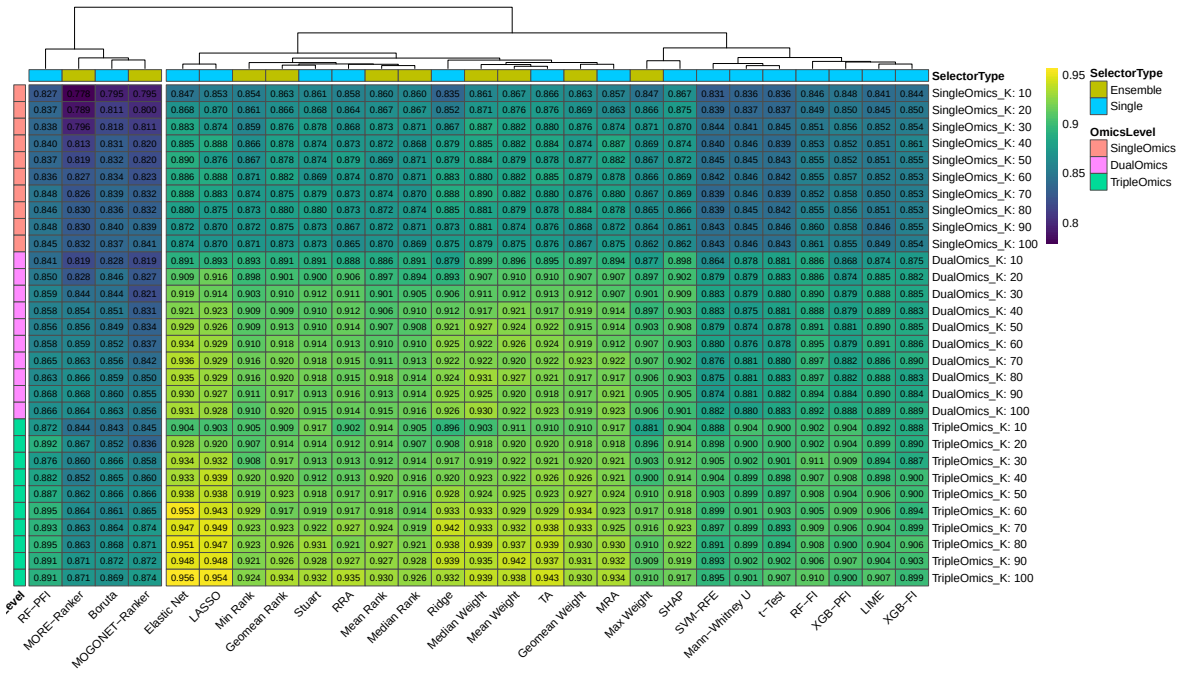


Figure 11: Heatmap of Maximum F1-Scores for Selector–Panel Size Combinations Across Omics Integration Levels. Maximum F1-scores are reported for selector–panel size–omics level combinations after averaging performance across cohorts. Rows correspond to different omics integration settings (single-omics, dual-omics, and triple-omics with panel sizes $K = 10$–100), while columns represent feature selection strategies (single rankers, ensemble aggregators, and embedded rankers from MORE and MOGONET). Each cell displays the best F1-score achieved at the model level for a given selector–omics configuration. Triple-omics integration consistently yields the strongest and most stable performance, particularly when combined with ensemble-based rank aggregation. Dual-omics configurations achieve moderate improvements, while single-omics setups remain more variable—highlighting the progressive benefit of multi-layer data fusion for biomarker discovery.

ducible 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, SPACA6) were consistently identified, some with previous links to AD pathology, while others remain poorly characterized. Importantly, CpG methylation sites (cg06690548, cg12981137, cg19832721, cg16003238, 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

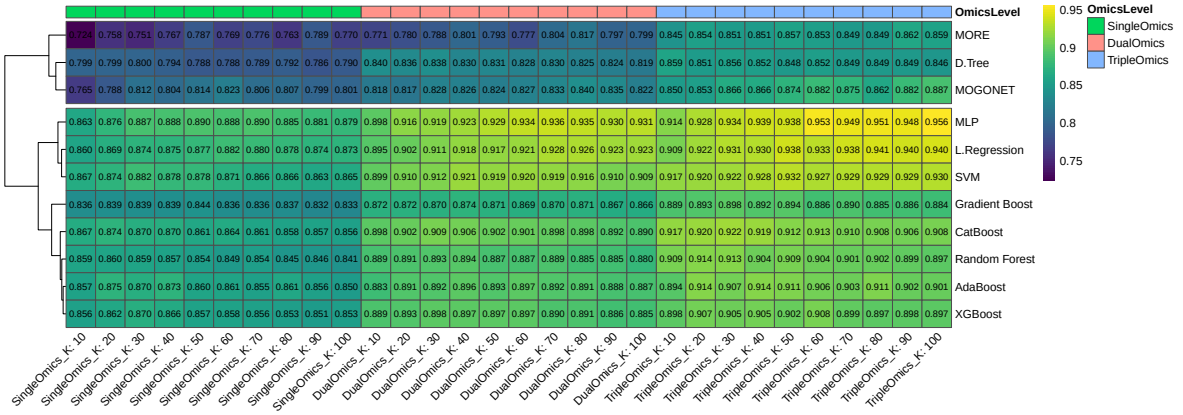


Figure 12: Heatmap of Maximum F1-Scores for Model–Omics Level–Panel Size Combinations. Maximum F1-scores are reported for each omics level–panel size–classifier combination after averaging performance across cohorts and selecting the maximum across feature selectors. Rows represent machine learning classifiers spanning linear (L.Regression), kernel-based (SVM), ensemble tree-based (Random Forest, XGBoost, CatBoost, AdaBoost, Gradient Boost), deep learning models (MLP, MORE, MOGONET), and a D.Tree baseline. Columns correspond to omics integration strategies (single-omics, dual-omics, and triple-omics at panel sizes $K = 10$–100). Each cell displays the best F1-score achieved for a given panel size–model–omics configuration. Triple-omics integration consistently yielded the strongest overall performance, with traditional classifiers (e.g., L.Regression, Random Forest, XGBoost) achieving robust results across panel sizes. Single-omics exhibited greater variability, whereas dual-omics occasionally reached peak scores but less consistently than triple-omics.

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 progressive supranuclear palsy disease pathology.

In BC, both miRNAs and mRNAs were robustly recovered, with strong overlap with established oncogenic drivers. Validated 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 breast cancer (PPP1R12B, LRRC3B, TMEM220, PPAPD1C1A, HSD17B6, ADAMTS5, SDPR, SPRY2) were reproducibly selected 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.

Table 6: 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

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.

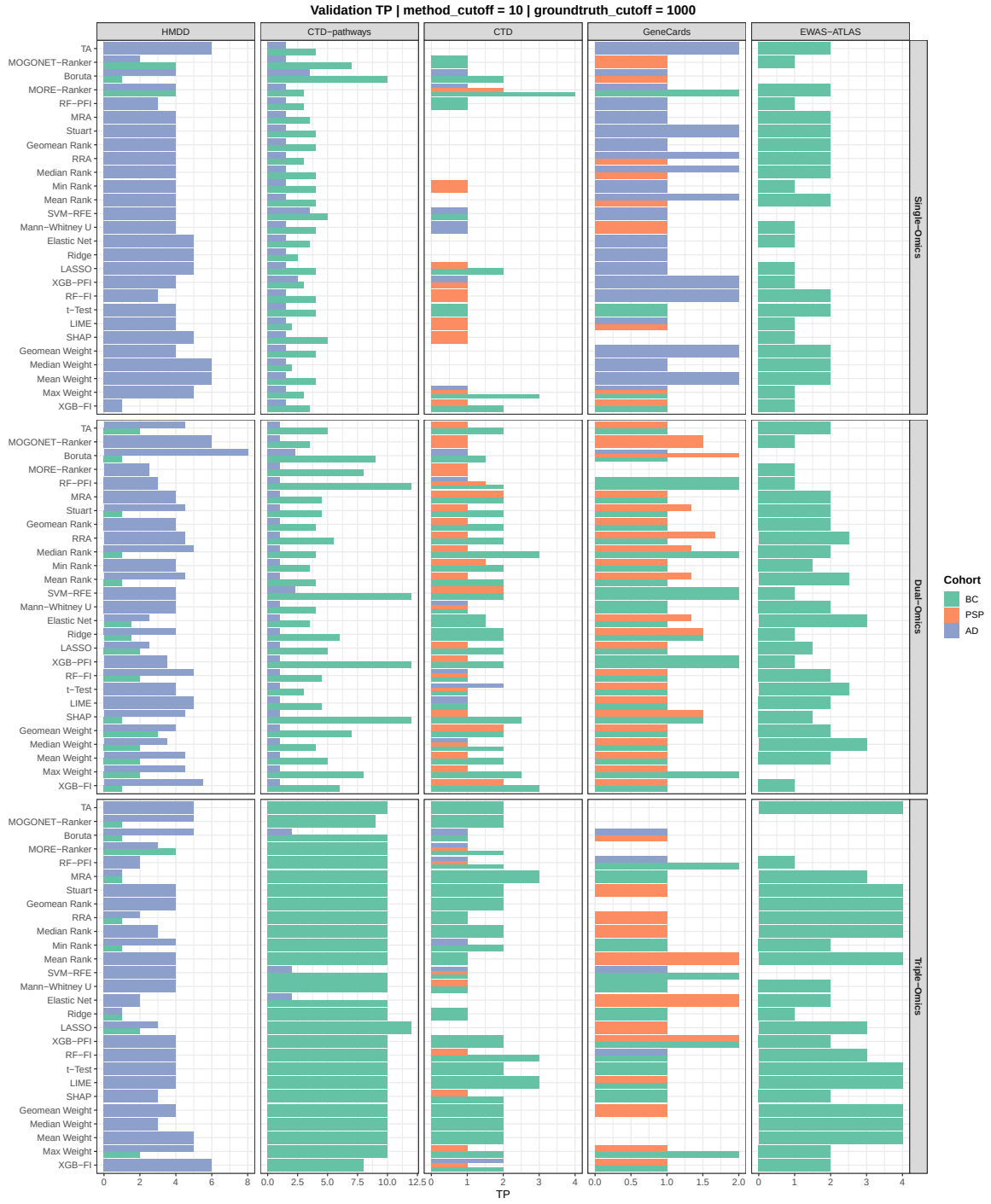


Figure 13: External validation (Top-10 cutoff). Bars show overlaps with curated disease-associated features in HMDD (miRNAs), CTD-pathways, CTD, GeneCards, and EWAS-ATLAS across single-, dual-, and triple-omics for AD, PSP, and BC. Integration increases validation, most clearly in CTD-pathways; ensembles are consistently strong.

4 Discussion

In this study, we systematically benchmarked 27 feature ranking methods, multiple classifiers, and diverse omics integration strategies across three heterogeneous cohorts: AD, PSP, and BC. By evaluating both predictive performance and biological validation, we provide a comprehensive assessment of how methodological choices influence the discovery and robustness of candidate biomarkers.

Our results demonstrate that model complexity

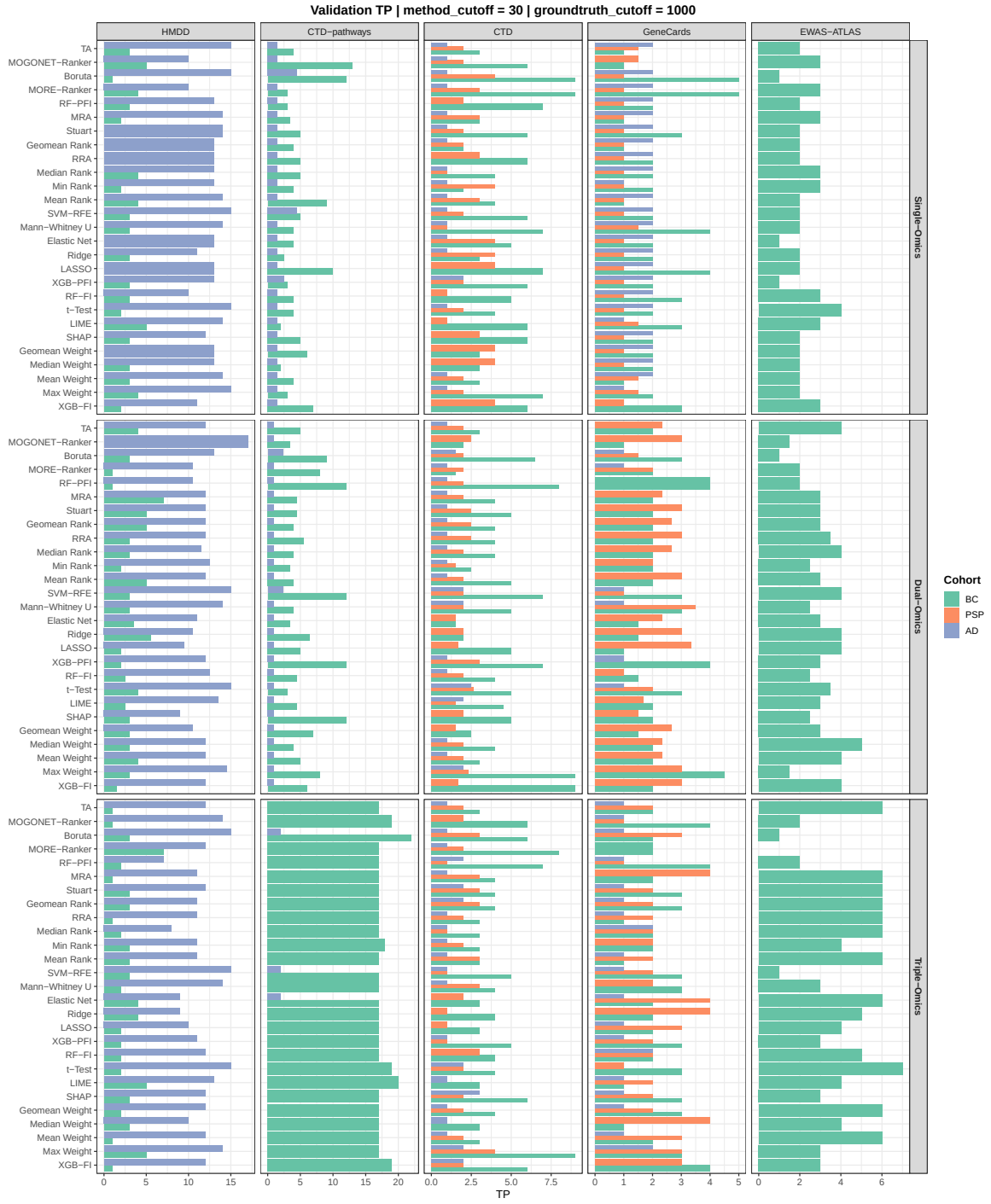


Figure 14: External validation (Top-30 cutoff). Dynamic range widens; CTD-pathways and EWAS-ATLAS show pronounced gains; qualitative ordering of methods persists with ensembles leading.

is not necessarily synonymous with improved predictive power. Traditional classifiers—including L.Regression, support vector machines, ensemble tree-based approaches (Random Forest, XGBoost, CatBoost), and multilayer perceptrons (MLPs)—consistently matched or outperformed multiview deep learning methods such as MORE and MOGONET (Figures 10). This was particularly striking in BC, where nearly all classifiers reached ceiling-level performance (Figure 6), indicating that high-quality single- or dual-omics signals can be effectively captured with relatively simple models. In contrast, more complex frameworks often exhibited lower medians and broader variability, underscoring the challenges of training deep models in high-dimensional, low-sample

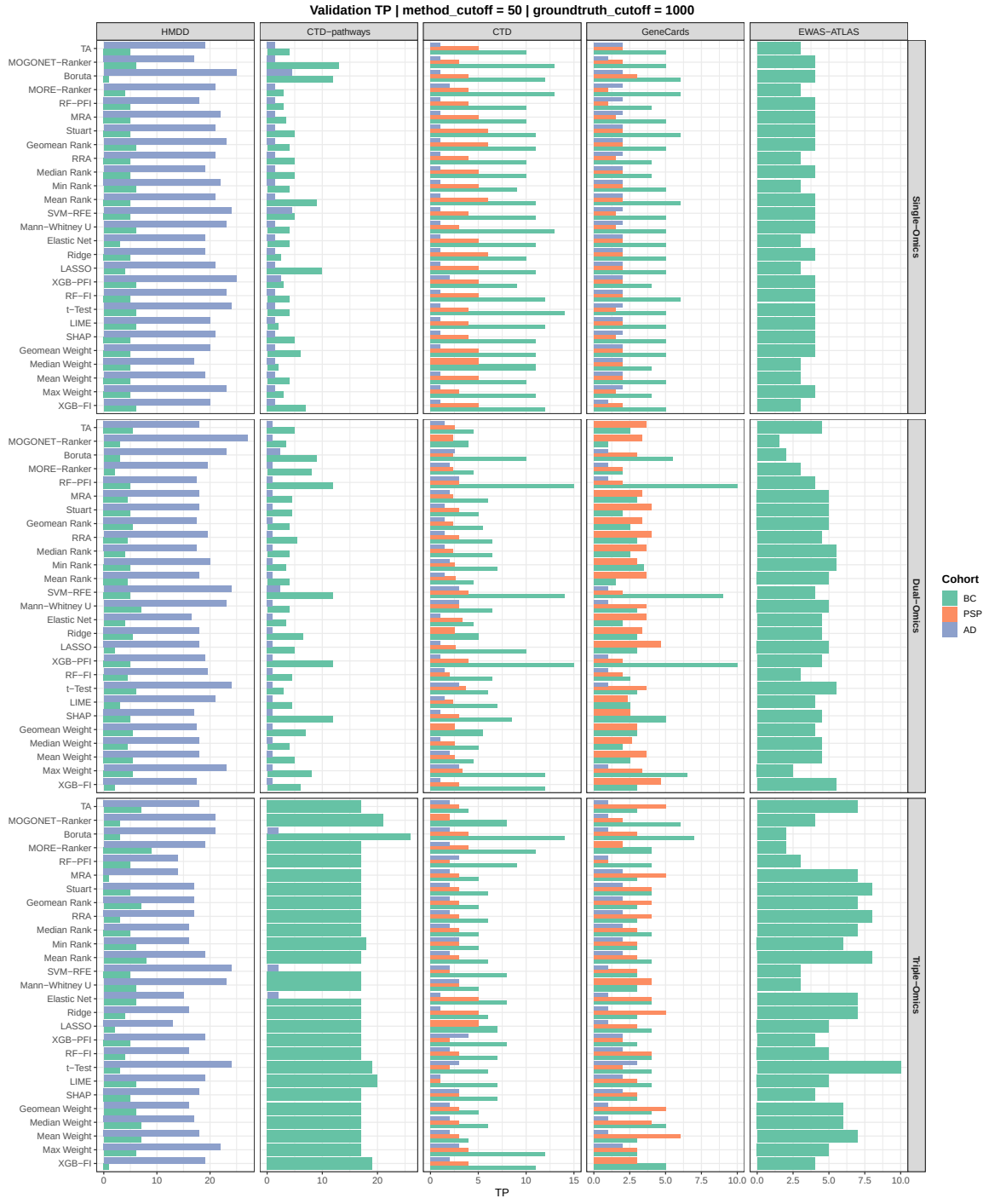


Figure 15: External validation (Top-50 cutoff). Broad uplift across cohorts; between-method gaps narrow slightly; dual/triple-omics remain strongest.

omics settings.

Feature selection emerged as a critical driver of performance and interpretability. Ensemble-based rank aggregation methods—including Geomean (Rank and Weight), Median and Mean Ranks, Median and Mean Weight, Stuart, TA, and RRA—consistently produced the most stable and accurate biomarker panels Figures 10–11). Importantly, several traditional selectors such as LASSO, Elastic Net, Ridge, and SHAP remained highly competitive, reinforcing their continued relevance in biomarker pipelines. By contrast, weaker selectors (e.g., Boruta, MOGONET-Ranker, MORE-Ranker) showed lower and less stable performance across cohorts and panel sizes. These findings highlight that robust ensemble

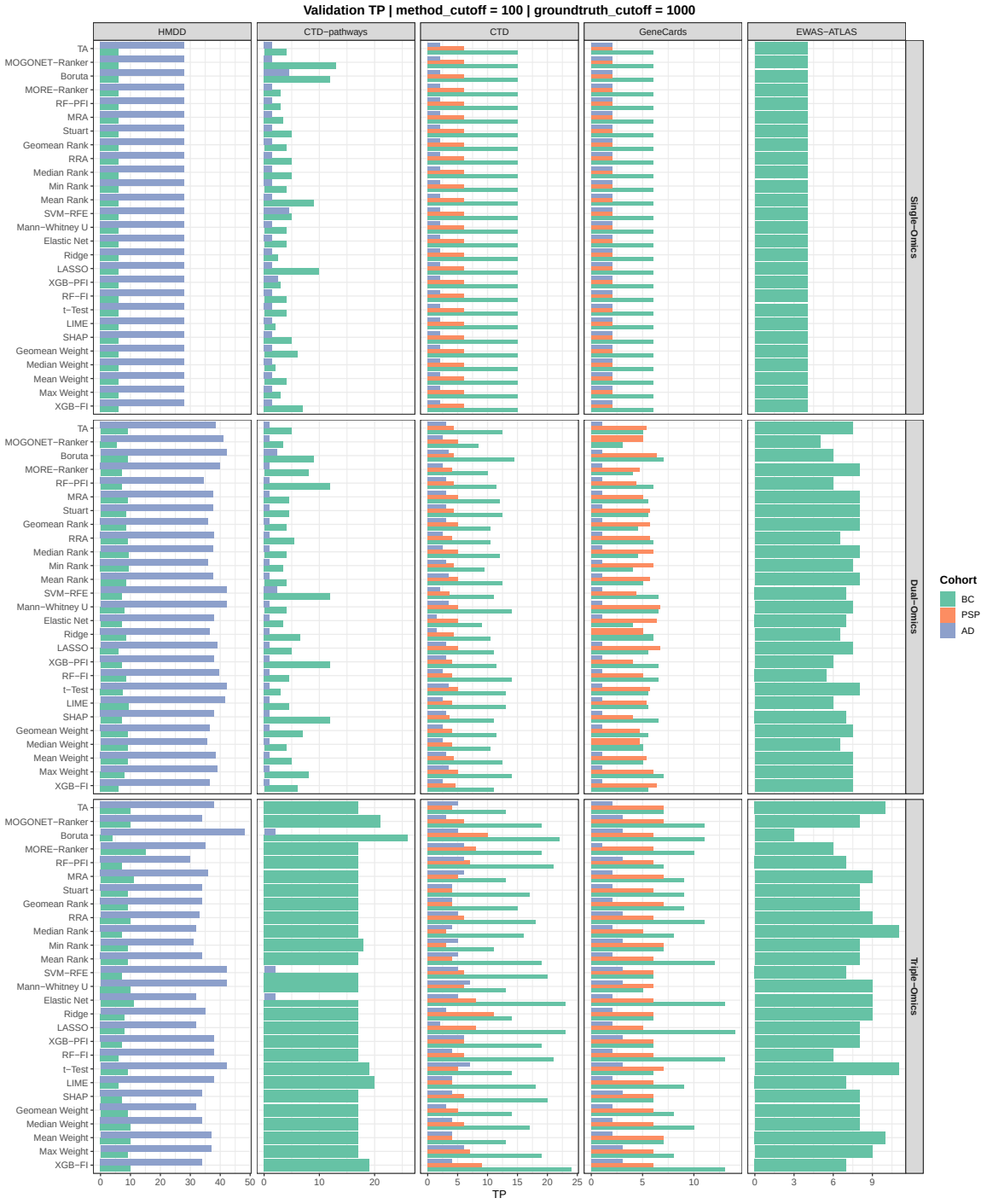


Figure 16: External validation (Top-100 cutoff). High absolute counts with partial saturation; integrative advantage and method hierarchy maintained.

ranking strategies not only mitigate the instability of single methods but also facilitate reproducibility across heterogeneous cohorts.

Omics integration strategies further shaped outcomes. Across median and maximum performance comparisons, dual-omics configurations frequently outperformed single-omics, providing the optimal balance of predictive accuracy and generalization. For example, combining miRNA with DNA methylation in AD and mRNA with proteomics in PSP improved discrimination relative to single-omics and approached the performance of corresponding triple-omics settings. Nonetheless, triple-omics retained the overall maximum performances across all feature selectors, panel sizes, and most models (Figures 10–11). These results indi-

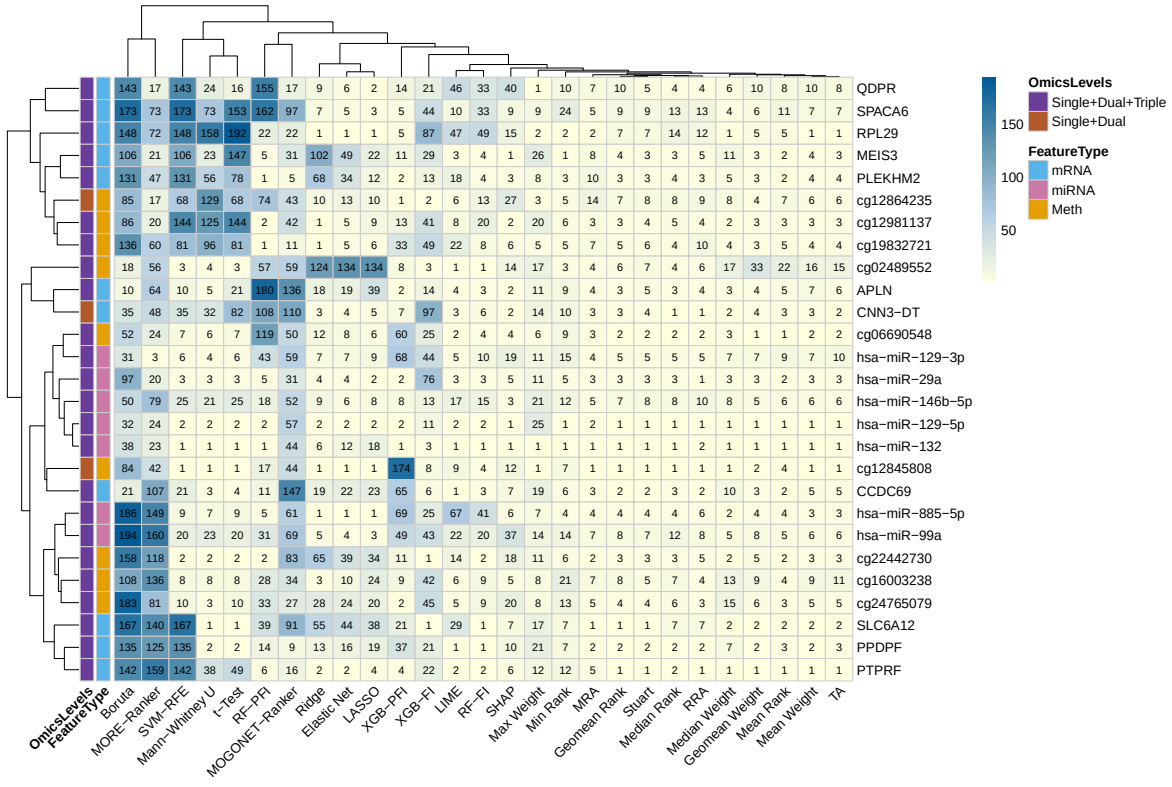


Figure 17: Rank heatmap for AD. Feature ranking stability across multiple selection methods in AD. Rows represent candidate microRNAs, genes, and CpG methylation sites, while columns denote rankers grouped into single and ensemble methods. Each cell displays the rank assigned by that ranker to that feature. Smaller numeric ranks indicate higher importance (top-ranked). Among the consistently top-ranked features, validated biomarkers such as hsa-miR-129-5p, hsa-miR-132, hsa-miR-146b-5p, ARRDC2, PLEKHM2, and PPDPF overlapped with known Alzheimer's disease associations. In contrast, novel candidates, including hsa-miR-885-5p, APLN, CCDC69, SPACA6, MEIS3, and CNN3-DT were robustly prioritized across rankers but remain underexplored in AD. Similarly, several CpG methylation sites (cg12981137, cg22442730, cg16003238, cg19832721, cg12864235) were repeatedly ranked highly despite limited prior evidence, suggesting underexplored epigenetic contributions. Together, the heatmap demonstrates the ability of ensemble ranking to recover known AD biomarkers while highlighting reproducible, potentially novel candidates across molecular layers.

Table 7: 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

cate that while multi-omics integration can yield synergistic gains, the benefits depend strongly on cohort-specific sample size balance and modality quality, echoing prior concerns that indiscriminate addition of modalities may introduce noise or reduce sample coverage, thereby hindering generalization. In contrast, BC was barely affected by the omics integration level, with nearly all configurations reaching ceiling-level performance.

Our results caution against over-engineering. Deep learning models such as MORE and MOGONET, while theoretically powerful, imposed heavy computational costs and often failed to outperform well-regularized traditional classifiers such as L.Regression, SVMs, or Random Forests. In many instances, ensembles of single rankers, including Mean/Median Weight/Rank, and RRA, provided greater stability than single deep-embedded rankers. Given the modest cohort sizes of typical omics studies, introducing architec-

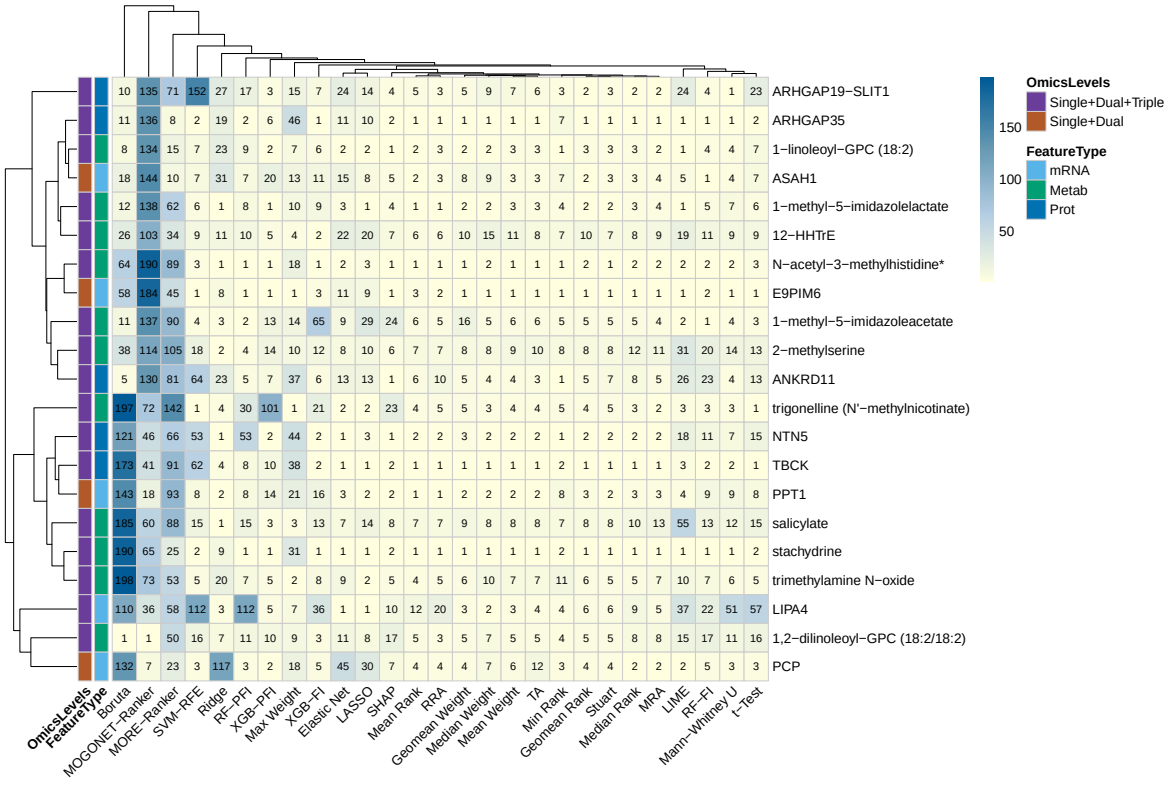


Figure 18: Rank heatmap for PSP. Feature ranking stability across multiple selection methods in PSP (progressive supranuclear palsy, PSP). Rows represent candidate mRNA, proteins, and metabolites, while columns denote rankers grouped into single and ensemble methods. Each cell displays the rank assigned by that ranker to that feature; smaller numeric ranks indicate higher importance (top-ranked). Consistently top-ranked features included validated metabolites and genes with prior neurological relevance, such as trimethylamine N-oxide (TMAO) (neuroinflammation), ANKRD11 (chromatin regulator), and PPT1 (lysosomal enzyme). At the same time, several novel candidates emerged, including genes like TBCK, ARHGAP19-SLIT1, ARHGAP35, and ASAH1, as well as metabolites such as stachydrine, trigonelline, salicylate, 1-methyl-5-imidazoleacetate, 2-methylserine, 12-HHTrE, and 1-linoleoyl–GPC (18:2), which were reproducibly prioritized but remain poorly studied in PSP. Together, the results highlight both the recovery of known disease-linked features and the discovery of reproducible, underexplored candidates across omics layers.

tural complexity without large-scale training data risks overfitting, reduced reproducibility, and diminished interpretability.

Our validation analyses consistently show that integrating multiple omic layers concentrates biologically meaningful signal. Moving from single-omics to dual- and then triple-omics yields a clear, stepwise increase in true-positive (TP) overlap with curated resources, even though the fixed per-panel quota forces each modality to contribute fewer features under integration. This pattern argues against simple quota or "more-of-one-omic" effects. Dual-omics can, in principle, benefit from repeat appearances of the same biology across two layers, but triple-omics does not—features appear once—yet it still achieves many of the strongest recoveries. The most parsimonious interpretation is that cross-omics prioritization filters out idiosyncratic, modality-specific noise and enriches for shared, disease-relevant biology.

Methodologically, the selector hierarchy is stable across databases, cohorts, and cutoffs $k \in \{10, 30, 50, 100\}$. Ensemble aggregations—both rank-based (Mean/Median/Geomean of ranks; Stuart; RRA) and weight-based (Mean/Median/Geomean of importances)—consistently occupy the top tier. Strong traditional rankers (e.g., t-Test, LASSO, Ridge/Elastic Net, SVM-RFE, RF-FI, XGB-FI) are often competitive and typically form the next band. Methods such as LIME, `TA`, MOGONET-Ranker, and MORE-Ranker tend to validate less across settings. Increasing k raises absolute TP counts and modestly narrows gaps among

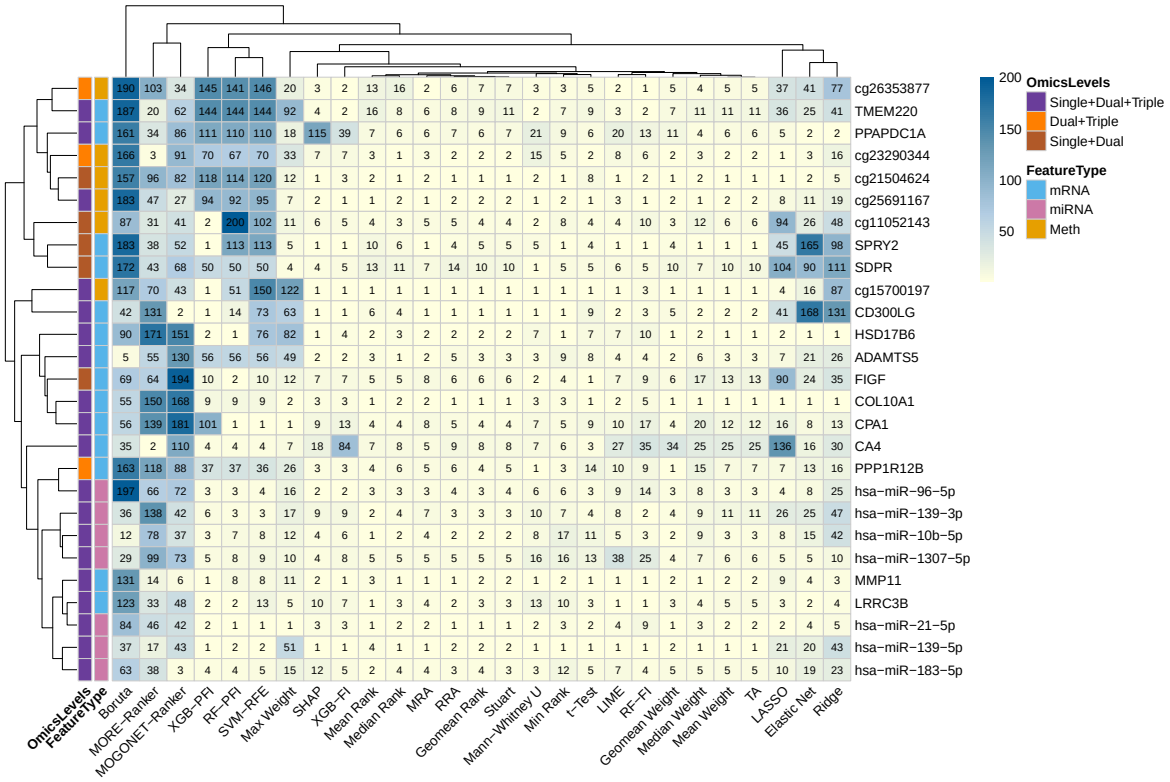


Figure 19: Rank heatmap for BC. Feature ranking stability across multiple selection methods in BC (breast cancer). Rows represent candidate features (microRNAs, genes, and CpG methylation sites), while columns denote individual rankers grouped into single and ensemble methods. Each cell displays the rank assigned by that ranker to that feature; smaller numeric ranks indicate higher importance (top-ranked). Among the consistently top-ranked features, validated biomarkers such as hsa-miR-21-5p, hsa-miR-139-5p/3p, hsa-miR-96-5p, COL10A1, MMP11, and FIGF overlapped with known breast cancer associations. In contrast, novel candidates including PPP1R12B, LRRC3B, TMEM220, PPAPDC1A, HSD17B6, ADAMTS5, SDPR, and SPRY2 were robustly prioritized across rankers but lack strong prior evidence in BC. Similarly, several CpG sites (cg23290344, cg26353877, cg25691167, cg11052143, cg15700197) were repeatedly ranked highly yet remain underexplored in breast cancer. Together, the heatmap demonstrates the ability of ensemble ranking to recover known BC biomarkers while highlighting reproducible, potentially novel candidates across molecular layers.

methods, but the qualitative ordering remains unchanged, indicating that conclusions are not sensitive to the exact cutoff.

The choice of external resource shapes both the dynamic range and interpretability of validation. CTD-pathways is the most discriminative readout: it shows the largest absolute gains under integration and cleanly separates methods, consistent with the idea that convergent signals across layers project strongly at the pathway level. HMDD (miRNA–disease) and GeneCards (gene associations) display steadier, narrower ranges yet still benefit from multi-omics, with fewer zeros and more selectors above baseline. CTD gene–disease associations are comparatively conservative—many methods score low under single-omics—but dual/triple-omics reduces zeros and lifts the mid-range, again pointing to cross-omics concentration rather than quota reshuffling. EWAS-ATLAS is cohort-limited yet consistently rises with integration, underscoring that methylation-anchored signals are not exceptions to the trend.

Cohort effects persist but follow a coherent pattern. BC generally achieves the highest absolute TP counts—especially in CTD-pathways and CpG validations—suggesting that disease context and data richness interact favorably with multi-omics integration. AD benefits most in miRNA-based validation (HMDD), aligning with the strengths of its available modalities, and still gains from integration elsewhere. PSP shows moderate but consistent improvements with fewer dramatic shifts. These differences likely reflect

Table 8: 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

cohort-specific biology, platform composition, and depth per modality rather than instability of the selection strategies.

The convergence of results across four cutoffs strengthens the robustness of our conclusions. At k=10, integration already confers clear advantages; by k=30 and k=50, dynamic ranges widen and method ordering becomes more evident; at k=100, absolute counts grow and some panels approach saturation, yet the integrative advantage and selector hierarchy persist. Thus, our findings hold across stringent to permissive thresholds, suggesting they are not artifacts of one particular operating point.

Two caveats merit mention. First, while dual-omics can occasionally benefit from cross-layer repeat hits, the strong performance of triple-omics—where features are unique—indicates that repeatability alone cannot explain the gains. Second, database coverage and curation depth vary by resource and disease area, so validation magnitudes should be interpreted as lower bounds on biological plausibility rather than exhaustive truth. Within these limits, the consistency across resources, cohorts, and k strongly supports the central claim.

Practically, these results recommend multi-omics pipelines that (i) integrate across layers, and (ii) employ ensemble aggregation to enhance stability and biological recovery. Such designs produce panels that are not only predictive but also align with established knowledge bases, increasing confidence that shortlisted markers capture shared disease mechanisms. Detailed examination of validated and putatively novel features—especially those repeatedly prioritized across modalities and databases—offers promising candidates for mechanistic follow-up and translational evaluation.

Possible explanations for higher counts in CTD-pathways. We note that the pathway reference used for CTD-pathways was constructed by performing pathway enrichment with `g:Profiler` on the transcriptomics cohort (when available). This procedure can yield a larger set of candidate pathways than a fixed, manually curated threshold, thereby increasing the opportunity for overlaps (higher TP counts). In addition, (i) enrichment often returns families of highly overlapping pathways (parent–child or near-duplicate gene sets), effectively multiplying match opportunities; (ii) pathway-level validation aggregates signal across multiple genes, so cross-omics prioritization may concentrate on shared mechanisms that map more readily to pathways than to individual genes; and (iii) cohort-specific choices in multiple-testing control (e.g., FDR), background gene universe, and database coverage can further expand the validated pathway set. Collectively, these factors can inflate absolute counts in the CTD-pathways panel relative to gene-level resources, without altering the qualitative ordering of methods or the observed integrative advantage.

In AD, our benchmark highlighted a set of reproducible miRNA markers, including miR-132, miR-129-5p, miR-146b-5p, and miR-29a, which have all been implicated in synaptic regulation, inflammation, and neurodegeneration [59][60][61][62][63][64][65][66][67][68][69][70]. These findings reinforce their credibility as diagnostic anchors. In parallel, miR-885-5p emerged as a robust but underexplored candidate, warranting follow-up. Several genes (APLN, PPDPF, MEIS3, CCDC69, SPACA6) and CpG loci (cg06690548, cg12981137, cg19832721, cg16003238, cg22442730) were also consistently ranked, with many lacking extensive prior study in AD, underscoring their novelty.

In PSP, metabolite features dominated. Trimethylamine N-oxide (TMAO) was prioritized and aligns with prior evidence linking it to cognitive decline and inflammation [71][72][73]. Trigonelline, stachydrine, and salicylate also surfaced, consistent with experimental reports of neuroprotective or anti-inflammatory activity [74][75][76][77][78]. Among proteins, PPT1—a lysosomal enzyme—confirmed known links to neuronal morphology and function [79] and the necessity of depalmitoylation in synaptic plasticity [80]. Meanwhile, ANKRD11 and TBCK represented less-characterized but reproducibly high-ranked signals. Genes such as ASAH1 and

ARHGAP family members (e.g., ARHGAP19-SLIT1, ARHGAP35) also emerged as novel leads, with mechanistic plausibility through lipid metabolism and neuronal signaling pathways.

In BC, familiar miRNA and mRNA drivers were consistently recovered. hsa-miR-21-5p is a well-known oncomiR that promotes proliferation, invasion, and immune evasion in breast tumors, in part via PTEN/TIMP3 pathways, and its high tumor or circulating levels are linked to poor prognosis [81]. Recent reports have further confirmed its detectability as a circulating biomarker in the blood of patients with breast cancer mutations [82], reinforcing its translational potential beyond tumor tissue. hsa-miR-139-5p/-3p generally acts as a tumor suppressor; both strands are downregulated in breast cancer and inhibit invasion/migration and metastatic programs when restored [83]. hsa-miR-96-5p functions as an oncogenic miRNA that drives proliferation and motility by repressing FOXO1/FOXO3a and related targets [84]. Among genes, COL10A1 is recurrently overexpressed in breast tumors and promotes malignant progression (e.g., via P4HB), with multiple studies linking it to worse outcomes [85]. MMP11 (stromelysin-3) shows high expression in tumor and stromal compartments and is associated with adverse clinicopathologic features and poorer survival [86]. FIGF (VEGF-D) is a lymphangiogenic growth factor that facilitates lymphatic spread and metastasis in breast cancer [87]. Alongside these, several promising but underexplored candidates were reproducibly prioritized. For instance, LRRC3B has been reported as a tumor suppressor in breast cancer cell models, and is frequently epigenetically silenced via promoter methylation in breast carcinoma [88]. The repeated ranking of CpG methylation sites such as cg23290344, cg26353877, cg25691167, cg11052143, and cg15700197 further emphasizes the importance of epigenetic signals, aligning with broader evidence that methylation profiles differ significantly between tumor and adjacent normal breast tissues [89].

Based on these results, we recommend a set of selector–classifier–panel size combinations for future users (Table 9). Across integration levels, F1-scores generally increased with panel size up to $K = 60$, after which performance plateaued. Triple-omics configurations achieved the strongest overall performance (0.91–0.95), particularly when combined with ensemble rankers (e.g., RF-PFI, XGB-PFI, mean- or median-based aggregation) and traditional classifiers such as L.Regression, SVM, or MLP. Dual-omics provided robust and reproducible results (0.87–0.92), with SHAP, Elastic Net, and rank-based ensembles paired with CatBoost or MLP emerging as consistent choices, especially for $K = 40$–60. Single-omics analyses yielded lower performance overall (0.83–0.89), but competitive results were obtained with SHAP, Elastic Net, or Ridge regression at $K = 30$–70, paired with interpretable classifiers such as L.Regression, SVM, or MLP. These findings suggest that mid-sized panels (30–60 features) optimized with ensemble rankers and traditional models strike the best balance between stability and predictive accuracy, while larger panels ($K \geq 70$) offer only marginal gains. Deep learning models did not consistently outperform simpler classifiers, reinforcing that their use should be reserved for larger datasets with independent validation.

While prior benchmarking efforts have advanced the field, they typically lacked the full tri-variate perspective we adopt—explicitly considering the interactions between feature selectors, classifiers, the variety of omics data types, and panel sizes across heterogeneous cohorts. For instance, Li et al. systematically compared feature selection strategies on multi-omics TCGA datasets [25], providing an important reference point for cancer studies, but their work did not evaluate how panel size or classifier choice modulate reproducibility. Urbanowicz et al. performed one of the most comprehensive evaluations of Relief-based algorithms [90], focusing on simulated and real bioinformatics data, but their scope was method-centric rather than integrative across classifiers and omics layers. Bommert et al. benchmarked filter-based feature selection methods for high-dimensional survival data [91], highlighting stability and ranking consistency in gene expression–driven prognosis, yet they did not test cross-omics generalization or multi-modal integration.

Other recent studies have pushed toward multi-omics and task-specific optimization. Labory et al. benchmarked feature selection and extraction combinations for omics classification [92], identifying synergies between dimensionality reduction and statistical filters, but their work emphasized single-omics scenarios. Łukaszuk et al. examined the stability of feature selection across cancer types using TCGA datasets [26]. Zappia et al. dissected the impact of feature selection choices on single-cell RNA-seq integration [93], revealing that the number and type of selected features profoundly affect downstream analysis. Spooner et al. benchmarked ensemble machine learning approaches for multi-omics integration [94], showing strong performance of hybrid ensembles, but they did not explicitly decompose contributions from selectors, classifiers, and feature set sizes.

Table 9: Recommended combinations of omics integration level, panel size (K), feature selectors, and models derived from benchmarking analyses. F1-scores were averaged across cohorts, and the top five performing selector–model combinations were used to define recommendations, with score ranges reported.

Omics Level	Panel Size (K)	Best Feature Selectors	Best Models	Performance Trend (F1)
Single-Omics	10	SHAP, Mean Weight, TA, Geomean Weight, Geomean Rank	CatBoost, SVM, MLP	0.863 – 0.867
	20	TA, Mean Weight, SHAP	MLP, AdaBoost, CatBoost, SVM	0.874 – 0.876
	30	Median Weight, Elastic Net, Mean Weight, TA	MLP, SVM	0.880 – 0.887
	40	LASSO, MRA, Elastic Net, Median Weight, TA	MLP	0.884 – 0.888
	50	Elastic Net, Median Weight, MRA, Ridge, RRA	MLP	0.879 – 0.890
	60	LASSO, Elastic Net, TA, Ridge	MLP, L.Regression	0.882 – 0.888
	70	Median Weight, Elastic Net, Ridge, LASSO, Mean Weight	MLP	0.882 – 0.890
	80	Ridge, Geomean Weight, Median Weight, Stuart, Geomean Rank	MLP	0.880 – 0.885
	90	Median Weight, TA, Geomean Rank, Mean Weight	MLP, L.Regression	0.874 – 0.881
	100	Median Weight, TA, MRA, Ridge, Mean Weight	MLP	0.875 – 0.879
Dual-Omics	10	Median Weight, shap, Geomean Weight, Mean Weight	SVM, MLP, CatBoost	0.896 – 0.899
	20	LASSO, TA, Mean Weight, Elastic Net	MLP, SVM	0.909 – 0.916
	30	Elastic Net, LASSO, TA, Geomean Weight, Mean Weight	MLP, SVM	0.912 – 0.919
	40	LASSO, Mean Weight, Elastic Net, Geomean Weight	MLP, SVM	0.919 – 0.923
	50	Elastic Net, Median Weight, LASSO, Mean Weight, TA	MLP	0.922 – 0.929
	60	Elastic Net, LASSO, Mean Weight, Ridge, TA	MLP	0.924 – 0.934
	70	Elastic Net, LASSO, Geomean Weight, Ridge	MLP, L.Regression	0.922 – 0.936
	80	Elastic Net, Median Weight, LASSO, Mean Weight	MLP, L.Regression	0.926 – 0.935
	90	Elastic Net, LASSO, Ridge, Median Weight	MLP, L.Regression	0.923 – 0.930
	100	Elastic Net, Median Weight, LASSO, Ridge, TA	MLP	0.923 – 0.931
Triple-Omics	10	MRA, Stuart, mean_rank	SVM, CatBoost, MLP	0.911 – 0.917
	20	Elastic Net, Mean Weight, LASSO, TA	MLP, L.Regression, SVM	0.920 – 0.928
	30	Elastic Net, LASSO	MLP, L.Regression, SVM	0.922 – 0.934
	40	LASSO, Elastic Net, TA	MLP, L.Regression, SVM	0.926 – 0.939
	50	LASSO, Elastic Net	MLP, L.Regression, SVM	0.929 – 0.938
	60	Elastic Net, LASSO, Geomean Weight, Median Weight, Ridge	MLP	0.933 – 0.953
	70	LASSO, Elastic Net, Ridge, TA	MLP, L.Regression	0.938 – 0.949
	80	Elastic Net, LASSO, Median Weight, TA	MLP, L.Regression	0.939 – 0.951
	90	LASSO, Elastic Net, Mean Weight, Ridge	MLP, L.Regression	0.939 – 0.948
	100	Elastic Net, LASSO, TA, Median Weight	MLP, L.Regression	0.939 – 0.956

Similarly, Claude et al. showed that hybrid ensemble selectors improved transcriptomic classification [20], while Li et al. explored block-omics combinations for survival prediction [26], both advancing multi-omics interpretability but without addressing systematic cross-cohort reproducibility. Classical works also remain highly relevant. Christin et al. critically assessed feature selection in metabolomics pipelines [95]; Pudjihartono et al. reviewed modern FS methods and their limitations [96]; and Siegismund et al. compared feature selection techniques in high-content screening data [97]. Recent large-scale reviews further underscore the need for reproducibility frameworks in multi-omics feature selection.

By contrast, our work spans Alzheimer's disease (AD), progressive supranuclear palsy (PSP), and breast cancer (BC) cohorts; integrates across molecular layers (mRNA, miRNA, methylation, proteins, and metabolites); and systematically benchmarks 27 feature selectors (including single, ensemble, and deep-learning embedded rankers

like MORE and MOGONET). We explicitly model selector × classifier × panel size interactions, emphasizing reproducibility across heterogeneous omics levels, thereby filling a critical methodological gap in integrative biomarker discovery.

There are several limitations of this study. Cohort sizes, particularly in PSP, remain modest, which restricts statistical power. Cross-cohort generalization is challenging due to differences in disease biology and cohort design. Technical sources of variability, including CpG mapping and platform-specific normalization, may also influence feature rankings. Additionally, while deep learning models underperformed here, larger datasets could reveal performance gains that we were unable to observe.

Future work should extend the benchmark to larger Alzheimer's cohorts such as ADNI and AMP-AD, integrate more comprehensive proteomics and metabolomics data, and experimentally validate novel candidates such as miR-885-5p in AD, LRRC3B and SDPR in BC, and ARHGAP19-SLIT1 or ASAH1 in PSP. Longitudinal modeling is also needed to track progression and treatment response, moving beyond case–control classification toward clinically actionable prediction.

Finally, to promote transparency and reproducibility, we developed **BioMark**, a web tool that enables users to analyze their own omics datasets with different rankers and classifiers. In parallel, we provide an **interactive benchmark explorer**, a lightweight web application for dynamic inspection of our results, including summary tables, heatmaps, and performance curves. Together, these tools make our benchmark not only a research contribution but also a resource for the community.

5 Conclusion

This study presents a comprehensive benchmarking framework for multi-omics biomarker discovery, addressing a critical question in bioinformatics and translational research: does increasing methodological complexity yield better biomarkers? By systematically evaluating 27 feature selection strategies and 11 predictive models across three diverse disease cohorts—Alzheimer's disease (AD), progressive supranuclear palsy (PSP), and breast cancer (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, support vector machines, and multilayer perceptrons 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 ¿ Single, and Triple ¿ Dual—triple-omics generally yields the best overall performance, with gains most evident when data quality and sample size are adequate. Importantly, compact biomarker panels (30–60 features) derived from ensemble selectors and interpretable models not only achieve high predictive accuracy but also facilitate downstream biological interpretation and clinical translation. The biological validation of identified biomarkers against curated databases (e.g., HMDD, CTD, GeneCards, EWAS-ATLAS) reinforces the clinical relevance of our approach. Moreover, the reproducible identification of both well-established and novel candidates across omics layers underscores the utility of ensemble strategies in uncovering robust disease signatures. To support the broader research community, we provide a web-based interactive explorer for visualizing performance metrics and biomarker rankings. 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 trade-offs between model sophistication and practical utility, especially in the context of translational applications.

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 Robust Rank Aggregation (RRA) algorithms were adapted from the `RobustRankAggreg` R package (version 1.2; Kolde & Laur) [55], released under the GPL-2 license. Algorithmic equivalence with

the original R implementation was verified using benchmark gene lists.

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.

Acknowledgements

A. Çakmak and C. Mesue Njume were supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) 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."

Biographical Note

Cyrille Mesue Njume is a graduate researcher in molecular biology and bioinformatics at Istanbul Technical University, focusing on multi-omics biomarker discovery and machine learning, and conducting his thesis under the supervision of Drs Ali Çakmak and Aslı Kumbasar within an international neurodegeneration consortium.

Irene Petracci is a researcher at the Molecular Markers Laboratory, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, working on molecular biomarkers and translational studies in neurodegenerative diseases, with particular interest in circulating microRNA signatures in dementia.

Sonia Bellini is a researcher at the Molecular Markers Laboratory, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, investigating molecular signatures, fluid biomarkers, and disease mechanisms underlying Alzheimer's disease, frontotemporal dementia, and related neurodegenerative conditions.

Katarzyna Goljanek-Whysall is a Senior Lecturer in Physiology at the University of Galway. Her research focuses on epigenetic and microRNA-mediated mechanisms in musculoskeletal and neuromuscular deterioration, exploring biomarker and therapeutic potential in ageing, sarcopenia, ALS, and muscle-wasting disorders.

Leo R. Quinlan is a Senior Lecturer in Physiology at the University of Galway and Principal Investigator in human physiology and medical devices, specializing in electrophysiology, neuromodulation, and integrative approaches to cardiovascular, neurological, and device-related clinical problems.

Agnieszka Fiszer is a researcher in the Department of Medical Biotechnology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, studying RNA biology, gene regulation, and molecular mechanisms of human disease, with a focus on RNA-based biomarkers and therapeutic strategies.

Barbara Borroni is a clinician-scientist at IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli and the University of Brescia, with expertise in dementia and movement disorders, integrating clinical, imaging, and molecular data to characterize neurodegenerative disease subtypes and progression.

Roberta Ghidoni heads the Molecular Markers Laboratory, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia. Her research focuses on fluid biomarkers, molecular pathways, and translational strategies in Alzheimer's disease, frontotemporal dementia, and related neurodegenerative conditions.

Aslı Kumbasar is an Associate Professor in the Department of Molecular Biology and Genetics at Istanbul Technical University. Her research addresses neurogenesis and gene expression regulation in neural stem cells, and she co-supervises Cyrille's in vitro validation work alongside Dr Ali Çakmak.

Ali Çakmak is a faculty member in the Department of Computer Engineering at Istanbul Technical University, specializing in bioinformatics, machine learning, and multi-omics integration, and serves as corresponding author and supervisor for the computational aspects of this consortium-driven biomarker discovery project.

References

[1] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, May 2017.

[2] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," *Bioinformatics and Biology Insights*, vol. 14, p. 117793221989905, Jan. 2020.

[3] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang, "Salmon: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers in Genetics*, vol. 10, Mar. 2019.

[4] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Athena: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network," *BioData Mining*, vol. 6, Dec. 2013.

[5] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, p. 30–37, Nov. 2006.

[6] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, vol. 8, p. 37–49, Jan. 2008.

[7] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, p. e124, Aug. 2005.

[8] A.-L. Boulesteix and M. Slawski, "Stability and aggregation of ranked gene lists," *Briefings in Bioinformatics*, vol. 10, p. 556–568, Aug. 2009.

[9] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, p. 5923–5928, Apr. 2006.

[10] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, p. 488–492, Feb. 2005.

[11] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solis, R. Duque, H. Bersini, and A. Nowe, "Batch effect removal methods for microarray gene expression data integration: a survey," *Briefings in Bioinformatics*, vol. 14, p. 469–490, July 2012.

[12] W. W. B. Goh, W. Wang, and L. Wong, "Why batch effects matter in omics data, and how to avoid them," *Trends in Biotechnology*, vol. 35, p. 498–507, June 2017.

[13] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *2012 IEEE 13th International Conference on Information Reuse amp; Integration (IRI)*, p. 356–363, IEEE, Aug. 2012.

[14] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, p. 95–116, Dec. 2006.

[15] B. Pes, "Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains," *Neural Computing and Applications*, vol. 32, p. 5951–5973, Feb. 2019.

[16] X. Song, L. R. Waitman, Y. Hu, A. S. L. Yu, D. Robins, and M. Liu, "Robust clinical marker identification for diabetic kidney disease with ensemble feature selection," *Journal of the American Medical Informatics Association*, vol. 26, p. 242–253, Jan. 2019.

[17] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, "On developing an automatic threshold applied to feature selection ensembles," *Information Fusion*, vol. 45, p. 227–245, Jan. 2019.

[18] A. Spooner, G. Mohammadi, P. S. Sachdev, H. Brodaty, and A. Sowmya, "Ensemble feature selection with data-driven thresholding for alzheimer's disease biomarker discovery," *BMC Bioinformatics*, vol. 24, Jan. 2023.

[19] P. Paplomatas, M. G. Krokidis, P. Vlamos, and A. G. Vrahatis, "An ensemble feature selection approach for analysis and modeling of transcriptome data in alzheimer's disease," *Applied Sciences*, vol. 13, p. 2353, Feb. 2023.

[20] E. Claude, M. Leclercq, P. Thébault, A. Droit, and R. Uricaru, "Optimizing hybrid ensemble feature selection strategies for transcriptomic biomarker discovery in complex diseases," *NAR Genomics and Bioinformatics*, vol. 6, July 2024.

[21] Y. Liang, A. Gharipour, E. Kelemen, and A. Kelemen, "Homogeneous ensemble feature selection for mass spectrometry data prediction in cancer studies," *Mathematics*, vol. 12, p. 2085, July 2024.

[22] Z. Yao, G. Zhu, J. Too, M. Duan, and Z. Wang, "Feature selection of omic data by ensemble swarm intelligence based approaches," *Frontiers in Genetics*, vol. 12, Mar. 2022.

[23] P. Le, X. Gong, L. Ung, H. Yang, B. P. Keenan, L. Zhang, and T. He, "A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data," *Frontiers in Systems Biology*, vol. 4, Mar. 2024.

[24] S. Budhraja, M. Doborjeh, B. Singh, S. Tan, Z. Doborjeh, E. Lai, A. Merkin, J. Lee, W. Goh, and N. Kasabov, "Filter and wrapper stacking ensemble (fwse): a robust approach for reliable biomarker discovery in high-dimensional omics data," *Briefings in Bioinformatics*, vol. 24, Sept. 2023.

[25] Y. Li, U. Mansmann, S. Du, and R. Hornung, "Benchmark study of feature selection strategies for multi-omics data," *BMC Bioinformatics*, vol. 23, Oct. 2022.

[26] T. Łukaszuk, J. Krawczuk, K. Żyła, and J. Kesik, "Stability of feature selection in multi-omics data analysis," *Applied Sciences*, vol. 14, p. 11103, Nov. 2024.

[27] A. Davis, T. Wiegers, D. Sciaky, F. Barkalow, M. Strong, B. Wyatt, J. Wiegers, R. McMorran, S. Abrar, and C. Mattingly, "Comparative toxicogenomics database's 20th anniversary: update 2025," *Nucleic Acids Research*, vol. 53, p. D1328–D1334, Oct. 2024.

[28] C. Cui, B. Zhong, R. Fan, and Q. Cui, "Hmdd v4.0: a database for experimentally supported human microrna-disease associations," *Nucleic Acids Research*, vol. 52, p. D1327–D1332, Aug. 2023.

[29] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "Genecards: a novel functional genomics compendium with automated data mining and query reformulation support.," *Bioinformatics*, vol. 14, p. 656–664, Jan. 1998.

[30] Z. Xiong, M. Li, Y. Ma, R. Li, and Y. Bao, "Gmqn: A reference-based method for correcting batch effects and probe bias in humanmethylation beadchip," *Frontiers in Genetics*, vol. 12, Jan. 2022.

[31] K. Mihajlović, G. Ceddia, N. Malod-Dognin, G. Novak, D. Kyriakis, A. Skupin, and N. Pržulj, "Multi-omics integration of scrna-seq time series data predicts new intervention points for parkinson's disease," *Scientific Reports*, vol. 14, May 2024.

[32] R. K. Tripathy, Z. Frohock, H. Wang, G. A. Cary, S. Keegan, G. W. Carter, and Y. Li, "Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks," *npj Systems Biology and Applications*, vol. 11, May 2025.

[33] J. Liang, X. Huang, W. Li, and Y. Hu, "Identification and external validation of the hub genes associated with cardiorenal syndrome through time-series and network analyses," *Aging*, vol. 14, p. 1351–1373, Feb. 2022.

[34] Y. Wang, Z. Wang, X. Yu, X. Wang, J. Song, D.-J. Yu, and F. Ge, "More: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification," *Briefings in Bioinformatics*, vol. 26, Nov. 2024.

[35] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "Mogonet

integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Communications*, vol. 12, June 2021.

[36] A. P. Pérez-González, A. L. García-Kroepfly, K. A. Pérez-Fuentes, R. I. García-Reyes, F. F. Solis-Roldan, J. A. Alba-González, E. Hernández-Lemus, and G. de Anda-Jáuregui, "The rosmap project: aging and neurodegenerative diseases through omic sciences," *Frontiers in Neuroinformatics*, vol. 18, Sept. 2024.

[37] M. Allen, M. M. Carrasquillo, C. Funk, B. D. Heavner, F. Zou, C. S. Younkin, J. D. Burgess, H.-S. Chai, J. Crook, J. A. Eddy, H. Li, B. Logsdon, M. A. Peters, K. K. Dang, X. Wang, D. Serie, C. Wang, T. Nguyen, S. Lincoln, K. Malphrus, G. Bisceglio, M. Li, T. E. Golde, L. M. Mangravite, Y. Asmann, N. D. Price, R. C. Petersen, N. R. Graff-Radford, D. W. Dickson, S. G. Younkin, and N. Ertekin-Taner, "Human whole genome genotype and transcriptome data for alzheimer's and other neurodegenerative diseases," *Scientific Data*, vol. 3, Oct. 2016.

[38] M. J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, J. Zhu, and D. Haussler, "Visualizing and interpreting cancer genomics data via the xena platform," *Nature Biotechnology*, vol. 38, p. 675–678, May 2020.

[39] L. Kolberg, U. Raudvere, I. Kuzmin, P. Adler, J. Vilo, and H. Peterson, "g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)," *Nucleic Acids Research*, vol. 51, p. W207–W212, May 2023.

[40] D. A. Bennett, J. A. Schneider, Z. Arvanitakis, and R. S. Wilson, "Overview and findings from the religious orders study," *Current Alzheimer Research*, vol. 9, p. 628–645, June 2012.

[41] P. L. De Jager, Y. Ma, C. McCabe, J. Xu, B. N. Vardarajan, D. Felsky, H.-U. Klein, C. C. White, M. A. Peters, B. Lodgson, P. Nejad, A. Tang, L. M. Mangravite, L. Yu, C. Gaiteri, S. Mostafavi, J. A. Schneider, and D. A. Bennett, "A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research," *Scientific Data*, vol. 5, Aug. 2018.

[42] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, p. 1113–1120, Sept. 2013.

[43] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, p. 749–760, Oct. 2018.

[44] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, p. 56–67, Jan. 2020.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[46] R. TIBSHIRANI, "The lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, p. 385–395, Feb. 1997.

[47] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, p. 55–67, Feb. 1970.

[48] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, 2011.

[49] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, p. 50–60, Mar. 1947.

[50] M. B. Kursa and W. R. Rudnicki, "Feature selection with theborutapackage," *Journal of Statistical Software*, vol. 36, no. 11, 2010.

[51] L. Breiman, "Random forests," *Machine Learning*, vol. 45, p. 5–32, Oct. 2001.

[52] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 785–794, ACM, Aug. 2016.

[53] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, p. 389–422, Jan. 2002.

[54] A. Spooner, G. Mohammadi, P. S. Sachdev, H. Brodaty, and A. Sowmya, "Ensemble feature selection with data-driven thresholding for alzheimer's disease biomarker discovery," *BMC Bioinformatics*, vol. 24, Jan. 2023.

[55] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, p. 573–580, Jan. 2012.

[56] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *Journal of Computer and System Sciences*, vol. 66, p. 614–656, June 2003.

[57] R. Mayeux, "Biomarkers: Potential uses and limitations," *NeuroRX*, vol. 1, p. 182–188, Apr. 2004.

[58] G. Grande, M. Valletta, D. Rizzuto, X. Xia, C. Qiu, N. Orsini, M. Dale, S. Andersson, C. Fredolini, B. Winblad, E. J. Laukka, L. Fratiglioni, and D. L. Vetrano, "Blood-based biomarkers of alzheimer's disease and incident dementia in the community," *Nature Medicine*, vol. 31, p. 2027–2035, Mar. 2025.

[59] C. Zeng, X. Meng, D. Mai, K. Xu, and S. Qu, "Overexpression of mir-132-3p contributes to neuronal protection in in vitro and in vivo models of alzheimer's disease," *Behavioural Brain Research*, vol. 417, p. 113584, Jan. 2022.

[60] N. Zhang, Y. Lyu, X. Pan, L. Xu, A. Xuan, X. He, W. Huang, and D. Long, "mir-146b-5p promotes the neural conversion of pluripotent stem cells by targeting smad4," *International Journal of Molecular Medicine*, vol. 40, p. 814–824, July 2017.

[61] E. Salta, A. Sierksma, E. Vanden Eynden, and B. De Strooper, "mir-132 loss derepresses itpkb and aggravates amyloid and tau pathology in alzheimer's brain," *EMBO Molecular Medicine*, vol. 8, p. 1005–1018, Aug. 2016.

[62] M. Liu, Y. Peng, Y. Che, M. Zhou, Y. Bai, W. Tang, S. Huang, B. Zhang, S. Deng, C. Wang, and Z. Yu, "Mir-146b-5p/traf6 axis is essential for ginkgo biloba l. extract gbe to attenuate lps-induced neuroinflammation," *Frontiers in Pharmacology*, vol. 13, Aug. 2022.

[63] S.-W. Han, J.-M. Pyun, P. J. Bice, D. A. Bennett, A. J. Saykin, S. Y. Kim, Y. H. Park, and K. Nho, "mir-129-5p as a biomarker for pathology and cognitive decline in alzheimer's disease," *Alzheimer's Research amp; Therapy*, vol. 16, Jan. 2024.

[64] Z. Mei, J. Liu, J. P. Schroeder, D. Weinshenker, D. M. Duong, N. T. Seyfried, Y. Li, P. Jin, A. P. Wingo, and T. S. Wingo, "Lowering hippocampal mir-29a expression slows cognitive decline and reduces beta-amyloid deposition in 5×fad mice," *Molecular Neurobiology*, vol. 61, p. 3343–3356, Nov. 2023.

[65] S. Nagaraj, C. Quintanilla-Sánchez, K. Ando, L. Lopez-Gutierrez, E. Doeraene, A.-C. Kosa, E. Aydin, J.-P. Brion, and K. Leroy, "Downregulation of hsa-mir-132 and hsa-mir-129: non-coding rna molecular signatures of alzheimer's disease," *Frontiers in Molecular Neuroscience*, vol. 17, June 2024.

[66] S. S. Hébert, K. Horré, L. Nicolaï, A. S. Papadopoulou, W. Mandemakers, A. N. Silahtaroglu, S. Kauppinen, A. Delacourte, and B. De Strooper, "Loss of microrna cluster mir-29a/b-1 in sporadic alzheimer's disease correlates with increased bace1/-secretase expression," *Proceedings of the National Academy of Sciences*, vol. 105, p. 6415–6420, Apr. 2008.

[67] L. Kaurani, R. Pradhan, S. Schröder, S. Burkhardt, A.-L. Schuetz, D. M. Krüger, T. Pena, P. Heutink, F. Sananbenesi, and A. Fischer, "A role for astrocytic mir-129-5p in frontotemporal dementia," *Translational Psychiatry*, vol. 15, Apr. 2025.

[68] H. Walgrave, S. Balusu, S. Snoeck, E. Vanden Eynden, K. Craessaerts, N. Thrupp, L. Wolfs, K. Horré, Y. Fourne, A. Ronisz, E. Silajdžić, A. Penning, G. Tosoni, Z. Callaerts-Vegh, R. D'Hooge, D. R. Thal, H. Zetterberg, S. Thuret, M. Fiers, C. S. Frigerio, B. De Strooper, and E. Salta, "Restoring mir-132 expression rescues adult hippocampal neurogenesis and memory deficits in alzheimer's disease," *Cell Stem Cell*, vol. 28, pp. 1805–1821.e8, Oct. 2021.

[69] Y.-M. Ma and L. Zhao, "Mechanism and therapeutic prospect of mirnas in neurodegenerative diseases," *Behavioural Neurology*, vol. 2023, p. 1–24, Nov. 2023.

[70] N. Xu, A.-D. Li, L.-L. Ji, Y. Ye, Z.-Y. Wang, and L. Tong, "mir-132 regulates the expression of synaptic proteins in app/ps1 transgenic mice through c1q," *European Journal of Histochemistry*, vol. 63, May 2019.

[71] Y. Zhang, G. Wang, R. Li, R. Liu, Z. Yu, Z. Zhang, and Z. Wan, "Trimethylamine n-oxide aggravated cognitive impairment from app/ps1 mice and protective roles of voluntary exercise," *Neurochemistry International*, vol. 162, p. 105459, Jan. 2023.

[72] W. Quan, C.-M. Qiao, G.-Y. Niu, J. Wu, L.-P. Zhao, C. Cui, W.-J. Zhao, and Y.-Q. Shen, "Trimethylamine n-oxide exacerbates neuroinflammation and motor dysfunction in an acute mptp mice model of parkinson's disease," *Brain Sciences*, vol. 13, p. 790, May 2023.

[73] V. E. Brunt, T. J. LaRocca, A. E. Bazzoni, Z. J. Sapinsley, J. Miyamoto-Ditmon, R. A. Gioscia-Ryan, A. P. Neilson, C. D. Link, and D. R. Seals, "The gut microbiome–derived metabolite trimethylamine n-oxide modulates neuroinflammation and cognitive function with aging," *GeroScience*, vol. 43, p. 377–394, Aug. 2020.

[74] Z. He, P. Li, P. Liu, and P. Xu, "Exploring stachydrine: from natural occurrence to biological activities and metabolic pathways," *Frontiers in Plant Science*, vol. 15, Aug. 2024.

[75] M. Lagraoui, G. Sukumar, J. R. Latoche, S. K. Maynard, C. L. Dalgard, and B. C. Schaefer, "Salsalate treatment following traumatic brain injury reduces inflammation and promotes a neuroprotective and neurogenic transcriptional response with concomitant functional recovery," *Brain, Behavior, and Immunity*, vol. 61, p. 96–109, Mar. 2017.

[76] M. Faizan, I. Jahan, M. Ishaq, A. Alhalmi, R. Khan, O. M. Noman, S. Hasson, and R. A. Mothana, "Neuroprotective effects of trigonelline in kainic acid-induced epilepsy: Behavioral, biochemical, and functional insights," *Saudi Pharmaceutical Journal*, vol. 31, p. 101843, Dec. 2023.

[77] Y. Liang, X. Dai, Y. Cao, X. Wang, J. Lu, L. Xie, K. Liu, and X. Li, "The neuroprotective and antidiabetic effects of trigonelline: A review of signaling pathways and molecular mechanisms," *Biochimie*, vol. 206, p. 93–104, Mar. 2023.

[78] L. Li, L. Sun, Y. Qiu, W. Zhu, K. Hu, and J. Mao, "Protective effect of stachydrine against cerebral ischemia-reperfusion injury by reducing inflammation and apoptosis through p65 and jak2/stat3 signaling pathway," *Frontiers in Pharmacology*, vol. 11, Feb. 2020.

[79] T. Sapir, M. Segal, G. Grigoryan, K. M. Hansson, P. James, M. Segal, and O. Reiner, "The interactome of palmitoyl-protein thioesterase 1 (ppt1) affects neuronal morphology and function," *Frontiers in Cellular Neuroscience*, vol. 13, Mar. 2019.

[80] K. P. Koster, E. Flores-Barrera, E. Artur de la Villarmois, A. Caballero, K. Y. Tseng, and A. Yoshii, "Loss of depalmitoylation disrupts homeostatic plasticity of ampars in a mouse model of infantile neuronal ceroid lipofuscinosis," *The Journal of Neuroscience*, vol. 43, p. 8317–8335, Oct. 2023.

[81] D. Bautista-Sánchez, C. Arriaga-Canon, A. Pedroza-Torres, I. A. De La Rosa-Velázquez, R. González-Barrios, L. Contreras-Espinosa, R. Montiel-Manríquez, C. Castro-Hernández, V. Fragoso-Ontiveros, R. M. Álvarez Gómez, and L. A. Herrera, "The promising role of mir-21 as a cancer biomarker and its importance in rna-based therapeutics," *Molecular Therapy - Nucleic Acids*, vol. 20, p. 409–420, June 2020.

[82] C. Alavanda, E. Dirimtekin, M. Mortoglou, E. Arslan Ates, A. I. Guney, and P. Uysal-Onganer, "Brca mutations and microrna expression patterns in the peripheral blood of breast cancer patients," *ACS Omega*, Apr. 2024.

[83] K. Krishnan, A. L. Steptoe, H. C. Martin, D. R. Pattabiraman, K. Nones, N. Waddell, M. Mariasegaram, P. T. Simpson, S. R. Lakhani, A. Vlassov, S. M. Grimmond, and N. Cloonan, "mir-139-5p is a regulator of metastatic pathways in breast cancer," *RNA*, vol. 19, p. 1767–1780, Oct. 2013.

[84] Y. Hong, H. Liang, U. ur Rehman, Y. Wang, W. Zhang, Y. Zhou, S. Chen, M. Yu, S. Cui, M. Liu, N. Wang, C. Ye, C. Zhao, Y. Liu, Q. Fan, C.-Y. Zhang, J. Sang, K. Zen, and X. Chen, "mir-96 promotes cell proliferation, migration and invasion by targeting ptpn9 in breast cancer," *Scientific Reports*, vol. 6, Nov. 2016.

[85] W. Yang, X. Wu, and F. Zhou, "Collagen type x alpha 1 (col10a1) contributes to cell proliferation, migration, and invasion by targeting prolyl 4-hydroxylase beta polypeptide (p4hb) in breast cancer," *Medical Science Monitor*, vol. 27, Dec. 2020.

[86] K.-W. Min, D.-H. Kim, S.-I. Do, J.-S. Pyo, K. Kim, S. W. Chae, J. H. Sohn, Y.-H. Oh, H. J. Kim, S. H. Choi, Y. J. Choi, and C. H. Park, "Diagnostic and prognostic relevance of mmp-11 expression in the stromal fibroblast-like cells adjacent to invasive ductal carcinoma of the breast," *Annals of Surgical Oncology*, vol. 20, p. 433–442, Nov. 2012.

[87] T. Karnezis, R. Shayan, C. Caesar, S. Roufail, N. Harris, K. Ardipradja, Y. Zhang, S. Williams, R. Farnsworth, M. Chai, T. Rupasinghe, D. Tull, M. Baldwin, E. Sloan, S. Fox, M. Achen, and S. Stacker, "Vegf-d promotes tumor metastasis by regulating prostaglandins produced by the collecting lymphatic endothelium," *Cancer Cell*, vol. 21, p. 181–195, Feb. 2012.

[88] G.-S. Li, G.-Y. Kong, and Y. Zou, "Protective role of lrrc3b in preventing breast cancer metastasis and recurrence post-bupivacaine," *Oncology Letters*, vol. 14, p. 5013–5017, Aug. 2017.

[89] S. R. Dennis, T. Tsukioki, G. Cottone, W. Zhou, P. A. Ganz, M. E. Sehl, Y. Luo, S. A. Khan, and S. Clare, "Dna methylation patterns in breast cancer, paired benign tissue from ipsilateral and contralateral breast, and healthy controls," *Breast Cancer Research*, vol. 27, June 2025.

[90] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, p. 168–188, Sept. 2018.

[91] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Briefings in Bioinformatics*, vol. 23, Sept. 2021.

[92] J. Labory, E. Njomgue-Fotso, and S. Bottini, "Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data," *Computational and Structural Biotechnology Journal*, vol. 23, p. 1274–1287, Dec. 2024.

[93] L. Zappia, S. Richter, C. Ramírez-Suástegui, R. Kfuri-Rubens, L. Vornholz, W. Wang, O. Dietrich, A. Frishberg, M. D. Luecken, and F. J. Theis, "Feature selection methods affect the performance of scrna-seq data integration and querying," *Nature Methods*, vol. 22, p. 834–844, Mar. 2025.

[94] A. Spooner, M. K. Moridani, B. Toplis, J. Behary, A. Safarchi, S. Maher, F. Vafaee, A. Zekry, and A. Sowmya, "Benchmarking ensemble machine learning algorithms for multi-class, multi-omics data integration in clinical outcome prediction," *Briefings in Bioinformatics*, vol. 26, Mar. 2025.

[95] C. Christin, H. C. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, and P. Horvatovich, "A critical assessment of feature selection methods for biomarker discovery in clinical proteomics," *Molecular amp; Cellular Proteomics*, vol. 12, p. 263–276, Jan. 2013.

[96] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, June 2022.

[97] D. Siegismund, M. Fassler, S. Heyse, and S. Steigele, "Benchmarking feature selection methods for compressing image information in high-content screening," *SLAS Technology*, vol. 27, p. 85–93, Feb. 2022.