Molecular Therapy
Nucleic Acids
Original Article

BRPtools: An AutoML-Powered web platform for multiclass disease prediction from bulk blood RNA-seq data

Yichen Guo,[1,4] Fengyuan Yang,[1,4] Xuelu Zhang,[1] Zhaoyu Zhai,[1] and Jianbo Pan[1,2,3]

[1]Basic Medicine Research and Innovation Center for Novel Target and Therapeutic Intervention, Ministry of Education, College of Pharmacy, Chongqing Medical University, Chongqing 400016, China; [2]Reproductive Medicine Center, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China; [3]Precision Medicine Center, The Second Affiliated Hospital of Chongqing Medical University, Chongqing 400010, China

Blood-based transcriptomic profiling provides a minimally invasive approach for disease diagnosis; however, the integration of large-scale, heterogeneous RNA sequencing (RNA-seq) datasets remains challenging. Here, we manually curated and uniformly reprocessed 134 publicly available human RNA-seq datasets (*n* = 9,872 samples), covering 88 distinct blood-related diseases from whole blood and peripheral blood mononuclear cell data. To enable robust and scalable multiclass prediction, we employed AutoGluon, an ensemble-based AutoML framework that automates model selection and hyperparameter tuning through multilayer stacking. Our models achieved high accuracy across most disease categories (>90%), demonstrating strong generalizability. To support biological interpretation, we also performed differential expression and pathway enrichment analyses, revealing disease-associated signatures enriched in relevant Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways. We implemented these capabilities in BRPtools, a web-based platform featuring two modules: (1) a search module for exploring gene- and disease-level expression profiles and (2) a predict module for disease classification using user-uploaded RNA-seq count matrices. The platform supports diagnostic inference and validation and is designed to be accessible to users without programming experience. BRPtools offers an integrated, interpretable, and user-friendly solution for transcriptome-based disease prediction, supporting applications in biomarker discovery, digital diagnostics, and precision medicine.

INTRODUCTION

The World Health Organization's International Classification of Diseases, 11th Revision (ICD-11) catalogs over 50,000 diseases and health-related conditions, many of which are associated with significant physiological dysfunction or increased mortality risk.[1] Despite substantial advances in biomedicine, the underlying molecular mechanisms of numerous complex diseases, such as Alzheimer's disease (AD), remain elusive. This knowledge gap underscores the urgent need for accurate and accessible biomarkers to facilitate early diagnosis, monitor disease progression, and guide therapeutic interventions. Traditionally, biomarker discovery has focused on invasive or semi-invasive sample types. For example, cerebrospinal fluid (CSF)-based assays for AD, although highly specific, require lumbar puncture procedures that are both technically challenging and physically uncomfortable for patients.[2] The invasiveness, cost, and infrastructure demand of such methods severely limit their suitability for routine clinical use, particularly in primary care or large-scale population screening scenarios. These practical limitations have intensified efforts to identify noninvasive biomarkers that are equally reliable yet more accessible.

Peripheral blood has emerged as a compelling alternative source for biomarker discovery because of its ease of collection, cost-effectiveness, and capacity to reflect systemic physiological changes.[3,4] In particular, transcriptomic profiling of whole blood (WB) or peripheral blood mononuclear cell (PBMC) data using RNA sequencing (RNA-seq) allows for high-throughput characterization of disease-specific gene expression patterns. Differentially expressed genes (DEGs) identified through these approaches have not only shown promise as candidate diagnostic markers but also contributed to elucidating the molecular basis of disease.[5,6] Building on this foundation, several studies have employed classical machine-learning models, such as support vector machines (SVMs), random forests (RFs), or gradient boosting to classify diseases from blood transcriptomes, achieving encouraging accuracy in infection and neurodegenerative disorders.[7,8] However, these models often suffer from limited cross-cohort generalizability and unstable feature selection, reflecting the difficulty of translating signatures across heterogeneous datasets. Effectively translating transcriptomic signatures into clinically actionable diagnostics requires robust computational strategies capable of handling the inherent complexity of RNA-seq data. Key

Received 28 September 2025; accepted 15 May 2026;
https://doi.org/10.1016/j.omtn.2026.102956.

[4]These authors contributed equally

Correspondence: Jianbo Pan, Basic Medicine Research and Innovation Center for Novel Target and Therapeutic Intervention, Ministry of Education, College of Pharmacy, Chongqing Medical University, Chongqing 400016, China
E-mail: panjianbo@cqmu.edu.cn

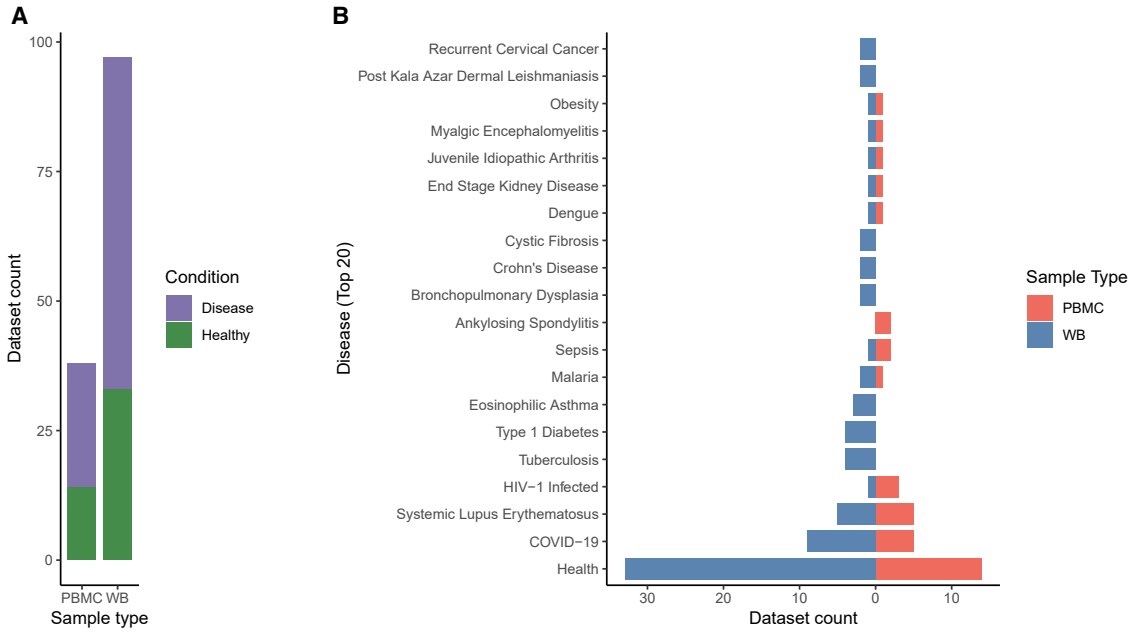


Figure 1. Distribution of curated RNA-seq datasets in BRPtools

(A) Stacked bar plot showing the number of datasets for each sample type, WB and PBMC, stratified by condition (disease vs. healthy control). (B) Horizontal bar plot displaying the dataset count for the top 20 most represented disease categories, separately for PBMCs and WBs.

challenges include technical variability across platforms, batch effects, and the high dimensionality of gene expression data, which often comprise tens of thousands of features measured across relatively few samples, making robust feature selection an essential requirement.

In this study, we assembled and reanalyzed transcriptomic data from 9,872 peripheral blood samples representing 88 disease types, sourced from the Gene Expression Omnibus (GEO) repository. To build a robust diagnostic model, we employed AutoGluon, an automated multilayer ensemble-learning framework.[9] This approach integrates predictions from diverse base learners and stacking layers, with automated hyperparameter optimization and model selection. By combining model ensembling with automated optimization, the framework effectively mitigates overfitting and feature-selection instability in high-dimensional RNA-seq data, thereby improving generalizability across heterogeneous datasets. Our model achieved >0.90 accuracy across most disease categories. In addition to predictive modeling, we performed differential expression and functional enrichment analyses using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) resources, which revealed enriched biological pathways and facilitated downstream interpretation and hypothesis generation.[10,11]

To increase accessibility and facilitate translational use, we developed BRPtools (http://www.inbirg.com/blood/home), an interactive, web-based platform that integrates large-scale gene expression data with real-time disease prediction capabilities. The platform provides two primary modules: a search module for exploring gene- and disease-specific expression profiles and a predict module that enables users to upload RNA-seq count matrices for diagnostic inference. By providing an accessible interface without the need for programming expertise, BRPtools helps bridge the gap between complex transcriptomic data and practical diagnostic utility.

Taken together, our work presents a scalable, interpretable, and clinically oriented framework for transcriptome-based disease prediction, laying the groundwork for future biomarker-guided applications in precision medicine.

RESULTS

Scheme of the online platform

BRPtools currently integrates transcriptomic data from 88 distinct blood disease types compiled from a total of 134 RNA-seq datasets. Among these, in addition to healthy controls, 73 disease categories are represented by WB samples derived from 97 datasets, whereas 28 categories are captured by PBMC samples drawn from 38 datasets. Notably, several conditions, including healthy controls, are included in several datasets. To ensure analytical consistency and minimize technical bias, all datasets were subjected to a standardized processing pipeline that included adapter trimming, quality assessment, genome alignment, read quantification, and normalization. A detailed summary of the dataset composition, including disease name, GEO accession, sample counts, and availability for prediction or differential expression analysis, is provided in Table S1. Figure 1A summarizes the distribution of datasets across sample types and

conditions, highlighting the proportion of disease vs. healthy control samples in both the PBMC and WB sources. Notably, whole-blood datasets are more prevalent and contain a greater proportion of healthy control samples than PBMCs. Figure 1B depicts the dataset frequency for the top 20 most represented disease types, stratified by sample source. This figure highlights the diversity and relative abundance of disease representations within the BRPtools database, offering insights into disease-specific transcriptomic coverage across blood-derived tissues.

Differential expression and enrichment analyses

To elucidate disease-specific transcriptional alterations, we performed differential gene expression analysis between case and healthy control samples for each dataset that included both conditions. Using the limma-voom pipeline, we identified DEGs at a false discovery rate (FDR) threshold of <0.05. Figure 2A shows a representative volcano plot for a sepsis dataset, highlighting significantly upregulated (red) and downregulated (blue) genes on the basis of $\log_2$fold change and adjusted p values. These patterns reflect substantial transcriptional regulation associated with the disease state.

To interpret the biological significance of these DEGs, we conducted GO and KEGG pathway enrichment analyses using the clusterProfiler package. Immune- and inflammation-related processes were strongly enriched in sepsis datasets, as shown in Figure 2B, consistent with the central role of innate immune activation in septic pathology.[12] These concordances with established disease biology confirm the biological validity of the underlying datasets and verify the accuracy of our standardized data processing pipeline. These precomputed results are fully integrated into the BRPtools search module, where users can interactively explore differential expression patterns, enrichment outputs, and gene-level annotations for each supported dataset.

Model assessment and selection

Prior to model evaluation, we confirmed that the multi-cohort integration was free from profound technical biases. Principal-component analysis (PCA) and the average silhouette width (ASW) metrics demonstrated that ComBat-seq effectively eliminated dataset-specific clustering, ensuring that downstream models were trained on conserved biological signals rather than residual platform artifacts (Figure S1). To assess the diagnostic performance of the individual models used within the AutoGluon framework, we compared the classification accuracy of all base learners and ensemble models using stratified 5-fold cross-validation within the training set, and then evaluated the final models on the held-out test set for both the PBMC and WB data. As shown in Figure 3A, the PBMC-based models exhibited consistent performance across folds, with the WeightedEnsemble_L2 model achieving the highest accuracy on both cross-validation and independent test evaluations. Figure 3B presents the corresponding results for the WB data, where a similar pattern was observed: the WeightedEnsemble_L2 model again outperformed all other individual learners, demonstrating superior generalizability across datasets. To further summarize performance on the held-out test sets, we computed overall accuracy, macro-precision, and macro-recall for the WeightedEnsemble_L2 model based on both PBMC and WB datasets (Table 1). These results underscore the robustness of ensemble-based approaches in handling heterogeneous blood transcriptome data, validating their selection as the default predictor within BRPtools.

To strictly assess cross-cohort generalizability as an external validation, we evaluated the models using the targeted leave-one-dataset-out (LODO) strategy. As expected for strictly external evaluations across independent laboratories and sequencing platforms, the LODO validation exhibited a moderate reduction in predictive performance compared to internal splits. Nevertheless, the models maintained a prominent diagonal classification trend across the targeted disease categories, indicating robust resilience (Figure S2).

To provide a comprehensive evaluation of the model's performance across diverse disease categories, we examined the classification details for each individual class. The confusion matrices for both PBMC and WB models demonstrated high diagonal consistency, indicating minimal cross-disease misclassification (Figure S3).

Model interpretation

To address the "black box" nature of the AutoML framework and identify key transcriptomic signatures driving the multiclass predictions, we extracted and analyzed the feature importance scores from the trained models. For the PBMC models, the framework converged on a highly sparse core signature of 10 predictive genes (Figure 4A), providing a potentially actionable biomarker panel for clinical diagnostics. In contrast, the WB models relied on a broader set of predictive features, from which the top 10 genes were identified (Figure 4B).

To bridge the gap between predictive performance and biological mechanisms, we performed GO and KEGG pathway enrichment analyses specifically on the top predictive features derived from the machine learning models. For the WB model, the analysis revealed significant enrichment in systemic immune responses and disease pathogenesis, including chemotaxis and chemokine-mediated signaling pathways (Figures 4C and 4D). These results demonstrate that the model relies on conserved, biologically relevant transcriptomic signals rather than uninformative noise, enhancing interpretability.

Application modules

BRPtools features two user-friendly analytical modules: search and predict.

Search module

The search module provides an interactive portal for exploring gene- and disease-level transcriptomic signatures. Users can perform fuzzy searches using gene symbols or disease names to retrieve a rich set of precomputed results.

When querying by gene, BRPtools returns a summary table with matching gene symbols, chromosomal coordinates, and gene types.

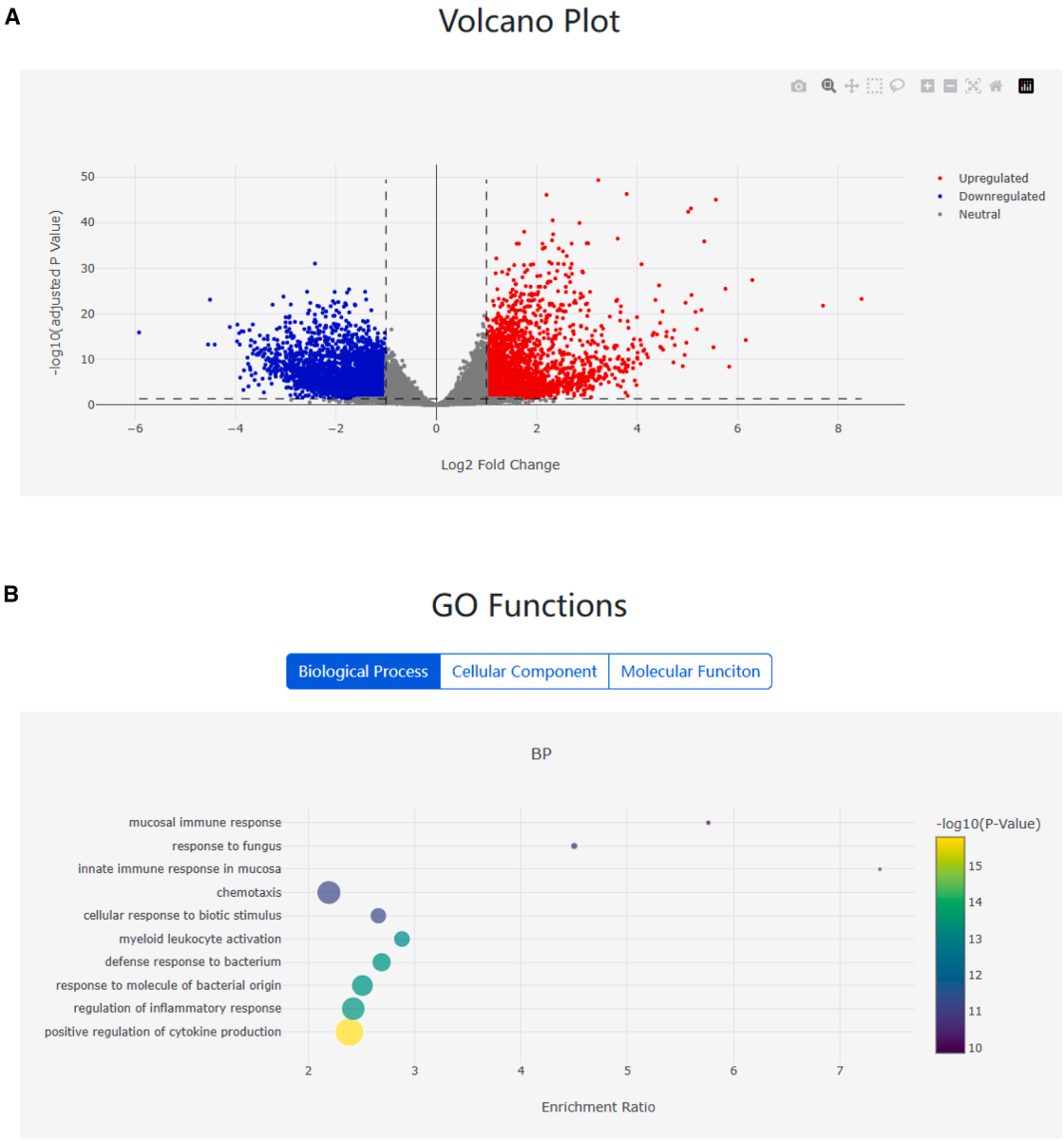


Figure 2. Differential expression and functional enrichment analysis

(A) Volcano plot showing differentially expressed genes (DEGs) identified from one representative dataset using the limma-voom pipeline. The red dots represent upregulated genes, the blue dots indicate downregulated genes, and the gray dots represent nonsignificant genes. Thresholds for significance (FDR < 0.05, $|\log_2$fold change$| > 1$) are marked with dashed lines. (B) GO enrichment analysis of DEGs in the "biological process" category. The dot plot displays the top enriched terms, with the dot size reflecting the gene count, the color representing statistical significance ($-\log_{10}$ adjusted *p* value), and the *x* axis showing the enrichment ratio.

Clicking on a gene symbol reveals detailed expression patterns across sample types (WB vs. PBMC) and disease categories. This feature enables quick inspection of gene dysregulation trends, which is particularly useful for validating candidate biomarkers. For example, querying "TP53" shows its basic annotations (Figures 5A and 5B), source-specific expression levels (Figure 5C), and relative activity across disease groups (Figure 5D).

Alternatively, users can initiate a query by entering the name of a disease to retrieve all relevant RNA-seq datasets curated within BRPtools. Upon a successful match, a comprehensive summary table is returned, listing all datasets associated with the queried disease. Each entry in the table includes key metadata such as the corresponding GEO accession ID, the biological sample source (WB or PBMC), and annotations indicating whether the dataset includes healthy control samples. The presence of control

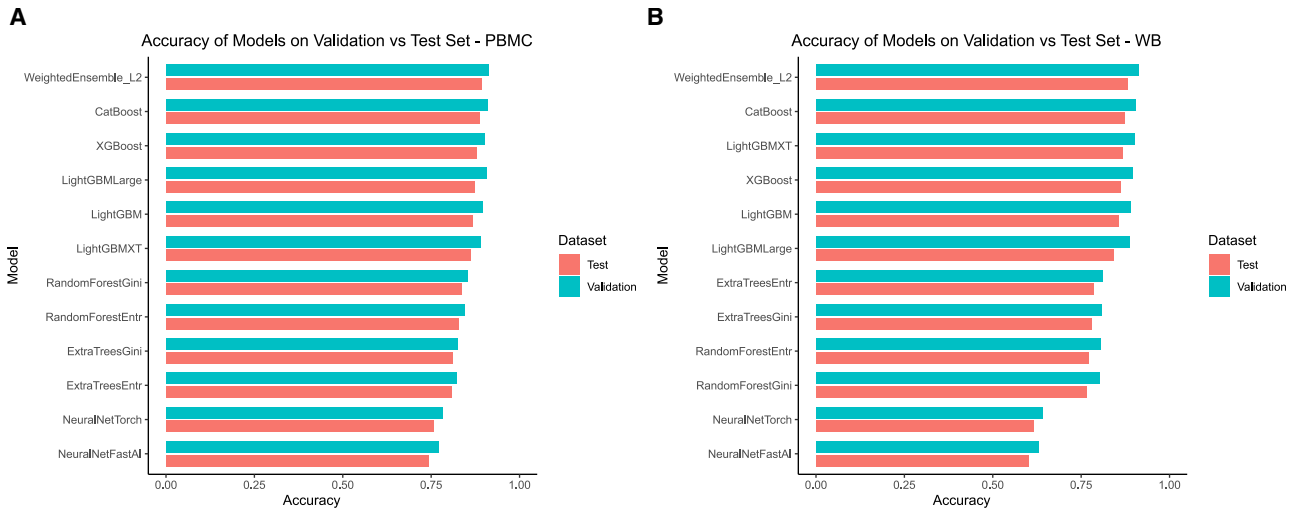


Figure 3. Accuracy comparison of AutoGluon models based on validation vs. test set

(A) PBMC-derived data. (B) WB-derived data. Each bar represents the classification accuracy of a specific model based on the validation vs. the test set.

samples is critical, as it determines the availability of precomputed results from differential expression analysis and functional enrichment. Additionally, the table denotes whether each dataset was incorporated into the final model training process, which is contingent on meeting minimum sample size requirements (i.e., $\geq$10 samples per class).

For example, a user query for the term "Sepsis" retrieves multiple associated datasets that vary in both sample type and metadata completeness. Users can click on the dataset ID in the summary table to access a detailed view of the associated analytical results. This includes a tabular list of DEGs, ranked by statistical significance or fold change, as well as graphical summaries such as volcano plots that visually display upregulated and downregulated genes (Figure 2A). Additional outputs include GO enrichment analysis and KEGG pathway enrichment analysis results (Figure 2B), helping users contextualize the DEGs within relevant biological processes and molecular pathways.

Predict module

The predict module enables real-time disease classification on the basis of RNA-seq gene expression data. Users can upload a count matrix in CSV format, select the appropriate sample type (WB or PBMC), and initiate prediction (Figure 5E). BRPtools then processes the data through a pretrained multiclass classification model, returning predicted disease labels for each sample (Figure 5F). The results can be downloaded or filtered by sample ID. To assess model reliability, users may upload a separate file containing true class labels. The system then computes validation metrics, including the confusion matrix, accuracy, macrorecall, and macroprecision (Figures 5G and 5H). These metrics offer immediate feedback on model performance for a given dataset.

Table 1. Classification performance of the BRPtools WeightedEnsemble_L2 multiclass predictor based on the PBMC and WB test sets

Metric	PBMC	WB
Accuracy	0.9391	0.9216
Macro-precision	0.9425	0.9199
Macro-recall	0.9451	0.9211

DISCUSSION

The primary aim of this study was to develop a scalable, interpretable, and user-accessible platform for transcriptome-based disease prediction using blood RNA-seq data. To this end, we curated and harmonized data from 134 publicly available human RNA-seq datasets, covering 88 distinct blood-related diseases, and integrated them into a unified framework for biomarker exploration and diagnostic inference. By leveraging AutoML and a user-friendly web interface, our goal was to bridge the gap between high-dimensional transcriptomic data and practical, real-world applications in precision medicine.

In this study, we developed and deployed BRPtools, an integrated web-based platform that enables exploration and prediction of human blood diseases on the basis of transcriptomic signatures. By implementing a standardized preprocessing pipeline across all datasets, we created a centralized and reproducible resource for transcriptome-driven biomarker discovery. The platform allows users to explore gene-level signatures interactively and generate disease predictions with minimal computational overhead.

Our diagnostic model is powered by the AutoGluon machine learning framework, an ensemble-based AutoML system that

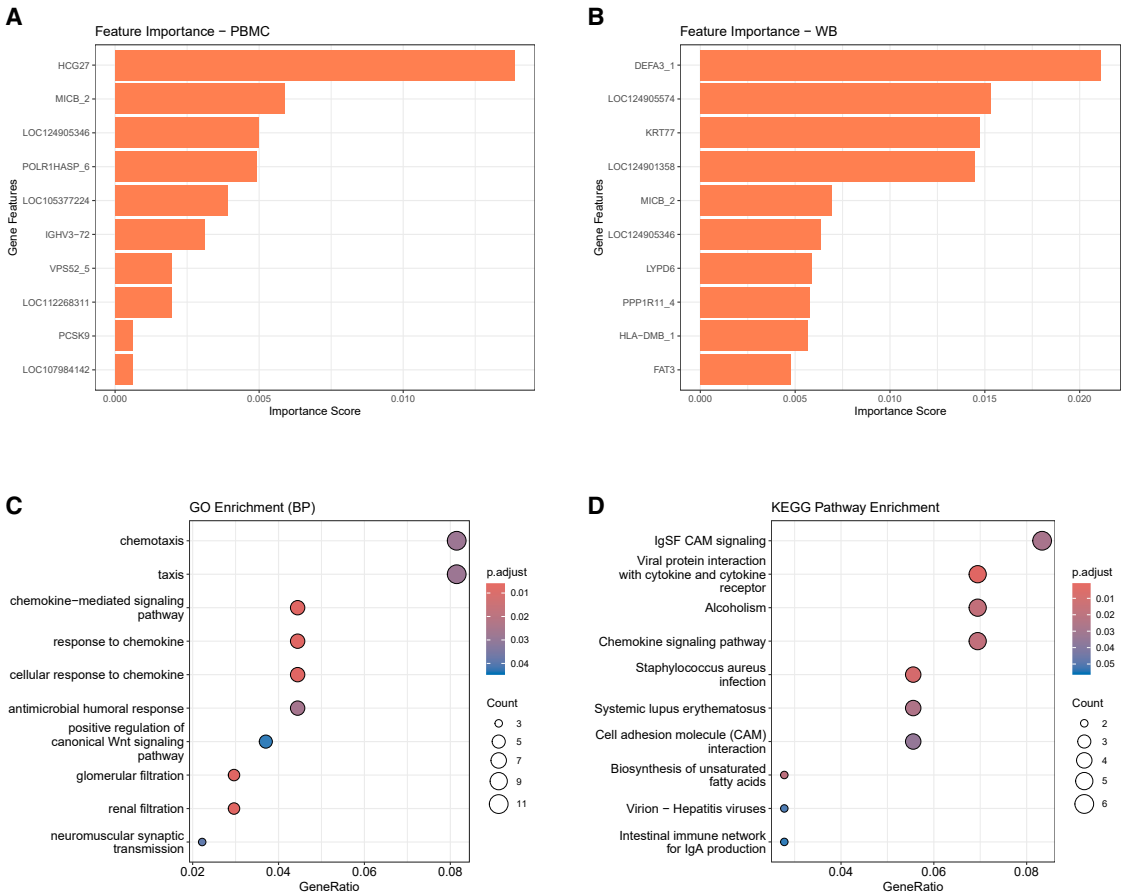


Figure 4. Machine learning feature importance and functional enrichment analysis for disease classification

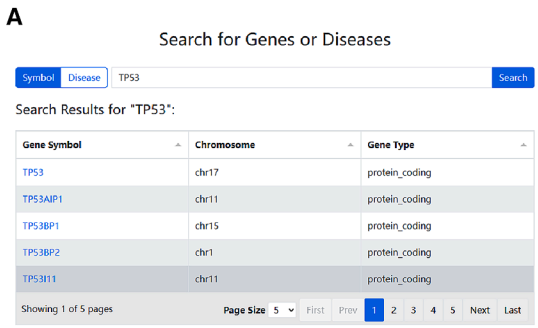
(A and B) Horizontal bar plots showing the top 10 predictive features sorted by their importance scores derived from the machine learning models for (A) peripheral blood mononuclear cell (PBMC) and (B) whole blood (WB) datasets. (C and D) Dot plots displaying the functional enrichment results for the identified top features from the WB model. (C) Top enriched Gene Ontology (GO) terms within the biological process category. (D) Top enriched KEGG pathways.

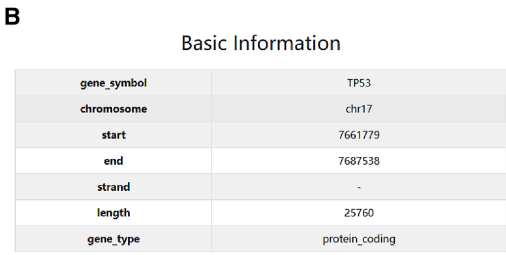
integrates diverse base models through multilayer stacking and automated hyperparameter optimization. This approach offers two major advantages over traditional single-model classifiers. First, it increases robustness to overfitting. Second, it automates the model selection and tuning process, improving the prediction performance while reducing the need for manual intervention.

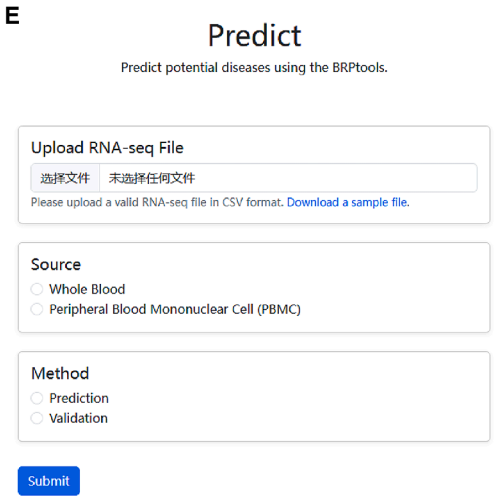
In benchmark evaluations, the AutoGluon-based multiclass classifier achieved high predictive accuracy across a wide range of diseases. This level of performance highlights the viability of AutoML ensemble strategies for large-scale disease prediction tasks, particularly in transcriptomic contexts where platform-specific biases and sample heterogeneity often compromise generalizability.

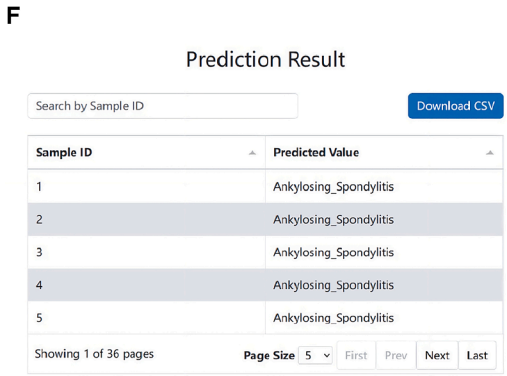
BRPtools addresses a key gap in the bioinformatics ecosystem: the lack of user-friendly platforms that bridge high-dimensional transcriptomic data with interpretable diagnostic tools. While repositories such as GEO offer access to raw and processed RNA-seq data, they typically lack interactive interfaces or embedded analytical capabilities. In contrast, BRPtools incorporates both a search module, which enables gene- and disease-centric exploration, and a predict module, which allows users to upload their own RNA-seq data and obtain disease predictions with optional validation metrics. The platform's intuitive design and web-based deployment make it accessible to a broad range of users, including clinicians, biomedical researchers, and diagnostic developers, even those without bioinformatics training.

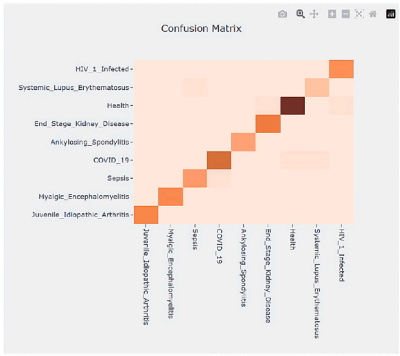
Beyond its technical implementation, the practical utility of BRPtools extends to several high-value clinical and research scenarios where prior diagnostic information is often ambiguous or incomplete. First, the platform facilitates the differential diagnosis of conditions with overlapping systemic presentations. For instance, distinguishing between severe infections (e.g., sepsis) and autoimmune flare-ups remains a significant clinical challenge due to their convergent transcriptomic signatures in blood; BRPtools can provide objective evidence to assist clinicians in narrowing down these possibilities. Second, BRPtools serves as a potent resource for

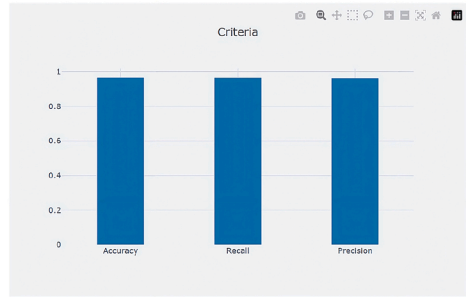












(legend on next page)

cross-disease mechanism discovery. Researchers analyzing uncharacterized or rare syndromes can utilize our platform to identify transcriptionally similar disease categories, offering immediate clues for shared pathophysiological pathways and potential drug repurposing strategies. Third, the multiclass nature of our model enables the detection of comorbidities or secondary complications. In patients with a known primary condition, significant deviations in the prediction probability distribution—where secondary labels show rising confidence—may highlight the onset of concurrent issues, such as secondary infections or immune-related adverse events. By addressing these complex scenarios, BRPtools transitions from a confirmatory tool to a dynamic engine for diagnostic inference and biological hypothesis generation.

Despite these strengths, several limitations warrant discussion. First, the model's predictive accuracy may be compromised in disease categories with limited sample representation, as ensemble models such as AutoGluon still depend on sufficient training data diversity to generalize well across disease classes. A detailed examination of the confusion matrices (Figure S3) reveals that most misclassifications occur between diseases with closely related clinical presentations or shared systemic inflammatory profiles. For example, certain infectious diseases were occasionally misclassified as other febrile illnesses, and some autoimmune conditions showed cross-prediction. These misclassifications suggest that the model is capturing overarching immune signatures that are conserved across similar pathological states. To improve model performance for underrepresented diseases, expanding sample sizes and enhancing class balance in future datasets are critical. Second, although we applied batch correction methods to mitigate technical variation, residual batch effects may still influence model training and prediction, especially when data from heterogeneous sources are integrated. Third, our analysis is restricted to bulk RNA-seq data from WBs and PBMCs, which, while readily accessible, may not capture the full complexity of cellular heterogeneity or early-stage disease signals. Additionally, while our cross-validation demonstrated high internal and cross-cohort generalizability, the platform currently lacks validation on an entirely independent, prospective clinical cohort. Moreover, the platform currently focuses on adult diseases, and its performance in pediatric or rare disease contexts remains to be systematically evaluated.

In the future, we envision several improvements. We plan to expand the underlying dataset library to include additional RNA-seq studies, particularly those related to rare and pediatric diseases. In parallel, we aim to explore alternative classification strategies, including deep learning-based architectures, to complement the current model and potentially enhance prediction performance for complex disease types. In addition, support for single-cell RNA-seq and multi-omics data (e.g., proteomics) will be considered in future versions to broaden the scope and resolution of diagnostic inference.

To facilitate real-world applications, we also plan to implement programmatic access via application programming interfaces (APIs) and provide containerized deployment (e.g., Docker images). Such infrastructure will enable the seamless and secure integration of the BRPtools pipeline into hospital information systems and local research environments. Furthermore, future updates will include exportable analysis reports, enhanced visualization modules, and customizable reference panels to adapt the prediction system to specific clinical or experimental contexts.

In summary, BRPtools offers a robust, extensible, and user-friendly solution for blood transcriptome-based disease prediction. By harmonizing data from 134 RNA-seq datasets across 88 disease types, implementing AutoML-driven modeling with AutoGluon, and deploying these capabilities in an accessible web interface, BRPtools supports real-time diagnostics, interactive data exploration, and biomarker discovery, without requiring programming expertise. With planned expansions into rare diseases, pediatric data, and multi-omics integration, BRPtools further advances the role of transcriptomics in precision diagnostics and translational medicine.

MATERIALS AND METHODS

Data description

All transcriptomic datasets were retrieved from the GEO database. A systematic search was conducted in March 2025 using the following criteria: (1) study type specified as "expression profiling by high-throughput sequencing," (2) tissue source restricted to "blood," and (3) organism designated as "*Homo sapiens*." This yielded a broad pool of candidate datasets, which we then manually curated to ensure quality and consistency.

Datasets were excluded if they (a) lacked sufficient clinical metadata (e.g., disease labels, sample type), (b) employed non-bulk RNA sequencing protocols (e.g., single-cell RNA-seq), or (c) contained fewer than five samples per class. After filtering, we retained 134 datasets, representing two major sample types, WB (n = 97 datasets) and PBMC (n = 38 datasets), with a total of 9,872 samples covering 88 disease types.

Figure 5. Interactive application modules in BRPtools

(A) Gene search results displaying a summary table of gene symbols matching the query, including chromosomal locations and gene types. (B) Basic annotations for the selected genes, including the Ensembl ID, official symbol, and functional category. (C) Boxplot showing the expression levels of the queried genes across different sample types: whole blood (WB) and peripheral blood mononuclear cells (PBMCs). (D) Bar plot representing the relative expression of the gene across multiple disease categories within a selected sample type. (E) User interface of the predict module, which allows users to upload RNA-seq count matrices (CSV format) and specify the sample type (WB or PBMC). (F) Output table displaying the predicted disease labels and confidence scores for each uploaded sample. (G) A confusion matrix is generated when users provide ground-truth labels, visualizing classification accuracy across disease categories. (H) Bar plots showing macroprecision and macrorecall values across all classes.

Metadata were harmonized across datasets, with sample attributes such as diagnosis, accession number, platform, and source (WB or PBMC) recorded in a structured table. These metadata served as the basis for downstream stratification in the training and visualization modules.

RNA-seq data analyses

To minimize technical heterogeneity, we implemented a uniform preprocessing pipeline across all datasets. Raw sequencing data (FASTQ format) were retrieved by converting GEO accession numbers (GSE) to Sequence Read Archive (SRA) run identifiers using the rentrez R package (v1.2.3) and then downloading the corresponding FASTQ files from the European Nucleotide Archive (ENA) via the Aspera ascp transfer protocol for speed and reliability.[13]

The analysis began with a quality control assessment of the raw sequencing data using FastQC (v0.12.1) with MultiQC (v1.27.1) for comprehensive quality reporting.[14,15] Adapter sequences and low-quality reads were trimmed using FastP (v0.24.0).[16] Cleaned reads were aligned to the human reference genome GRCh38.p14 using HISAT2 (v2.2.1).[17] Following alignment, SAM files were converted to the sorted BAM format using SAMtools (v1.21).[18] Gene-level read counts were quantified using featureCounts (v2.0.8), and individual sample count matrices were combined to create a unified expression matrix for downstream analysis.[19]

For differential expression analysis, we implemented an established limma-voom (v3.62.1) and edgeR (v4.4.0) pipeline.[20,21] The analysis proceeded through several key steps: (1) filtering lowly expressed genes using edgeR's filterByExpr() function, (2) normalizing for library size differences using the trimmed mean of M values (TMM) method, (3) modeling mean-variance relationships in the $\log_2$-CPM transformed data using voom(), (4) fitting linear models and applying empirical Bayes moderation, and (5) identifying DEGs at an FDR-adjusted p value threshold of 0.05.

Functional analyses

To explain the biological functions of DEGs, we conducted comprehensive functional enrichment analysis using clusterProfiler (v4.14.0).[22] Gene symbols were first converted to Entrez IDs using the org.Hs.e.g.,.db (v3.20.0) annotation database through AnnotationDbi (v1.68.0).[23,24] We then performed both GO term enrichment and KEGG pathway analysis using the enrichGO() and enrichKEGG() functions, respectively. Significant enrichment results were determined at an FDR < 0.05 with Benjamini-Hochberg adjustment. The top enriched biological terms and pathways were subsequently visualized to facilitate interpretation.

Model construction

To enable accurate and scalable multiclass disease prediction, we implemented an automated machine learning (AutoML) framework using the AutoGluon (v1.3.0) Python package. AutoGluon is a high-level, flexible AutoML toolkit that automates the full pipeline of model selection, hyperparameter optimization, and ensembling. This toolkit supports a broad range of base learners, such as gradient boosting machines, neural networks, and decision trees, and employs a multilayer stacked ensembling strategy to improve predictive robustness. In this architecture, predictions from base-level models are fed into subsequent meta-models (higher stacking layers), which learn to optimally combine the outputs of earlier layers. This ensemble design improves generalizability across noisy or heterogeneous datasets, making it particularly well-suited for transcriptomic data from multiple sources.

Given the heterogeneity across datasets, we applied separate training workflows for the WB and PBMC samples. To ensure model reliability while accommodating rare diseases, only disease categories with at least 10 samples were retained for model development (Table S2). Because this inclusive threshold inherently introduces severe class imbalance, we incorporated sample weighting strategies during the AutoGluon model training phase.[25] By assigning higher weights to minority classes (rare diseases) and lower weights to majority classes, the model penalizes misclassifications of small-sample diseases more heavily, thereby preventing majority-class dominance. Raw gene expression count matrices were first corrected for batch effects using ComBat-Seq from the sva package. This method adjusts for known batch variables while preserving the negative binomial distribution inherent in count data, allowing downstream compatibility with standard RNA-seq modeling techniques.

To evaluate the efficacy of batch-effect removal, we performed PCA for visual assessment and computed the ASW to quantitatively measure the degree of sample mixing across datasets before and after correction.

Following batch correction, the expression matrices were filtered using the filterByExpr() function from the edgeR package to remove genes with consistently low expression across samples. To effectively isolate the most relevant biological signals from high-dimensional transcriptomic noise, we calculated the global variance for each gene and retained the top 3,000 highly variable genes as input features.[26] The remaining genes were normalized using the TMM method via calcNormFactors(), which corrects for compositional differences across libraries. The normalized counts were then converted into transcripts per million (TPM) values to enable comparability across samples with differing library sizes and sequencing depths.

The cleaned and normalized data were then randomly partitioned into training (80%) and testing (20%) subsets, stratified by disease class. All model development, including hyperparameter tuning and 5-fold cross-validation, was restricted to the training set. The independent test set was withheld until final evaluation, ensuring that performance estimates reflect out-of-sample generalizability. The classification accuracy was used as the primary optimization metric during training, but secondary metrics such as macroprecision and macrorecall were also tracked to account for potential class imbalance. Accuracy was defined as the proportion of correctly classified

samples:

$$Accuracy = \frac{TP_1 + TP_2 + \ldots + TP_k}{N} = \frac{\sum_{i=1}^{k} TP_i}{N} \quad \text{(Equation 1)}$$

where k is the number of classes, TP_i is the number of true positive predictions for class k, and N is the total number of samples.

Macroprecision and macrorecall were calculated by averaging the classwise precision and recall, respectively.

$$Macro - Precision = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{TP_i}{TP_i + FP_i}\right) \quad \text{(Equation 2)}$$

$$Macro - Recall = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{TP_i}{TP_i + FN_i}\right) \quad \text{(Equation 3)}$$

where FP_i and FN_i denote the number of false-positive and false-negative predictions for class i, respectively.

While the stratified 5-fold cross-validation provides a robust internal evaluation, random sample splitting across integrated datasets may inadvertently introduce information leakage and yield overoptimistic results due to dataset-specific batch biases. To rigorously assess true cross-cohort generalizability and simulate real-world external validation, we additionally implemented a targeted LODO cross-validation strategy.

Because several disease categories in our curated database are exclusively represented by a single GEO dataset, a global LODO evaluation across all diseases is mathematically infeasible, as leaving out the sole dataset would eliminate that entire class from the training process. To overcome this, we restricted the LODO validation specifically to disease categories that are represented by multiple independent GEO datasets. In each LODO iteration, one entire GEO dataset was strictly held out as an unseen external test set, while the remaining datasets were used to train the AutoGluon models. This approach ensures that the test samples originate from a completely independent laboratory, processing pipeline, and time period, thereby confirming that the models are driven by conserved biological signatures rather than residual technical artifacts.

Model performance was evaluated based on held-out test sets that were completely independent of the training and validation phases. Following model selection, the best-performing predictors—identified as the WeightedEnsemble_L2 models for both the PBMC and WB datasets—were retrained using the entire dataset to maximize their exposure to all available transcriptomic patterns. These finalized, fully trained models were then deployed into the BRPtools platform to support real-time inference.

In the deployed system, uploaded RNA-seq count matrices are processed using the same preprocessing pipeline (filtering, normalization, and TPM conversion) and then passed into the trained model. For each input sample, BRPtools returns the predicted disease class, enabling users to perform rapid and reproducible diagnostics on the basis of transcriptomic data.

To evaluate the contribution of individual genes to the classification, we utilized the built-in feature importance mechanism of the AutoGluon ensemble framework. Feature importance scores were calculated to quantify the impact of each gene on model performance. Functional enrichment analyses (GO and KEGG) were then performed on the top-ranked predictive genes using the clusterProfiler package to interpret the biological context of the identified signatures.

Online prediction platform construction

We implemented the BRPtools web application using the Django (v5.1.6) Python web framework, which is hosted on the Alibaba Cloud. Back-end services handled dataset retrieval, model inference, and result rendering. The input matrices were parsed and validated using pandas, and predictions were performed by calling the pretrained R models.

The front end was built using Bootstrap (v5.3.3) for responsive design. Data tables were rendered with Tabulator (v6.3.0), supporting column filters and fuzzy search. All the interactive plots (e.g., volcano plots and enrichment charts) were generated using Plotly (v2.35.3), ensuring browser compatibility and user-friendly visualization.

DATA AND CODE AVAILABILITY

BRPtools is freely available online at http://www.inbirg.com/blood/home, and there is no login requirement.

The source code for the data processing pipeline and model training is available at https://github.com/Polarixus/BRPtools.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-MSX0289) and the National Natural Science Foundation of China (32470699).

AUTHOR CONTRIBUTIONS

J.P. designed the research. Y.G., F.Y., and X.Z. performed the research. Y.G. wrote the manuscript. J.P. obtained funding and is responsible for this article. Z.Z. provided valuable suggestions. All authors read and approved the manuscript. Y.G. and F.Y. contributed equally to this work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2026.102956.

REFERENCES

1. Harrison, J.E., Weber, S., Jakob, R., and Chute, C.G. (2021). ICD-11: an international classification of diseases for the twenty-first century. BMC Med. Inf. Decis. Making *21*, 206.

2. Fagan, A.M., Shaw, L.M., Xiong, C., Vanderstichele, H., Mintun, M.A., Trojanowski, J.Q., Coart, E., Morris, J.C., and Holtzman, D.M. (2011). Comparison of Analytical Platforms for Cerebrospinal Fluid Measures of β-Amyloid 1-42, Total tau, and P-tau181 for Identifying Alzheimer Disease Amyloid Plaque Pathology. Arch. Neurol. *68*, 1137–1144.
3. Mattsson, N., Cullen, N.C., Andreasson, U., Zetterberg, H., and Blennow, K. (2019). Association Between Longitudinal Plasma Neurofilament Light and Neurodegeneration in Patients With Alzheimer Disease. JAMA Neurol. *76*, 791–799.
4. Janelidze, S., Mattsson, N., Palmqvist, S., Smith, R., Beach, T.G., Serrano, G.E., Chai, X., Proctor, N.K., Eichenlaub, U., Zetterberg, H., et al. (2020). Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. Nat. Med. *26*, 379–386.
5. Wang, H., Li, Y., Ryder, J.W., Hole, J.T., Ebert, P.J., Airey, D.C., Qian, H.-R., Logsdon, B., Fisher, A., Ahmed, Z., et al. (2018). Genome-wide RNAseq study of the molecular mechanisms underlying microglia activation in response to pathological tau perturbation in the rTg4510 tau transgenic animal model. Mol. Neurodegener. *13*, 65.
6. Liu, T., Yu, N., Ding, F., Wang, S., Li, S., Zhang, X., Sun, X., Chen, Y., and Liu, P. (2015). Verifying the markers of ovarian cancer using RNA-seq data. Mol. Med. Rep. *12*, 1125–1130.
7. Kelly, J., Moyeed, R., Carroll, C., Luo, S., and Li, X. (2023). Blood biomarker-based classification study for neurodegenerative diseases. Sci. Rep. *13*, 17191.
8. Sweeney, T.E., Wong, H.R., and Khatri, P. (2016). Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Sci. Transl. Med. *8*, 346ra91.
9. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. Preprint at arXiv. https://doi.org/10.48550/arXiv.2003.06505.
10. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
11. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.
12. Taha, S., Bindayna, K., Aljishi, M., Sultan, A., and Almansour, N. (2025). Transcriptomic Profiling Reveals Distinct Immune Dysregulation in Early-Stage Sepsis Patients. Int. J. Mol. Sci. *26*, 6647.
13. Winter, D.J. (2017). rentrez: An R package for the NCBI eUtils API. https://doi.org/10.7287/peerj.preprints.3179v2.
14. Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data. Babraham.ac.uk. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
15. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.
16. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.
17. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.
18. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
19. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.
20. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
21. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
22. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS *16*, 284–287.
23. Carlson, M. Hs.eg.db. Bioconductor. https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.
24. Pagès, H., Carlson, M., Falcon, S., and Li, N. (2023). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. Bioconductor. https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html.
25. Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., and Wei, G.-W. (2025). A review of machine learning methods for imbalanced data challenges in chemistry. Chem. Sci. *16*, 7637–7658.
26. Castillo-Secilla, D., Gálvez, J.M., Carrillo-Perez, F., Verona-Almeida, M., Redondo-Sánchez, D., Ortuno, F.M., Herrera, L.J., and Rojas, I. (2021). KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. Comput. Biol. Med. *133*, 104387.