Universidad Europea

Máster en Bioinformática

Análisis de firmas moleculares para la estratificación y la predicción en la enfermedad de Parkinson

Autor: Santos Antequera Fernández

Tutor: Jordi Martorell Marugán

Curso 2024-2025

AGRADECIMIENTOS

Quiero agradecer y dedicar este trabajo a todas las personas que han estado a mi lado durante su realización.

A mis padres, Santos y Manuela, por haberme dado la oportunidad de llegar hasta aquí, por su apoyo incondicional y por seguir animándome incluso en los momentos más complicados (y por no perder la fe en que todo este esfuerzo sirva para algo).

A mi familia, y en especial a mis abuelos, tanto a los que están como a los que ya se han ido, con el deseo de que investigaciones como esta puedan contribuir algún día a mejorar su bienestar y su salud.

A mis amigos, por su compañía, por los ánimos en los días más duros y por recordarme que siempre se puede seguir adelante.

A mi pareja, Leila, por estar a mi lado en cada momento, por su paciencia infinita y por conseguir que todo pareciera un poco más fácil cuando el trabajo se hacía cuesta arriba, incluso cuando parecía que no había mucha salida.

Y también a mi gatita, que en algún momento intentó sabotearme el trabajo pulsando justo la tecla que no debía, pero que, aun así, consiguió hacerme sonreír en los momentos de estrés.

Gracias a todos, y especialmente a los profesores del máster y a mi tutor, por su guía, su apoyo y por brindarme la oportunidad de realizar este estudio.

En memoria de mi abuelo Juan Antequera Ballesteros, que padeció Parkinson y a quien dedico con todo mi cariño este trabajo.

RESUMEN

El presente trabajo tiene como objetivo explorar la aplicabilidad de diferentes estrategias de pathway scoring y técnicas de machine learning para la identificación de firmas moleculares en sangre asociadas a la enfermedad de Parkinson idiopática (IPD), utilizando datos del estudio original de Shamir et al. (2017). Se evaluó el rendimiento de modelos supervisados para la clasificación de pacientes IPD frente a controles, así como la coherencia biológica y la capacidad de estratificación no supervisada de los perfiles moleculares derivados.

Objetivos:

Evaluar y comparar el desempeño de múltiples métodos de cálculo de scores funcionales (singscore, Z-score, GSVA, ssGSEA, PLAGE y norm_FGSEA) combinados con distintas bases de datos (KEGG, Reactome, GO:BP, GO:MF, GO:CC, DisGeNET y HPO), mediante el entrenamiento de modelos de clasificación binaria e identificación de patrones moleculares característicos en pacientes con Parkinson idiopático. Asimismo, analizar la existencia de posibles subgrupos moleculares mediante clustering no supervisado (M3C).

Material y métodos:

Los datos de expresión génica del conjunto GSE99039 se procesaron para generar scores moleculares por vía biológica y base de datos funcional. Se aplicó un filtrado por varianza, conservando los 750 scores con más varianza por combinación. Para la clasificación IPD vs CONTROL, se empleó pathMED con un esquema de validación cruzada anidada: el bucle interno optimizó los hiperparámetros y el externo estimó las métricas finales, calculadas como la media de los resultados obtenidos en los conjuntos de prueba. Se entrenaron los algoritmos Random Forest, KNN, XGBoost, GLM y LDA, evaluando su rendimiento mediante Accuracy, Balanced Accuracy, MCC, Recall, Specificity, Precision y F1.

Además, se realizaron visualizaciones complementarias (curvas ROC y PR, matrices de confusión, calibration plots, heatmaps de scores y gráficos de importancia de rutas). Finalmente, se aplicó el algoritmo M3C para identificar subtipos moleculares y evaluar su coherencia clínica.

Resultados:

Los modelos mostraron métricas de rendimiento moderadas (Accuracy, Balanced Accuracy, Precision y F1 entre 0,5 y 0,75), coherentes con la variabilidad observada en el análisis PCA. La combinación ssGSEA–GO:CC con XGBoost alcanzó el mejor desempeño global (Accuracy = 0,670; Balanced Accuracy = 0,665; MCC = 0,670), aunque las diferencias entre algoritmos fueron limitadas. El clustering no supervisado reveló agrupamientos consistentes entre muestras IPD y controles, y la caracterización funcional de las rutas más relevantes señaló procesos relacionados con metabolismo

mitocondrial, degradación proteica y respuesta inflamatoria, en concordancia con la fisiopatología de la enfermedad.

Conclusiones:

El uso combinado de pathway scoring y modelos de machine learning permite capturar señales moleculares sutiles asociadas al Parkinson idiopático, aunque con una capacidad predictiva moderada. La integración de múltiples bases de datos funcionales y la paralelización de tareas facilitaron un análisis exhaustivo y eficiente. Los resultados apoyan la relevancia de rutas metabólicas y proteostáticas en la patogénesis de la enfermedad, y sugieren que estrategias multiómicas y cohortes más amplias podrían mejorar la robustez de las clasificaciones moleculares.

Palabras clave:

Machine learning, Parkinson, molecular scoring, clasificación supervisada, clustering no supervisado

ABSTRACT

This study aimed to explore the applicability of different pathway scoring strategies and machine learning approaches for identifying blood-based molecular signatures associated with idiopathic Parkinson's disease (IPD), using data from the original study by Shamir et al. (2017). The main objectives were to evaluate the performance of supervised classification models distinguishing IPD patients from controls and to assess the biological coherence and unsupervised stratification potential of the derived molecular profiles.

Objectives:

To compare the performance of several functional scoring methods (singscore, Z-score, GSVA, ssGSEA, PLAGE, and norm_FGSEA) combined with different functional databases (KEGG, Reactome, GO:BP, GO:MF, GO:CC, DisGeNET, and HPO) through binary classification models, and to explore potential molecular subgroups using unsupervised clustering (M3C).

Materials and methods:

Gene expression data from GSE99039 were processed to obtain molecular scores per pathway and database. The 750 most variable scores were retained per combination. Supervised classification (IPD vs. CONTROL) was performed using pathMED with a nested cross-validation design, where the inner loop optimized model hyperparameters and the outer loop provided unbiased performance estimates averaged across test folds. Random Forest, KNN, XGBoost, GLM, and LDA algorithms were trained and evaluated using Accuracy, Balanced Accuracy, MCC, Recall, Specificity, Precision, and F1 metrics.

Complementary visualizations included ROC and PR curves, confusion matrices, calibration plots, molecular heatmaps, and pathway importance charts. Unsupervised clustering with M3C was also applied to identify and characterize potential molecular subtypes.

Results:

Overall, model performance was moderate (Accuracy, Balanced Accuracy, Precision, and F1 mostly between 0.5 and 0.75), consistent with PCA variability patterns. The ssGSEA–GO:CC combination with XGBoost achieved the best overall performance (Accuracy = 0.670; Balanced Accuracy = 0.665; MCC = 0.670), while differences among algorithms were relatively small. Unsupervised analyses revealed consistent IPD–control clustering, and pathway importance analyses highlighted mitochondrial metabolism, protein degradation, and inflammatory response processes linked to Parkinson's pathophysiology.

Conclusions:

Combining pathway scoring with machine learning enables the detection of subtle molecular differences between IPD patients and controls, yielding moderate but biologically meaningful predictive power. Task parallelization improved computational efficiency, and the functional interpretation of key pathways supports known mechanisms of Parkinson's disease. Further integration of multi-omics data and larger cohorts could enhance model robustness and clinical applicability.

Keywords:

Machine learning, Parkinson's disease, pathway scoring, supervised classification, unsupervised clustering

1. Introducción

Enfermedad de Parkinson

La enfermedad de Parkinson (EP) es la segunda patología neurodegenerativa más frecuente a nivel global, estimándose en 2019 más de 8,5 millones de personas viviendo con EP. Esta cifra es impulsada por el envejecimiento poblacional y presenta una tendencia ascendente en las próximas décadas (World Health Organization, 2023). En el plano clínico, es un trastorno neurodegenerativo progresivo con fenotipo mixto, constituyendo síntomas motores como la bradicinesia, la rigidez o el temblor, además de las manifestaciones no motoras, que pueden anticipar el diagnóstico y reflejan la heterogeneidad clínica y biológica de la enfermedad (Ryan et al., 2025; Roodveldt et al., 2024).

Durante las últimas décadas, los avances en el manejo de la EP han sido notables, destacando la mejora de los tratamientos farmacológicos, la cirugía funcional y las terapias de rehabilitación. Sin embargo, las estrategias terapéuticas disponibles únicamente se limitan a aliviar los síntomas sin modificar el curso progresivo de la enfermedad. En consecuencia, la búsqueda de tratamientos capaces de frenar o revertir los procesos neurodegenerativos continúa siendo una prioridad en la investigación actual. A medida que la enfermedad progresa, los pacientes desarrollan discapacidad irreversible y pérdida de autonomía, lo que hace especialmente urgente el desarrollo de estrategias verdaderamente modificadoras de la enfermedad para mejorar la calidad de vida de las personas afectadas (Sun & Guo, 2025).

Neuropatológicamente, la EP combina la pérdida de neuronas dopaminérgicas en la pars compacta de la sustancia negra con la acumulación de α-sinucleína en forma de cuerpos de Lewy. Estos agregados alteran la homeostasis celular, provocando estrés oxidativo, disfunción mitocondrial o autofagia, entre otros, y actúan como señales de peligro para células inmunes del sistema nervioso central (Ryan et al., 2025; Roodveldt et al., 2024). Esta alteración se considera el sello distintivo de la enfermedad y se asocia directamente con los síntomas motores. No obstante, estos defectos pueden estar precedidos durante más de una década por manifestaciones no motoras, lo que evidencia que la neurodegeneración comienza mucho antes de la aparición de los signos clínicos más visibles. La edad avanzada constituye el principal factor de riesgo, afectando aproximadamente al 1 % de la población mayor de 60 años y alcanzando su máxima prevalencia en individuos de más de 80 años. Además, se han identificado otros factores de riesgo, como el sexo masculino, la predisposición genética, los antecedentes familiares y la exposición a determinadas toxinas ambientales (Roodveldt et al., 2024).

Mecanismos moleculares e inmunológicos en Parkinson

Mecanismos moleculares

La contribución genética a la EP combina variantes raras de gran efecto (p. ej., SNCA, LRRK2, PRKN/PINK1, VPS35) y un fondo poligénico de baja penetrancia. Los meta-análisis GWAS sitúan la heredabilidad atribuible a variantes comunes en torno al 22 % y han consolidado decenas de loci implicados en sinapsis, autofagia e inmunidad (Nalls et al., 2019). En España, un GWAS específico mostró señales robustas en SNCA, LRRK2, MAPT/KANSL1 y HLA-DQB1, un enriquecimiento de PARK2 para edad de inicio y particularidades de haplotipos útiles para "fine-mapping" (Bandrés-Ciga et al., 2019). En esa cohorte, LRRK2 p.G2019S apareció en aproximadamente en un 2,8 % de los casos y en torno al 0,3 % de controles, subrayando su peso en subpoblaciones ibéricas (Bandrés-Ciga et al., 2019).

Dentro del eje lisosomal, GBA1 destaca por su impacto en riesgo y fenotipo, teniendo una asociación consistente en múltiples centros y justifica integrar estatus GBA en la estratificación clínica y de investigación (Sidransky et al., 2009).

En paralelo, la neuroinflamación opera como modulador de progresión, siendo así que la exposición de la microglía o los astrocitos a α-sinucleína induce disfunción mitocondrial y activa NLRP3, con la liberación de IL-1β/TNF-α. Además, el daño mitocondrial también puede disparar la vía cGAS–STING, conectando estrés organelar con respuestas de tipo interferón y neurodegeneración dopaminérgica (Roodveldt et al., 2024). Más aún, se han descrito alteraciones periféricas (citoquinas/quimiocinas, reactividad T frente a α-sinucleína) que apoyan un eje centro-periferia con permeabilidad de BHE e infiltración linfoide (Roodveldt et al., 2024).

Por otro lado, la presencia de mutaciones en LRRK2 parece potenciar estados microgliales proinflamatorios y defectos lisosomales, reforzando su papel como nodo inmunometabólico. Existen datos experimentales recientes que muestran hiperactivación o inflamación microglial y astroglial dependiente de LRRK2, con rutas NF-κB/NLRP3 y tráfico lisosomal alterado (Yao et al., 2023; Pajarillo et al., 2023).

De cara a biomarcadores longitudinales, la integración multimodal en PPMI sugiere que la "señal" informativa cambia con el tiempo. SNPs y huellas transcriptómicas / metilómicas en sangre ayudan en fases tempranas, mientras proteínas de LCR y DNAm ganan peso más tarde. Combinar modalidades mejora la estratificación respecto a escalas clínicas puras (Ryan et al., 2025).

En conjunto, las causas genéticas (monogénica y poligénica), liosomales e inmunes (centrales y periféricos) se entrelazan conformando la heterogeneidad de la EP. Esto refuerza la utilidad de los paneles genéticos dirigidos (LRRK2, GBA1, PRKN/PINK1) según procedencia/edad de inicio, de los biomarcadores multimodales sensibles al estadio y, de las dianas inmunometabólicas (NLRP3, cGAS–STING, LRRK2) para ensayos de terapia modificadora (Bandrés-Ciga et al., 2019; Nalls et al., 2019; Roodveldt et al., 2024; Ryan et al., 2025).

Mecanismos inmunológicos

En los últimos años se ha prestado una atención creciente al papel del sistema inmunitario en la fisiopatología y progresión de la EP. La evidencia integrada de pacientes y modelos preclínicos indica que procesos inflamatorios contribuyen a la degeneración dopaminérgica y a la heterogeneidad clínica, como se comentó en la sección anterior de "Mecanismos moleculares" (Roodveldt et al., 2024). En el SNC, la microglía y los astrocitos muestran activación sostenida con señalización TLR2/TLR4 para NF-κB/MAPK y activación del inflamasoma NLRP3, lo que favorece la liberación de IL-1β, TNF-α e IL-6 y la amplificación del daño. A la vez, la disfunción mitocondrial puede activar cGAS–STING, reforzando respuestas tipo interferón (Roodveldt et al., 2024). En cerebros de EP y en modelos con α-sin se ha observado un aumento de TSPO-PET compatible con microgliosis, incluso en fases prodrómicas como el trastorno de conducta del sueño MOR (iRBD), lo que respalda la cronología temprana de la neuroinflamación (Stokholm et al., 2017; Roodveldt et al., 2024).

La inmunidad periférica también está alterada, observándose cambios en subpoblaciones de monocitos y linfocitos T/B, incremento de quimiocinas y citoquinas circulantes y asociación con progresión motora y cognitiva, lo que sugiere un eje centro–periferia con permeabilidad de la barrera hematoencefálica (BHE) e infiltración linfoide (Roodveldt et al., 2024). En metaanálisis, las personas con EP presentan niveles sanguíneos más altos de IL-6, TNF-α, IL-1β, IL-2, IL-10, CRP y RANTES/CCL5, configurando un entorno proinflamatorio sistémico con relevancia pronóstica (Qin et al., 2016).

En el plano adaptativo, varios trabajos han demostrado reactividad T específica frente a péptidos de α-sin (CD4+ y CD8+) con mayor frecuencia en EP temprana, lo que aporta un mecanismo autoinmune plausible y conecta con asociaciones genéticas en HLA (Sulzer et al., 2017; Roodveldt et al., 2024).

Por último, también convergen datos genéticos y experimentales en torno al eje intestino–inmunidad–cerebro. Dicho así, las variantes LRRK2 muestran efectos compartidos sobre el riesgo de EP y enfermedad de Crohn, y en modelos con LRRK2-G2019S la colitis experimental exacerba neuroinflamación, patología de α-sin y pérdida dopaminérgica (Hui et al., 2018; Roodveldt et al., 2024). Este entramado inmunometabólico posiciona nodos como NLRP3, cGAS–STING y LRRK2 como dianas y respalda el desarrollo de biomarcadores inmunes con sensibilidad temporal (Roodveldt et al., 2024).

Evidencia desde otras enfermedades neurodegenerativas

La alteración de rutas metabólicas, inflamatorias y de plasticidad sináptica observada en la enfermedad de Parkinson también se documenta en otras patologías neurodegenerativas. En Alzheimer, por ejemplo, se ha desarrollado un modelo no invasivo basado en la firma metabólica sanguínea, donde los cambios en fosforilación oxidativa, metabolismo lipídico (citocromos P450, esfingolípidos) y estrés oxidativo

actúan como indicadores tempranos de deterioro cognitivo (Feng et al., 2023). Estas observaciones subrayan la convergencia entre disfunción mitocondrial y alteraciones inmunometabólicas en la neurodegeneración (Poddar et al., 2021).

De forma análoga, estudios de artrosis (OA) y Alzheimer han identificado genes compartidos vinculados con el estrés celular, la inflamación y el metabolismo energético, lo que sugiere mecanismos comunes de envejecimiento y disfunción celular más allá de la región o tejido afectado (Liu et al., 2025). Esta coincidencia en rutas redox y señalización inflamatoria aporta un contexto evolutivo al deterioro neuronal.

Por otro lado, aunque la esquizofrenia no se considera una enfermedad neurodegenerativa clásica, los análisis transcriptómicos integrados han identificado genes esenciales de respuesta a la enfermedad (DREGs) implicados en plasticidad sináptica, regulación inmune, neurotransmisión y desarrollo neuronal, mostrando una red biológica estrechamente interconectada (Ni et al., 2025). Estas alteraciones moleculares refuerzan la idea de una vulnerabilidad compartida del cerebro frente a procesos inflamatorios y de disrupción de la conectividad sináptica.

En conjunto, la evidencia procedente de otras patologías respalda la hipótesis de que la interacción entre metabolismo, inmunidad y sinapsis constituye un eje transversal en la neurodegeneración. Este paralelismo contextualiza los mecanismos inmunometabólicos descritos en la EP dentro de un marco común de vulnerabilidad celular, aunque el presente trabajo se centra principalmente en sus particularidades patogénicas.

Justificación del enfoque computacional y del uso de Machine Learning

En el contexto de la medicina personalizada, el análisis de datos a nivel individual se ha convertido en un pilar para la comprensión de las enfermedades neurodegenerativas. Los métodos clásicos de puntuación génica suelen requerir cohortes amplias y modelos paramétricos, lo que limita su aplicabilidad en estudios con muestras reducidas. En contraste, los métodos de scoring molecular no paramétrico, como Singscore, ofrecen una alternativa robusta que permite generar puntuaciones estables y comparables incluso en conjuntos pequeños. Este enfoque transforma perfiles de expresión génica en valores únicos interpretables, que reflejan la actividad de rutas biológicas o fenotipos específicos, resultando especialmente útil en patologías heterogéneas como la enfermedad de Parkinson (Foroutan et al., 2018).

Durante los últimos años, los enfoques computacionales basados en bioinformática y aprendizaje automático (Machine Learning, ML) han adquirido relevancia en el estudio del Parkinson. A diferencia de los métodos estadísticos tradicionales, los algoritmos de ML integran grandes volúmenes de información clínica, molecular y ómica, identificando patrones no lineales que escapan al análisis convencional y mejorando la clasificación y predicción de la enfermedad (Fan et al., 2025). Esta capacidad de capturar interacciones complejas entre variables ha permitido abordar con mayor precisión la heterogeneidad molecular del Parkinson y superar las limitaciones de los modelos lineales.

Entre los avances recientes, Wang et al. (2025) demostraron que la combinación de algoritmos Least Absolute Shrinkage and Selection Operator (LASSO) y Support Vector Machine – Recursive Feature Elimination (SVM-RFE) permite construir modelos diagnósticos de alta precisión y reproducibilidad en cohortes externas, basados en la identificación de genes clave relacionados con el metabolismo de purinas. De forma complementaria, Li et al. (2025) integraron múltiples algoritmos de ML con análisis bioinformático para detectar biomarcadores inflamatorios en Parkinson, mostrando que la combinación de LASSO, SVM-RFE y Random Forest mejora la sensibilidad y estabilidad del modelo predictivo. Estos trabajos subrayan la utilidad del ML para la detección temprana, la estratificación de pacientes y el seguimiento longitudinal.

Asimismo, Fan et al. (2025) desarrollaron modelos predictivos y pronósticos basados en datos poblacionales (NHANES 1999–2018), utilizando enfoques explicables mediante Shapley Additive Explanations (SHAP). Estas herramientas de interpretabilidad aportan transparencia al proceso decisional del modelo, facilitando su integración en la práctica clínica y mejorando la confianza en la predicción personalizada.

Integración de Scoring Molecular con Enfoques de Machine Learning

La integración del scoring molecular con modelos de ML representa una evolución hacia sistemas predictivos más interpretables y personalizados. Este enfoque permite incorporar la información derivada de la puntuación génica en algoritmos supervisados, potenciando la identificación de biomarcadores relevantes para el diagnóstico y la progresión del Parkinson. Los modelos resultantes, basados en técnicas como SVM, LASSO o redes neuronales, pueden detectar patrones complejos y mejorar la sensibilidad frente a las variaciones individuales (Li et al., 2025; Wang et al., 2025).

Además, la combinación de ambos enfoques optimiza la clasificación de pacientes según perfiles moleculares y epigenéticos, favoreciendo la estratificación longitudinal y la caracterización de subtipos biológicos. En este contexto, los scores moleculares obtenidos mediante distintos métodos de enriquecimiento (como Singscore, ssGSEA o GSVA) se emplean como variables de entrada en los modelos de aprendizaje automático, permitiendo capturar la actividad coordinada de conjuntos génicos y mejorar la interpretabilidad sin comprometer el poder predictivo (Foroutan et al., 2018).

En conjunto, esta estrategia integra la cuantificación funcional de genes con algoritmos de predicción capaces de adaptarse a datos heterogéneos, ofreciendo una aproximación sólida y reproducible para el estudio del Parkinson. Los resultados reportados por Fan et al. (2025) refuerzan esta idea al demostrar que los modelos explicables basados en ML no solo alcanzan una precisión superior a los métodos clínicos tradicionales, sino que también revelan la contribución de variables biológicas críticas, como la edad, el perfil metabólico o biomarcadores séricos en la predicción individual.

Estado del arte en la estratificación de pacientes de Enfermedad de Parkinson

La presencia de heterogeneidad clínica y biológica en la EP, como son las diferencias en el inicio, la velocidad de progresión y los dominios afectados, ha impulsado la búsqueda de estrategias de estratificación que permitan definir subgrupos con trayectorias distinguibles y, en última instancia, personalizar el manejo y los ensayos clínicos (Wüllner et al., 2023; Marek et al., 2018). La disponibilidad de cohortes longitudinales con múltiples modalidades, como PPMI, ha sido decisiva para pasar de tipologías basadas solo en clínica a marcos multimodales que integran clínica, imagen, genética, epigenética y proteómica. (Marek et al., 2018; Wüllner et al., 2023).

Los primeros enfoques de subtipificación se apoyaron en perfiles clínicos. Un trabajo de referencia definió criterios prácticos (mild-motor predominant, intermediate y diffuse-malignant) a partir de baterías motoras y no motoras, mostrando diferencias pronósticas y de biomarcadores en LCR e imagen entre subtipos (Fereshtehnejad et al., 2017). Validaciones posteriores han reproducido trayectorias diferenciadas y ritmos de avance (por ejemplo, enfoques SuStaIn y validaciones a 2 años), reforzando el valor de una clasificación clínica sistemática como base de la estratificación (Zhou et al., 2023; Johansson et al., 2023).

En paralelo, la estratificación impulsada por datos ha crecido con modelos de aprendizaje automático. En un estudio longitudinal, Severson y col. identificaron estados de enfermedad y patrones de progresión integrando series temporales clínicas mediante ML, con rendimiento superior a aproximaciones lineales convencionales (Severson et al., 2021). Este tipo de modelos aporta marcos explicables y comparables entre cohortes, y ha abierto la puerta a combinar variables clínicas con señal biológica (p. ej., LCR, imagen, ómicas) en pipelines únicos.

La integración multimodal ha dado un salto cualitativo con aproximaciones recientes sobre PPMI. Ryan y col. (2025) aplican un marco flexible e integrativo (PSN + GCN) y demuestran que la combinación de modalidades informativas cambia con el tiempo: en EP genética, SNPs + metilación del ADN (DNAm) discriminan de forma robusta PD/PL/HC, mientras que en EP idiopática gana peso la proteómica de LCR en etapas más tardías; además, observan señales epigenéticas compartidas entre EP genética e idiopática (Ryan et al., 2025). Estos resultados coinciden con la literatura que subraya el valor pronóstico/diagnóstico del LCR (incluida α-sin y paneles proteómicos) para mejorar la subtipificación más allá de la clínica (Parnetti et al., 2019).

Más allá de PPMI, emergen marcos multiescala que enlazan transcriptómica cerebral, neuroimagen longitudinal y factores comportamentales para inferir mecanismos subyacentes y proponer subtipos biológicos con posible valor terapéutico (Adewale et al., 2025). En el terreno metodológico, modelos de redes neuronales en grafos que fusionan ómicas e imagen han mostrado utilidad para clasificación y subtipado, aportando representaciones compactas de conectomas junto a firmas moleculares (Chan et al., 2022). En conjunto, el campo converge en que la estratificación efectiva

requiere modelos longitudinales, multimodales y explicables, capaces de adaptarse a la evolución temporal de la señal biológica y a la diversidad etiopatogénica de la EP.

Los componentes que más consistentemente aportan para estratificar son: (1) clínica estandarizada longitudinal (para anclar trayectorias), (2) DNAm y otras ómicas sanguíneas (sensibles y relativamente accesibles), y (3) proteómica de LCR (mayor cercanía al SNC). La combinación equilibrada de estos ejes, con validación externa y técnicas XAI, es el estándar emergente para subtipar pacientes y seleccionar cohortes enriquecidas en ensayos de terapias modificadoras. (Ryan et al., 2025; Parnetti et al., 2019; Severson et al., 2021).

2. Hipótesis y Objetivos

Hipótesis: El estudio de las alteraciones moleculares en sangre periférica de pacientes con enfermedad de Parkinson puede revelar patrones transcriptómicos específicos asociados con la enfermedad, que permitan estratificar a los pacientes en subgrupos con perfiles clínicos diferenciados y contribuyan a una predicción más precisa de su progresión.

Objetivos: El objetivo principal es evaluar diversos enfoques de estratificación de la enfermedad del Parkinson, basados en el análisis de características moleculares, utilizando una variedad de metodologías y recursos de datos. Para lograrlo, se plantearon los siguientes objetivos específicos:

A. Calcular scores moleculares con diferentes metodologías y bases de datos.

B. Predecir el diagnóstico de Parkinson y otras variables clínicas.

C. Estratificar los pacientes en subgrupos moleculares y caracterizar los subgrupos encontrados.

D. Evaluar qué metodologías y bases de datos funcionan mejor para la predicción y estratificación de esta enfermedad.

3. Metodología

Obtención de datos

El estudio parte de un conjunto de datos públicos presentes en NCBI Gene Expression Omnibus (GEO) bajo el identificador GSE99039 (Enlace en Anexo). El conjunto incluye perfiles de expresión génica en sangre de 205 pacientes con enfermedad de Parkinson idiopática y 233 individuos sanos, generado y publicado por Shamir et al., 2017. Las muestras fueron analizadas mediante la plataforma Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), un microarray de expresión de oligonucleótidos que incluye 54.675 sondas agrupadas en conjuntos (probe sets) para medir la expresión de más de

47.000 transcritos humanos, lo que permite una cobertura amplia de genes codificantes y predichos (NCBI GEO, GPL570).

El proceso de descarga y lectura de los datos se realizó mediante el paquete "GEOquery" en R, facilitando la obtención y manejo de la información directamente desde GEO. Según la información disponible en el repositorio y los valores observados en la matriz de expresión (rango 1.51–14.88), los datos se encontraban previamente normalizados y transformados a escala $\log_2$ por los autores del estudio. Por tanto, el preprocesamiento posterior se centró en la curación, control de calidad y homogeneización de las anotaciones, evitando aplicar una nueva normalización sobre datos ya procesados. Después de la descarga, se llevó a cabo un exhaustivo control de calidad, que incluyó la revisión y filtrado de datos para eliminar ruido que pudiera afectar los resultados del análisis.

Durante esta fase se realizó la corrección de los identificadores génicos, ya que originalmente las filas de la matriz estaban etiquetadas con los códigos de sondas del microarray (por ejemplo, "1053_at"). Utilizando la información contenida en los metadatos de la plataforma, dichos identificadores se sustituyeron por los símbolos génicos oficiales (Gene Symbol), como "TP53". Además, dado que varias sondas pueden mapear a un mismo gen, las entradas duplicadas se agruparon por símbolo génico empleando la mediana de las sondas correspondientes, al ser una medida robusta frente a valores atípicos. Este procedimiento permitió obtener una única medida representativa por gen, garantizando la coherencia y consistencia del conjunto de datos para los análisis posteriores.

Cálculo de scores moleculares con diferentes metodologías

El scoring molecular consiste en asignar valores cuantitativos a cada muestra con el fin de reflejar su grado de concordancia con múltiples conjuntos génicos (gene sets) previamente definidos. Este procedimiento permite transformar datos de expresión de miles de genes en una matriz de scores, donde cada muestra obtiene un valor por cada ruta biológica o proceso funcional considerado. De este modo, se cuantifica hasta qué punto los genes asociados a un determinado mecanismo (rutas de KEGG, términos de Gene Ontology o bases de datos como Reactome, DisGeNET o HPO) se encuentran coordinadamente activados o reprimidos en cada muestra.

En este trabajo se empleó el paquete pathMED, un paquete desarrollado por el grupo de investigación del Centro Pfizer–Universidad de Granada–Junta de Andalucía de Genómica e Investigación Oncológica (GENYO), orientado al análisis multi-ómico y a la aplicación de modelos de aprendizaje automático para estudios de medicina de precisión. pathMED permite calcular scores de conjuntos génicos utilizando distintos métodos de enriquecimiento y posteriormente entrenar modelos predictivos o de estratificación a partir de ellos (Toro-Domínguez et al., 2022). A continuación, se detallan los criterios empleados en este estudio:

- *singscore*: Método que calcula la posición relativa de los genes de una firma en el ranking de expresión de una muestra, produciendo un valor entre -1 y 1 que indica si los genes están sobre- o sub-expresados, respectivamente (Foroutan et al., 2018).
- *GSVA*: Funciona transformando la matriz de expresión (gen-muestra) en una matriz de enriquecimiento (vías- muestra), usando estimaciones no paramétricas de densidad y un estadístico tipo Kolmogorov-Smirnov con el fin de calcular la actividad relativa de las rutas en cada muestras. Suele ser muy útil en estudios heterogéneos y con RNA-Seq (Hänzelmann, Castelo, & Guinney, 2013).
- *ssGSEA*: Consiste en una adaptación de GSEA a nivel individual, es decir, para cada muestra compara la distribución de los genes de una firma frente al resto mediante un estadístico de tipo running sum. De esta manera, proporciona un score por muestra y vía, normalizado respecto a todas las vías y muestras (Barbie et al., 2009).
- *Z-score*: Calcula un score para cada muestra como la media de los z-scores de los genes en la firma (habiendo estandarizado respecto a controles). Es un método simple, rápido y útil como base, aunque asume independencia entre genes (Lee et al., 2008).
- *PLAGE*: Método paramétrico y sensible a correlaciones entre genes, que estandariza la expresión génica a Z-scores y aplica una descomposición en valores singulares (SVD). De este modo, el primer componente principal resume la actividad de la ruta en cada muestra (Tomfohr, Lu & Kepler 2005).
- *norm_FGSEA*: Consiste en la implementación optimizada del algoritmo GSEA que permite realizar miles de permutaciones en segundos, produciendo valores normalizados (NES) comparables entre rutas. Es un método muy usado en pipelines modernos por su eficiencia y reproducibilidad (Korotkevich et al., 2021)

Estas metodologías se utilizaron consultando las siguientes bases de datos:

- *Reactome*: Base de datos curada que detalla procesos biológicos humanos, incluyendo enfermedades y fármacos. Su versión más reciente (v94, que está disponible desde septiembre de 2025) amplió la cobertura de reacciones y vías, resultando especialmente útil para integrar datos ómicos y analizar mutaciones somáticas en cáncer, entre otros trastornos complejos (Milacic et al., 2024).
- *KEGG*: Ofrece una representación detallada de las vías metabólicas y las interacciones moleculares, lo cual es crucial para analizar la expresión génica en enfermedades. Especialmente en cáncer. Sin embargo, se recomienda integrarlo

con otras bases como Reactome para obtener resultados más robustos (Kanehisa et al., 2017).

- *GO (BP, CC* y *MF)*: GO es una ontología estándar que describe las funciones de los genes en tres áreas: procesos biológicos, funciones moleculares y componentes celulares. Aunque es útil para análisis genéticos, su alcance limitado se complementa bien con otras ontologías como HPO para mejorar los estudios de enfermedades complejas (Gene Ontology Consortium, 2021).
- *HPO*: Ontología que describe anormalidades fenotípicas en enfermedades humanas hereditarias y se utiliza para identificar similitudes fenotípicas entre enfermedades y se ha integrado con otras bases de datos como GO y Reactome para mejorar el análisis de enfermedades genéticas y fenotípicas (Gargano et al., 2024).
- *DisGeNet*: Plataforma que integra asociaciones genéticas de enfermedades, combinando datos curados y extraídos de la literatura, siendo ideal para la investigación de enfermedades raras y complejas, ya que cuenta con información de más de 24.000 enfermedades y 17.000 genes (Piñero et al., 2020).

Tras la obtención de los distintos scores moleculares, se utilizó otra función de pathMED, con el propósito de traducir los identificadores internos de cada base de datos en sus correspondientes descripciones biológicas (Tabla 1). Este paso permite interpretar de forma directa los resultados, asociando cada identificador con la ruta, proceso biológico o término funcional al que pertenece.

Posteriormente, estas anotaciones se emplearon más adelante en el análisis para identificar los términos biológicos más relevantes en los análisis de importancia de variables, sustituyendo los identificadores numéricos por descripciones comprensibles desde el punto de vista biológico.

Tabla 1. *Ejemplo del formato de anotación esperado para las distintas bases de datos: KEGG (identificadores hsa…), Reactome (R-HSA-…) y GO:BP (GO:…). En cada caso, el identificador aparece en la columna izquierda y el término funcional asociado en la columna derecha.*

ID	Término
hsa00010	Glycolysis / Gluconeogenesis
hsa00020	Citrate cycle (TCA cycle)
hsa00030	Pentose phosphate pathway
hsa00040	Pentose and glucuronate interconversions
hsa00051	Fructose and mannose metabolism
R-HSA-1059683	Interleukin-6 signaling
R-HSA-109581	Apoptosis
R-HSA-109582	Hemostasis

R-HSA-109606	Intrinsic Pathway for Apoptosis
R-HSA-109703	PKB-mediated events
GO:0000002	mitochondrial genome maintenance
GO:0000003	reproduction
GO:0000012	single strand break repair
GO:0000017	alpha-glucoside transport
GO:0000018	regulation of DNA recombination

Uso de algoritmos de Machine Learning para predecir variables clínicas

En primer lugar, se realizaron una serie de modificaciones sobre los datos de scores para reducir la dimensionalidad de los datos y así reducir los tiempos de ejecución. De este modo, se realizó un filtrado de scores por varianza, dejando únicamente los 750 scores con mayor varianza en caso de que hubiera una mayor cantidad de ellos por cada uno de las bases de datosque teníamos. Se decidió mantener el límite de 750, ya que, además de esa mejora en la velocidad de ejecución, en algunas pruebas se observaron resultados similares a cuando se usaban hasta 2000 genesets. Además, en este filtro, también se eliminaron scores que tuvieran valores NA o infinitos, para evitar problemas en la siguiente fase de entrenamiento de modelos.

Basándose en el estudio original (Shamir et al., 2017), se escogieron las categorías CONTROL e IPD (Idiopathic Parkinson's Disease) para la clasificación del modelo, siendo IPD la variable respuesta positiva. Aunque el conjunto GSE99039 incluye otras categorías diagnósticas (como GPD, MSA, PSP o PDD), estas no se consideraron, ya que el objetivo del análisis fue centrarse en la comparación entre pacientes con Parkinson idiopático y sujetos sanos.

Tras ello, se realizó un PCA a modo exploratorio, con el fin de observar el estado de los datos y el grado de dificultad para poder generar modelos fiables.

Posteriormente, realizó clasificación supervisada, entrenando y validando modelos de predicción binaria (sano/enfermo de EP) sobre los datos del estudio utilizando la función *methodsML* para preparar la lista de algoritmos de Machine Learning que serán entrenados y testeados, que en nuestro caso son los siguientes:

1. **Random Forest (RF):** Algoritmo basado en árboles de decisión que utiliza un enfoque de bagging (Bootstrap Aggregating), donde múltiples árboles de decisión son entrenados sobre diferentes subconjuntos aleatorios de los datos. La predicción final es una votación o promedio de los resultados de todos los árboles. Este algoritmo es conocido por su robustez, capacidad de manejar datos de alta dimensión y resistencia al sobreajuste (Breiman, 2001; O'Connell et al., 2025).
2. **K-Nearest Neighbors (KNN):** Consiste en un algoritmo de clasificación supervisada basado en la proximidad. Dado un conjunto de datos de entrenamiento, KNN clasifica una nueva muestra según la clase mayoritaria de

sus K vecinos más cercanos. Este modelo es fácil de implementar, pero su rendimiento puede verse afectado por la dimensionalidad de los datos y el valor de K (Cover & Hart, 1967; Halder et al., 2024).

3. **xgbTree (XGBoost):** Es un algoritmo de boosting que utiliza árboles de decisión como base. Se entrena de forma secuencial, donde cada nuevo árbol corrige los errores cometidos por los árboles anteriores. Este modelo ha demostrado un rendimiento excepcional en tareas de clasificación y es altamente eficaz para manejar grandes volúmenes de datos con alta dimensionalidad y desbalanceo en las clases (Chen & Guestrin, 2016; Egbo et al., 2025).
4. **Generalized Linear Model (GLM):** Extensión de la regresión lineal que permite modelar variables dependientes con distribuciones no normales (como poisson, binomial, etc.). Es un modelo sencillo y muy utilizado para tareas de predicción en las que la relación entre las variables de entrada y la variable de salida se puede aproximar mediante una función lineal (McCullagh & Nelder, 1989; Alaqeeli & Alturki, 2023).
5. **Linear Discriminant Analysis (LDA):** Técnica de reducción de dimensionalidad que también puede utilizarse para clasificación. Busca proyectar los datos en un espacio de menor dimensión donde la separación entre clases sea maximizada. Se utiliza especialmente cuando las clases son linealmente separables y puede ser útil en modelos con datos de alta dimensionalidad (Tharwat et al., 2017; Qiao, 2023).

Además, mediante el parámetro *tuneLength* se realizó un CV interno de 10 iteraciones con distintas combinaciones de hiperparámetros, con el fin de seleccionar la configuración que ofreciera el mejor rendimiento, manteniendo un tiempo de ejecución razonable y evitando saturar la memoria RAM. De igual manera, otra estrategia empleada para reducir significativamente el tiempo de ejecución, sin comprometer el uso de memoria RAM, fue la utilización de la librería *doParallel*, que permite ejecutar operaciones en paralelo aprovechando varios núcleos de la CPU.

Posteriormente, se ajustaron y evaluaron múltiples algoritmos de aprendizaje automático, seleccionando automáticamente aquel que ofrece el mejor rendimiento para predecir una variable objetivo. Sus argumentos de entrada son *inputData*, que equivale a nuestra matriz de scores generada anteriormente; y el parámetro *var2predict*, que especifica el nombre de la columna en *metadata* que contiene la variable que se desea predecir. Finalmente, devuelve un objeto con los resultados completos (stats) y el modelo con mejor rendimiento, accesible. Esto permite seleccionar de forma directa los algoritmos más adecuados para cada combinación de método de scoring y base de datos.

Finalmente, devuelve un objeto con los resultados completos (stats) y el modelo con mejor rendimiento, accesible. Esto permite seleccionar de forma directa los algoritmos más adecuados para cada combinación de método de scoring y base de datos.

Para evaluar el rendimiento de los modelos y garantizar la fiabilidad de los resultados, se aplicó un esquema de validación cruzada anidada. En este diseño, un bucle interno se empleó para optimizar los hiperparámetros de cada modelo, mientras que un bucle externo se utilizó para estimar su capacidad predictiva de manera independiente. Se utilizaron cinco particiones externas y tres internas, repitiendo las validaciones internas tres veces con el objetivo de obtener estimaciones más estables y minimizar el riesgo de sobreajuste.

La división de las muestras se realizó de forma estratificada, manteniendo el equilibrio entre las clases IPD y CONTROL. Con el fin de asegurar la reproducibilidad, se fijó una semilla aleatoria constante y se mantuvo la misma estructura de partición en todas las combinaciones de métodos y bases de datos.

Visualización de los datos obtenidos

Para cada combinación de método de scoring (singscore, Z-score, GSVA, ssGSEA, PLAGE y norm_FGSEA) y base de datos funcional (KEGG, Reactome, GO:BP, GO:MF, GO:CC, DisGeNET y HPO), se generó un resumen con las principales métricas de rendimiento de los modelos entrenados. Las métricas finales se calcularon a partir de las predicciones obtenidas durante la evaluación de los modelos, proporcionando una estimación global de su capacidad de clasificación. Estas métricas incluyeron:

*Nota: TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative)

1. **Accuracy:** Proporción de predicciones correctas (verdaderos positivos y verdaderos negativos).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Balanced Accuracy:** Promedio de sensibilidad y especifidad, útil cuando las clases están desbalanceadas.

$$Balanced\ Accuracy\ = \frac{Recall\ (Sensibilidad) + Specificity}{2}$$

3. **Matthews Correlation Coefficient (MCC):** Métrica que tiene en cuenta verdaderos positivos, negativos, falsos positivos y falsos negativos, proporcionando una medida equilibrada de la calidad del modelo, especialmente con clases desbalanceadas. Para facilitar la visualización y comparación de resultados, se aplicó una transformación lineal MCC_norm = (MCC + 1)/2 para llevarlo al rango de 0 a 1, ya que inicialmente era de -1 a 1.

$$MCC\ = \frac{TP\ \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4. **Recall (Sensibilidad):** Proporción de verdaderos positivos sobre los casos que realmente pertenecen a la clase positiva.

$$Recall = \frac{TP}{TP + FN}$$

5. **Specificity (Especifidad):** Proporción de verdaderos negativos sobre los casos que no pertenecen a la clase positiva.

$$Specificity = \frac{TN}{TN + FP}$$

6. **Precision (Precisión):** Proporción de predicciones positivas correctas.

$$Precision = \frac{TP}{TP + FP}$$

7. **F1-score:** Media armónica entre la precisión y la sensibilidad, proporcionando un balance entre ambas.

$$Recall = 2 \times \frac{Precision \times Recall}{Presicion + Recall}$$

Con el objetivo de evaluar el rendimiento de los modelos entrenados y facilitar la interpretación biológica de los resultados, se generó una serie de visualizaciones complementarias derivadas de los distintos métodos de scoring y bases de datos empleados.

Estas representaciones incluyeron una tabla resumen con las métricas y un heatmap de métricas. Además, se generaron curvas ROC y PR, matrices de confusión, gráficos de calibración, mapas de calor de scores moleculares y gráficos de importancia de rutas funcionales, entre otras figuras específicas.

En conjunto, estas visualizaciones permitieron comparar el comportamiento de los algoritmos de aprendizaje automático, analizar su calibración y rendimiento, y examinar la relevancia biológica de las rutas más influyentes en la clasificación de los pacientes.

Evaluación comparativa de los métodos de scoring y bases de datos

Como se ha comentado anteriormente, se visualizaron los datos mediante una tabla resumen de todos los valores alcanzados en cada métrica por modelo y de forma más visual, un heatmap de métricas, donde el color representa la magnitud de la métrica para cada combinación. Junto con esta representación se permite identificar de forma global qué métodos y bases de datos proporcionaron los mejores resultados, qué métricas tienen los mejores valores, así como patrones de comportamiento entre aproximaciones similares.

No obstante, las métricas más informativas para comparar los modelos son MCC y Balanced Accuracy, al ser menos sensibles al desequilibrio de clases y ofrecer una evaluación más equilibrada del rendimiento global (Chicco & Jurman, 2020; Brodersen et al., 2010).

Curvas ROC y PR

Posteriormente, se evaluó la capacidad discriminante de cada modelo mediante las curvas ROC (Receiver Operating Characteristic) y PR (Precision-Recall), calculadas a partir de las predicciones out-of-fold almacenadas en "trainedModel$subsample.preds".

Para cada modelo se obtuvo el valor de área bajo la curva (AUC y AUPRC, respectivamente), que es indicador de la calidad de la clasificación independientemente del umbral de decisión.

Estas curvas permiten comprobar si un modelo mantiene un buen equilibrio entre sensibilidad y especificidad, y resultan especialmente útiles en contextos con clases desbalanceadas, como suele ocurrir en datos clínicos (Saito & Rehmsmeier, 2015).

Matrices de confusión

Para una inspección más detallada de las predicciones, se construyeron matrices de confusión a partir de las probabilidades OOF (out-of-fold). En ellas se representa el número de predicciones correctas e incorrectas para cada clase, con el fin de evaluar el tipo de errores cometidos (falsos positivos o falsos negativos). La generación de estos gráficos se realizó a través del paquete caret, empleando un umbral de clasificación estándar de 0.5.

La correcta interpretación de estas matrices es esencial para valorar y entender si los modelos tienden a sobreestimar o infravalorar la clase positiva (pacientes con Parkinson idiopática).

Calibration plot

La calibración de los modelos se analizó mediante gráficos que comparan la probabilidad predicha frente a la frecuencia observada de la clase positiva. En estos plots, los puntos situados debajo de la diagonal indican sobreestimación de la probabilidad (el modelo tiende a predecir valores más altos que la realidad), mientras que los puntos por encima de la diagonal representan subestimación.

Además, se calculó el Brier score como medida global de calibración, donde los valores más bajos (< 0,2) son indicativos de una mejor correspondencia entre probabilidades predichas y observadas (Steyerberg & Vergouwe, 2014).

Heatmap pathways

Se generaron heatmaps de los scores moleculares para cada modelo entrenado, organizando las muestras según la clase (IPD y CONTROL). Las rutas se representaron en filas, y las muestras en columnas, aplicando una normalización tipo z-score por fila para resaltar las diferencias relativas de expresión.

Los gráficos se realizaron tanto con todas las rutas (FULL) como con las 100 de mayor varianza (TOP-100), permitiendo visualizar patrones de activación molecular característicos de cada grupo.

Importancia de las rutas e interpretación biológica de los resultados

Por último, se calculó la importancia de las rutas moleculares para cada modelo mediante el análisis de la disminución de la precisión (variable importance, función varImp del paquete caret), que estima el peso relativo de cada predictor en la clasificación (Kuhn, 2008).

Sin embargo, en el caso de los modelos k-NN, donde no se dispone de medidas internas de importancia, esta se obtuvo mediante permutación de características, evaluando la pérdida de rendimiento al alterar aleatoriamente cada variable. Este enfoque permite identificar las rutas que más contribuyen a la discriminación entre grupos y se considera una estrategia robusta para la interpretación de modelos predictivos en bioinformática (Fisher, Rudin, & Dominici, 2019).

Las rutas con mayor valor de importancia se identificaron y resumieron en tablas (Top-10 por base de datos para cada método de cálculo de scores). Posteriormente, se elaboró una tabla adicional destinada a recoger los términos potencialmente relacionados con la enfermedad de Parkinson, como aquellos asociados al metabolismo mitocondrial, la respuesta inflamatoria o la degradación proteica.

Estratificación de pacientes con los scores moleculares y caracterización clínica de los subgrupos

A partir de los scores moleculares obtenidos para cada combinación de método de enriquecimiento y base de datos funcional, se realizó una estratificación no supervisada de los pacientes. Para cada una de las 42 combinaciones analizadas, se aplicó el algoritmo M3C (Monte Carlo Consensus Clustering), que determina el número óptimo de subgrupos (K) mediante un proceso iterativo de consenso basado en validación cruzada y simulaciones de Monte Carlo, seleccionando el valor de K que maximiza el Índice de Estabilidad Relativa del Clúster (RCSI) (John et al., 2020). Dado el elevado número de combinaciones evaluadas, se implementó una estrategia de paralelización

por tareas independientes, que permitió ejecutar simultáneamente varios análisis de clustering, reduciendo de forma significativa el tiempo de procesamiento y optimizando el uso de los recursos computacionales.

Posteriormente, se generó una tabla resumen comparativa para evaluar sistemáticamente la calidad de las estratificaciones obtenidas. Las métricas clave consideradas fueron el coeficiente de Silhouette promedio, que cuantifica la cohesión y separación entre clústeres (Rousseeuw, 1987), y el p-valor de asociación entre los subgrupos moleculares y el diagnóstico clínico (IPD/CONTROL), calculado mediante el test exacto de Fisher o el χ^2 de independencia, según correspondiera.

Para las combinaciones más prometedoras, se llevó a cabo una caracterización en profundidad de los subgrupos identificados. Esta incluyó la visualización de patrones moleculares mediante mapas de calor (heatmaps) y análisis de componentes principales (PCA) de los scores, junto con el análisis de composición clínica (proporción de pacientes IPD y controles en cada clúster). Asimismo, se identificaron las rutas funcionales diferenciales que definían cada subgrupo mediante el test de Wilcoxon, aplicando corrección por FDR para el control del error tipo I (Fay & Proschan, 2010). Las rutas más significativas se representaron mediante gráficos tipo dotplot, destacando la dirección y magnitud de las diferencias entre grupos.

Este enfoque integral permitió detectar subtipos moleculares de pacientes con perfiles diferenciados y evaluar comparativamente la coherencia y relevancia clínica de cada aproximación funcional empleada.

4. Resultados

Análisis exploratorio

El análisis exploratorio mediante PCA (Figura 1) muestra una alta superposición entre las muestras de pacientes con enfermedad de Parkinson idiopática (IPD) y los controles. Los dos primeros componentes principales explican aproximadamente el 37 % y el 7,7 % de la varianza total, respectivamente, sin observarse una separación clara entre grupos. Este patrón sugiere que, a nivel global de expresión génica agregada en términos de rutas moleculares (GSVA sobre GO:MF), las diferencias entre IPD y controles son sutiles, lo que anticipa una mayor dificultad para la clasificación supervisada posterior.

Se tomó como ejemplo este el modelo anteriormente comentado pero este comportamiento se repite de forma similar en las restantes combinaciones de métodos y bases de datos empleadas en el estudio (Más ejemplos en el Anexo, junto al resto de visualizaciones), indicando que la señal biológica discriminante es débil y probablemente dependiente de procesos específicos más que de patrones globales.

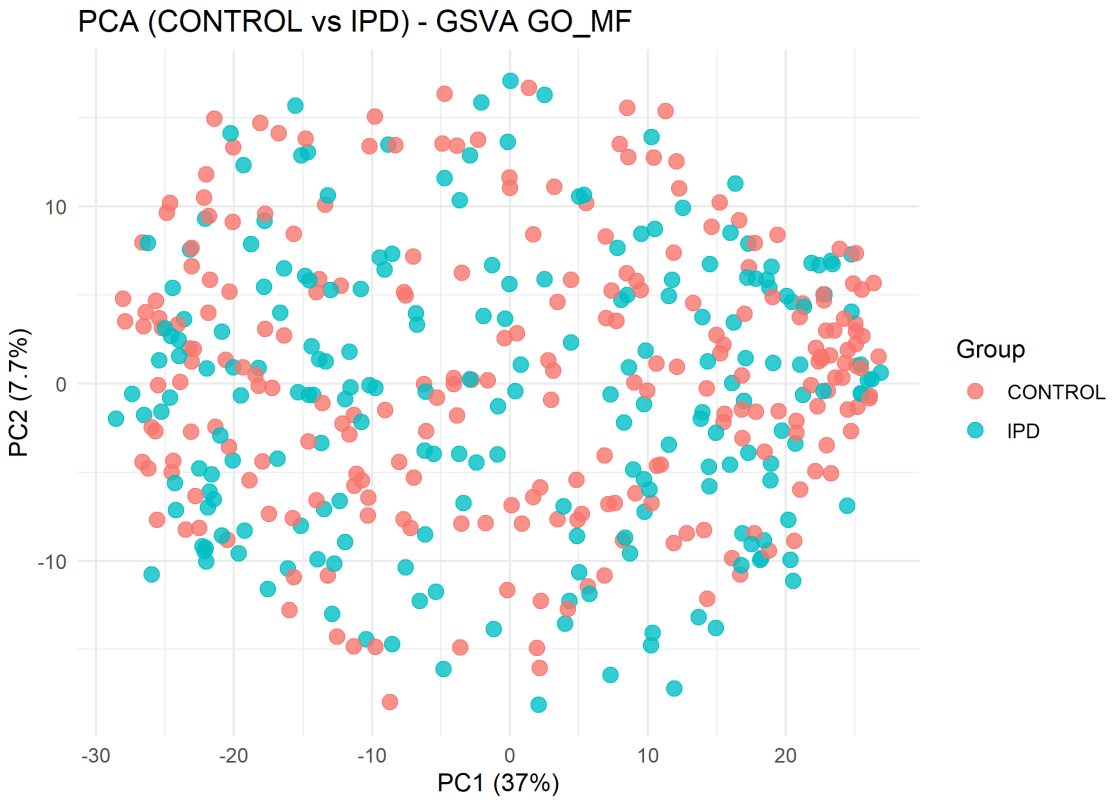


Figura 1. *Análisis de componentes principales (PCA) de los scores de rutas obtenidos mediante GSVA–GO:MF. Cada punto representa una muestra (rojo = CONTROL, azul = IPD). La falta de separación clara entre grupos sugiere una alta similitud global en los perfiles transcriptómicos a nivel de funciones moleculares.*

Resumen de los valores de las métricas de todos los modelos

En general, las métricas de rendimiento obtenidas para los mejores modelos de cada combinación de método de cálculo de scores y base de datos funcional (Tabla 2) corresponden a la media de los resultados obtenidos en los conjuntos de test del bucle externo de la validación cruzada anidada empleada por pathMED. Este procedimiento permitió optimizar los hiperparámetros en el bucle interno y estimar el rendimiento real de los modelos en el bucle externo, garantizando así una evaluación más robusta y libre de sobreajuste.

Los valores obtenidos muestran un rendimiento moderado, en concordancia con lo observado en el análisis exploratorio por PCA (Figura 1). Las métricas de Accuracy, Balanced Accuracy, Precision y F1 se sitúan en la mayoría de los casos entre 0,5 y 0,75, lo que indica un rendimiento aceptable pero limitado en la discriminación entre pacientes con enfermedad de Parkinson idiopática (IPD) y controles.

Entre todas las combinaciones evaluadas, el modelo basado en ssGSEA – GO:CC junto con el algoritmo XGBoost (xgbTree) alcanzó las puntuaciones más altas de forma global, con valores de Accuracy = 0,670, Balanced Accuracy = 0,665, MCC = 0,670, Precision =

0,683 y F1 = 0,623, consolidándose como el mejor clasificador en el conjunto de pruebas. El resto de métricas también mostraron resultados elevados para esta combinación, aunque los valores máximos individuales se alcanzaron en otros casos, como en singScore – GO:MF con Random Forest (RF) para Specificity (0,794), y en norm_FGSEA – GO:BP con XGBoost para Recall (0,591).

En conjunto, estos resultados confirman que, si bien los modelos logran captar cierta señal discriminante, las diferencias moleculares entre grupos son sutiles y la capacidad predictiva general es modesta, en línea con lo anticipado por el análisis exploratorio.

Tabla 2. *Resumen de las métricas obtenidas para cada uno de los modelos que obtuvieron mejores métricas para cada combinación de cálculo de scores y bases de datos, donde podemos ver marcado en color amarillo los valores máximos para cada una de las métricas.*

ScoreMethod	Database	BestModel	Accuracy	BalancedAcc	MCC	Recall	Specificity	Precision	F1
singscore	kegg	knn	0,596	0,587	0,593	0,448	0,725	0,601	0,510
singscore	reactome	xgbTree	0,603	0,597	0,599	0,507	0,687	0,588	0,543
singscore	go_bp	knn	0,621	0,617	0,618	0,555	0,678	0,600	0,575
singscore	go_mf	rf	0,641	0,631	0,640	0,467	0,794	0,672	0,547
singscore	go_cc	rf	0,644	0,637	0,641	0,536	0,739	0,642	0,582
singscore	disgenet	xgbTree	0,610	0,606	0,607	0,533	0,678	0,593	0,559
singscore	hpo	knn	0,605	0,603	0,604	0,561	0,644	0,584	0,568
Zscore	kegg	rf	0,619	0,614	0,615	0,541	0,687	0,603	0,570
Zscore	reactome	rf	0,601	0,596	0,598	0,508	0,683	0,587	0,541
Zscore	go_bp	rf	0,612	0,607	0,610	0,523	0,690	0,604	0,558
Zscore	go_mf	xgbTree	0,607	0,604	0,605	0,537	0,670	0,589	0,560
Zscore	go_cc	rf	0,635	0,628	0,631	0,535	0,721	0,623	0,573
Zscore	disgenet	rf	0,602	0,597	0,599	0,521	0,673	0,584	0,550
Zscore	hpo	xgbTree	0,591	0,586	0,588	0,493	0,679	0,575	0,528
GSVA	kegg	knn	0,596	0,587	0,591	0,453	0,721	0,587	0,511
GSVA	reactome	knn	0,587	0,578	0,582	0,435	0,721	0,580	0,494
GSVA	go_bp	rf	0,587	0,583	0,584	0,522	0,644	0,564	0,542
GSVA	go_mf	rf	0,639	0,635	0,636	0,565	0,704	0,625	0,593
GSVA	go_cc	rf	0,618	0,610	0,614	0,492	0,728	0,616	0,543
GSVA	disgenet	rf	0,625	0,620	0,622	0,527	0,713	0,615	0,565
GSVA	hpo	rf	0,614	0,610	0,611	0,546	0,674	0,596	0,569
ssGSEA	kegg	knn	0,623	0,615	0,623	0,475	0,755	0,640	0,538
ssGSEA	reactome	rf	0,628	0,619	0,625	0,469	0,769	0,643	0,540
ssGSEA	go_bp	knn	0,628	0,623	0,625	0,556	0,690	0,614	0,582
ssGSEA	go_mf	rf	0,648	0,640	0,646	0,503	0,777	0,668	0,573
ssGSEA	go_cc	xgbTree	0,670	0,665	0,670	0,581	0,749	0,683	0,623
ssGSEA	disgenet	knn	0,632	0,627	0,629	0,551	0,703	0,621	0,581
ssGSEA	hpo	knn	0,628	0,625	0,626	0,567	0,682	0,611	0,587
Plage	kegg	xgbTree	0,594	0,591	0,591	0,541	0,640	0,571	0,555

Plage	reactome	rf	0,614	0,608	0,611	0,512	0,703	0,611	0,556
Plage	go_bp	knn	0,594	0,588	0,590	0,494	0,682	0,573	0,529
Plage	go_mf	rf	0,598	0,591	0,593	0,479	0,702	0,588	0,528
Plage	go_cc	rf	0,614	0,606	0,609	0,491	0,722	0,600	0,535
Plage	disgenet	rf	0,607	0,601	0,605	0,503	0,699	0,602	0,544
Plage	hpo	rf	0,614	0,608	0,613	0,513	0,704	0,617	0,554
norm_FGSEA	kegg	rf	0,598	0,591	0,594	0,483	0,700	0,586	0,528
norm_FGSEA	reactome	rf	0,621	0,613	0,617	0,493	0,733	0,621	0,549
norm_FGSEA	go_bp	xgbTree	0,639	0,636	0,637	0,591	0,682	0,619	0,604
norm_FGSEA	go_mf	xgbTree	0,587	0,583	0,583	0,508	0,657	0,562	0,532
norm_FGSEA	go_cc	xgbTree	0,623	0,621	0,622	0,567	0,675	0,607	0,583
norm_FGSEA	disgenet	xgbTree	0,598	0,594	0,595	0,537	0,652	0,575	0,555
norm_FGSEA	hpo	rf	0,589	0,580	0,583	0,444	0,717	0,579	0,502

Los resultados comentados anteriormente con la Tabla 2, se representan de forma más visual en la Figura 2, donde se muestra un heatmap de métricas de rendimiento para cada combinación de método de scoring y base de datos funcional.

En conjunto, se observa que la Specificity presenta los valores más altos entre todas las métricas evaluadas, indicando que los modelos tienden a clasificar correctamente a los individuos sanos con mayor frecuencia. Por el contrario, la Recall (sensibilidad) muestra los valores más bajos, lo que sugiere una menor capacidad para identificar correctamente los casos de enfermedad de Parkinson idiopática. En cuanto al resto de métricas (Accuracy, Balanced Accuracy, Precision, F1 y MCC) se mantienen en torno a valores intermedios, generalmente cercanos a 0,6, lo que refleja un rendimiento global moderado.

De manera coherente con los resultados numéricos previos, la combinación ssGSEA–GO:CC con el algoritmo XGBoost (xgbTree) destaca como la que obtiene un comportamiento más equilibrado y valores superiores en la mayoría de las métricas. Por este motivo, dicho modelo se seleccionó como referencia para la presentación e interpretación de las gráficas y análisis posteriores.

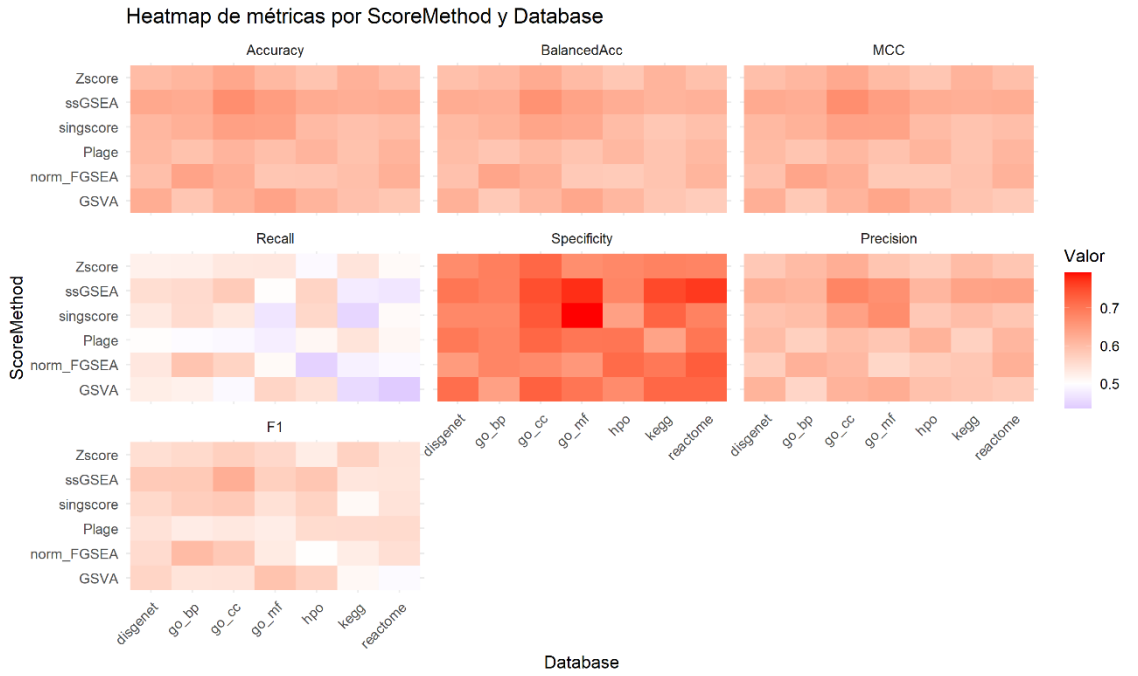
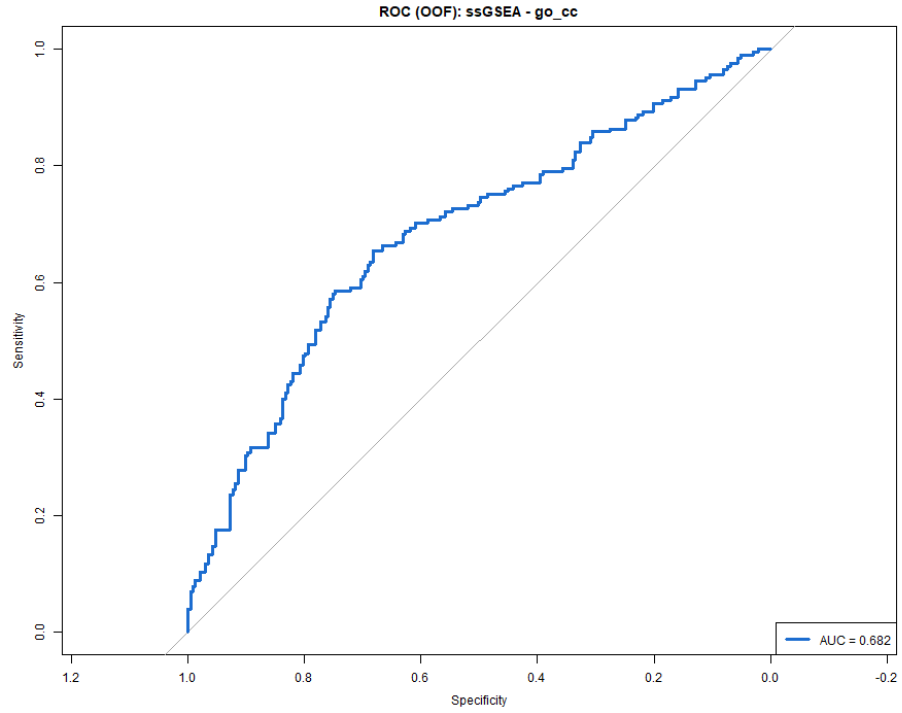


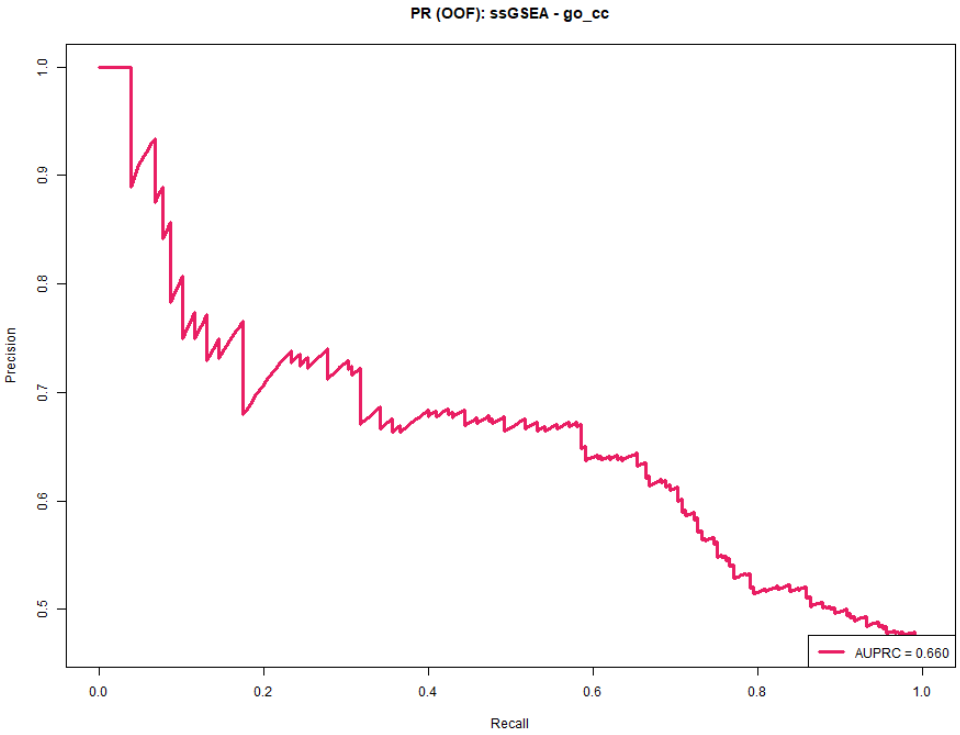
Figura 2. *Heatmap de las principales métricas de rendimiento para cada uno de los modelos obtenidos. Los tonos más intensos representan valores más altos en la métrica correspondiente.*

Curvas ROC y PR

La evaluación del rendimiento de los modelos mediante las curvas ROC y PR (Figuras 3 y 4) confirma la ausencia de una capacidad discriminante elevada. En ambos casos, las curvas no presentan la forma ideal (una trayectoria próxima a la esquina superior izquierda en la ROC y a la esquina superior derecha en la PR) que caracterizaría a un clasificador con un rendimiento óptimo.

Los valores de área bajo la curva (AUC) obtenidos se sitúan, en general, entre 0,55 y 0,69, lo que indica un comportamiento moderado o ligeramente superior al azar. En el caso concreto del modelo ssGSEA–GO:CC (Figuras 3 y 4), que destacó como el de mejor rendimiento global, se obtuvo un AUC = 0,682 en la curva ROC y un AUPRC = 0,660 en la curva PR. Estos resultados sugieren que, si bien el modelo es capaz de identificar cierto grado de patrón diferenciador entre pacientes con Parkinson y controles, la separación entre clases es limitada, reflejando la naturaleza sutil y heterogénea de las diferencias transcriptómicas observadas.





Figuras 3 y 4. *Gráficos ROC (3) y PR (4) de ssGSEA go_cc, donde se puede ver la distribución de las curvas en ambos casos acompañadas del valor AUC en el caso de ROC y AUPRC en el caso de PR.*

Matriz de confusión

Para evaluar el desempeño del modelo ssGSEA–GO:CC en la clasificación de muestras entre los grupos CONTROL e IPD, se utilizó una matriz de confusión (Figura 5). De este modo, se observó que el modelo realizó predicciones correctas para ambas clases, con 174 aciertos en la clase CONTROL y 119 aciertos en la clase IPD, superando a las predicciones erróneas (59 para CONTROL y 86 para IPD).

Estos resultados indican que el modelo logra una tasa de acierto superior a la de error en ambas categorías, mostrando un mejor rendimiento en la identificación de controles que en la de pacientes con enfermedad de Parkinson idiopática.

En conjunto, la matriz refleja un comportamiento aceptable, aunque sigue teneindo un margen de mejora grande.

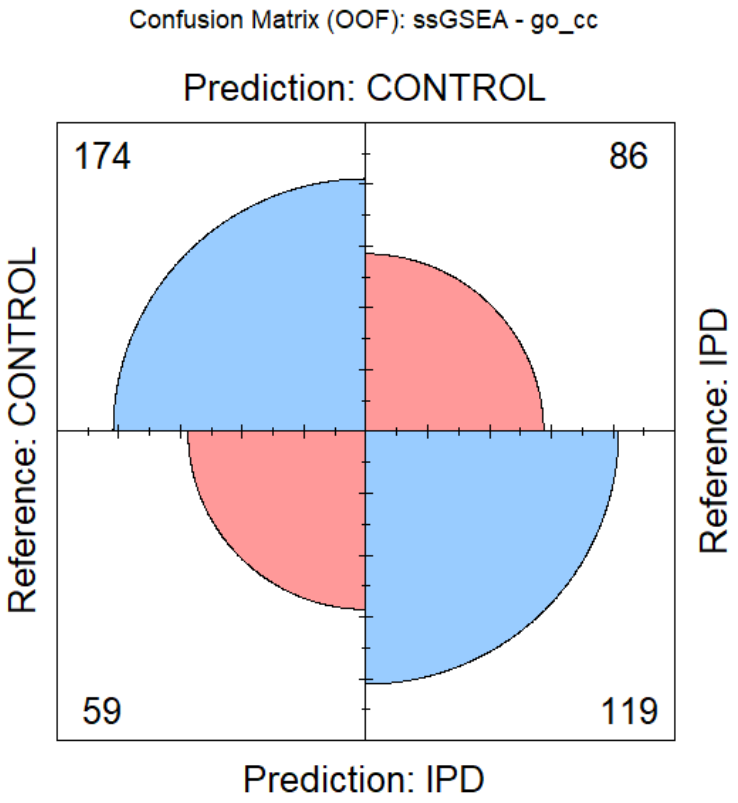


Figura 5. *Matriz de confusión para ssGSEA go_cc, que muestra el desempeño del modelo al clasificar las muestras como CONTROL o IPD. La diagonal principal en azul indica las predicciones correctas, mientras que los valores fuera de la diagonal, en rojo, representan los errores para cada una.*

Gráfico de calibración

El modelo muestra una calibración moderada (Figura 6), con un Brier score de 0,2378 que, aunque se encuentra ligeramente por encima del umbral recomendado (≈0,20), indica un rendimiento aceptable en la estimación de probabilidades.

Por otro lado, la curva de calibración sigue la tendencia general esperada, ya que cuanto mayor es la probabilidad predicha, mayor es la tasa observada de IPD. No obstante, se aprecian algunas desviaciones respecto a la línea ideal, especialmente en los extremos de la probabilidad predicha (eje X). Siendo así, que en las probabilidades más bajas (≈0,20–0,30) el modelo subestima el riesgo, mientras que entre aproximadamente 0,30 y 0,50 tiende a sobreestimarlo. Posteriormente, alrededor de 0,55 vuelve a subestimar, y en los valores más altos (a partir de ≈0,65) muestra una sobreestimación sostenida.

En conjunto, el modelo refleja una tendencia coherente, aunque presenta márgenes de mejora en la calibración, especialmente si se pretende un uso predictivo más preciso o clínicamente aplicable.

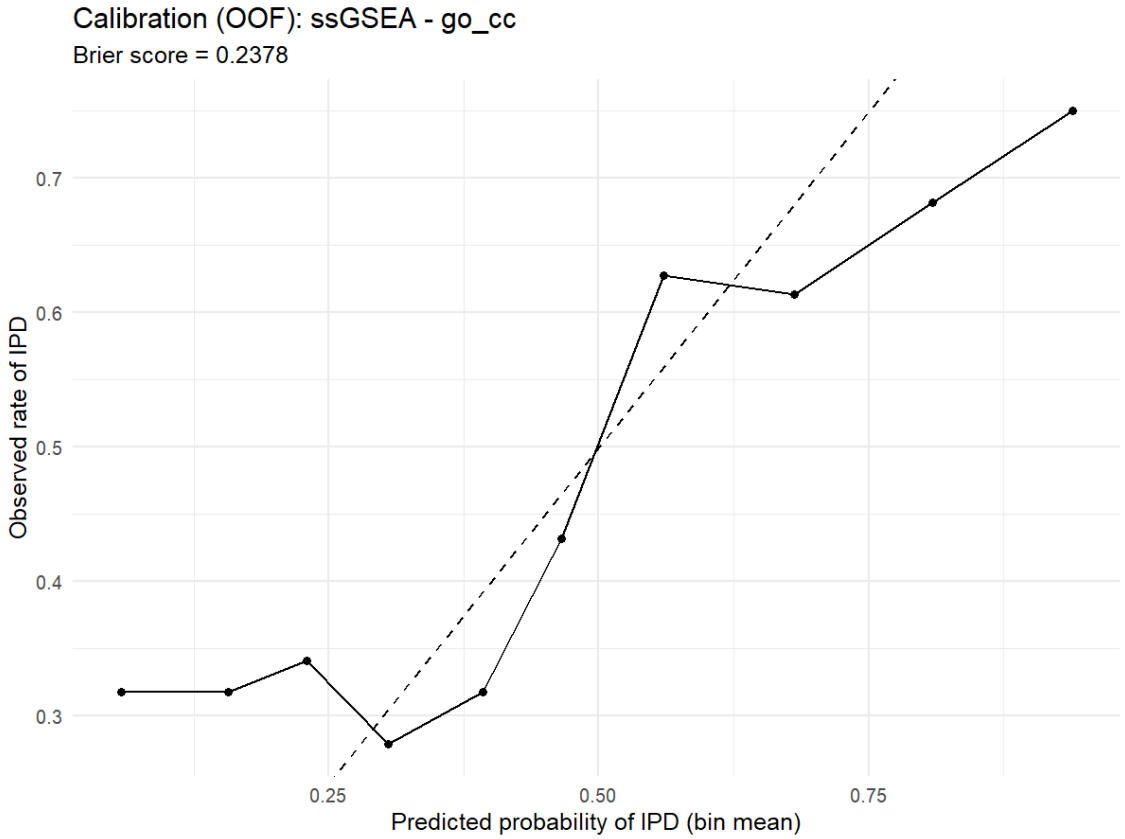


Figura 6. *Calibration plot de ssGSEA go_cc, donde la línea diagonal indica que las probabilidades predichas por el modelo coinciden con las observadas, acompañada del Brier score.*

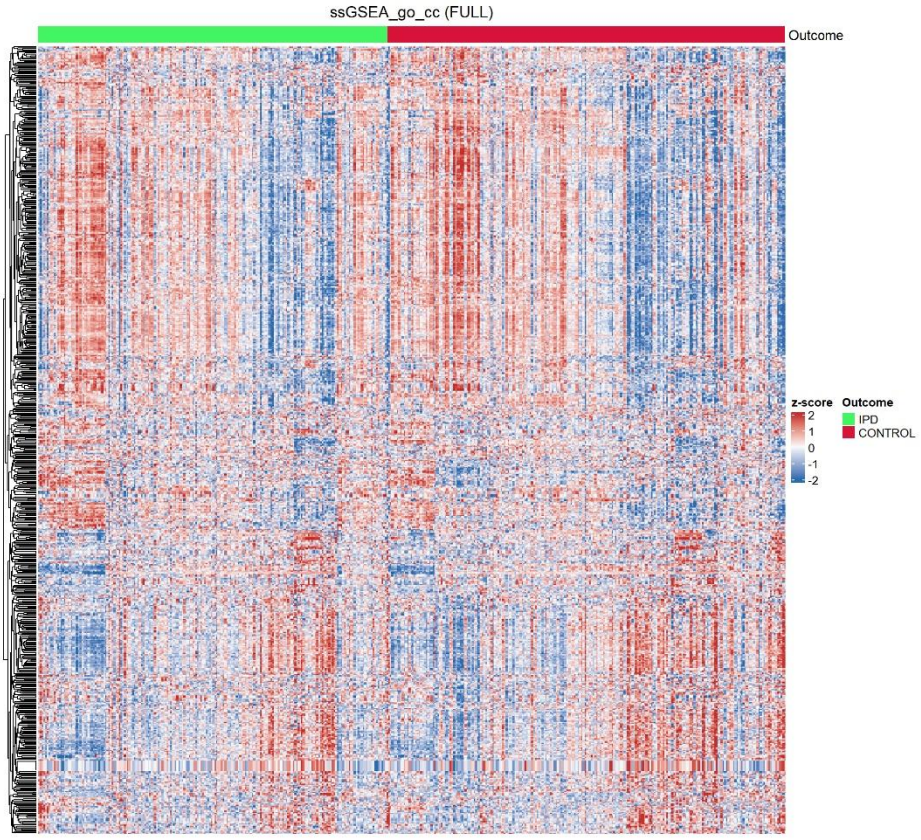
Heatmap de scores para clasificación supervisada

Las Figuras 7 y 8 muestran los mapas de calor generados mediante el método ssGSEA utilizando la base de datos Gene Ontology – Cellular Component (GO:CC), considerando tanto la matriz completa de rutas (FULL) como el subconjunto de las 100 rutas con mayor varianza entre las muestras (Top-100 var).

En ambos casos, los valores fueron estandarizados mediante z-score y se agruparon las muestras en función del tipo de paciente (Outcome: IPD vs. Control).

A nivel global, no se observan agrupaciones claramente diferenciadas entre los grupos IPD y control, ya que los patrones de activación y represión de las rutas parecen repetirse entre ambos conjuntos. Esto sugiere que, para esta base de datos en concreto (GO:CC), las diferencias funcionales entre condiciones son sutiles y distribuidas de manera heterogénea, sin una separación clara por clase.

Estos resultados indican que, al menos, bajo el enfoque ssGSEA con la ontología GO:CC, las señales moleculares asociadas a la enfermedad no son lo suficientemente fuertes como para generar una estratificación clara, probablemente debido a que las rutas celulares reflejan funciones comunes a ambos tipos de muestras, o a que las alteraciones relevantes para el Parkinson se expresan más intensamente en otras ontologías (por ejemplo, GO:BP o Reactome).



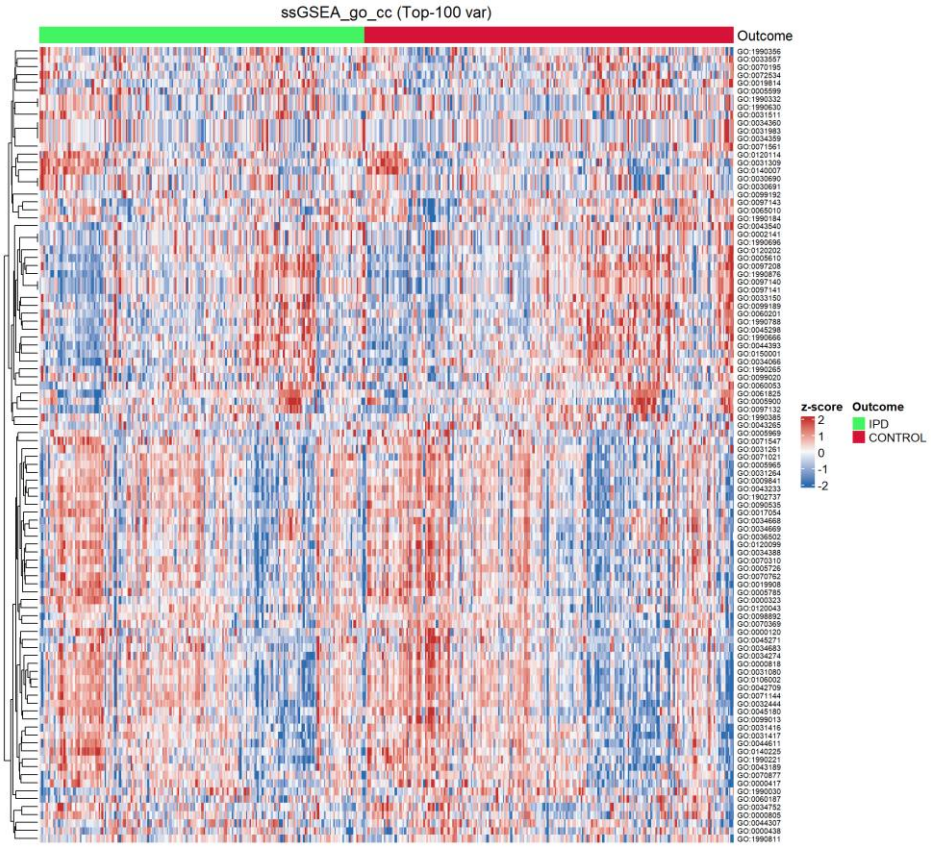


Figura 7 y 8. *Heatmap de top pathways vs muestras de ssGSEA go_cc, que muestra los scores estandarizados de los pathways más relevantes por muestra. Las filas representan pathways y las columnas muestras, agrupadas por similitud. La separación de los clusters de muestras refleja diferencias entre las clases (IPD/CONTROL).*

Importancia y términos más importantes para Parkinson

Para cada combinación de método de scoring y base de datos funcional, se identificaron los 10 términos con mayor relevancia predictiva según la importancia estimada por los modelos (Tabla 3). Por lo general, los modelos basados en árboles de decisión (Random Forest y XGBoost) mostraron valores de importancia más elevados y consistentes, mientras que los modelos de tipo k-NN presentaron valores más bajos y homogéneos. Esta diferencia se debe al método utilizado para el cálculo de la importancia (basado en la permutación de variables), que en modelos no paramétricos como k-NN tiende a generar distribuciones más uniformes. En estos algoritmos, la predicción depende del conjunto global de características utilizadas para calcular distancias entre muestras, de modo que el impacto individual de cada variable es menos pronunciado que en los modelos jerárquicos basados en árboles.

KNN – ssGSEA (DisGeNET y HPO)

En las bases de datos DisGeNET y HPO, los modelos k-NN identificaron rutas con importancias moderadas, aunque sin una señal dominante. En DisGeNET destacaron *X*

inactivation, familial skewed, Ciliary dyskinesia, primary y Macular degeneration, age-related, todas con importancias inferiores a 0,02.

En HPO, las rutas más relevantes fueron *Fusion of midphalangeal joints y Platelet-activating factor acetylhydrolase deficiency*, también con valores reducidos (<0,01), lo que indica una contribución limitada de las características fenotípicas a la clasificación en este caso.

KNN – ssGSEA (GO:BP y KEGG)

En los modelos basados en GO:BP y KEGG se observó un patrón similar. En GO:BP, los procesos más relevantes fueron la biosíntesis de péptidos antibacterianos activos contra bacterias Gram-negativas y la muerte mediada por neutrófilos de hongos, mientras que en KEGG destacaron *tryptophan metabolism* y *steroid biosynthesis.*

A pesar de su relevancia biológica potencial, las importancias permanecieron en valores bajos (<0,005), lo que sugiere que la combinación k-NN–ssGSEA en estas ontologías no captó diferencias moleculares robustas entre grupos.

XGBoost – ssGSEA (GO:CC)

El modelo XGBoost aplicado sobre GO:CC mostró una mejor capacidad de diferenciación. Entre las rutas con mayor peso se encontraron *cytoplasmic periphery of the nuclear pore complex, basal cortex y cation-transporting ATPase complex*, asociadas con procesos de organización estructural y transporte celular. Estas rutas reflejan posibles alteraciones funcionales en estructuras subcelulares implicadas en el mantenimiento de la homeostasis neuronal.

Random Forest – ssGSEA (GO:MF y Reactome)

Los modelos Random Forest obtuvieron los valores de importancia más altos.

En GO:MF, las rutas destacadas fueron *IgA receptor activity, H3-methyl-lysine-4 demethylase activity y cysteine-tRNA ligase activity*, vinculadas con actividad enzimática y regulación epigenética.

Por su parte, en Reactome se identificaron rutas como *Interleukin-33 signaling y The NLRP1 inflammasome*, asociadas con procesos inmunitarios e inflamatorios, junto con otras relacionadas con el transporte neuronal y la reparación celular (TCF7L2 signaling, SLC22A18 y SLC17A8).

Tabla 3. *Selección de los 10 términos con mayor importancia en la predicción, obtenidos mediante el método de scoring ssGSEA para cada base de datos funcional analizada.*

Base de datos	Término	Importancia	ID
HPO (kNN)	Fusion of midphalangeal joints	0,00776256	HP:0006187
	Platelet-activating factor acetylhydrolase deficiency	0,00776256	HP:0040175
	Hypoplastic spinal processes	0,00730594	HP:0008460

	Reflex asystolic syncope	0,00639269	HP:0500173
	Setting-sun eye phenomenon	0,00639269	HP:0012470
	Congenital stapes ankylosis	0,00639269	HP:0007943
	Abnormality of T cell physiology	0,00593607	HP:0011840
	Cone-shaped epiphyses of the distal phalanges of the hand	0,00593607	HP:0010248
	Irregular tarsal ossification	0,00547945	HP:0008134
	Multiple palmar creases	0,00547945	HP:0006114
DISGENET (kNN)	X INACTIVATION FAMILIAL SKEWED 1 (disorder)	0,01826484	C1848138
	CILIARY DYSKINESIA PRIMARY 28	0,01826484	C3809706
	MACULAR DEGENERATION AGE-RELATED 4 (disorder)	0,01735160	C1853147
	CORTICAL DYSPLASIA COMPLEX WITH OTHER BRAIN MALFORMATIONS 5	0,01643836	C3810407
	Amylo-1 6-glucosidase deficiency	0,01415525	C2936915
	Pelizaeus-Merzbacher-like disease autosomal recessive 2	0,01369863	C1850053
	MUSCULAR DYSTROPHY-DYSTROGLYCANOPATHY (CONGENITAL WITHOUT MENTAL RETARDATION) TYPE B 4	0,01324201	C2751052
	Epileptic Encephalopathy Early Infantile 4	0,01324201	C2677326
	SMITH-MCCORT DYSPLASIA 2	0,01278539	C3714896
	Cyclic neutropenia	0,01278539	C0221023
GO BP (kNN)	biosynthetic process of antibacterial peptides active against Gram-negative bacteria	0,01369863	GO:0002812
	benzene-containing compound metabolic process	0,01324201	GO:0042537
	neutrophil-mediated killing of fungus	0,01141553	GO:0070947
	negative regulation of retinoic acid biosynthetic process	0,01141553	GO:1900053
	deoxyribonucleotide metabolic process	0,01004566	GO:0009262
	cellular response to prostaglandin stimulus	0,00958904	GO:0071379
	positive regulation of fructose 1 6-bisphosphate metabolic process	0,00958904	GO:0060552
	proteasome core complex assembly	0,00913242	GO:0080129
	negative regulation of timing of catagen	0,00913242	GO:0051796
	response to indole-3-methanol	0,00867580	GO:0071680
KEGG (kNN)	Non-homologous end-joining	0,00502283	hsa03450
	Steroid biosynthesis	0,00456621	hsa00100
	Tryptophan metabolism	0,00456621	hsa00380
	Inflammatory mediator regulation of TRP channels	0,00365297	hsa04750
	Cholinergic synapse	0,00365297	hsa04725
	Metabolic pathways	0,00365297	hsa01100
	Ether lipid metabolism	0,00319635	hsa00565
	Long-term depression	0,00319635	hsa04730
	Prion disease	0,00319635	hsa05020
	Pantothenate and CoA biosynthesis	0,00273973	hsa00770
GO MF (Random forest)	IgA receptor activity	5,25974315	GO:0019766
	sodium:proton antiporter activity involved in regulation of cardiac muscle cell membrane potential	4,12046145	GO:0086040
	histone H3-methyl-lysine-4 demethylase activity	2,56978855	GO:0032453

	3'-5' RNA helicase activity	2,28411242	GO:0034458
	cysteine-tRNA ligase activity	2,05469217	GO:0004817
	sulfite oxidase activity	1,80046423	GO:0008482
	poly(C) RNA binding	1,26535806	GO:0017130
	pyruvate carboxylase activity	1,21281013	GO:0004736
	iodide peroxidase activity	1,20432706	GO:0004447
	C-rich strand telomeric DNA binding	1,14770711	GO:0061730
REACTOME (Random forest)	Defective SLC17A8 causes autosomal dominant deafness 25 (DFNA25)	4,56924653	R-HSA-5619076
	Signaling by TCF7L2 mutants	3,21122657	R-HSA-5339700
	Interleukin-33 signaling	1,72698744	R-HSA-9014843
	The NLRP1 inflammasome	1,47259751	R-HSA-844455
	Defective SLC22A18 causes lung cancer (LNCR) and embryonal rhabdomyosarcoma 1 (RMSE1)	1,45436519	R-HSA-5619066
	Defective UGT1A1 causes hyperbilirubinemia	1,39868008	R-HSA-5579002
	Proton-coupled monocarboxylate transport	1,29644632	R-HSA-433692
	Defective GALK1 can cause Galactosemia II (GALCT2)	1,10644959	R-HSA-5609976
	Severe congenital neutropenia type 4 (G6PC3)	1,07372881	R-HSA-3282872
	vRNA Synthesis	1,02990944	R-HSA-192814
GO CC (xgBoost)	cytoplasmic periphery of the nuclear pore complex	0,02960818	GO:1990723
	epidermal lamellar body	0,02944907	GO:0097209
	basal cortex	0,02813825	GO:0045180
	acetolactate synthase complex	0,02634132	GO:0005948
	cation-transporting ATPase complex	0,02612021	GO:0090533
	inactive sex chromosome	0,02471670	GO:0098577
	macrophage migration inhibitory factor receptor complex	0,01924112	GO:0035692
	galectin complex	0,01900258	GO:1990724
	tertiary granule	0,01639767	GO:0070820
	extrinsic component of lysosome membrane	0,01542273	GO:0032419

Partiendo de los resultados obtenidos en la Tabla 3, se realizó una revisión específica de los términos que podían ser más relevantes para la enfermedad de Parkinson (EP), seleccionando de esta manera un total de 15 términos (Tabla 4). Estos fueron elegidos por su coherencia con mecanismos previamente descritos en la literatura, así como por su presencia recurrente entre distintos modelos y bases de datos funcionales.

Entre los términos destacados se encuentran procesos relacionados con la respuesta inflamatoria e inmunitaria (*Interleukin-33 signaling, The NLRP1 inflammasome*), la disfunción mitocondrial y el metabolismo energético (*Tryptophan metabolism, Ether lipid metabolism*), y la actividad sináptica y neurotransmisión (*Cholinergic synapse, Long-term depression*). Asimismo, se identificaron rutas vinculadas con la degradación proteica y el control de calidad celular *(proteasome core complex assembly, extrinsic component of lysosome membrane*), junto con procesos estructurales y de transporte

intracelular (*cytoplasmic periphery of the nuclear pore complex, cation-transporting ATPase complex*).

En conjunto, estos resultados sugieren que los mecanismos funcionales más relevantes para la EP abarcan procesos inflamatorios, metabólicos y sinápticos, en consonancia con la fisiopatología conocida de la enfermedad.

Tabla 4. *Términos con mayor importancia en la predicción y con relevancia biológica en la enfermedad de Parkinson. Esta tabla se basa en los resultados de la Tabla 3, considerando únicamente los términos de mayor peso obtenidos mediante ssGSEA.*

Base de datos	Término	Importancia	Tipo de relevancia	Descripción
GO BP (kNN)	negative regulation of retinoic acid biosynthetic process	0,01141553	Neurodesarrollo dopaminérgico	La señalización del ácido retinoico regula la diferenciación y mantenimiento de neuronas dopaminérgicas y se ha implicado en la fisiopatología de la PD (Marie, 2021).
	cellular response to prostaglandin stimulus	0,00958904	Neuroinflamación / Mediadores lipídicos	PGE_2 vía EP4 modula neuroinflamación y protege frente a la pérdida dopaminérgica en modelos de PD (Teismann et al., 2003).
	proteasome core complex assembly	0,00913242	Proteostasis	Disfunción del sistema ubiquitina-proteasoma favorece acumulación de α-sinucleína y contribuye a la neurodegeneración dopaminérgica en PD (Cook, 2009).
KEGG (kNN)	Tryptophan metabolism	0,00456621	Metabolismo / neuroinflamación	Alteraciones en la vía del triptófano/quinurenina se han descrito en PD y pueden modular neuroinflamación y toxicidad (Venkatesan, 2020).
	Inflammatory mediator regulation of TRP channels	0,00365297	Estrés oxidativo / Canales iónicos / Neuroinflamación	Canales TRPM2 y TRPV1 se han implicado como moduladores de síntomas motores y no motores y de la respuesta inflamatoria en PD (Vaidya, 2020).
	Cholinergic synapse	0,00365297	Neurotransmisión / Degeneración colinérgica	Hay degeneración del sistema colinérgico basal y alteraciones corticales en PD, relacionadas con déficits cognitivos y de atención (Slater et al., 2024).
	Ether lipid metabolism	0,00319635	Metabolismo lipídico	Reducción de plasmalógenos en cerebro y plasma de pacientes con PD. En un caso, su suplementación mejoró niveles y algunos síntomas (Mawatari et al., 2020).
	Long-term depression	0,00319635	Plasticidad cortico - estriatal	Alteraciones en la LTD corticoestriatal contribuyen a los déficits motores de PD (Zhai et al., 2018).
	Prion disease	0,00319635	Mecanismo patológico tipo prion	La α-sinucleína presenta propagación tipo prion-like, un mecanismo propuesto para la progresión de la patología en PD (Brundin & Melki, 2017).
REACTOME (Random forest)	Interleukin-33 signaling	1,72698744	Astrocitos-microglía / inmunidad	IL-33 regula microglía; su deficiencia agrava neurodegeneración dopaminérgica (Shen et al., 2025).
	The NLRP1 inflammasome	1,47259751	Neuroinflamación / Inflammasoma microgliales	La activación de inflammasomas en microglía, especialmente NLRP3 y también NLRP1, desencadena la liberación de IL-1β y la muerte dopaminérgica en modelos de Parkinson. La inhibición de NLRP3 reduce la acumulación de α-sinucleína y protege neuronas nigroestriatales. Se considera un eje inflamatorio compartido en la familia NLRP (Gordon et al., 2018).
GO CC (xgBoost)	cytoplasmic periphery of the nuclear pore complex	0,02960818	Transporte núcleo-citoplasmático	Se han descrito alteraciones del poro nuclear y del transporte nucleocitoplásmico en neurodegeneración; hay evidencia emergente en PD (Riaz et al., 2024).

	cation-transporting ATPase complex	0,02612021	Genética / Biología lisosomal	ATP13A2/PARK9, ATPasa lisosomal de esta familia, está mutada en parkinsonismo de inicio precoz; su pérdida causa deficiencia lisosomal y neurodegeneración (Dehay et al., 2012).
	galectin complex	0,01900258	Microglía / Neuroinflamación	Galectina-3 promueve activación microglial por α-sinucleína y se propone como diana terapéutica en PD. (Boza-Serrano et al., 2014).
	extrinsic component of lysosome membrane	0,01542273	Lisosomas / Degradación de α-sinucleína	La chaperone-mediated autophagy (CMA) degrada α-sinucleína; LAMP2A y hsc70 están reducidas en cerebros de PD, apoyando un defecto lisosomal (Alvarez-Erviti, 2010).

Clustering no supervisado

Tras realizar el clustering no supervisado para las 42 combinaciones de métodos de scoring y bases de datos funcionales, se obtuvo la Tabla 5, que resume las métricas clave de rendimiento y significancia clínica de cada modelo. El análisis reveló un contraste notable entre la calidad estructural de los clústeres y su relevancia biológica.

La combinación Zscore_reactome alcanzó el coeficiente de Silhouette promedio más alto (0,3982), lo que indica una buena separación matemática entre grupos. No obstante, su p-valor de separación no fue significativo (p = 0,9642), sugiriendo que los clústeres formados no se correspondían con diferencias clínicas relevantes.

Por el contrario, la combinación Zscore_hpo destacó como la más prometedora desde una perspectiva clinicobiológica, al presentar el p-valor más bajo (p = 0,0047), lo que evidencia una asociación significativa entre la estratificación molecular y el diagnóstico de los pacientes. Aunque su coeficiente de Silhouette fue más moderado (0,2731), se encuentra dentro de un rango aceptable que indica la existencia de una estructura subyacente coherente. Este modelo identificó un número óptimo de tres clústeres (K = 3), reflejando una heterogeneidad biológica mayor que una simple división binaria IPD-control, pero manteniendo una complejidad interpretable. De forma preliminar, uno de los subgrupos concentró aproximadamente el 63 % de los pacientes con IPD, siendo uno de los valores más altos entre las combinaciones evaluadas, siendo la combinación ssGSEA_GO:CC la que alcanzó el máximo con un 65,85 %. En cuanto al índice RCSI, la combinación Zscore_hpo alcanzó un valor de 0,7421, indicativo de una estabilidad moderadamente alta de la partición. Si bien no se encuentra entre los valores máximos del conjunto, supera con amplitud a la mayoría de combinaciones evaluadas, situándose dentro de un rango considerado estable y reproducible para análisis de clustering basados en datos de expresión molecular.

Por estos motivos, se seleccionó la combinación Zscore_hpo como modelo de referencia para los análisis de clustering posteriores, cuyos resultados se detallan en las siguientes secciones.

Tabla 5. *Resumen de los resultados del análisis de clustering en formato tabular. Los valores más óptimos para cada métrica se señalan en color amarillo, cuando corresponde. Las combinaciones se encuentran ordenadas de menor a mayor según el p-valor obtenido en la comparación entre las condiciones clínicas (CONTROL vs IPD).*

Combination	Optimal_K	Max_RCSI	Silhouette_Avg	Separation_PValue	Max_Prop_IPD
Zscore_hpo	3	0,7421	0,2731	0,004666917	0,6364
norm_FGSEA_reactome	6	0,554	0,0472	0,005387272	0,6154
singscore_go_cc	6	0,2362	0,0626	0,021851073	0,65
ssGSEA_go_cc	4	0,359	0,1039	0,046598594	0,6585
singscore_go_bp	4	0,0802	0,1231	0,095179181	0,641
singscore_reactome	2	0,847	0,2045	0,099674862	0,509
ssGSEA_reactome	2	0,7937	0,2009	0,099674862	0,509
norm_FGSEA_go_bp	5	0,1894	0,0313	0,138502696	0,5797
norm_FGSEA_hpo	2	1,4001	0,1697	0,207407484	0,4913
Plage_go_bp	2	0,4112	0,3765	0,209199381	0,496
Plage_disgenet	2	0,9622	0,3547	0,261815945	0,4957
Plage_hpo	3	0,4114	0,3021	0,30916032	0,4954
Plage_reactome	4	0,7183	0,2496	0,336803565	0,5205
GSVA_disgenet	2	2,1267	0,3249	0,344458866	0,4931
ssGSEA_hpo	2	0,619	0,226	0,365289662	0,4896
singscore_hpo	2	0,4961	0,2248	0,365289662	0,4896
norm_FGSEA_disgenet	6	0,1735	0,0341	0,366337699	0,5849
Plage_go_mf	2	0,3363	0,3375	0,404595527	0,4865
Zscore_kegg	2	0,6203	0,3108	0,444020568	0,4859
GSVA_kegg	2	1,5445	0,3687	0,483235751	0,4866
norm_FGSEA_kegg	2	3,9455	0,2273	0,502908175	0,4877
GSVA_hpo	2	1,8086	0,2614	0,506560865	0,4862
GSVA_go_bp	2	1,9046	0,299	0,53266255	0,4846
ssGSEA_disgenet	4	0,4149	0,1108	0,535192081	0,4878
norm_FGSEA_go_cc	2	0,4405	0,1763	0,544644935	0,4797
GSVA_go_mf	2	1,6066	0,2513	0,57388602	0,4839
Plage_go_cc	2	1,8126	0,3504	0,583723974	0,4808
Plage_kegg	5	0,128	0,3229	0,600332815	0,5
Zscore_go_cc	2	0,4738	0,3363	0,60217734	0,4805
singscore_disgenet	4	0,4082	0,1109	0,614924953	0,4878
Zscore_go_bp	2	0,339	0,3802	0,620640098	0,4802
Zscore_go_mf	2	0,6609	0,3367	0,666705862	0,4786
norm_FGSEA_go_mf	3	0,7112	0,0828	0,67507772	0,4937
GSVA_reactome	2	2,166	0,237	0,719885654	0,4791
Zscore_disgenet	2	0,4503	0,344	0,731852213	0,4781
singscore_go_mf	3	0,2844	0,1418	0,776014089	0,487
ssGSEA_go_mf	3	0,3976	0,1385	0,776014089	0,487
ssGSEA_kegg	2	1,1504	0,3037	0,875643528	0,4727
GSVA_go_cc	2	2,2135	0,2424	0,886573395	0,4739

ssGSEA_go_bp	3	0,2247	0,1494	0,895954438	0,4803
singscore_kegg	2	1,5298	0,3022	0,959288708	0,4708
Zscore_reactome	2	1,1974	0,3982	0,96421045	0,4718

El gráfico del índice RCSI (Figura 9) muestra un mínimo local en K = 2 y un máximo pronunciado en K = 3, como se ha observado anteriormente en la Tabla 5, seguido de un descenso progresivo al incrementar el número de clústeres. Este patrón indica la existencia de tres subgrupos moleculares potenciales dentro del conjunto de muestras. No obstante, los valores absolutos del índice fueron moderados, lo que sugiere una estabilidad limitada y cierto solapamiento entre clústeres, coherente con la complejidad biológica esperada en estudios de expresión génica en EP.

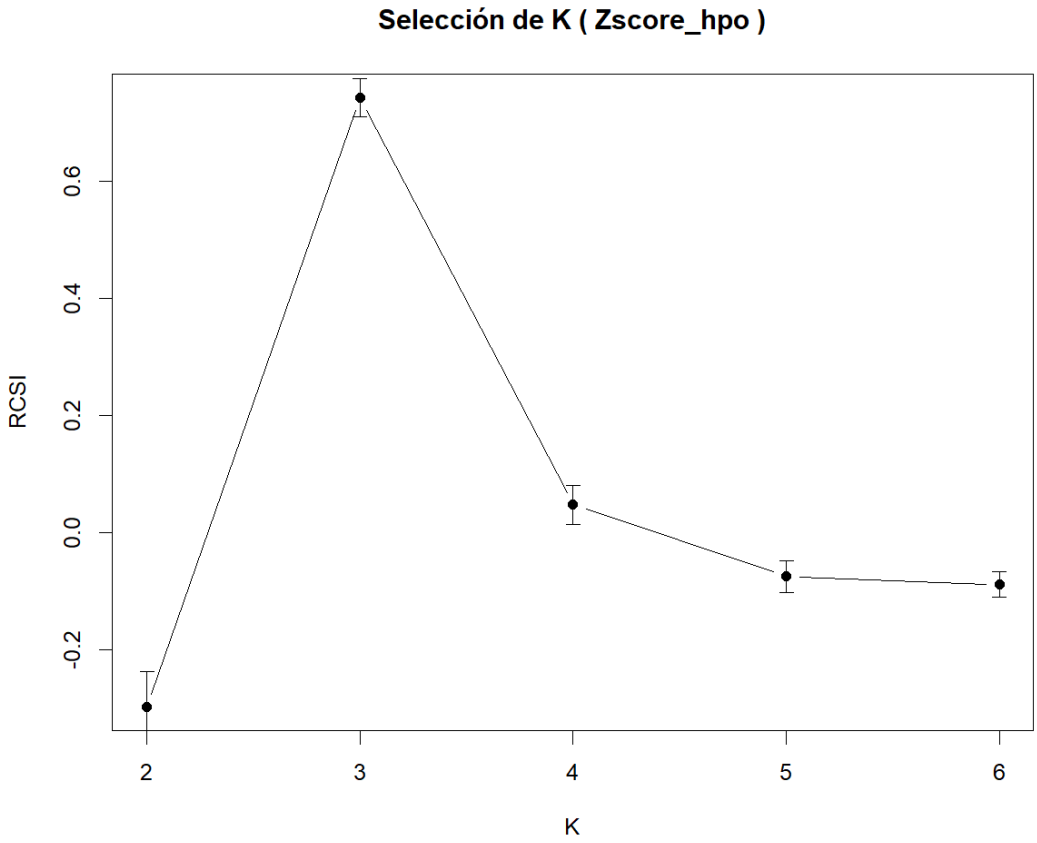


Figura 9. *Gráfico del índice RCSI (Relative Cluster Stability Index) obtenido por el algoritmo M3C a partir de los scores de Zscore_hpo. El pico máximo se alcanzó en K = 3, indicando la presencia de tres posibles grupos con estabilidad moderada.*

El análisis de silueta (Figura 10) evidenció una cohesión interna moderada dentro de los clústeres y una separación limitada entre ellos. Los valores individuales de la anchura de silueta oscilaron aproximadamente entre 0,15 y 0,45, con una media global de 0,273, lo

que indica que una parte de las muestras se sitúa en regiones intermedias entre los grupos definidos.

En conjunto, estos resultados sugieren una estructura de clústeres débilmente definida, más compatible con una variabilidad continua que con subgrupos claramente separados.

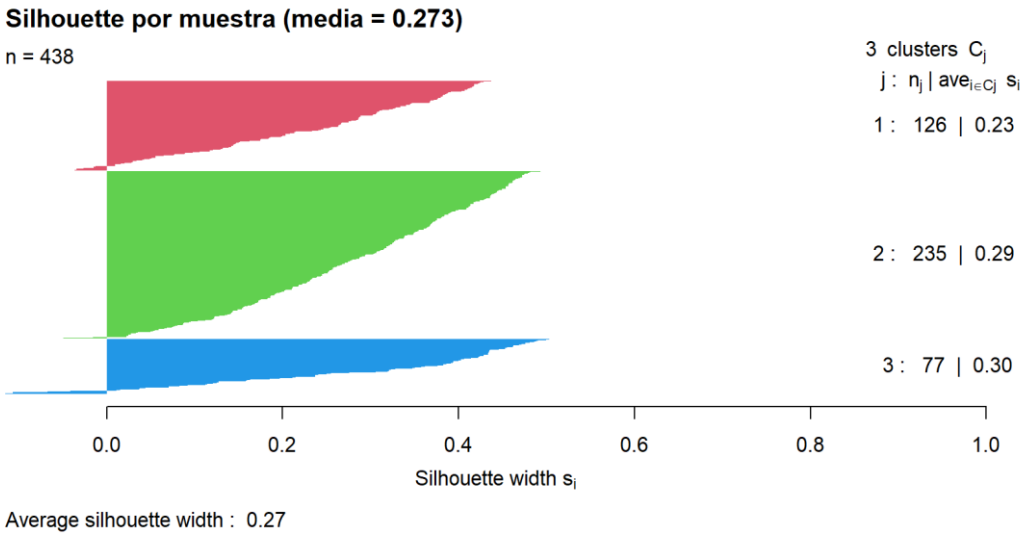


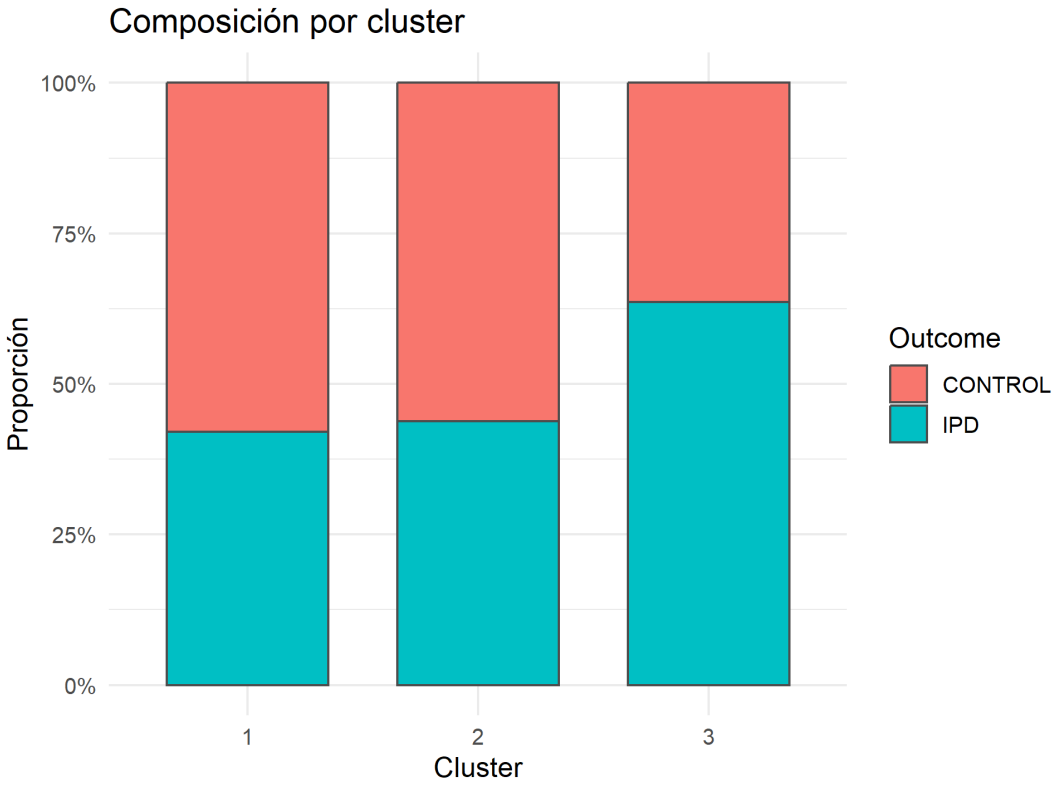
Figura 10. *Gráfico de anchura de silueta para K = 3 clusters obtenidos con M3C. El valor medio fue de 0,273, indicando una cohesión interna y separación entre grupos moderada.*

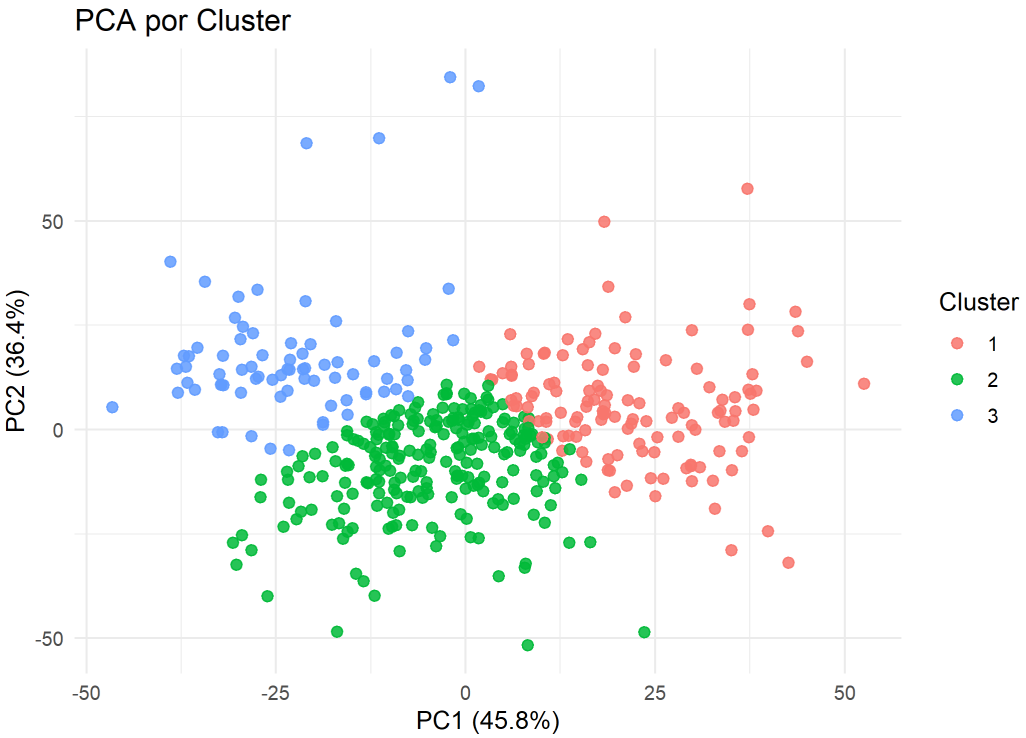
El gráfico de composición por clúster (Figura 11) muestra proporciones similares de pacientes IPD y controles en los clústeres 1 y 2, mientras que el clúster 3 concentra una mayor proporción de individuos con IPD, en concordancia con el resultado previamente observado en la Tabla 5. Aunque las diferencias no son extremas, esta distribución sugiere que los subgrupos identificados podrían reflejar gradientes moleculares asociados a la enfermedad, más que divisiones discretas entre pacientes y controles.

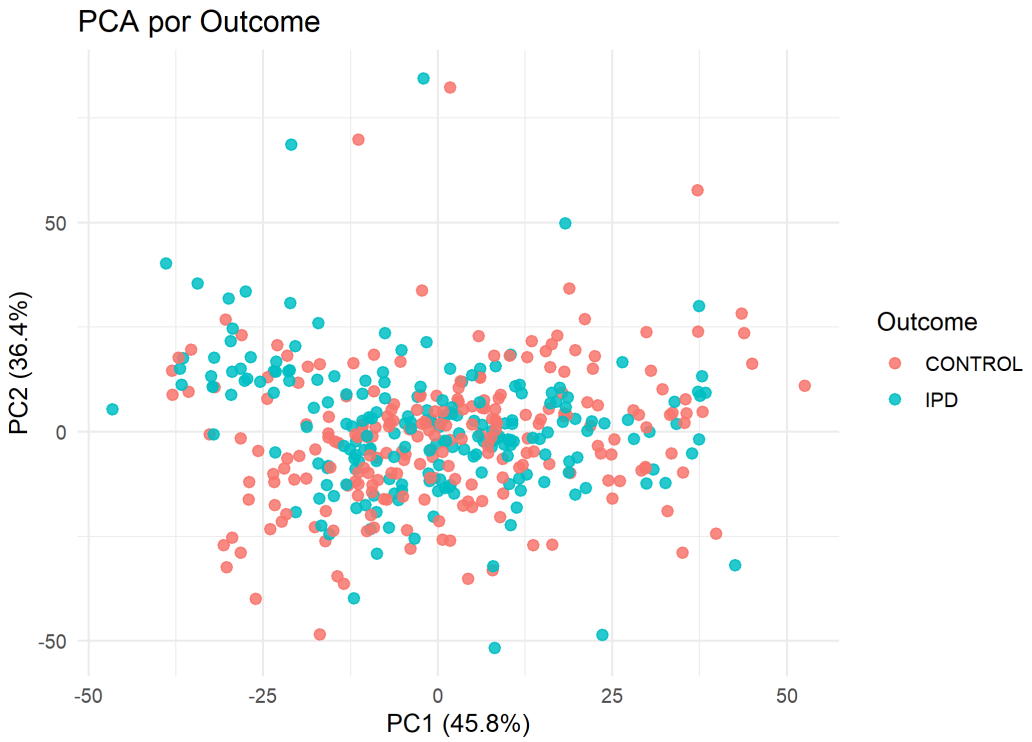
En el análisis de componentes principales (PCA) coloreado por clúster (Figura 12), se aprecia una tendencia leve a la separación entre grupos, especialmente para el clúster 3, que tiende a ocupar una región más definida del espacio de componentes. Sin embargo, el solapamiento entre los clústeres es considerable, lo que concuerda con los resultados del análisis de silueta y refuerza la idea de una estructura parcialmente difusa, típica de fenómenos biológicos complejos y multifactoriales.

Por su parte, el PCA coloreado por la variable clínica (Figura 13) revela una distribución prácticamente indistinguible entre pacientes y controles, lo que indica que la estratificación obtenida no está determinada directamente por el diagnóstico clínico, sino que probablemente responde a variaciones moleculares internas dentro del conjunto de muestras.

Este resultado respalda la validez exploratoria del clustering, sugiriendo que los subgrupos identificados podrían representar perfiles biológicos complementarios al fenotipo clínico tradicional.







Figuras 11, 12 y 13. *Para la combinación Zscore_hpo: (11) Distribución de muestras IPD y controles dentro de cada clúster. (12) Análisis PCA coloreado por clúster, mostrando una separación parcial. (13) PCA coloreado por condición clínica (IPD/Control) para comparación con la variable fenotípica.*

El mapa de calor obtenido a partir del modelo Zscore_hpo (Figura 14) muestra tres patrones globales de activación molecular que se corresponden con los clústeres identificados en el análisis anterior. Cada grupo presenta un perfil característico de rutas funcionales, con bloques definidos de activación (rojo) y represión (azul), lo que indica que las muestras dentro de cada clúster comparten tendencias moleculares comunes. Sin embargo, las transiciones graduales entre zonas rojas y azules sugieren la existencia de solapamiento entre subgrupos, coherente con los resultados del análisis de silueta y PCA.

En conjunto, el heatmap respalda la validez estructural del modelo de tres clústeres (K = 3), evidenciando una organización molecular heterogénea pero reproducible dentro del conjunto de pacientes.

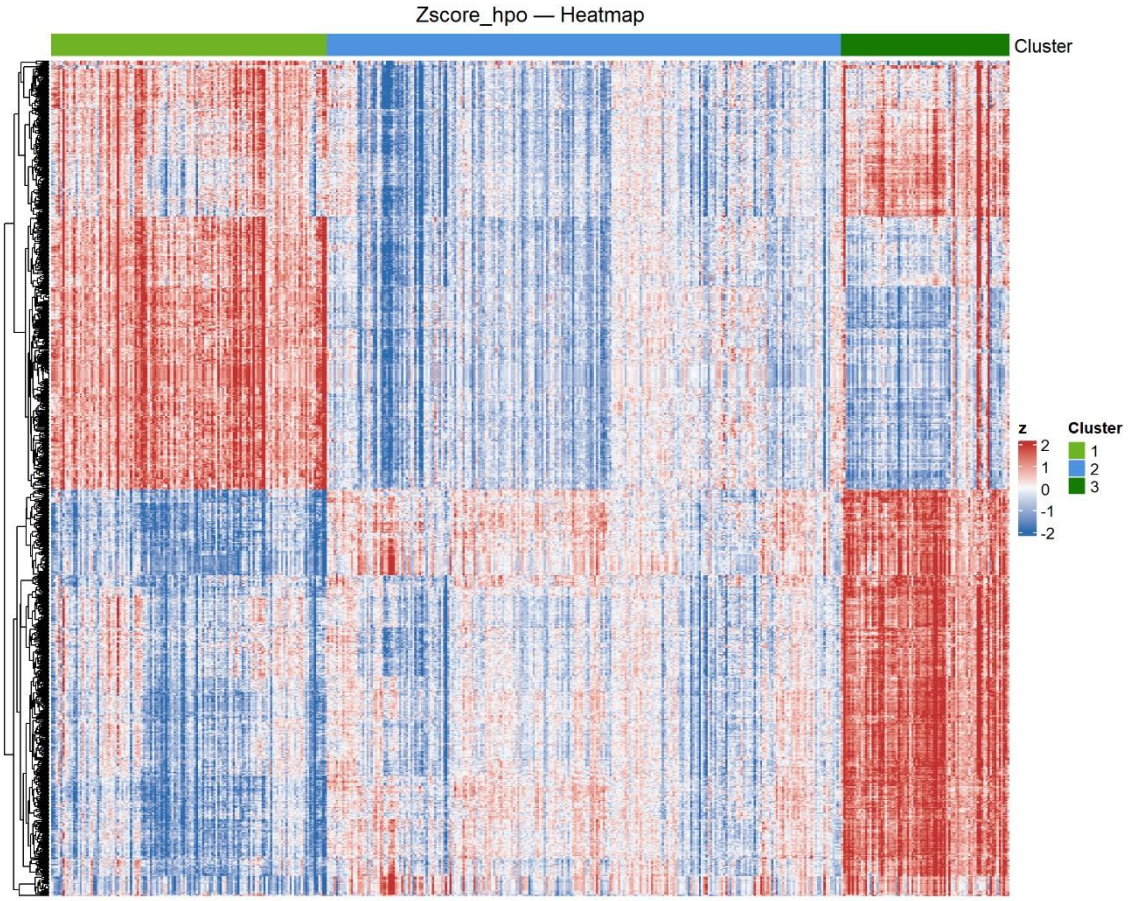


Figura 14. *Mapa de calor de los scores Zscore_hpo, estandarizados (z-score) y ordenados según la pertenencia a clúster. Se representan las rutas en filas y las muestras en columnas.*

Los resultados del test de chi-cuadrado (Tabla 6) confirmaron una asociación significativa entre los clústeres identificados y la condición clínica (IPD o control) (χ^2 = 10.735, gl = 2, p = 0.0047). Este resultado indica que la distribución de pacientes y controles difiere significativamente entre los tres clústeres, lo que sugiere que la estratificación obtenida con el modelo Zscore–HPO captura patrones moleculares parcialmente asociados con el diagnóstico clínico.

Tabla 6. *Tabla de contingencia (Clúster vs. Condición clínica) y resultados del test χ^2 de independencia para Zscore_hpo.*

Clúster	Control	IPD
1	75	53
2	132	103
3	28	49
Pearson's Chi-squared test / X-squared = 10.735, df = 2, p-value = 0.004667		

El análisis de rutas diferenciales entre los subgrupos identificados mediante el modelo Zscore_hpo reveló diferencias biológicas significativas entre las combinaciones de clústeres (Figuras 15–17).

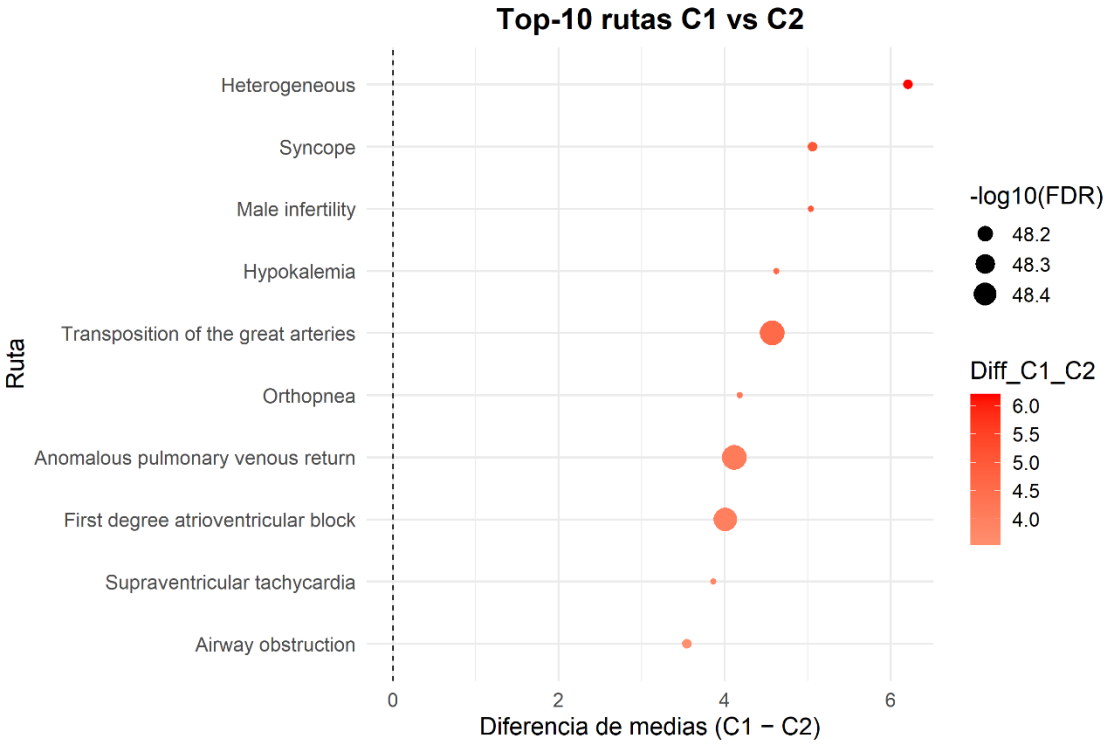
En la comparación C1 vs C2, se observaron rutas asociadas principalmente con funciones cardiorrespiratorias y circulatorias, incluyendo airway obstruction, first degree atrioventricular block, anomalous pulmonary venous return y transposition of the great arteries, predominantemente sobreexpresadas en el clúster 1.

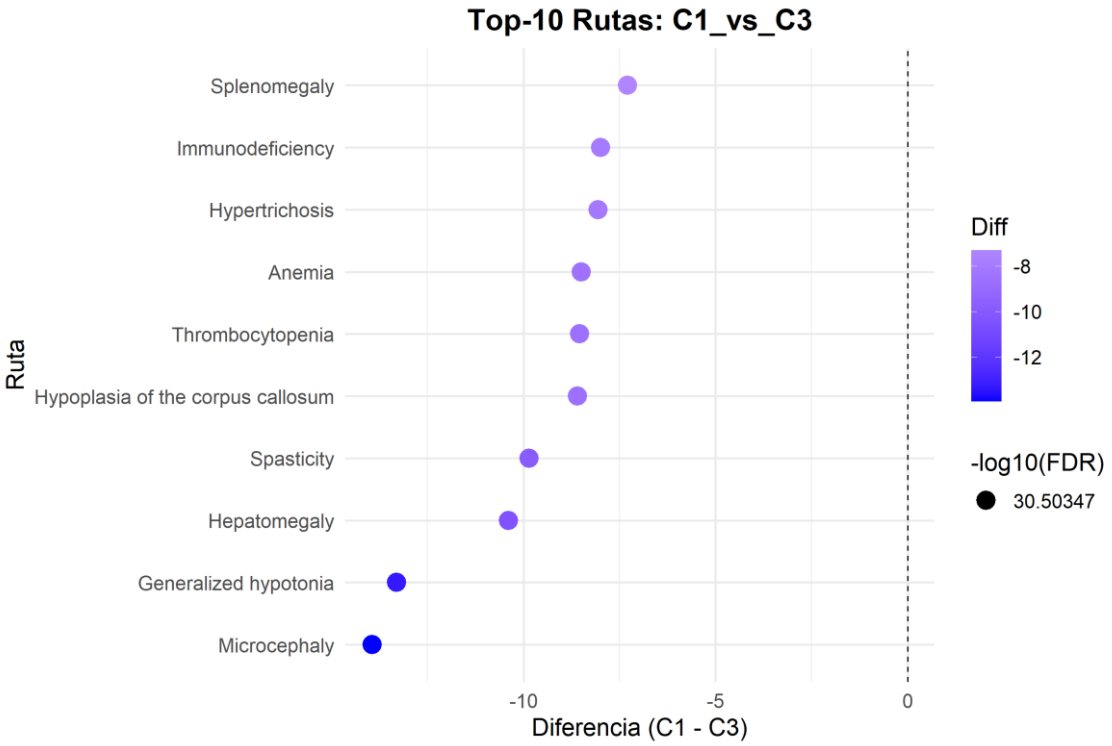
Por otro lado, la comparación C1 vs C3 mostró diferencias en rutas vinculadas a alteraciones hematológicas e inmunológicas, tales como anemia, thrombocytopenia, immunodeficiency o splenomegaly, además de procesos del desarrollo como hypoplasia of the corpus callosum y microcephaly, con mayor activación en el clúster 3.

Finalmente, en la comparación C2 vs C3, destacaron términos asociados a trastornos neurológicos y del desarrollo, incluyendo microcephaly, global developmental delay e infantile onset, que se encontraron principalmente sobreexpresados en el clúster 3.

En conjunto, estos resultados evidencian que los clústeres definidos presentan perfiles moleculares funcionalmente diferenciados, coherentes con los patrones observados en el heatmap, donde se apreciaban bloques de activación y represión específicos. Ello respalda la validez funcional del modelo de estratificación, indicando que los subgrupos obtenidos no solo muestran una estructura estadística reproducible, sino también diferencias biológicas detectables a nivel de rutas moleculares.

La interpretación fisiopatológica detallada de estos hallazgos y su posible relación con los mecanismos del Parkinson o con procesos sistémicos asociados se aborda en la sección de Discusión.


Top-10 rutas C1 vs C2
Heterogeneous
Syncope
Male infertility
Hypokalemia
Transposition of the great arteries
Orthopnea
Anomalous pulmonary venous return
First degree atrioventricular block
Supraventricular tachycardia
Airway obstruction
Ruta
0
2
4
6
Diferencia de medias (C1 − C2)
-log10(FDR)
48.2
48.3
48.4
Diff_C1_C2
6.0
5.5
5.0
4.5
4.0


Top-10 Rutas: C1_vs_C3
Splenomegaly
Immunodeficiency
Hypertrichosis
Anemia
Thrombocytopenia
Hypoplasia of the corpus callosum
Spasticity
Hepatomegaly
Generalized hypotonia
Microcephaly
Ruta
-10
-5
0
Diferencia (C1 - C3)
Diff
-8
-10
-12
-log10(FDR)
30.50347

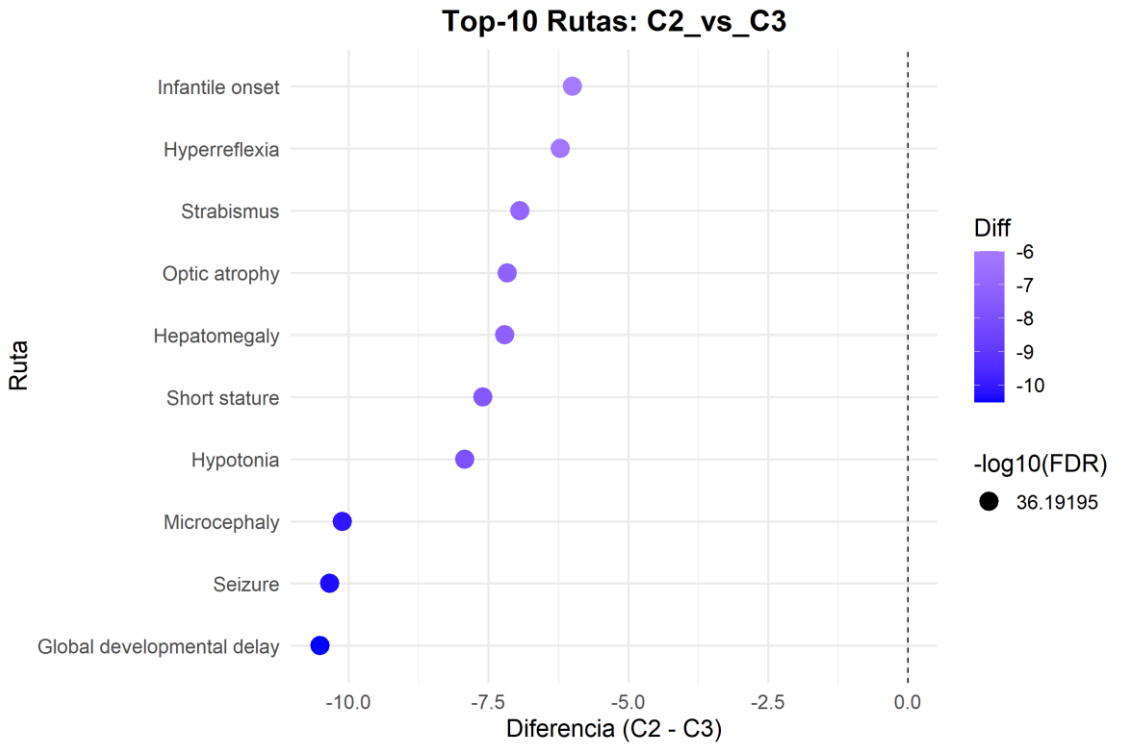


Figura 15 - 17. *Representación de las 10 rutas con mayor diferencia de puntuación media entre los clústeres, obtenidas mediante la combinación Zscore–HPO.El eje X muestra la diferencia entre las medias de los scores moleculares para cada comparación (C1 − C2, C1 − C3 y C2 − C3).Los valores positivos indican una mayor activación relativa en el primer clúster de la comparación, mientras que los valores negativos corresponden a una mayor activación en el segundo.El color refleja la magnitud de la diferencia media y el tamaño de los puntos representa el nivel de significación estadística (−log_{10} FDR).*

5. Discusión

Tal y como se observó en el PCA exploratorio inicial (Figura 1), la elevada dispersión de las muestras refleja una marcada heterogeneidad. Este comportamiento anticipa la dificultad para obtener modelos predictivos robustos, tanto en la clasificación supervisada (pathMED) como en la estratificación no supervisada (M3C).

En consonancia con ello, las métricas de rendimiento obtenidas para las 42 combinaciones de modelos (Tabla 2, Figura 2) mostraron valores moderados, sin que ninguno destacara de forma sobresaliente. El modelo con mejor desempeño fue el que combinó el algoritmo XGBoost con el método de cálculo de scores ssGSEA y la base de datos GO Cellular Components, que alcanzó los valores más altos en la mayoría de métricas, excepto en Recall y Specificity. En general, los modelos lograron captar cierta señal discriminante, aunque las diferencias entre pacientes y controles fueron sutiles. La mayor Specificity sugiere que los modelos tienden a reconocer mejor a los controles

que a los pacientes con enfermedad de Parkinson idiopática (IPD), coherente con la complejidad y variabilidad de la expresión génica en esta patología.

Las curvas ROC y PR (Figuras 3–4) confirmaron la capacidad discriminante limitada, con valores de AUC/AUPRC entre 0,55 y 0,69. Incluso el modelo seleccionado presentó AUC = 0,682 y AUPRC = 0,660, indicando una separación modesta pero no aleatoria. Las matrices de confusión y las curvas de calibración (Figuras 5–6) mostraron que, aunque los modelos superan el azar y presentan una calibración razonable, persisten márgenes de mejora. Los heatmaps (Figuras 7–8) también evidencian una ausencia de agrupaciones nítidas, lo que refuerza la idea de diferencias moleculares difusas entre condiciones.

En el análisis de importancia de variables (Tablas 3–4), se identificaron rutas con posible relevancia para la fisiopatología del Parkinson. Entre ellas destacan procesos relacionados con la respuesta inflamatoria (Interleukin-33 signaling, NLRP1 inflammasome), el metabolismo y la función mitocondrial (Tryptophan metabolism, Ether lipid metabolism), la neurotransmisión y plasticidad sináptica (Cholinergic synapse, Long-term depression), así como mecanismos de degradación proteica y transporte iónico (proteasome core complex assembly, cation-transporting ATPase complex). Estas rutas, coherentes con hallazgos previos en la literatura, sugieren que los modelos logran capturar señales biológicamente plausibles, aun cuando su rendimiento predictivo global sea limitado.

En cuanto al clustering no supervisado (Tablas 5–6; Figuras 9–15), los resultados fueron también moderados. El modelo con mejor equilibrio entre rendimiento estadístico y coherencia biológica fue el correspondiente a la combinación Zscore–HPO, que obtuvo el p-valor más bajo (0,0047) y un coeficiente de Silhouette de 0,2731, indicando una estructura débilmente definida pero existente. Este modelo identificó tres clústeres (K = 3), lo que refleja la heterogeneidad intrínseca de las muestras, manteniendo al mismo tiempo una complejidad interpretable. El tercer clúster concentró cerca del 66 % de pacientes con IPD, siendo uno de los valores más altos observados. El índice RCSI = 0,7421 indicó una estabilidad moderadamente alta, consistente con la dispersión visualizada en los gráficos de silueta y PCA. No obstante, al analizar la relación de estos clústeres con las variables clínicas, no se observó una correspondencia directa entre la partición molecular y el diagnóstico clínico, lo que sugiere que la estratificación obtenida captura variabilidad biológica no explicada por el fenotipo clínico clásico.

Las rutas diferenciales entre los clústeres 1 y 2 (Figura 15) mostraron alteraciones principalmente relacionadas con el sistema cardiovascular y respiratorio (airway obstruction, atrioventricular block, anomalous pulmonary venous return), lo que indica que parte de la señal transcripcional podría reflejar efectos sistémicos o comorbilidades presentes en la cohorte.

Por su parte, la comparación C1 vs C3 (Figura 16) evidenció diferencias centradas en procesos inmunitarios, hematológicos y del desarrollo, con términos como anemia, thrombocytopenia, immunodeficiency, splenomegaly o hypoplasia of the corpus

callosum, que sugieren una activación diferencial de mecanismos de respuesta inmune y maduración celular en el clúster 3.

Finalmente, en la comparación C2 vs C3 (Figura 17) destacaron rutas vinculadas a fenotipos neurológicos y del desarrollo (microcephaly, global developmental delay o infantile onset), también enriquecidas en el clúster 3. Aunque estos términos proceden de la ontología fenotípica (HPO) y no implican la presencia clínica de tales rasgos, su aparición apunta a una posible convergencia molecular con procesos neurodegenerativos o de desarrollo neuronal, lo que podría reflejar una diferenciación funcional relevante dentro de los subgrupos de pacientes con enfermedad de Parkinson.

En conjunto, estos resultados evidencian que los clústeres definidos presentan perfiles moleculares funcionalmente diferenciados, coherentes con los patrones observados en el heatmap, donde se apreciaban bloques de activación y represión específicos. Ello respalda la validez funcional del modelo de estratificación, indicando que los subgrupos obtenidos no solo muestran una estructura estadística reproducible, sino también diferencias biológicas detectables a nivel de rutas moleculares.

Esta diferenciación funcional se alinea con los hallazgos globales del trabajo, en los que tanto los enfoques supervisados como los no supervisados revelan una heterogeneidad molecular marcada y la coexistencia de mecanismos inmunitarios, metabólicos y neuronales en la fisiopatología de la enfermedad.

Estos resultados son coherentes con estudios recientes que apuntan a la complejidad y heterogeneidad biológica de la enfermedad de Parkinson. De manera similar a lo observado por Ryan, Marioni y Simpson (2025) en su enfoque longitudinal de estratificación, nuestros análisis muestran que, incluso al emplear distintos métodos de enriquecimiento y clasificación, los subgrupos moleculares tienden a solaparse, reflejando una variabilidad continua más que categorías discretas. Esta falta de separación clara sugiere que los perfiles transcriptómicos periféricos captan gradientes funcionales asociados al proceso patológico, pero también influencias externas, como el estado inmunitario o metabólico del individuo.

En este contexto, las rutas diferenciales identificadas (Interleukin-33 signaling, NLRP1 inflammasome, proteasome core complex assembly, cation-transporting ATPase complex o tryptophan metabolism) resultan especialmente relevantes. Estas vías se relacionan con procesos inflamatorios, homeostasis proteica y estrés celular, en línea con lo descrito en estudios que implican la activación inmunitaria periférica y la disfunción lisosomal en la fisiopatología del Parkinson (Sulzer et al., 2017; Pajarillo et al., 2023; Sidransky et al., 2009; Nalls et al., 2019). En conjunto, los hallazgos de este estudio refuerzan la hipótesis de que la enfermedad de Parkinson presenta una base molecular multifactorial, donde mecanismos inmunológicos, genéticos y metabólicos coexisten y modulan la progresión clínica.

De este modo, las líneas futuras de trabajo deberían incluir la incorporación de las etiquetas Genetic Parkinson's Disease (GPD) y Genetic Unaffected de los datos

originales, que podrían aportar contraste genético adicional; aplicar filtros de genes específicos (por ejemplo, eliminar genes de hemoglobina) para reducir el ruido y mejorar la sensibilidad de los modelos; y ampliar la muestra o validar los resultados con otras cohortes transcriptómicas.

6. Conclusiones

En conjunto, los resultados de este trabajo muestran un rendimiento moderado tanto en los enfoques supervisados como en los no supervisados, lo que refleja la complejidad molecular de la enfermedad de Parkinson y las limitaciones inherentes al uso de muestras de sangre periférica, un tejido altamente heterogéneo influido por factores como la edad, el sexo o el estado inmunitario.

Aun así, los modelos lograron capturar señales biológicamente coherentes, destacando rutas relacionadas con la inflamación, el metabolismo del triptófano, la función sináptica y la degradación proteica, procesos ampliamente implicados en la fisiopatología de la enfermedad. La combinación de métodos de enriquecimiento funcional y algoritmos de aprendizaje automático permitió integrar información molecular compleja, identificando patrones de expresión reproducibles y con potencial valor biológico.

El modelo no supervisado más robusto, basado en la combinación Zscore–HPO, permitió identificar tres subgrupos moleculares con perfiles diferenciados, lo que sugiere la existencia de heterogeneidad biológica más allá del diagnóstico clínico tradicional. Las comparaciones entre clústeres revelaron diferencias funcionales en vías inmunológicas, cardiovasculares y neurológicas, reforzando la hipótesis de que mecanismos múltiples e interconectados contribuyen a la progresión del Parkinson.

En términos metodológicos, este estudio demuestra la viabilidad de aplicar scoring molecular y machine learning sobre datos transcriptómicos de sangre periférica, aportando un marco reproducible para futuros análisis de estratificación.

Finalmente, estos resultados subrayan la necesidad de avanzar hacia enfoques integrativos, combinando datos multi-ómicos, información genética y variables clínicas, para mejorar la estratificación, caracterización y predicción de la progresión en la enfermedad de Parkinson.

7. Bibliografía Formato APA

Ryan, B., Marioni, R., & Simpson, T. I. (2025). An integrative network approach for longitudinal stratification in Parkinson's disease. *PLoS Computational Biology, 21*(3), e1012857. https://doi.org/10.1371/journal.pcbi.1012857

Roodveldt, C., Bernardino, L., Oztop-Cakmak, O., Dragic, M., Fladmark, K. E., Ertan, S., Aktas, B., Pita, C., Ciglar, L., Garraux, G., Williams-Gray, C., Pacheco, R., & Romero-Ramos, M. (2024). The immune system in Parkinson's disease: what we know so far. *Brain : a journal of neurology, 147*(10), 3306–3324. https://doi.org/10.1093/brain/awae177

Sun, R.-X., & Guo, Y. (2025). Gene signatures and immune correlations in Parkinson's disease Braak stages. *European Journal of Medical Research, 30*(1), 278. https://doi.org/10.1186/s40001-025-02554-y

World Health Organization. (2023, August 9). Parkinson disease (fact sheet). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/parkinson-disease

Bandres-Ciga, S., Ahmed, S., Sabir, M. S., Blauwendraat, C., Adarmes-Gómez, A. D., Bernal-Bernal, I., Bonilla-Toribio, M., Buiza-Rueda, D., Carrillo, F., Carrión-Claro, M., Gómez-Garre, P., Jesús, S., Labrador-Espinosa, M. A., Macias, D., Méndez-Del-Barrio, C., Periñán-Tocino, T., Tejera-Parrado, C., Vargas-González, L., Diez-Fairen, M., … Singleton, A. (2019). The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight. *Movement Disorders : Official Journal of the Movement Disorder Society, 34*(12), 1851–1863. https://doi.org/10.1002/mds.27864

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D. A., Noyce, A. J., Xue, A., Bras, J., Young, E., von Coelln, R., Simón-Sánchez, J., Schulte, C., Sharma, M., Krohn, L., Pihlstrøm, L., Siitonen, A., Iwaki, H., … International Parkinson's Disease Genomics Consortium (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet. Neurology, 18*(12), 1091–1102. https://doi.org/10.1016/S1474-4422(19)30320-5

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., Chen, C. M., Clark, L. N., Condroyer, C., De Marco, E. V., Dürr, A., Eblan, M. J., Fahn, S., Farrer, M. J., Fung, H. C., Gan-Or, Z., … Ziegler, S. G. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *The New England journal of medicine, 361*(17), 1651–1661. https://doi.org/10.1056/NEJMoa0901281

Yao, X. Y., Guan, L. N., Chen, Q., & Ren, C. (2023). LRRK2 G2019S and Parkinson's disease: insight from Neuroinflammation. *Postgraduate medical journal, 100*(1179), 4–11. https://doi.org/10.1093/postmj/qgad080

Pajarillo, E., Kim, S., Digman, A., Dutton, M., Son, D. S., Aschner, M., & Lee, E. (2023). The role of microglial LRRK2 kinase in manganese-induced inflammatory neurotoxicity via NLRP3 inflammasome and RAB10-mediated autophagy dysfunction. *The Journal of biological chemistry, 299*(7), 104879. https://doi.org/10.1016/j.jbc.2023.104879

Stokholm, M. G., Iranzo, A., Østergaard, K., Serradell, M., Otto, M., Svendsen, K. B., Garrido, A., Vilas, D., Borghammer, P., Santamaria, J., Møller, A., Gaig, C., Brooks, D. J., Tolosa, E., & Pavese, N. (2017). Assessment of neuroinflammation in patients with idiopathic rapid-eye-movement sleep behaviour disorder: a case-control study. *The Lancet. Neurology, 16*(10), 789–796. https://doi.org/10.1016/S1474-4422(17)30173-4

Qin, X. Y., Zhang, S. P., Cao, C., Loh, Y. P., & Cheng, Y. (2016). Aberrations in Peripheral Inflammatory Cytokine Levels in Parkinson Disease: A Systematic Review and Meta-analysis. *JAMA neurology, 73*(11), 1316–1324. https://doi.org/10.1001/jamaneurol.2016.2742

Sulzer, D., Alcalay, R. N., Garretti, F., Cote, L., Kanter, E., Agin-Liebes, J., Liong, C., McMurtrey, C., Hildebrand, W. H., Mao, X., Dawson, V. L., Dawson, T. M., Oseroff, C., Pham, J., Sidney, J., Dillon, M. B., Carpenter, C., Weiskopf, D., Phillips, E., Mallal, S., … Sette, A. (2017). T cells from patients with Parkinson's disease recognize α-synuclein peptides. *Nature, 546*(7660), 656–661. https://doi.org/10.1038/nature22815

Hui, K. Y., Fernandez-Hernandez, H., Hu, J., Schaffner, A., Pankratz, N., Hsu, N. Y., Chuang, L. S., Carmi, S., Villaverde, N., Li, X., Rivas, M., Levine, A. P., Bao, X., Labrias, P. R., Haritunians, T., Ruane, D., Gettler, K., Chen, E., Li, D., Schiff, E. R., … Peter, I. (2018). Functional variants in the LRRK2 gene confer shared effects on risk for Crohn's disease and Parkinson's disease. *Science translational medicine, 10*(423), eaai7795. https://doi.org/10.1126/scitranslmed.aai7795

Feng, Y., Chen, X., Zhang, X. D., & Huang, C. (2023). Metabolic Pathway Pairwise-Based Signature as a Potential Non-Invasive Diagnostic Marker in Alzheimer's Disease Patients. *Genes, 14*(6). https://doi.org/10.3390/genes14061285

Poddar, M. K., Banerjee, S., Chakraborty, A., & Dutta, D. (2021). Metabolic disorder in Alzheimer's disease. *Metabolic Brain Disease, 36*(5), 781–813. https://doi.org/10.1007/s11011-021-00673-z

Liu, N., Deng, Q., Peng, Z., Mao, D., Huang, Y., Meng, F., Zhang, X., Shen, J., Li, Z., Yan, W., & Peng, J. (2025). Characterization of gene expression profiles in Alzheimer's disease and osteoarthritis: A bioinformatics study. *PloS One, 20*(2), e0316708. https://doi.org/10.1371/journal.pone.0316708

Ni, T., Sun, Y., Li, Z., Tan, T., Han, W., Li, M., Zhu, L., Xiao, J., Wang, H., Zhang, W., Ma, Y., Wang, B., Wen, D., Chen, T., Tubbs, J., Zeng, X., Yan, J., Gui, H., Sham, P., & Guan, F. (2025). Integrated Transcriptome Analysis Reveals Novel Molecular Signatures for Schizophrenia Characterization. *Advanced Science (Weinheim, Baden-Wurttemberg, Germany), 12*(2), e2407628. https://doi.org/10.1002/advs.202407628

Wang, Y., Wu, D., Zheng, M., & Yang, T. (2025). An integrated bioinformatics and machine learning approach to identifying biomarkers connecting parkinson's disease with purine metabolism-related genes. *BMC Neurology, 25*(1), 161. https://doi.org/10.1186/s12883-025-04167-8

Fan, J., Cao, S., Peng, H., Zhi, Y., Zhan, S., & Li, R. (2025). Explainable machine learning-driven models for predicting Parkinson's disease and its prognosis: obesity patterns associations and models development using NHANES 1999-2018 data. *Lipids in Health and Disease, 24*(1), 241. https://doi.org/10.1186/s12944-025-02664-w

Li, Y., Jia, W., Chen, C., Chen, C., Chen, J., Yang, X., & Liu, P. (2025). Identification of biomarkers associated with inflammatory response in Parkinson's disease by bioinformatics and machine learning. *PloS One, 20*(5), e0320257. https://doi.org/10.1371/journal.pone.0320257

Foroutan, M., Bhuva, D. D., Lyu, R., Horan, K., Cursons, J., & Davis, M. J. (2018). Single sample scoring of molecular phenotypes. *BMC bioinformatics, 19*(1), 404. https://doi.org/10.1186/s12859-018-2435-4

Adewale, Q., Khan, A. F., Lin, S. J., Baumeister, T. R., Zeighami, Y., Carbonell, F., Ferreira, D., & Iturria-Medina, Y. (2025). Patient-centered brain transcriptomic and multimodal imaging determinants of clinical progression, physical activity, and treatment needs in Parkinson's disease. *NPJ Parkinson's disease, 11*(1), 29. https://doi.org/10.1038/s41531-025-00878-4

Chan, Y. H., Wang, C., Soh, W. K., & Rajapakse, J. C. (2022). Combining Neuroimaging and Omics Datasets for Disease Classification Using Graph Neural Networks. *Frontiers in neuroscience, 16*, 866666. https://doi.org/10.3389/fnins.2022.866666

Fereshtehnejad, S. M., Zeighami, Y., Dagher, A., & Postuma, R. B. (2017). Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain : a journal of neurology, 140*(7), 1959–1976. https://doi.org/10.1093/brain/awx118

Johansson, M. E., van Lier, N. M., Kessels, R. P. C., Bloem, B. R., & Helmich, R. C. (2023). Two-year clinical progression in focal and diffuse subtypes of Parkinson's disease. *NPJ Parkinson's disease, 9*(1), 29. https://doi.org/10.1038/s41531-023-00466-4

Zhou, C., Wang, L., Cheng, W., Lv, J., Guan, X., Guo, T., Wu, J., Zhang, W., Gao, T., Liu, X., Bai, X., Wu, H., Cao, Z., Gu, L., Chen, J., Wen, J., Huang, P., Xu, X., Zhang, B., Feng, J., … Zhang, M. (2024). Author Correction: Two distinct trajectories of clinical and neurodegeneration events in Parkinson's disease. *NPJ Parkinson's disease, 10*(1), 22. https://doi.org/10.1038/s41531-024-00635-z

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q., Shaw, L. M., Seibyl, J., Schuff, N., Singleton, A., Kieburtz, K., Toga, A. W., Mollenhauer, B., Galasko, D., Chahine, L. M., Weintraub, D., … Parkinson's Progression Markers Initiative (2018). The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Annals of*

clinical and translational neurology, 5(12), 1460–1477. https://doi.org/10.1002/acn3.644

Parnetti, L., Gaetani, L., Eusebi, P., Paciotti, S., Hansson, O., El-Agnaf, O., Mollenhauer, B., Blennow, K., & Calabresi, P. (2019). CSF and blood biomarkers for Parkinson's disease. *The Lancet. Neurology, 18*(6), 573–586. https://doi.org/10.1016/S1474-4422(19)30024-9

Severson, K. A., Chahine, L. M., Smolensky, L. A., Dhuliawala, M., Frasier, M., Ng, K., Ghosh, S., & Hu, J. (2021). Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. *The Lancet. Digital Health, 3*(9), e555–e564. https://doi.org/10.1016/S2589-7500(21)00101-1

Wüllner, U., Borghammer, P., Choe, C. U., Csoti, I., Falkenburger, B., Gasser, T., Lingor, P., & Riederer, P. (2023). The heterogeneity of Parkinson's disease. *Journal of neural transmission (Vienna, Austria : 1996), 130*(6), 827–838. https://doi.org/10.1007/s00702-023-02635-4

Shamir, R., Klein, C., Amar, D., Vollstedt, E. J., Bonin, M., Usenovic, M., Wong, Y. C., Maver, A., Poths, S., Safer, H., Corvol, J. C., Lesage, S., Lavi, O., Deuschl, G., Kuhlenbaeumer, G., Pawlack, H., Ulitsky, I., Kasten, M., Riess, O., Brice, A., … Krainc, D. (2017). *Analysis of blood-based gene expression in idiopathic Parkinson disease. Neurology, 89*(16), 1676–1683. https://doi.org/10.1212/WNL.0000000000004516

NCBI GEO (GPL570). (2003). *[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array*. *National Center for Biotechnology Information.* https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

Marie, A., Darricau, M., Touyarot, K., Parr-Brownlie, L. C., & Bosch-Bouju, C. (2021). Role and Mechanism of Vitamin A Metabolism in the Pathophysiology of Parkinson's Disease. *Journal of Parkinson's disease, 11*(3), 949–970. https://doi.org/10.3233/JPD-212671

Cook, C., & Petrucelli, L. (2009). A critical evaluation of the ubiquitin-proteasome system in Parkinson's disease. *Biochimica et biophysica acta, 1792*(7), 664–675. https://doi.org/10.1016/j.bbadis.2009.01.012

Venkatesan, D., Iyer, M., Narayanasamy, A., Siva, K., & Vellingiri, B. (2020). Kynurenine pathway in Parkinson's disease-An update. *eNeurologicalSci*, *21*, 100270. https://doi.org/10.1016/j.ensci.2020.100270

Vaidya, B., & Sharma, S. S. (2020). Transient Receptor Potential Channels as an Emerging Target for the Treatment of Parkinson's Disease: An Insight Into Role of Pharmacological Interventions. *Frontiers in cell and developmental biology, 8*, 584513. https://doi.org/10.3389/fcell.2020.584513

Slater, N. M., Melzer, T. R., Myall, D. J., Anderson, T. J., & Dalrymple-Alford, J. C. (2024). Cholinergic Basal Forebrain Integrity and Cognition in Parkinson's Disease: A Reappraisal

of Magnetic Resonance Imaging Evidence. *Movement disorders : official journal of the Movement Disorder Society, 39*(12), 2155–2172. https://doi.org/10.1002/mds.30023

Mawatari, S., Ohara, S., Taniwaki, Y., Tsuboi, Y., Maruyama, T., & Fujino, T. (2020). Improvement of Blood Plasmalogens and Clinical Symptoms in Parkinson's Disease by Oral Administration of Ether Phospholipids: A Preliminary Report. *Parkinson's disease*, *2020*, 2671070. https://doi.org/10.1155/2020/2671070

Brundin, P., & Melki, R. (2017). Prying into the Prion Hypothesis for Parkinson's Disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 37*(41), 9808–9818. https://doi.org/10.1523/JNEUROSCI.1788-16.2017

Shen, C., Liu, Z., Chen, F., Zhu, X., Fan, G., Ning, P., Li, Q., Zhang, X., Dong, H., Liu, Y., Yu, M., Fei, J., & Huang, F. (2025). Interleukin-33 promotes dopaminergic neuron survival and inhibits glial activation in Parkinson's disease models. *Brain, behavior, and immunity, 129*, 787–800. https://doi.org/10.1016/j.bbi.2025.07.010

Gordon, R., Albornoz, E. A., Christie, D. C., Langley, M. R., Kumar, V., Mantovani, S., Robertson, A. A. B., Butler, M. S., Rowe, D. B., O'Neill, L. A., Kanthasamy, A. G., Schroder, K., Cooper, M. A., & Woodruff, T. M. (2018). Inflammasome inhibition prevents α-synuclein pathology and dopaminergic neurodegeneration in mice. *Science translational medicine, 10*(465), eaah4066. https://doi.org/10.1126/scitranslmed.aah4066

Dehay, B., Ramirez, A., Martinez-Vicente, M., Perier, C., Canron, M. H., Doudnikoff, E., Vital, A., Vila, M., Klein, C., & Bezard, E. (2012). Loss of P-type ATPase ATP13A2/PARK9 function induces general lysosomal deficiency and leads to Parkinson disease neurodegeneration. Proceedings of the National *Academy of Sciences of the United States of America, 109*(24), 9611–9616. https://doi.org/10.1073/pnas.1112368109

Boza-Serrano, A., Reyes, J. F., Rey, N. L., Leffler, H., Bousset, L., Nilsson, U., Brundin, P., Venero, J. L., Burguillos, M. A., & Deierborg, T. (2014). The role of Galectin-3 in α-synuclein-induced microglial activation. *Acta neuropathologica communications, 2*, 156. https://doi.org/10.1186/s40478-014-0156-0

Alvarez-Erviti, L., Rodriguez-Oroz, M. C., Cooper, J. M., Caballero, C., Ferrer, I., Obeso, J. A., & Schapira, A. H. (2010). Chaperone-mediated autophagy markers in Parkinson disease brains. *Archives of neurology, 67*(12), 1464–1472. https://doi.org/10.1001/archneurol.2010.198

Teismann, P., Tieu, K., Choi, D. K., Wu, D. C., Naini, A., Hunot, S., Vila, M., Jackson-Lewis, V., & Przedborski, S. (2003). Cyclooxygenase-2 is instrumental in Parkinson's disease neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America, 100*(9), 5473–5478. https://doi.org/10.1073/pnas.0837397100

Zhai, S., Tanimura, A., Graves, S. M., Shen, W., & Surmeier, D. J. (2018). Striatal synapses, circuits, and Parkinson's disease. *Current opinion in neurobiology, 48*, 9–16. https://doi.org/10.1016/j.conb.2017.08.004

Riaz, Z., Richardson, G. S., Jin, H., Zenitsky, G., Anantharam, V., Kanthasamy, A., & Kanthasamy, A. G. (2024). Nuclear pore and nucleocytoplasmic transport impairment in oxidative stress-induced neurodegeneration: relevance to molecular mechanisms in Pathogenesis of Parkinson's and other related neurodegenerative diseases. *Molecular neurodegeneration, 19*(1), 87. https://doi.org/10.1186/s13024-024-00774-0

Toro-Domínguez, D., Martorell-Marugán, J., Martinez-Bueno, M., López-Domínguez, R., Carnero-Montoro, E., Barturen, G., Goldman, D., Petri, M., Carmona-Sáez, P., & Alarcón-Riquelme, M. E. (2022). Scoring personalized molecular portraits identify Systemic Lupus Erythematosus subtypes and predict individualized drug responses, symptomatology and disease progression. *Briefings in bioinformatics*, *23*(5), bbac332. https://doi.org/10.1093/bib/bbac332

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics, 14*, 7. https://doi.org/10.1186/1471-2105-14-7

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., … Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature, 462*(7269), 108–112. https://doi.org/10.1038/nature08460

Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS computational biology, 4*(11), e1000217. https://doi.org/10.1371/journal.pcbi.1000217

Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics, 6*, 225. https://doi.org/10.1186/1471-2105-6-225

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis [Preprint]. *bioRxiv*. https://doi.org/10.1101/060012

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Stephan, R., Tiwari, K., Varusai, T., Weiser, J., Wright, A., … D'Eustachio, P. (2024). The Reactome Pathway Knowledgebase 2024. *Nucleic acids research*, *52*(D1), D672–D678. https://doi.org/10.1093/nar/gkad1025

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research, 45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research, 49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Gargano, M. A., Matentzoglu, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, A. V., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G., Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., Čady, J., … Robinson, P. N. (2024). *The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic acids research, 52*(D1), D1333–D1346. https://doi.org/10.1093/nar/gkad1005

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research, 48*(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Breiman, L. (2001) Random Forests. *Machine Learning 45(1)*, 5–32. https://doi.org/10.1023/A:1010933404324

Cover, T. & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory, 13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings *of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall/CRC.

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications, 30*(2), 169–190. https://doi.org/10.3233/AIC-170729

O'Connell, N. S., Jaeger, B. C., Bullock, G. S., & Speiser, J. L. (2025). A comparison of random forest variable selection methods for regression modeling of continuous outcomes. *Briefings in bioinformatics, 26*(2), bbaf096. https://doi.org/10.1093/bib/bbaf096

Halder, R. K., Uddin, M. N., Uddin, M. A., & Aryal, S. (2024). Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data, 11*, 113. https://doi.org/10.1186/s40537-024-00973-y

Egbo, B., Nigmetolla, Z., Khan, N. A., & Jamwal, P. K. (2025). Explainable machine learning for early detection of Parkinson's disease in aging populations using vocal biomarkers. *Frontiers in Aging Neuroscience, 17*, 1672971. https://doi.org/10.3389/fnagi.2025.1672971

Alaqeeli, O., & Alturki, R. (2023). Evaluating the Performance of the Generalized Linear Model (glm) R Package Using Single-Cell RNA-Sequencing Data. *Applied Sciences, 13*(20), 11512. https://doi.org/10.3390/app132011512

Qiao M. (2023). Factorized discriminant analysis for genetic signatures of neuronal phenotypes. *Frontiers in neuroinformatics, 17*, 1265079. https://doi.org/10.3389/fninf.2023.1265079

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)* (pp. 3121–3124). IEEE. https://doi.org/10.1109/ICPR.2010.764

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal*, *35*(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but *Many* are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research : JMLR*, *20*, 177.

John, C. R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., & Barnes, M. (2020). M3C: Monte Carlo reference-based consensus clustering. *Scientific reports*, *10*(1), 1816. https://doi.org/10.1038/s41598-020-58766-1

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, *4*, 1–39. https://doi.org/10.1214/09-SS051

1. Anexos

Reflexión sobre sostenibilidad y ODS

Este trabajo contribuye al ODS 3: Salud y bienestar, ya que promueve el uso de herramientas de bioinformática para mejorar la comprensión molecular de enfermedades neurodegenerativas como el Parkinson.

Además, el uso de datos públicos y análisis reproducibles apoya el ODS 9: Industria, innovación e infraestructura, fomentando la ciencia abierta y el uso ético de la información biomédica.

En términos metodológicos, el diseño de un flujo de trabajo eficiente y reproducible responde a principios de sostenibilidad en la investigación, al reducir la necesidad de recursos experimentales y favorecer la reutilización de datos, en línea con las políticas de ciencia FAIR y los valores de investigación responsable.

Enlace a GitHub, donde se encuentra el script del trabajo:

https://github.com/SantosAF/TFM-Parkinson-Analisis-Molecular

Enlace a Google Drive, donde se encuentra el entorno completo de R y la carpeta con todos los resultados obtenidos del trabajo:

https://drive.google.com/drive/folders/1Kq5Eq2wMYmAZWtA6EXZG3hvUjDOj1PuF?usp=sharing

Enlace de los datos que se utilizan en este TFM:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99039

```
> sessionInfo()
R version 4.5.1 (2025-06-13 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

Matrix products: default
  LAPACK version 3.12.1

locale:
[1] LC_COLLATE=Spanish_Spain.utf8  LC_CTYPE=Spanish_Spain.utf8
[3] LC_MONETARY=Spanish_Spain.utf8 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.utf8

time zone: Europe/Madrid
tzcode source: internal
```

```
attached base packages:
[1] grid      parallel  stats     graphics  grDevices utils     datase
ts  methods
[9] base

other attached packages:
 [1] progressr_0.16.0      cluster_2.1.8.1       M3C_1.30.0
 [4] circlize_0.4.16       ComplexHeatmap_2.24.1 stringr_1.5.2
 [7] dplyr_1.1.4           PRROC_1.4             rlang_1.1.6
[10] pROC_1.19.0.1         reshape2_1.4.4        MASS_7.3-65
[13] xgboost_1.7.11.1      kknn_1.4.1            randomForest_4.7-1.2
[16] doParallel_1.0.17     iterators_1.0.14      foreach_1.5.2
[19] caret_7.0-1           lattice_0.22-7        ggplot2_4.0.0
[22] fgsea_1.34.2          GEOquery_2.76.0       Biobase_2.68.0
[25] BiocGenerics_0.54.1   generics_0.1.4        pathMED_1.0.2

loaded via a namespace (and not attached):
  [1] magrittr_2.0.4              clue_0.3-66
  [3] GetoptLong_1.0.5            matrixStats_1.5.0
  [5] compiler_4.5.1              png_0.1-8
  [7] vctrs_0.6.5                 pkgconfig_2.0.3
  [9] shape_1.4.6.1               crayon_1.5.3
 [11] fastmap_1.2.0               XVector_0.48.0
 [13] rmarkdown_2.30              tzdb_0.5.0
 [15] prodlim_2025.04.28          UCSC.utils_1.4.0
 [17] purrr_1.1.0                 xfun_0.53
 [19] GenomeInfoDb_1.44.3         jsonlite_2.0.0
 [21] recipes_1.3.1               DelayedArray_0.34.1
 [23] BiocParallel_1.42.1         R6_2.6.1
 [25] stringi_1.8.7               RColorBrewer_1.1-3
 [27] reticulate_1.43.0           limma_3.64.3
 [29] parallelly_1.45.1           rpart_4.1.24
 [31] GenomicRanges_1.60.0        lubridate_1.9.4
 [33] Rcpp_1.1.0                  SummarizedExperiment_1.38.1
 [35] knitr_1.50                  future.apply_1.20.0
 [37] snow_0.4-4                  readr_2.1.5
 [39] IRanges_2.42.0              igraph_2.1.4
 [41] rentrez_1.2.4               Matrix_1.7-3
 [43] splines_4.5.1               nnet_7.3-20
 [45] timechange_0.3.0            tidyselect_1.2.1
 [47] rstudioapi_0.17.1           abind_1.4-8
 [49] yaml_2.3.10                 timeDate_4041.110
 [51] codetools_0.2-20            listenv_0.9.1
 [53] tibble_3.3.0                plyr_1.8.9
 [55] withr_3.0.2                 S7_0.2.0
 [57] askpass_1.2.1               Rtsne_0.17
 [59] evaluate_1.0.5              future_1.67.0
 [61] survival_3.8-3              xml2_1.4.0
 [63] pillar_1.11.1               MatrixGenerics_1.20.0
 [65] stats4_4.5.1                hms_1.1.3
 [67] S4Vectors_0.46.0            scales_1.4.0
 [69] globals_0.18.0              class_7.3-23
 [71] glue_1.8.0                  tools_4.5.1
 [73] data.table_1.17.8           RSpectra_0.16-2
 [75] ModelMetrics_1.2.2.2        gower_1.0.2
 [77] XML_3.99-0.19               fastmatch_1.1-6
```

```
 [79] cowplot_1.2.0              tidyr_1.3.1
 [81] matrixcalc_1.0-6           umap_0.2.10.0
 [83] ipred_0.9-15               colorspace_2.1-2
 [85] nlme_3.1-168               GenomeInfoDbData_1.2.14
 [87] patchwork_1.3.2            cli_3.6.5
 [89] S4Arrays_1.8.1             lava_1.8.1
 [91] corpcor_1.6.10             doSNOW_1.0.20
 [93] gtable_0.3.6               digest_0.6.37
 [95] SparseArray_1.8.1          rjson_0.2.23
 [97] farver_2.1.2               htmltools_0.5.8.1
 [99] caretEnsemble_4.0.1        lifecycle_1.0.4
[101] hardhat_1.4.2              httr_1.4.7
[103] statmod_1.5.1              GlobalOptions_0.1.2
[105] openssl_2.3.4
```