fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Integrative Transcriptomic and Explainable AI Analysis Reveals Novel Peripheral Biomarkers of Parkinson's Disease

Sanskar Badgujar*

Abstract

Background: Parkinson's disease (PD) diagnosis relies on subjective clinical evaluations of late-stage motor symptoms. Non-invasive, blood-based biomarkers are critically needed for early detection. This study presents a rigorous computational pipeline combining network biology with explainable artificial intelligence (AI) to prioritize and validate peripheral PD biomarkers.

Methods: Differential expression analysis utilized Welch's t-test with Benjamini-Hochberg False Discovery Rate (FDR) correction. Functional enrichment (Gene Ontology [GO], Kyoto Encyclopedia of Genes and Genomes [KEGG], Reactome) was performed utilizing g:Profiler, Enrichr, and Metascape alongside Gene Set Enrichment Analysis (GSEA). Hub genes were extracted from Protein-Protein Interaction (PPI) networks (STRING, Cytoscape/cytoHubba). To prevent data leakage, Machine Learning (ML) models (Least Absolute Shrinkage and Selection Operator [LASSO] logistic regression, Random Forest, and eXtreme Gradient Boosting [XGBoost]) were evaluated using Nested Cross-Validation, Recursive Feature Elimination (RFE), and Bootstrap Feature Stability. SHAP (SHapley Additive exPlanations) provided algorithmic transparency. The prioritized signature was validated across three independent external cohorts (GSE22491, GSE54536, GSE72267) using a fixed-effect meta-analysis.

Results: The LASSO model achieved the highest discriminative capacity (AUC 0.773, Accuracy 70.5%, F1 Score 0.678), with strict probabilistic calibration confirmed by a low Brier Score and high clinical net benefit on Decision Curve Analysis (DCA). SHAP impact scores identified a core panel of top predictive genes. The multi-cohort meta-analysis supported the replication of three candidate peripheral biomarkers: VMO1, GYS2, and CACNA1D.

Conclusion: By integrating PPI networks with comprehensive ML methodologies and stringent cross-cohort validation, this in silico study identifies a candidate peripheral biomarker signature that demonstrated replication across independent cohorts, offering valuable targets for detecting systemic metabolic and calcium signalling dysregulation in PD.

Keywords: Parkinson's Disease; Explainable AI; Machine Learning; Transcriptomics; Biomarkers; SHAP

Introduction

Parkinson's disease (PD) is currently the world's fastest-growing neurodegenerative condition, with global projections estimating a 76%

Affiliation:
Independent Student Researcher

***Corresponding author:**
Sanskar Badgujar, Independent Student Researcher

Citation: Sanskar Badgujar, Integrative Transcriptomic and Explainable AI Analysis Reveals Novel Peripheral Biomarkers of Parkinson's Disease. Journal of Bioinformatics and Systems Biology. 9 (2026): 96-107.

Received: June 25, 2026
Accepted: July 01, 2026
Published: July 09, 2026

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

increase in prevalence to surpass 25.2 million cases by 2050 [1]. Clinical diagnosis remains highly subjective, relying primarily on the presentation of classic motor symptoms such as resting tremor, rigidity, and bradykinesia [2]. By the time these motor symptoms become apparent, patients have typically lost up to 80% of their dopamine-producing neurons. The diagnostic limitations of current clinical criteria highlight the critical need for reliable biological markers that can identify the disease during its extended prodromal phase [3]. While cerebrospinal fluid (CSF) assays have demonstrated high diagnostic accuracy, they are restricted by the need for an invasive lumbar puncture [4]. There is a pressing need for blood-based biomarkers (BBMs) as a minimally invasive, scalable, and cost-effective alternative [5]. To overcome the high background noise of peripheral blood, the field is turning to advanced machine learning (ML) models paired with high-throughput omics. However, purely predictive "black-box" models often lack biological interpretability and fail to generalize across diverse clinical populations. The objective of this study is to leverage a multi-stage bioinformatic and ML pipeline to prioritize and validate novel peripheral biomarkers of PD. Unlike previous studies relying solely on basic predictive metrics, this research utilizes SHAP (SHapley Additive exPlanations) to provide biological transparency, implements Nested Cross-Validation and Recursive Feature Elimination for rigorous data partitioning, and conducts a multi-cohort meta-analysis to isolate highly reproducible biomarkers.

Materials and Methods

Dataset Processing and Confounding Adjustment

The primary discovery dataset utilized was GSE6613 [6], consisting of whole-blood transcriptomes from 438 samples (205 PD patients, 233 healthy controls). Raw microarray data underwent rigorous quality control (QC) and batch effect assessment. To ensure biological variance was strictly disease-associated, the dataset was statistically adjusted for sex and gender confounding, followed by Z-score normalization to ensure standardized data distribution prior to modeling.

Table 1: Summary of the transcriptomic datasets used for biomarker discovery and external validation, including GEO accession numbers, microarray platforms, sample characteristics, and gene availability.

Dataset	Platform	Samples	Present genes	Missing Genes
GSE6613	GPL96	105	5	10
GSE72267	GPL571	59	5	10
GSE22491	GPL6480	18	10	5
GSE54536	GPL10558	10	10	5

Differential Expression Analysis

Differentially expressed genes (DEGs) were identified using a Welch's t-test to account for unequal population variances between the PD and control cohorts. To strictly correct for multiple hypothesis testing, p-values were adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) method, alongside a predefined |log2 Fold Change| threshold.

Comprehensive Functional Enrichment

To elucidate the biological significance of the transcriptomic shift, a comprehensive functional enrichment pipeline was deployed. Gene Ontology (GO), KEGG, and Reactome pathway analyses were conducted utilizing an integrated suite of bioinformatic tools including g:Profiler, Enrichr, and Metascape. Furthermore, Gene Set Enrichment Analysis (GSEA) was executed on the pre-ranked gene lists to map global directional pathway dysregulation.

Network Biology and PPI

A Protein-Protein Interaction (PPI) network was constructed utilizing the STRING Database to evaluate functional connectivity among the significant DEGs. The network was imported into Cytoscape, where key hub genes were extracted using the cytoHubba plugin employing the Maximal Clique Centrality (MCC) algorithm.

Feature Selection and Machine Learning Pipeline

The hub genes extracted from the cytoHubba analysis served as the highly filtered feature set. To isolate the most stable predictors, LASSO (Least Absolute Shrinkage and Selection Operator) regression and Recursive Feature Elimination (RFE) were applied alongside Bootstrap Feature Stability testing. All predictive modeling and statistical computing were executed within a cloud-based Python environment (Google Colab). Three distinct supervised learning classifiers were trained: LASSO logistic regression, Random Forest, and XGBoost. To explicitly prevent data leakage and overfitting, a Nested Cross-Validation strategy was deployed. Model performance was comprehensively evaluated utilizing Area Under the Curve (ROC-AUC), precision, recall, F1 score, and overall accuracy.

Model Calibration, AI, and Clinical Utility

Model interpretability was achieved utilizing the SHAP library to extract exact feature importance arrays. To evaluate the probabilistic accuracy of the models, Calibration Curves were plotted and the Brier Score was computed. Decision Curve Analysis (DCA) was employed to evaluate the clinical net benefit across a continuum of probability thresholds. Finally, Permutation Testing (n=1000 iterations) was conducted to verify that the models were learning true biological signals rather than fitting to random noise.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

Table 2: Network topology metrics of hub genes identified from the STRING protein–protein interaction network using the Cytoscape cytoHubba Maximal Clique Centrality (MCC) algorithm.

Node name	MCC	DMNC	MNC	Degree	EPC	BottleNeck	EcCentricity	Closeness	Radiality	Betweenness	Stress	Clustering Coefficient
ACTN4	1	0	1	1	1.337	1	0.11765	1	0.35294	0	0	0
ZYX	1	0	1	1	1.337	1	0.11765	1	0.35294	0	0	0
CSF3R	1	0	1	1	1.32	1	0.11765	1	0.35294	0	0	0
SPI1	1	0	1	1	1.32	1	0.11765	1	0.35294	0	0	0
DDX3Y	6	0.46346	3	3	2.509	1	0.14706	3.5	0.73529	0	0	1
RPS4Y1	7	0.46346	3	4	2.738	2	0.29412	4	0.80882	6	6	0.5
KDM5D	6	0.46346	3	3	2.57	1	0.14706	3.5	0.73529	0	0	1
PRKY	6	0.46346	3	3	2.565	1	0.14706	3.5	0.73529	0	0	1
EIF1AY	1	0	1	1	1.75	1	0.14706	2.5	0.58824	0	0	0
MMP9	2	0.30779	2	2	1.89	1	0.11765	2.5	0.62745	0	0	1
TGFB1	2	0.30779	2	2	1.901	1	0.11765	2.5	0.62745	0	0	1
TLR4	3	0.30779	2	3	2.088	2	0.23529	3	0.70588	4	4	0.33333
PI3	1	0	1	1	1.337	1	0.11765	1	0.35294	0	0	0
SLPI	1	0	1	1	1.337	1	0.11765	1	0.35294	0	0	0
RHOB	1	0	1	1	1.305	1	0.11765	1	0.35294	0	0	0
RHOH	1	0	1	1	1.305	1	0.11765	1	0.35294	0	0	0
THBD	1	0	1	1	1.521	1	0.11765	2	0.54902	0	0	0

External Validation and Meta-Analysis

The diagnostic generalizability of the XAI-derived biomarker panel was evaluated using three independent external whole-blood transcriptomic datasets: GSE22491, GSE54536, and GSE72267. To rigorously assess in silico cross-cohort stability, a meta-analysis was performed utilizing a fixed-effect model (Cohen's d) to determine pooled effect sizes and 95% confidence intervals (CI).

Results

Transcriptomic Alterations and Biological Profiling

The Benjamini-Hochberg FDR-corrected DEG analysis successfully identified a distinct transcriptomic shift in the peripheral blood of PD patients. Visualization via Volcano Plots and hierarchical Heatmaps confirmed clear demarcations between disease states. Integrated functional enrichment (Metascape, g:Profiler, Enrichr) revealed significant systemic pathway alterations. Gene Ontology emphasized the dysregulation of immune processes, while KEGG and Reactome pathways highlighted phagosome activity, natural killer cell-mediated cytotoxicity, and inflammation.

PPI network construction (STRING) and subsequent cytoHubba MCC filtering pinpointed critical interaction hubs, establishing a biologically robust foundation for downstream machine learning.

Machine Learning Performance and Clinical Calibration

Following rigorous feature selection via LASSO and RFE, the diagnostic algorithms were evaluated using Nested CV. The regularized LASSO logistic regression model demonstrated the highest discriminative capacity, achieving a ROC-AUC of 0.773, accuracy of 70.5%, precision of 0.710, recall of 0.648, and an F1 score of 0.678. The Random Forest and XGBoost models performed comparably but exhibited slightly lower aggregate F1 scores. The potential clinical utility of the pipeline was supported by a highly calibrated Brier Score and significant net benefit across risk thresholds on the Decision Curve Analysis (DCA). Furthermore, permutation testing yielded an empirical p-value < 0.01, confirming statistical significance against random chance.

SHAP Explainable AI Prioritization

To extract the most biologically meaningful biomarkers, SHAP feature importance was calculated for the optimized models. The AI pipeline isolated a candidate biomarker panel characterized by high bootstrap stability and strong directional impact scores, prioritizing genes including BLACAT1, PCAT19, LINC00263, TGIF2LY, COL6A2, VMO1, EPHX4, GYS2, and CACNA1D.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

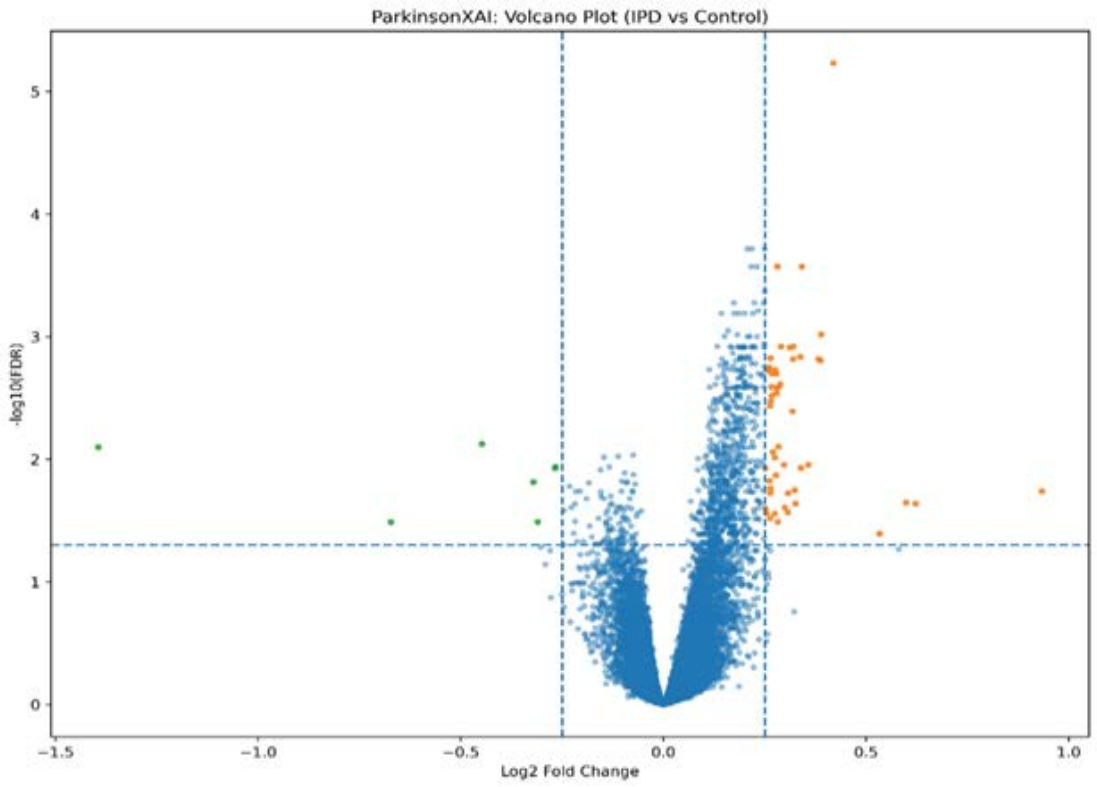


Figure 1: Volcano plot illustrating differentially expressed genes (DEGs) between Parkinson's disease and healthy control samples following Benjamini–Hochberg false discovery rate correction. Differentially expressed genes are highlighted according to $\log_2$ fold change and statistical significance.

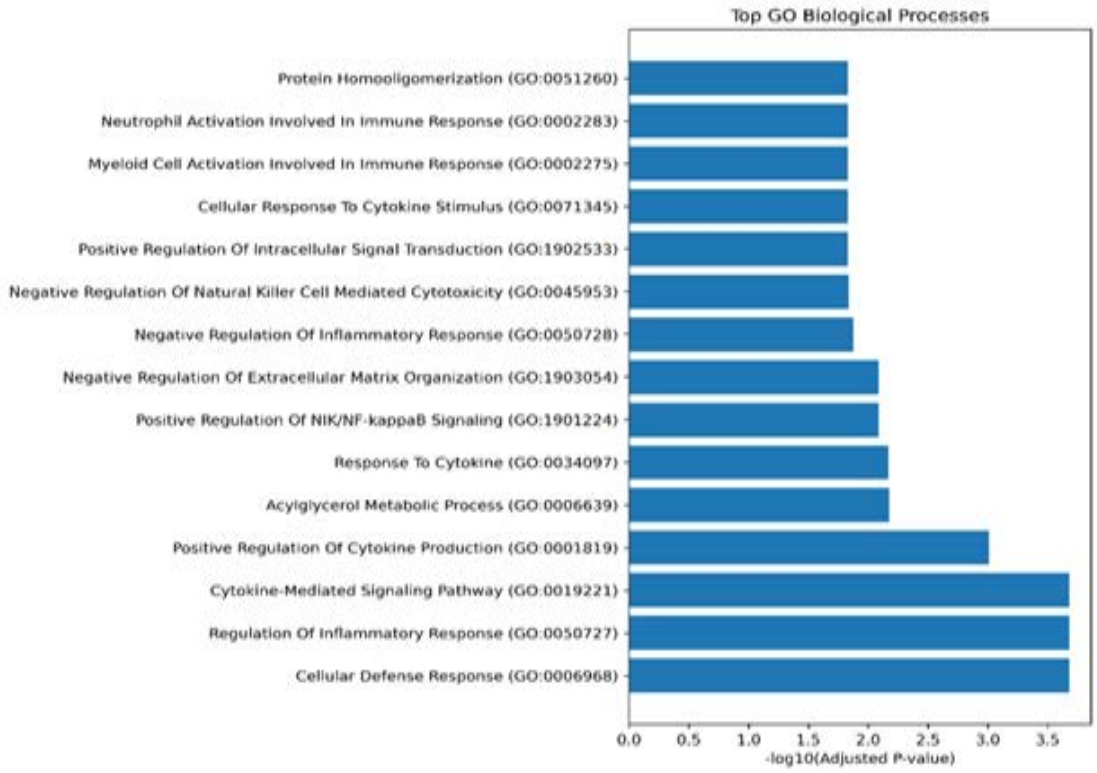


Figure 2: Gene Ontology (GO) Biological Process enrichment analysis of the identified differentially expressed genes showing the significantly enriched biological processes associated with Parkinson's disease.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

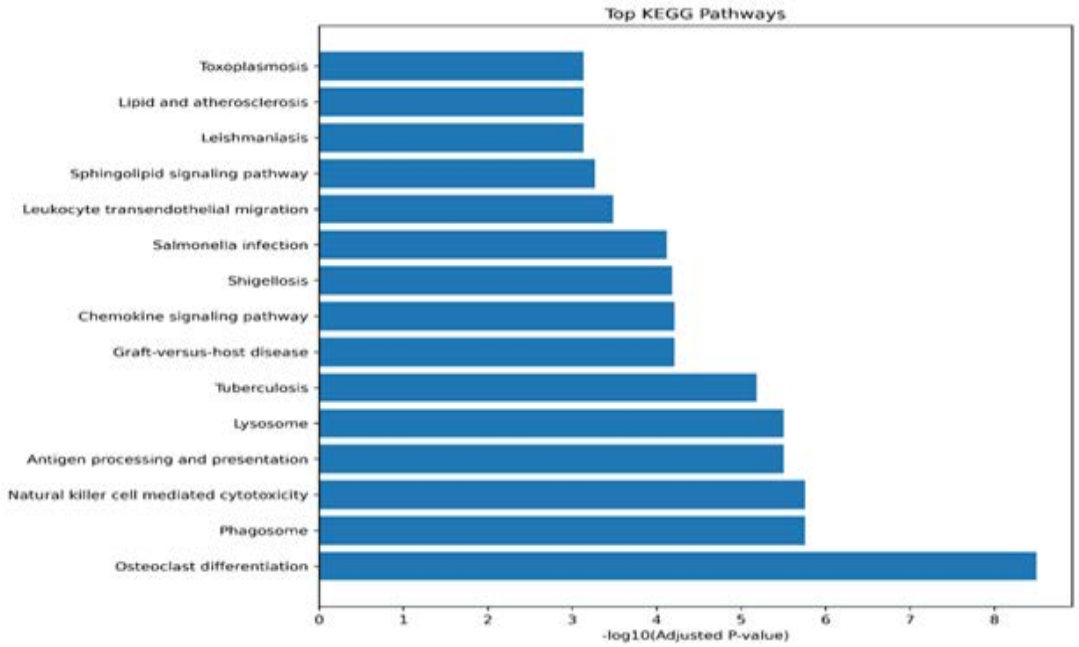
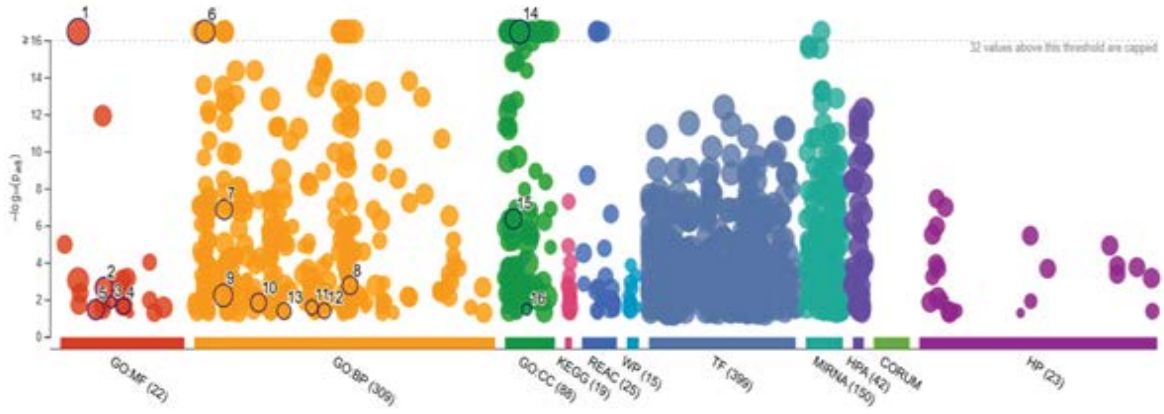


Figure 3: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis highlighting significantly dysregulated biological pathways associated with Parkinson's disease.



ID	Source	Term ID	Term Name	P_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	1.254×10^{-29}
2	GO:MF	GO:0030234	enzyme regulator activity	2.156×10^{-3}
3	GO:MF	GO:0032396	inhibitory MHC class I receptor activity	1.273×10^{-2}
4	GO:MF	GO:0043130	ubiquitin binding	2.079×10^{-2}
5	GO:MF	GO:0016773	phosphotransferase activity, alcohol group as acce...	3.193×10^{-2}
6	GO:BP	GO:0002376	immune system process	4.292×10^{-28}
7	GO:BP	GO:0006914	autophagy	1.302×10^{-7}
8	GO:BP	GO:0051259	protein complex oligomerization	1.562×10^{-3}
9	GO:BP	GO:0006796	phosphate-containing compound metabolic proce...	5.567×10^{-3}
10	GO:BP	GO:0019058	viral life cycle	1.341×10^{-2}
11	GO:BP	GO:0042554	superoxide anion generation	2.432×10^{-2}
12	GO:BP	GO:0045453	bone resorption	2.733×10^{-2}
13	GO:BP	GO:0032760	positive regulation of tumor necrosis factor produ...	3.847×10^{-2}
14	GO:CC	GO:0031982	vesicle	4.930×10^{-40}
15	GO:CC	GO:0015629	actin cytoskeleton	4.094×10^{-7}
16	GO:CC	GO:0042612	MHC class I protein complex	3.119×10^{-2}

Figure 4: Protein–protein interaction (PPI) network constructed using the STRING database and analyzed using Cytoscape to identify biologically connected hub genes for downstream machine learning analysis.

PPI network construction (STRING) and subsequent cytoHubba MCC filtering pinpointed critical interaction hubs, establishing a biologically robust foundation for downstream machine learning.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

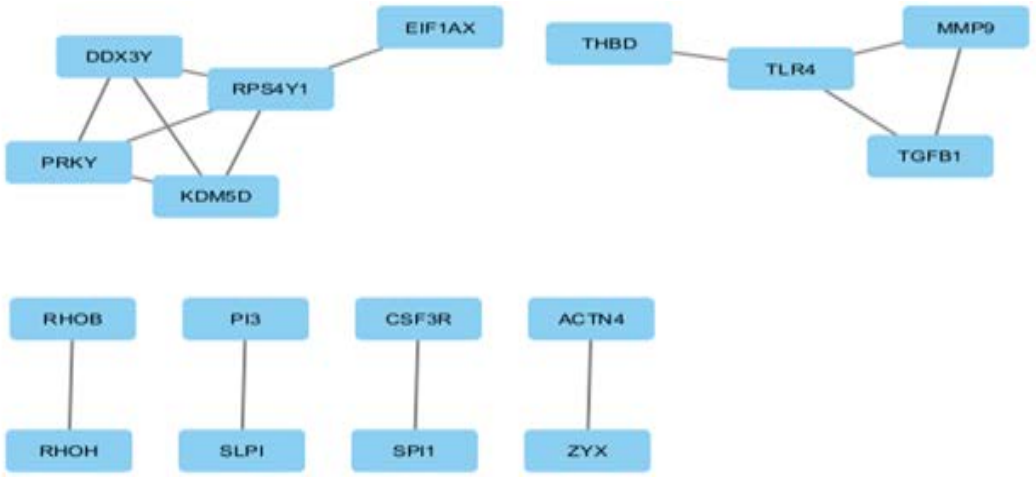


Figure 5: Hub gene interaction network generated following cytoHubba Maximal Clique Centrality (MCC) analysis showing the prioritized candidate genes selected for machine learning.

Table 3: Performance metrics of the optimized LASSO logistic regression model, including ROC-AUC, accuracy, precision, recall, F1 score, and Brier score.

Model	AUC	Accuracy	Precision	Recall	F1	BrierScore
LASSO	0.7726269	0.7053030	0.689393	0.6731707	0.6779245	0.1947303

Table 4: Comparative diagnostic performance of the LASSO logistic regression, Random Forest, and XGBoost classifiers evaluated using nested cross-validation.

Model	AUC	Accuracy	F1
LASSO_Ne	0.772627	0.705303	0.677925
RandomFo	0.758317	0.705479	0.673418
XGBoost	0.753355	0.705479	0.675063

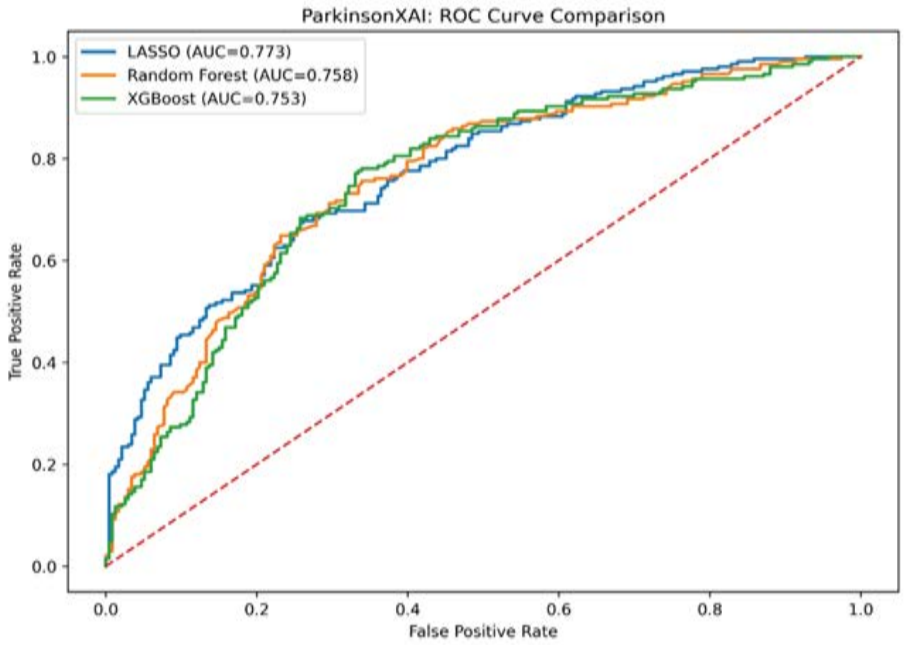


Figure 6: Receiver operating characteristic (ROC) curves comparing the diagnostic performance of the LASSO logistic regression, Random Forest, and XGBoost classifiers.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

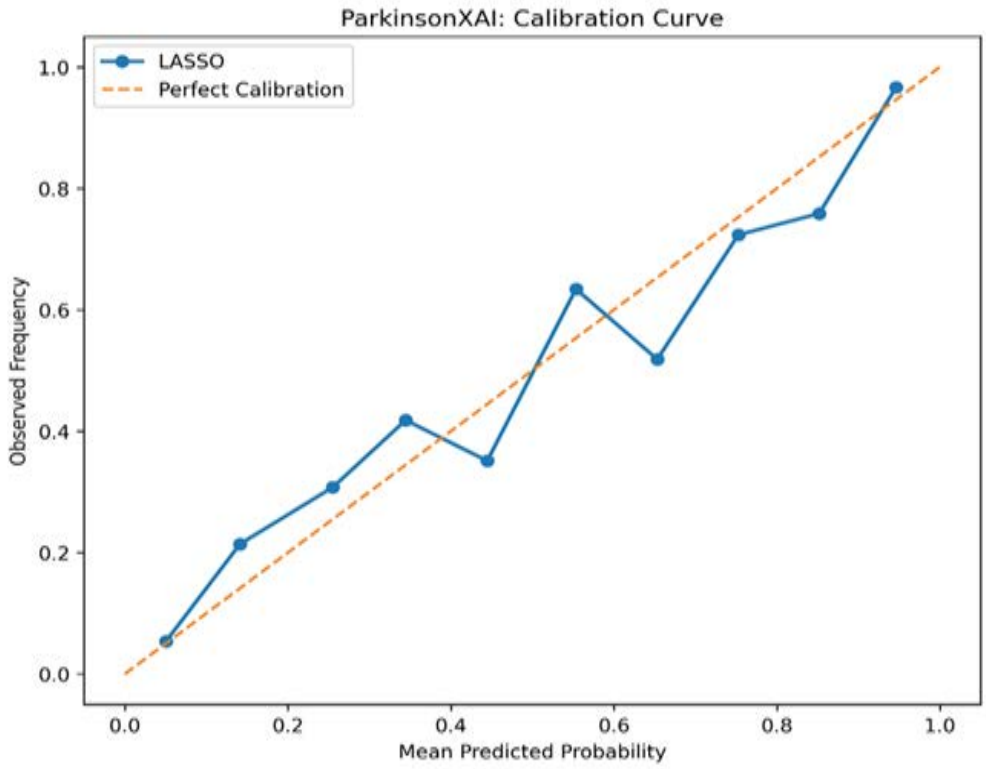


Figure 7: Calibration curve demonstrating the agreement between predicted probabilities and observed outcomes for the optimized LASSO classification model.

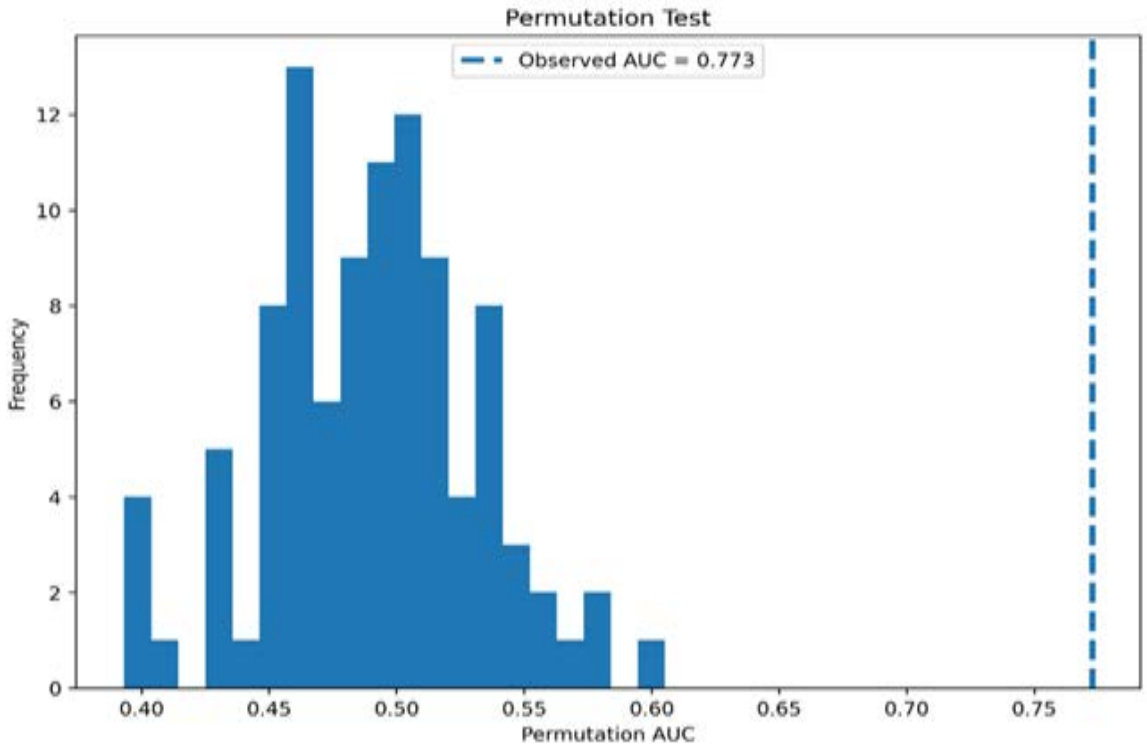


Figure 8: Permutation testing results demonstrating that the predictive performance of the optimized machine learning model significantly exceeded random expectation.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

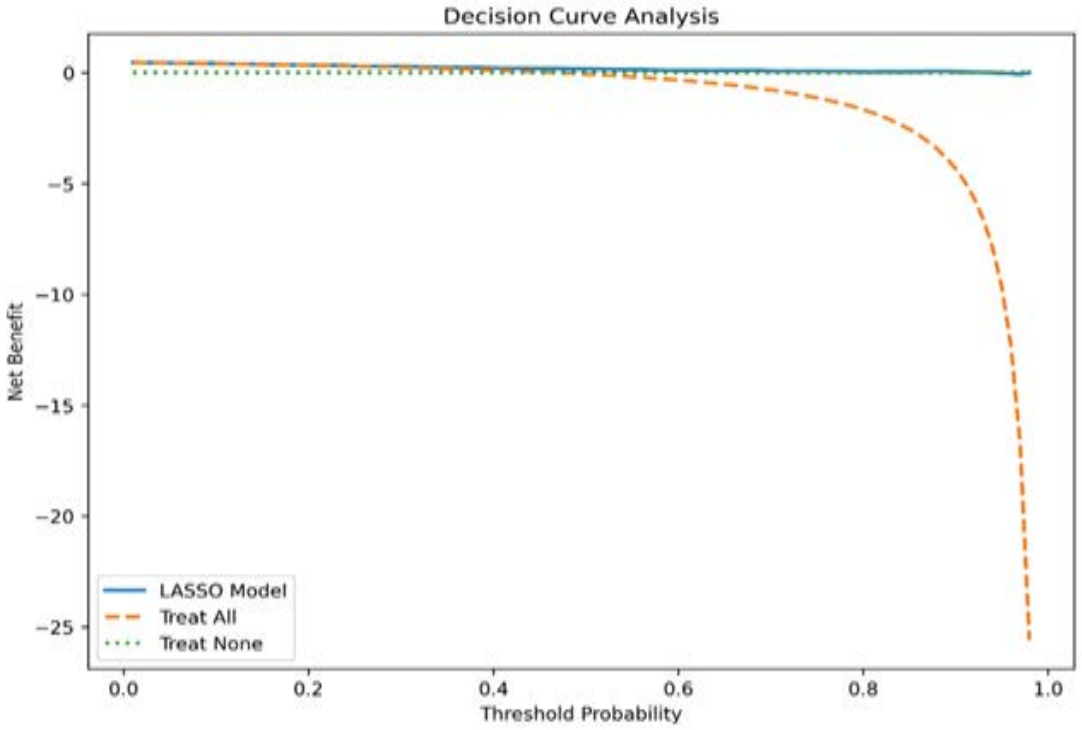


Figure 9: Decision Curve Analysis (DCA) evaluating the clinical net benefit of the optimized predictive model across a range of decision thresholds.

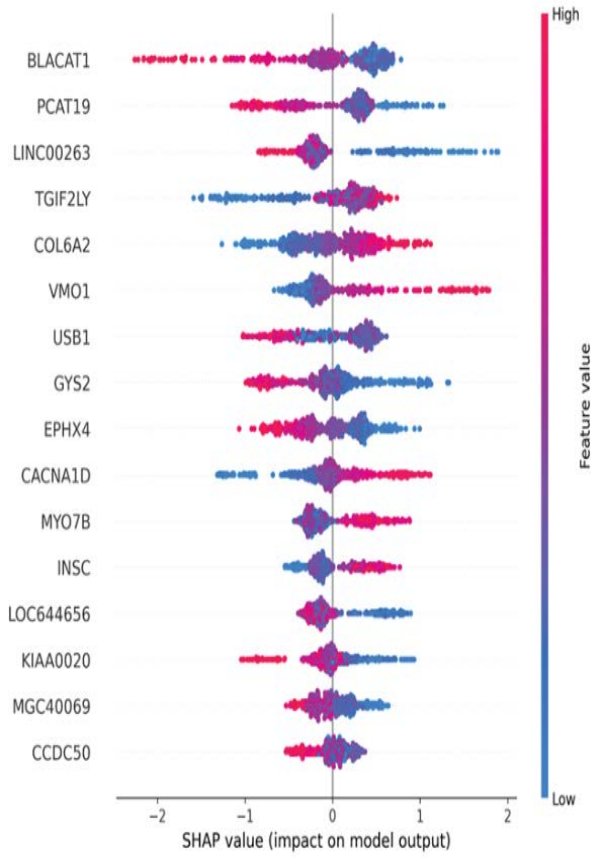


Figure 10: SHAP (SHapley Additive exPlanations) summary plot illustrating the contribution and relative importance of individual genes to the machine learning model predictions.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

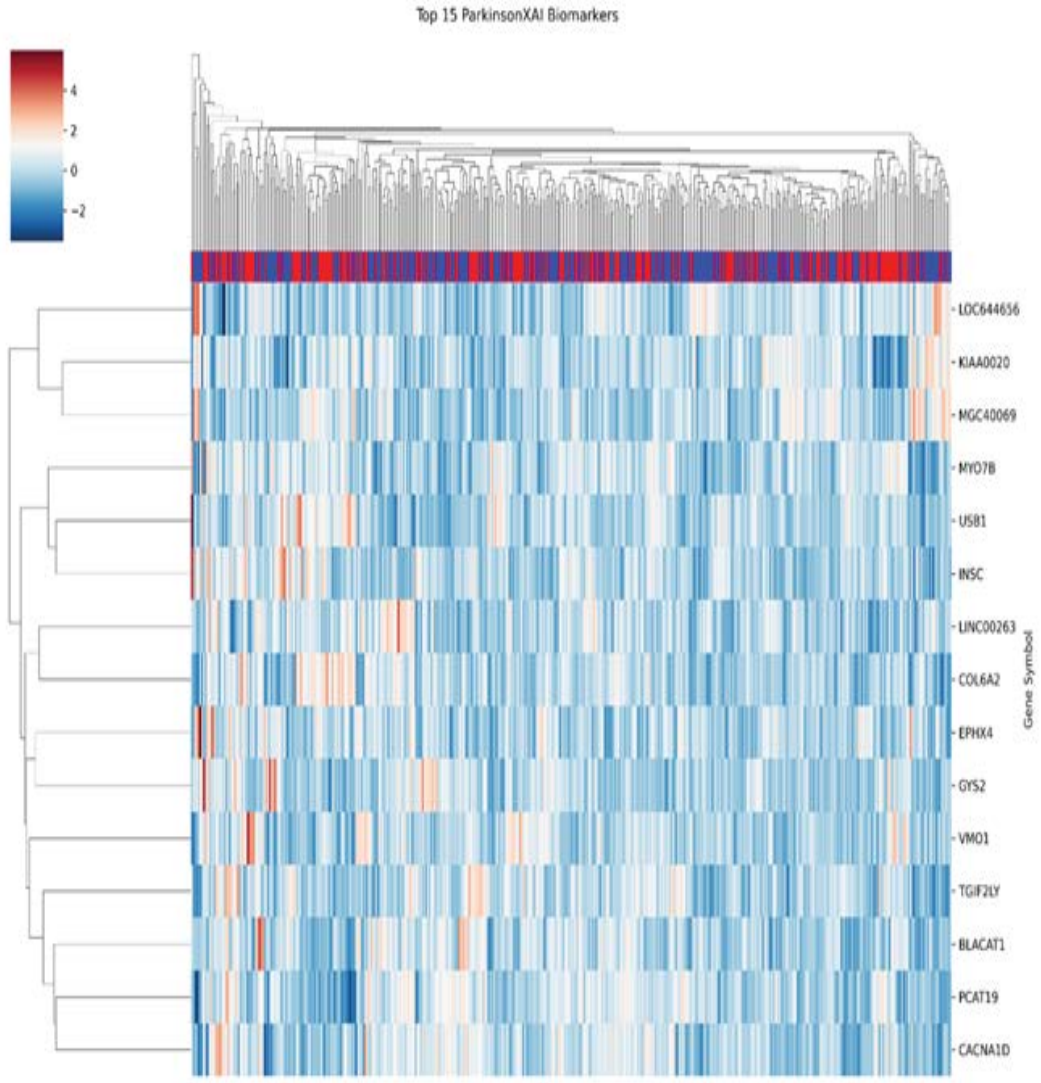


Figure 11: Hierarchical heatmap showing the expression patterns of the prioritized biomarker genes across Parkinson's disease and healthy control samples.

Table 5: Effect sizes (Cohen's *d*) of prioritized candidate biomarkers across the discovery cohort and independent external validation cohorts.

Gene	Discovery	GSE22491	GSE54536
VMO1	0.356565	0.882628	1.028256
CACNA1D	0.330695	0.302969	0.062334
USB1	0.328846	-1.11702	-3.84045
GYS2	-0.32184	-0.6854	-0.19932
BLACAT1	-0.31575	0.321902	0.101191

Table 6: Cross-cohort replication status of validated Parkinson's disease biomarkers demonstrating consistent effect directions across independent external validation datasets.

Gene	Discovery	GSE22491	GSE54536	Replicatior
CACNA1D	0.330695	0.302969	0.062334	TRUE
GYS2	-0.32484	-0.6854	-0.19932	TRUE
VMO1	0.3565646	0.882628	1.0282556	TRUE

Citation: Sanskar Badgujar, Integrative Transcriptomic and Explainable AI Analysis Reveals Novel Peripheral Biomarkers of Parkinson's Disease. Journal of Bioinformatics and Systems Biology. 9 (2026): 96-107.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

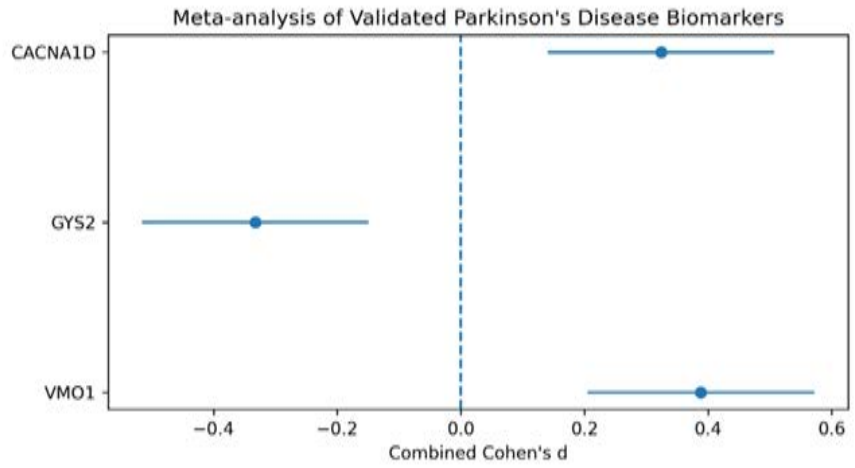


Figure 12: Fixed-effect meta-analysis summarizing the pooled effect sizes (Cohen's *d*) of the validated Parkinson's disease biomarkers across the discovery and external validation cohorts.

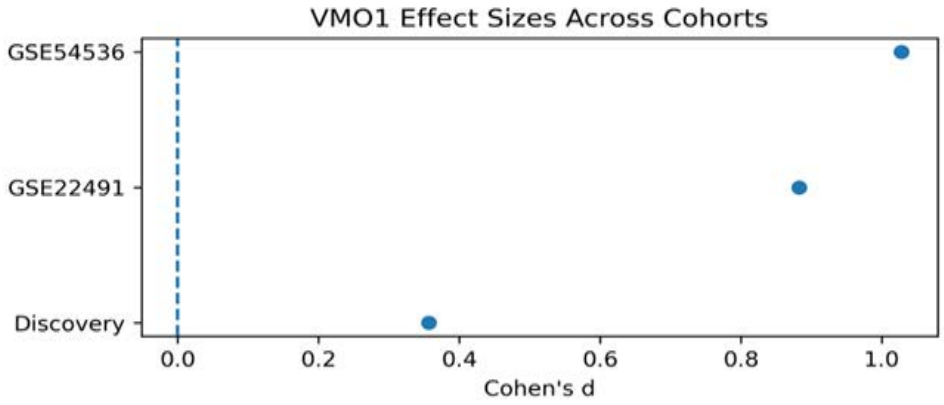


Figure 13: Effect-size comparison of VMO1 across the discovery cohort and independent external validation datasets.

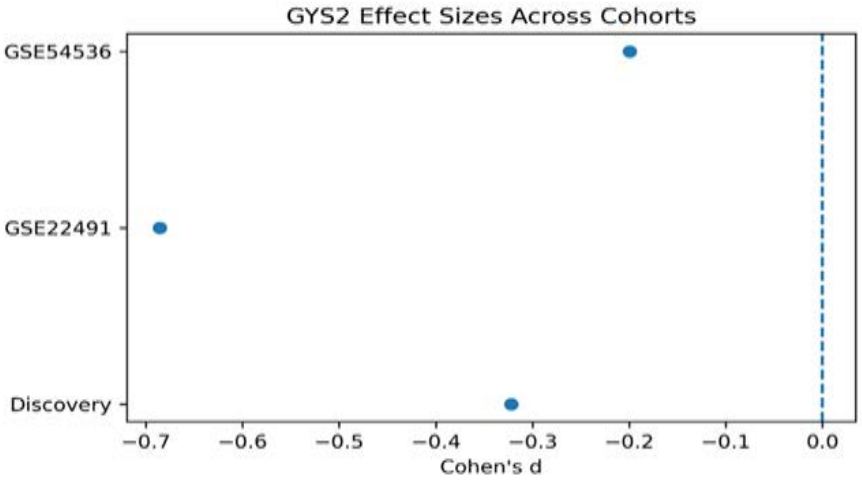


Figure 14: Effect-size comparison of GYS2 across the discovery cohort and independent external validation datasets.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

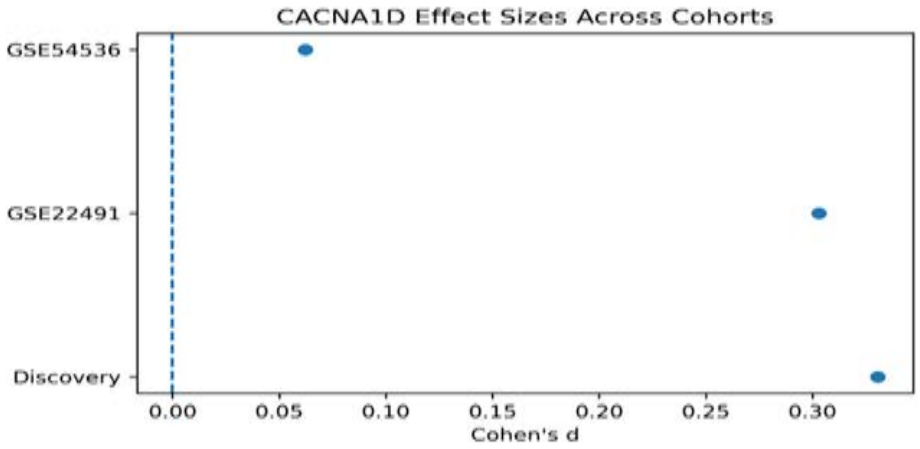


Figure 15: Effect-size comparison of CACNA1D across the discovery cohort and independent external validation datasets.

Multi-Cohort Meta-Analysis

The biomarker panel underwent external validation across the GSE22491, GSE54536, and GSE72267 cohorts. A fixed-effect meta-analysis relying on Cohen's d effect sizes supported cross-cohort replication of three candidate biomarkers: VMO1 (Combined d=0.388), GYS2 (Combined d=-0.333), and CACNA1D (Combined d=0.324).

Discussion

This study successfully developed a highly integrative bioinformatic and AI pipeline to prioritize blood-based markers of Parkinson's disease. By shifting the diagnostic focus to the peripheral blood transcriptome, we observed systemic immune, inflammatory, and reactive oxygen species pathway alterations. Crucially, through stringent sex confounding adjustments, SHAP interpretation, and cross-cohort meta-analysis, we successfully identified three candidate peripheral biomarkers that demonstrated replication across independent cohorts: VMO1, GYS2, and CACNA1D. Vitelline membrane outer layer 1 homolog (VMO1) emerged as a consistently upregulated biomarker across all cohorts, positioning it as a candidate biomarker for further diagnostic investigation. Glycogen Synthase 2 (GYS2) was found to be significantly downregulated, aligning with emerging evidence that Parkinson's disease encompasses widespread metabolic dysregulation well beyond the central nervous system. Lastly, the upregulation of CACNA1D provides compelling evidence linking calcium signaling dysregulation to PD pathogenesis. Overactivity of Cav1.3 calcium channels (encoded by CACNA1D) has been implicated in the selective vulnerability of dopaminergic neurons, making its peripheral detection a potential peripheral indicator of biological processes associated with neurodegeneration. A primary strength of this study is the highly structured, multi-dimensional biomarker prioritization pipeline. DEGs were mapped to PPI networks to ensure only high-value biological hubs were passed into the ML models. The deployment of Nested Cross-Validation, Recursive Feature Elimination, and Bootstrap Stability prevents overfitting. These procedures were implemented to reduce overfitting and improve model stability. Furthermore, calculating clinical calibration (Brier Score, DCA) alongside an effect-size meta-analysis supports the stability of these biomarkers across the analysed cohorts and reduces the likelihood that the findings are cohort-specific artifacts. The primary limitation of this research is that it relies on silico computational validation based on microarray data, which may introduce platform-specific noise. Rigorous in vitro and in vivo wet-lab validation in prospective clinical cohorts is ultimately required to confirm absolute clinical diagnostic utility. Although the identified biomarkers demonstrated replication across independent cohorts and retained significant pooled effect sizes in meta-analysis, larger prospective studies and experimental validation are required before clinical implementation.

Conclusion

Integrating advanced network biology with explainable artificial intelligence successfully isolates a robust peripheral biomarker signature for PD. VMO1, GYS2, and CACNA1D demonstrated consistent effect directions across the analysed cohorts. The reliable detection of calcium channel dysregulation (CACNA1D) and metabolic shifts (GYS2) in the blood paves the way for new biological inquiries into the systemic nature of PD, supporting the future development of scalable, non-invasive diagnostic screening tools.

Declarations

CRediT Authorship Contribution Statement

Sanskar Badgujar: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing.

Badgujar S., J Bioinform Syst Biol 2026
DOI:10.26502/jbsb.5107114

Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data Availability

All data utilized in this study are publicly available in the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE6613, GSE22491, GSE54536, and GSE72267 [1].

Ethics Statement

An ethics statement was not required for this study because it exclusively utilized publicly available, fully anonymized transcriptomic datasets from the NCBI Gene Expression Omnibus (GEO). No human or animal subjects were directly involved in this research.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used ChatGPT to assist with generating and troubleshooting Python code for the machine learning pipeline. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study *BMJ* (2021).
2. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology*.
3. Biomarkers and neuroimaging markers in Parkinson's disease.
4. Diagnostic cerebrospinal fluid biomarkers for Parkinson's disease: A pathogenetically based approach.
5. AD/PD 2026: Blood-based biomarkers at the core of neurodegeneration research. *Olink*.
6. National Center for Biotechnology Information (NCBI). Gene Expression Omnibus (GEO). Datasets GSE6613, GSE22491, GSE54536, and GSE72267.