



RIDAA
Repositorio Institucional
Digital de Acceso Abierto de la
Universidad Nacional de Quilmes



Universidad
Nacional
de Quilmes

Alonso, Macarena

Descifrando las interacciones proteína-proteína : implicancias en enfermedades neurodegenerativas abordadas con inteligencia artificial



Esta obra está bajo una Licencia Creative Commons Argentina.
Atribución - No Comercial - Sin Obra Derivada 2.5
<https://creativecommons.org/licenses/by-nc-nd/2.5/ar/>

Documento descargado de RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes de la Universidad Nacional de Quilmes

Cita recomendada:

Alonso, M. (2026). *Descifrando las interacciones proteína-proteína: implicancias en enfermedades neurodegenerativas abordadas con inteligencia artificial. (Tesis de doctorado). Universidad Nacional de Quilmes, Bernal, Argentina. Disponible en RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes <http://ridaa.unq.edu.ar/handle/20.500.11807/6576>*

Puede encontrar éste y otros documentos en: <https://ridaa.unq.edu.ar>

Descifrando las interacciones proteína-proteína: implicancias en enfermedades neurodegenerativas abordadas con inteligencia artificial

TESIS DOCTORAL

Macarena Alonso

macarenaalonso94@gmail.com

Resumen

La esclerosis lateral amiotrófica (ELA) es una enfermedad neurodegenerativa caracterizada por la pérdida progresiva de las neuronas motoras, con una supervivencia de tres a cinco años tras los primeros síntomas. Solo el 10% de los casos sigue un patrón de herencia autosómico dominante, con el resto siendo esporádicos. Las causas subyacentes de esta enfermedad siguen siendo poco comprendidas e involucran una compleja interacción de factores genéticos y ambientales.

A su vez, esta patología se estudia cada vez más junto con la demencia frontotemporal (DFT) como extremos de un espectro continuo, dada la superposición de características genéticas, sintomatológicas y neuropatológicas.

Ambas enfermedades carecen de biomarcadores específicos, lo que dificulta la estratificación precisa de pacientes, el diagnóstico y el seguimiento terapéutico. En este contexto, los enfoques computacionales capaces de integrar distintos tipos de datos, como los transcriptómicos e interactómicos, representan una vía prometedora para identificar potenciales dianas terapéuticas o nuevos ejes de estudio.

En este estudio se aplicaron Redes Neuronales Convolucionales de Grafos combinadas con Propagación de Relevancia por Capas de Grafos, un modelo de inteligencia artificial explicativa, para analizar perfiles de expresión génica en sangre total de individuos con y sin ELA en el contexto de redes de interacción proteína-proteína. Este enfoque permitió identificar genes codificantes de proteínas clave para la clasificación y las vías biológicas potencialmente implicadas en la patología de la ELA. Posteriormente, el análisis se extendió a pacientes con DFT para explorar sistemáticamente similitudes y diferencias entre ambas condiciones, ofreciendo nuevas perspectivas sobre sus mecanismos patológicos.

Abstract

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease characterized by the progressive loss of motor neurons, with a survival period of three to five years following the onset of initial symptoms. Only 10% of cases follow an autosomal dominant inheritance pattern, with the rest being sporadic. The underlying causes of this disease remain poorly understood and involve a complex combination of genetic and environmental factors.

Furthermore, this pathology is being increasingly studied alongside frontotemporal dementia (FTD) as the ends of a continuous spectrum, given the overlapping genetic, symptomatological, and neuropathological features of both conditions.

Both diseases lack specific biomarkers, which hinders the precise stratification of patients, diagnosis, and therapeutic monitoring. In this context, computational approaches capable of integrating different types of data, such as transcriptomic and interactomic data, represent a promising avenue for identifying potential therapeutic targets or novel research directions.

In this study, we applied Graph Convolutional Neural Networks combined with Graph Layer-Wise Relevance Propagation, an explainable artificial intelligence model, to analyze whole blood gene expression profiles from ALS and healthy individuals within protein-protein interaction networks. This approach enabled the identification of relevant protein-coding genes for classification and biological pathways potentially implicated in ALS pathology. The analysis was subsequently extended to FTD patients to systematically explore similarities and differences between both conditions, offering new insights into their underlying pathological mechanisms.

TESIS DOCTORAL

Descifrando las interacciones proteína-proteína: implicancias en enfermedades neurodegenerativas abordadas con inteligencia artificial

Lic. Macarena Alonso



Directora: Dra. Cristina Marino Buslje
Instituto Leloir (FIL)/Instituto de Investigaciones Bioquímicas
de Buenos Aires (IBBA)

Abril, 2026

UNIVERSIDAD NACIONAL DE QUILMES

DEPARTAMENTO DE CIENCIA Y TECNOLOGÍA

Descifrando las interacciones proteína-proteína:
implicancias en enfermedades neurodegenerativas
abordadas con inteligencia artificial

Lic. Macarena Alonso

*Tesis presentada para optar al título de
Doctor en Ciencia y Tecnología de la
Universidad Nacional de Quilmes*

Directora: Dra. Cristina Marino Buslje

*Realizado en el laboratorio de Bioinformática del Instituto
Leloir (FIL)/Instituto de Investigaciones Bioquímicas de
Buenos Aires (IBBA) - Buenos Aires, Argentina*

Abril, 2026

Agradecimientos

A Andi, mi compañero de vida, por su paciencia, compañía y amor incommensurable a lo largo de todo este recorrido.

A mis amigos, por las alegrías compartidas y su apoyo constante y leal.

A mi familia, por su cariño, por escucharme y por haberme incentivado, desde siempre, a seguir el camino de la ciencia.

A Alejo y Berlioz, quienes, sin saberlo y aun en silencio, han sido de las mejores compañías.

A mis compañeros de laboratorio y del área de Bioinformática del Instituto Leloir, por su guía, enseñanza y paciencia durante todo el desarrollo de este trabajo, además de los muchos mates compartidos.

A Cristina, por guiarme durante todo este camino, por su confianza y, sobre todo, por brindarme la oportunidad de cumplir uno de mis mayores sueños: dedicarme a investigar aquello que me apasiona.

A la universidad pública, gratuita y de calidad, por su compromiso y excelencia, y con el deseo de que continúe formando profesionales de alto nivel.

"Todo el día y la noche entera, el cuerpo interviene [...] La criatura interior puede mirar sólo a través del cristal [...] no puede separarse del cuerpo ni un solo instante; debe pasar por toda la interminable procesión de cambios, calor y frío, confort e incomodidad, hambre y satisfacción, salud y enfermedad, hasta que adviene la inevitable catástrofe: el cuerpo se rompe en pedazos y el alma (se dice) escapa."

— Virginia Woolf, *Sobre estar enferma* (1926)

Índice

Abreviaturas	10
Introducción	12
1.1 Conceptos generales de las enfermedades neurodegenerativas	13
1.2 La Esclerosis Lateral Amiotrófica	13
1.3 Relación entre la Esclerosis Lateral Amiotrófica y la Demencia Frontotemporal	15
1.4 Redes biológicas	17
1.5 Redes neuronales convolucionales y su aplicación a redes biológicas	18
1.6 Interpretación de las redes neuronales convolucionales en grafos	22
Hipótesis y Objetivos	24
2.1 Hipótesis	25
2.2 Objetivo general	25
2.3 Objetivos específicos	25
Materiales y Métodos	26
3.1 Recopilación de datos	27
3.1.1 Perfiles de expresión de pacientes de ELA	27
3.1.2 Perfiles de expresión de pacientes de DFT	29
3.2 Construcción de redes PPI	30
3.2.1 IntAct	32
3.2.2 BioGRID	33
3.2.3 HIPPIE	34

3.3 Modelos de Machine Learning	34
3.3.1 GCNN-GLRP	37
3.3.2 Regresión Logística	43
3.3.3 Random Forest	46
3.3.4 XGBoost (eXtreme Gradient Boosting)	48
3.4 Cálculo del ranking de los genes según su importancia en la clasificación	50
3.5 Expresión diferencial	51
3.6 Métricas de centralidad	52
3.7 Análisis de enriquecimiento de vías biológicas	53
Resultados	54
4.1 El Análisis de Componentes Principales no distingue a los pacientes con ELA de los controles sanos	55
4.2 Ciertos genes presentan expresión génica diferencial entre grupos	56
4.3 Los modelos GCNN y de Regresión Logística son clasificadores eficientes para casos y controles	57
4.4 Los genes predictivos más relevantes muestran una presencia estable tanto en GCNN-GLRP como en la Regresión Logística.	60
4.5 Variabilidad de los genes más relevantes entre modelos y redes PPI	62
4.6 Los genes más relevantes obtenidos con GCNN muestran una mayor centralidad en todas las redes PPI en comparación con los de la Regresión Logística	64
4.7 Se identificaron vías enriquecidas en los genes más relevantes identificados por GCNN-GLRP	65
4.8 Se observan características compartidas entre la ELA y la DFT	66
Discusión	71
Conclusiones	78
Conclusión general	79
Trabajos complementarios	80

7.1 Control de patrones de señal en las redes PPI durante la clasificación	81
7.2 Unión de redes de interacción proteína-proteína	83
7.3 Organelas sin membrana y ELA	84
7.4 AnswerALS, expresión génica obtenida mediante RNA-Seq	86
Publicaciones con referato	90
Bibliografía	106
Material suplementario	115
10.1 Tablas de rendimiento de la clasificación con redes PPI de MLOs	116
10.1.1 IntAct	116
10.1.2 BioGRID	117
10.1.3 HIPPIE	118
10.2 Lista de genes más relevantes para la clasificación de pacientes de ELA con GCNN y RL	118
10.3 Lista de genes más relevantes para la clasificación de pacientes BH-MT y LG-MT con GCNN y RL	120
10.4 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con ELA	121
10.5 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con subtipos de DFT (BH vs MT)	122
10.6 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con subtipos de DFT (LG vs MT)	124

Abreviaturas

BvFTD: Variante Comportamental de la Demencia Frontotemporal (del inglés *Behavioral Variant Frontotemporal Dementia*)

CBS: Síndrome Corticobasal (del inglés *Corticobasal syndrome*)

CNN: Redes Neuronales Convolucionales (del inglés *Convolutional Neural Networks*)

DEG: Gen Diferencialmente Expresado (del inglés *Differentially Expressed gene*)

DFT: Demencia Frontotemporal

ELA: Esclerosis Lateral Amiotrófica

GCNN: Redes Neuronales Convolucionales de Grafos (del inglés *Graph Convolutional Neural Network*)

GEO: Gene Expression Omnibus

GLRP: Propagación de Relevancia por Capas de Grafos (del inglés *Graph Layer-wise Relevance Propagation*)

IA: Inteligencia Artificial

LRP: Propagación de Relevancia por Capas (del inglés *Layer-wise Relevance Propagation*)

ML: Aprendizaje automático (del inglés *Machine Learning*)

NFvPPA: Afasia progresiva primaria no fluida (del inglés *Nonfluent variant Primary Progressive Aphasia*)

PCA: Análisis de Componentes Principales (del inglés *Principal Component Analysis*)

PPI: Interacción proteína-proteína (del inglés *Protein-protein Interaction*)

PSP: Parálisis supranuclear progresiva (del inglés *Progressive Supranuclear Palsy*)

RF: *Random Forest*

RL: Regresión Logística

SvPPA: Afasia progresiva primaria semántica (del inglés *Semantic variant Primary Progressive Aphasia*)

TLR: Receptor tipo Toll (del inglés *Toll-like Receptor*)

Introducción

1.1 Conceptos generales de las enfermedades neurodegenerativas

Las enfermedades neurodegenerativas son un conjunto de patologías caracterizadas por la muerte de tipos específicos de neuronas y causantes de afecciones tales como deterioro cognitivo, disfunción motora y cambios en el comportamiento, entre otras.

Éstas afectan actualmente cerca del 15 % de la población mundial (1). Se prevé que su carga sanitaria a niveles económicos, de mortalidad y morbilidad se vea al menos duplicada en las próximas décadas.

Si bien estas tendencias se explican por el crecimiento y envejecimiento de la población mundial, comprender su etiología se ha vuelto más complejo debido a los rápidos cambios en las características demográficas, así como a las variaciones en los factores de riesgo (p. ej., sobrepeso y obesidad). Esto es especialmente importante para los países de ingresos bajos y medios, entre los cuales se encuentra Argentina, ya que son, a su vez, los que sufren la mayor carga (2).

1.2 La Esclerosis Lateral Amiotrófica

Una de estas enfermedades es la esclerosis lateral amiotrófica (ELA), un trastorno caracterizado por la muerte progresiva de neuronas motoras superiores (NMS) e inferiores (NMI).

Las NMS activan el movimiento voluntario, originándose principalmente en la corteza cerebral y transmitiendo información desde el cerebro hasta la médula espinal y el tronco encefálico, sin salir del sistema nervioso central (SNC).

Las NMI, en cambio, son las encargadas de estimular directamente los músculos para que se contraigan, existiendo tanto en el SNC como en el sistema nervioso periférico (SNP). Sus cuerpos celulares se encuentran en el asta ventral de la médula espinal y en los núcleos de los nervios craneales, e inervan directamente

el músculo esquelético. La **Figura 1** muestra el circuito usual de ambos tipos de neuronas.

La vía para el movimiento voluntario puede dividirse en el tracto corticoespinal, el cual permite el movimiento de las extremidades, y el tracto corticobulbar, que modula los músculos de la cara, cabeza y cuello. De esta manera, se puede categorizar a la ELA como de inicio “espinal” si los primeros síntomas son debilidad y parálisis muscular en las extremidades. En cambio, si se presenta inicialmente dificultad para hablar, masticar o tragar, junto con debilidad o espasmos en los músculos de la cara, la mandíbula y la garganta, su inicio se denomina “bulbar” (3).

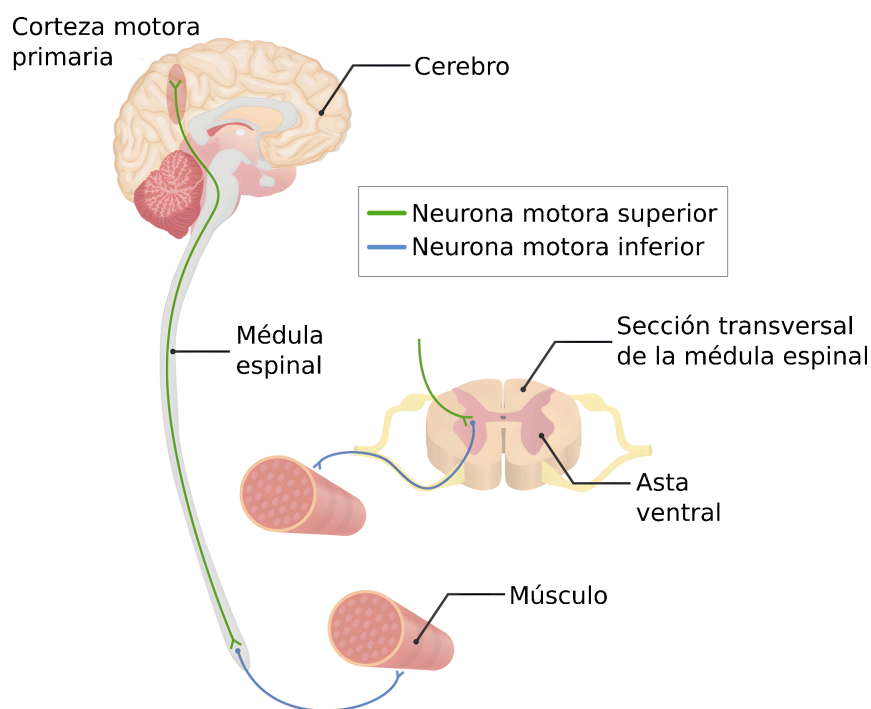


Figura 1. Esquema general de los circuitos de neuronas motoras superiores (NMS) e inferiores (NMI) en humanos. Creada con BioRender.

En la mayoría de los casos, estos síntomas se detectan en adultos y tienden a empeorar rápidamente a medida que la enfermedad progresa, afectando la totalidad de sus movimientos. El periodo de supervivencia tras la aparición de los síntomas iniciales oscila entre tres y cinco años, y en la mayoría de los casos la muerte se produce por insuficiencia respiratoria (4,5).

Aproximadamente el 90% de los casos de ELA son esporádicos, sin vínculo familiar, mientras que el 10% restante sigue un patrón de herencia autosómico dominante (6). Sin embargo, a pesar de las numerosas investigaciones realizadas, las causas subyacentes de la ELA siguen sin ser claras y se cree que son diversas, interviniendo factores tanto genéticos como ambientales (7–11).

1.3 Relación entre la Esclerosis Lateral Amiotrófica y la Demencia Frontotemporal

En algunos casos, los pacientes con ELA son también diagnosticados con demencia frontotemporal (DFT), que implica la degeneración de los lóbulos frontal y temporal anterior. Esta última se caracteriza por provocar cambios de comportamiento, deterioro de la función ejecutiva y/o alteraciones del lenguaje.

Existen cinco variantes de DFT, dependiendo de las regiones cerebrales y áreas corporales específicas afectadas, especialmente al inicio de la enfermedad. Por un lado, los pacientes que padecen la variante conductual (BvFTD, del inglés *Behavioural variant Frontotemporal Dementia*) representan aproximadamente la mitad de todos los casos de DFT. Éstos presentan cambios de personalidad, apatía, deterioro del juicio, reducción del autocontrol y pérdida de empatía. Por otro lado, en la parálisis supranuclear progresiva (PSP, del inglés *Progressive Supranuclear Palsy*) y el síndrome corticobasal (CBS, del inglés *Corticobasal Syndrome*) los pacientes comienzan con problemas de movimiento. En la PSP, los pacientes suelen presentar inestabilidad temprana al caminar y parálisis de la mirada vertical, mientras que el CBS se caracteriza por síntomas motores asimétricos y signos corticales. Por último, existen dos variantes en las que se ve afectado el lenguaje: la afasia primaria progresiva no fluente (NFvPPA, del inglés *Nonfluent Variant Primary Progressive Aphasia*) y afasia primaria progresiva semántica (SvPPA, del inglés *Semantic Variant Primary Progressive Aphasia*). La primera se caracteriza por un habla agramática y forzada, mientras que la segunda implica un habla fluida con profundos déficits de comprensión y denominación de palabras. Las principales

diferencias entre estas dos afecciones radican en las áreas cerebrales específicas afectadas y en la naturaleza de la alteración del lenguaje.

Si bien el panorama de la DFT es diverso y complejo, en los últimos años, la ELA y la DFT se han estudiado como los extremos de un continuo debido a sus patrones genéticos, clínicos y patológicos coincidentes (12) (**Figura 2**).

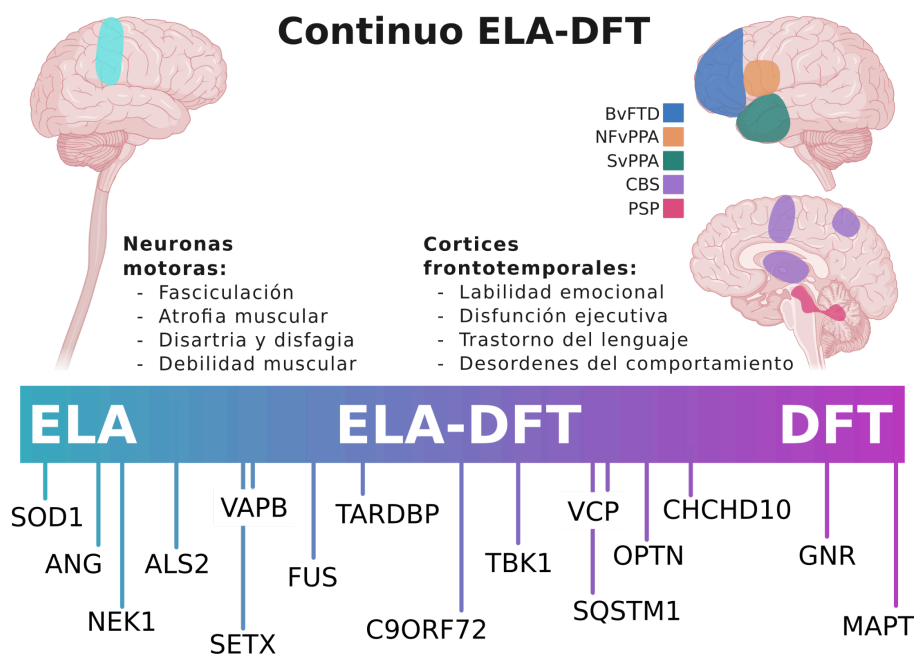


Figura 2. Esquema del espectro compartido por la ELA y la DFT, tanto en regiones cerebrales afectadas, síntomas y genes involucrados en las patologías. Creada con BioRender.

Un rasgo común es, por ejemplo, la presencia de formas anormales y mal funcionamiento de la proteína TDP-43, un sello distintivo en la ELA que se encuentra en el 97% de todos los pacientes y en el 50% de los pacientes con DFT (13–15). TDP-43 es una proteína ubicua de unión ADN/ARN que forma parte integral de numerosas funciones de procesamiento del ARN, incluyendo la regulación transcripcional, el empalme del ARN y el transporte del ARN. La pérdida de su función normal conduce a defectos de empalme, incluyendo la retención aberrante de intrones, alterando en última instancia la expresión génica y las redes de interacción proteína-proteína (PPI, del inglés *Protein-protein Interaction*) (16–18).

Otras coincidencias son que ambas enfermedades carecen de biomarcadores específicos, planteando así importantes retos a la hora de definir el diagnóstico, el pronóstico, la estratificación de cohortes y la progresión de la terapia.

Por lo tanto, sería de utilidad descubrir biomarcadores distintivos, capaces de identificar subpoblaciones de pacientes con ELA o DFT y ofrecer posibles terapias basadas en dianas específicas, producto de las diferencias en el transcriptoma y el interactoma de pacientes e individuos sanos.

1.4 Redes biológicas

Como se dijo anteriormente, es sabido que la red PPI en las neuronas motoras de pacientes con ELA se ve afectada, siendo, por ejemplo, uno de los motivos (entre otros) el mal funcionamiento de la TDP-43 mutada. Esto hace que considerar la red de interacción entre proteínas sea un buen punto de partida para comprender mejor las diferencias entre las personas que padecen la enfermedad y las personas sanas.

En la teoría de grafos, una estructura de red está conformada por nodos, que son la unidad fundamental de la red (o grafo) y representan una o más entidades (como personas, ubicaciones o conceptos) conectadas a otros nodos mediante enlaces (o aristas), que representan las relaciones entre ellos (**Figura 3**). En biología, los nodos son moléculas o procesos biológicos relacionados entre sí, y el criterio de interés define el tipo de relación y, por ende, el tipo de enlace.

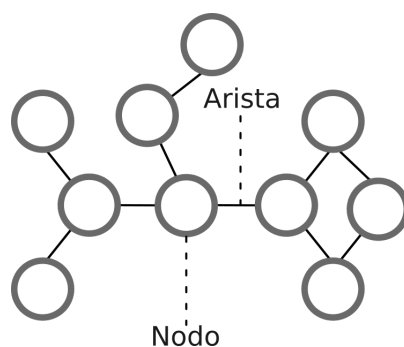


Figura 3. Representación de una red no dirigida según la teoría de grafos.

En el caso de las redes PPI, los nodos son proteínas y los enlaces representan si estas interactúan entre sí. La interacción puede definirse con distintos grados de confianza, dependiendo de los experimentos y bibliografía utilizados para evaluarla.

Existen diversas bases de datos cuyo objetivo es representar de la forma más fiel posible los proteomas de distintos organismos. Dentro de las más conocidas se hallan IntAct (19), HIPPIE (20), BioGRID (21), STRING (22), IID (23) y HPRD (24), entre muchas otras. Cada una de ellas cuenta con criterios específicos para definir una PPI en su base de datos, tales como experimentos de alto o bajo rendimiento, bibliografía específica, pasos de curado a seguir o incluso en qué organismos está presente.

El estudio de estas redes es de suma importancia para entender los complejos sistemas que conviven dentro de las células, pudiendo profundizar en distintos procesos celulares, en las funciones de las proteínas (ya que de muchas aún se desconoce su rol biológico) y en procesos evolutivos. Otra utilidad muy relevante es la de comprender los mecanismos de patogénesis de múltiples enfermedades, pudiendo acercarnos a tratamientos terapéuticos específicos, siendo las PPI un entorno crucial para orientar el descubrimiento de fármacos.

Estudiar redes de este tipo requiere de análisis específicos de redes, sin olvidar su contexto biológico; por lo tanto, las redes convolucionales aplicadas a grafos podrían captar las particularidades de estas, sin dejar de lado dicho contexto.

1.5 Redes neuronales convolucionales y su aplicación a redes biológicas

En los últimos años, el aprendizaje profundo (o en inglés *Deep Learning*) se ha aplicado con éxito en diversos ámbitos, logrando tareas generales tales como reconocer y detectar objetos visualmente, reconocer el habla o procesar el lenguaje

natural. También ha tenido éxito en tareas más específicas, como el descubrimiento de fármacos y estudios genómicos (25,26).

Los métodos de aprendizaje profundo buscan aprender automáticamente las representaciones abstractas de datos brutos, de modo que puedan ser utilizadas para realizar predicciones, clasificaciones o reconocimientos de patrones. Usualmente, estos modelos se componen de varias capas de procesamiento, y en muchas aplicaciones se utilizan arquitecturas de redes neuronales con múltiples capas ocultas (en inglés *Hidden Layers*), de ahí denominado aprendizaje "profundo".

Dentro de este conjunto de modelos se hallan las redes neuronales convolucionales (CNN, del inglés *Convolutional Neural Networks*) (27). Éstas generan resultados a partir de datos de estructura euclidiana, pudiendo tener éstos distinta dimensionalidad espacial. Por ejemplo, 1D las señales y secuencias, 2D las imágenes y 3D los vídeos. Las CNN son capaces de capturar patrones de señal espaciales locales y luego fusionarlas en abstracciones de alto nivel.

Originalmente fueron creadas con el objetivo de analizar imágenes, pero en la actualidad se han convertido en herramientas fundamentales para el procesamiento de múltiples tipos de matrices, no sólo de píxeles. Sus arquitecturas más comunes suelen contener tres tipos de capas (**Figura 4**):

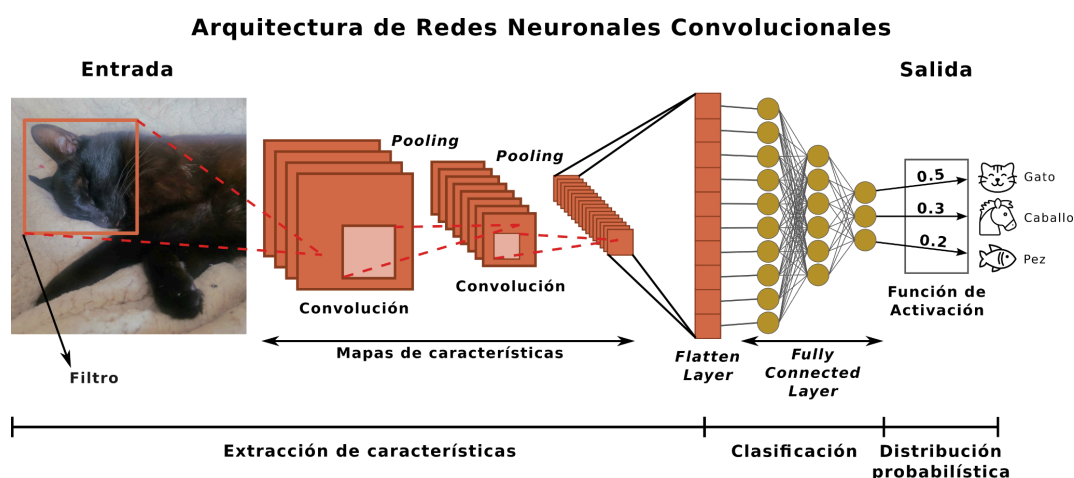


Figura 4. Esquema de una arquitectura típica de redes neuronales convolucionales.

- Capas convolucionales (*Convolutional Layers*, en inglés). La convolución es la operación matemática en la cual se combinan dos funciones para dar como resultado una tercera función. En las capas de convolución se combinan los datos de entrada con distintos filtros (también denominados *kernels*, en inglés). Los filtros son matrices más pequeñas de pesos y sesgos determinados por los datos de entrada. Éstos son aprendidos por el modelo, y cada uno de ellos tendrá la capacidad de captar características únicas de los datos. Una capa convolucional puede tener decenas de filtros, aprendiendo múltiples características a la vez.

En el caso de datos matriciales, como las imágenes o los grafos, la operación de convolución aplica el filtro sobre regiones locales de los datos de entrada y calcula un producto escalar que mide la presencia de un patrón específico en distintas ubicaciones. Así, detecta características locales independientemente de su posición.

Luego, se aplica una función de activación que introduce no linealidad en cada neurona de la red, determinando si esta se activa o no en función de su entrada, lo que permite al modelo aprender representaciones complejas que no podrían capturarse mediante combinaciones lineales.

De esta manera, cada filtro crea un mapa de características (*feature map*, en inglés), cuyos valores indican tanto la ubicación como la intensidad con la que la característica aprendida aparece en los datos de entrada.

- Capas de agrupación (*Pooling Layers*, en inglés): El proceso de *pooling* consiste en la condensación de los datos de salida obtenidos luego de la convolución. Esto permite reducir su dimensionalidad, condensar la información y favorecer la invariabilidad frente a desplazamientos pequeños de los filtros. Las dos estrategias más utilizadas son las de *Max* y *Average Pooling*, calculando el máximo o el promedio de los valores de los datos en un sector local, respectivamente.

Mientras que la capa convolucional detecta señales locales, la capa de agrupación fusiona características similares. Varias capas convolucionales y de

agrupación sucesivas permiten el aprendizaje de características jerárquicas, en las que las características de nivel superior se obtienen componiendo las de nivel inferior. Por ejemplo, en las imágenes, las combinaciones locales de bordes crean motivos, los motivos se ensamblan en partes y las partes forman objetos.

- Capas densas o totalmente conectadas (*Fully Connected Layer*, en inglés): Las capas totalmente conectadas reciben las representaciones generadas por las capas convolucionales y de agrupación. Con estas representaciones proporcionan valores de probabilidad que clasifican/predicen a qué categoría corresponden los datos de entrada. Las capas densas actúan como una ‘caja negra’: al estar todas las neuronas conectadas entre sí, se crean miles o millones de pesos que interactúan entre sí de forma no lineal, lo que hace imposible rastrear de manera intuitiva qué combinación de entradas provocó qué salida

De este modo, a diferencia de las capas convolucionales que están diseñadas para detectar patrones locales en los datos de entrada, las capas densas establecen conexiones entre cada neurona y la totalidad de las neuronas de la capa precedente. Esto les permite integrar de manera global todas las representaciones extraídas por las capas convolucionales previas. Así, mientras en las capas convolucionales las neuronas se conectan únicamente a una región restringida de la capa anterior, implicando que solo se procese un subconjunto de la información disponible, las capas completamente conectadas tienen como función consolidar dicha información para producir una representación que sirva de base para la clasificación o predicción final del modelo.

Con el paso de los años, las CNN se han vuelto más complejas, dando lugar a nuevos enfoques capaces de extender la técnica a dominios no euclidianos, como las redes biológicas. Esto se logró mediante la generalización de las CNN a grafos, ofreciendo así valiosas herramientas para el análisis bioinformático, como las *Graph Convolutional Neural Networks* (GCNN). Estas últimas se han utilizado exitosamente, por ejemplo, para clasificar pacientes con cáncer de mama, distinguiendo entre metastásicos y no metastásicos (28). Básicamente, lo que se ha logrado es asociar a cada nodo de una red molecular un valor de expresión génica,

realizando en última instancia una tarea de clasificación de grafos de señales únicos de cada paciente.

1.6 Interpretación de las redes neuronales convolucionales en grafos

Como los modelos de aprendizaje profundo son considerados una ‘caja negra’, los diseñadores de estas inteligencias artificiales (IA) no son capaces de definir cuál fue el proceso exacto por el cual los algoritmos arriban a respuestas específicas. Para intentar interpretar como razonan estos sistemas se han desarrollado métodos denominados IA explicativa (*Explainable AI*, en inglés), que apuntan a proveer al usuario entendimiento y transparencia respecto a estos algoritmos. Las soluciones obtenidas mediante redes neuronales deben ser debidamente explicadas antes de poder ser consideradas en el ámbito clínico (29). A modo de ejemplo, el Reglamento General de Protección de Datos de la Unión Europea restringe la toma de decisiones automatizada producida por algoritmos (30). Por ende, resulta imperativo interpretar correctamente las decisiones propuestas por el aprendizaje profundo para que puedan ser aplicadas en la práctica clínica.

La propagación de relevancia por capas (*Layer-wise Relevance Propagation*, *LRP*, en inglés) (31) es un método de IA explicativa a través del cual se logra una propagación hacia atrás dentro de toda la arquitectura de la red neuronal. Comenzando por los valores probabilísticos finales a los que se arribó durante la clasificación, se le otorga a cada una de las neuronas individuales de la red un valor de relevancia para dicha clasificación. En última instancia la información original es regenerada, teniendo cada una de las características de entrada un valor de relevancia para la clasificación. Aplicado en grafos se obtiene GLRP (de inglés *Graph Layer-wise Relevance Propagation*) capaz de identificar cuáles son los nodos cuyo contexto y señales en una red irregular son críticos para la clasificación precisa del grafo de señales.

En esta trabajo se utilizó GCNN-GLRP (28) para estudiar perfiles de expresión génica de muestras de sangre total de individuos con ELA y controles sanos, integrados a redes PPI, con el fin de identificar genes codificadores de proteínas clave para la distinción entre grupos y sus posibles roles en la ELA. Con estos genes se realizaron análisis de enriquecimiento en vías biológicas, buscando la posible vinculación de distintos procesos biológicos a la patología de la enfermedad. Por último, se repitió el análisis con pacientes con DFT para explorar diferencias y puntos en común con la ELA. En la **Figura 5** se muestra el flujo de trabajo que se siguió a lo largo de esta tesis.

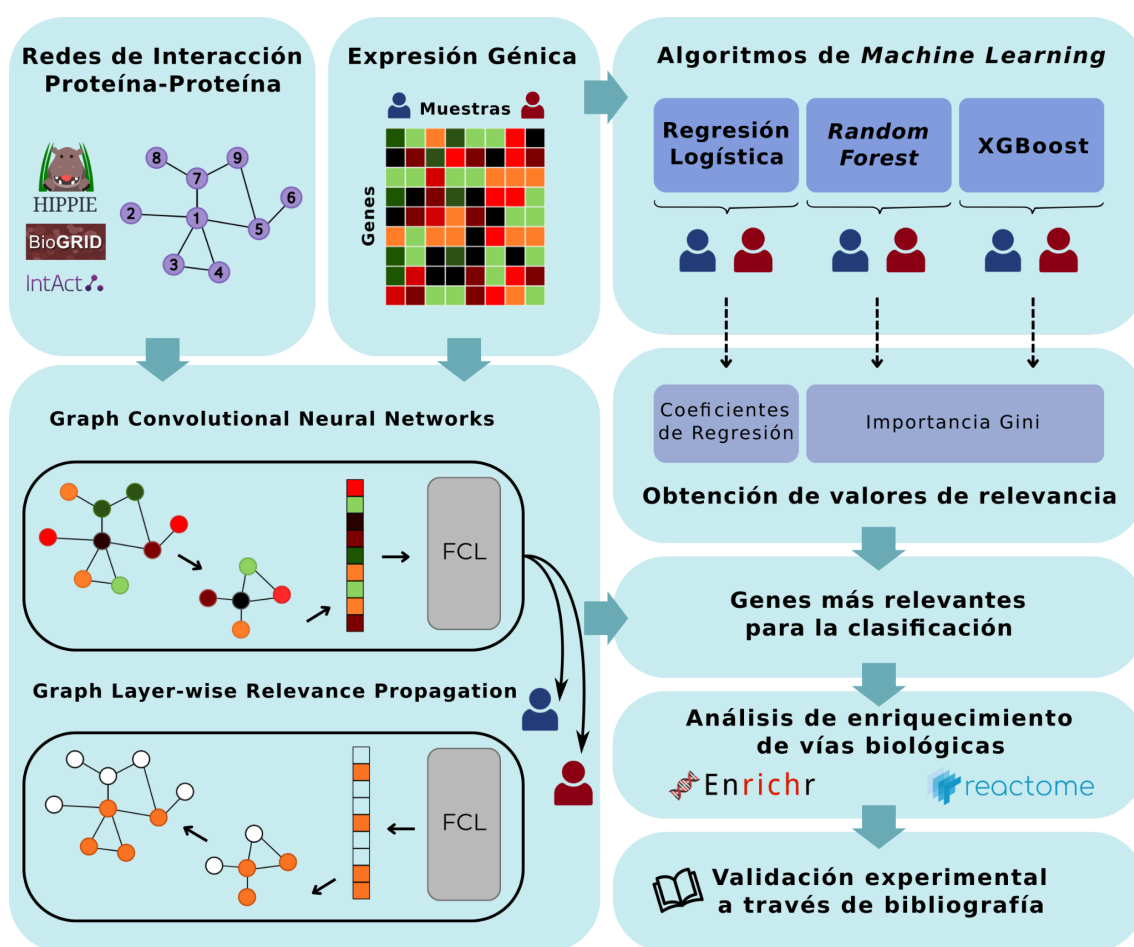


Figura 5. Diagrama con el flujo de trabajo seguido para la realización de este trabajo.

Hipótesis y Objetivos

2.1 Hipótesis

La integración de redes de interacción proteína-proteína y perfiles de expresión génica en pacientes con esclerosis lateral amiotrófica es fundamental para la identificación de mecanismos biológicos relevantes en la patología.

2.2 Objetivo general

El objetivo general de la tesis fue, por un lado, aplicar enfoques de inteligencia artificial para clasificar pacientes con esclerosis lateral amiotrófica y demencia frontotemporal a partir de su expresión génica, evaluando el impacto de incorporar redes de interacción proteína-proteína en la clasificación misma.

Por otro lado, identificar los genes relevantes para la clasificación, para dilucidar la posible relación biológica con los mecanismos moleculares de ambas enfermedades.

2.3 Objetivos específicos

- Clasificar pacientes con esclerosis lateral amiotrófica y controles sanos a partir de su expresión génica en muestras de sangre total, utilizando distintos modelos de inteligencia artificial, tanto incorporando como sin incorporar información de redes de interacción proteína-proteína.
- Estimar y comparar la estabilidad y la selección de genes relevantes para la clasificación de pacientes realizada por el modelo GCNN y otros algoritmos de inteligencia artificial.
- Evaluar la posible relación biológica entre los genes identificados como más relevantes para la clasificación y la esclerosis lateral amiotrófica.
- Reproducir el análisis en pacientes con demencia frontotemporal y comparar los resultados con los obtenidos en esclerosis lateral amiotrófica.

Materialles y Métodos

3.1 Recopilación de datos

3.1.1 Perfiles de expresión de pacientes de ELA

Los dos conjuntos de datos utilizados en este trabajo se obtuvieron del repositorio *NCBI Gene Expression Omnibus* (GEO) (32). Ambos se construyeron a partir de muestras de sangre de pacientes con ELA y controles sanos (33,34). El primer conjunto (GEO *accession*: GSE112676, de ahora en más referido como ALS741) se encontraba compuesto por perfiles de expresión génica pertenecientes a 233 pacientes con ELA y 508 controles, los cuales fueron obtenidos utilizando la plataforma de expresión Illumina HumanHT-12 BeadChip V3.0. El segundo conjunto (GEO *accession*: GSE112680, referido como ALS301) incluía perfiles de expresión génica pertenecientes a 164 pacientes con ELA y 137 individuos sanos, los cuales fueron generados usando la plataforma de expresión Illumina HumanHT-12 BeadChip V4.0.

Las plataformas utilizadas son *microarrays* de ADN: *chips* sólidos pequeños en los cuales se encuentra fijada de forma ordenada una colección de sondas de ADN, correspondientes a un gen en particular. Esta colección dependerá de lo que se desea analizar en la muestra. Si bien cada *chip* posee características distintivas, los pasos usuales para realizar el experimento son comunes: Se debe extraer el ARN, para luego convertirlo en ADN complementario (ADNc) que es marcado con fluorescencia. Se lo coloca sobre el *chip*, buscando que hibridice con la sonda complementaria. Finalmente, un láser detecta la fluorescencia y mide la intensidad en cada punto, lo que indica el nivel de expresión de cada gen (**Figura 6**). En este caso, ambas plataformas utilizadas contenían muestras del transcriptoma completo (*whole-genome expression*), siendo 48.804 el total de sondas para el Illumina HumanHT-12 BeadChip V3.0, y de 47.231 para la V4.0.

En ambos conjuntos de datos los valores de intensidad de señal se hallaban normalizados siguiendo el mismo *pipeline*, dando un total de 29.830 genes. Adicionalmente, como se trataban de dos conjuntos diferentes, para poder

compararlos se los normalizó calculando el ‘puntaje z’ (*z-score*, en inglés) de la siguiente manera:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

Donde Z_i es el **Z-score** del gen i , X_i su valor de expresión, μ_i la media y σ_i la desviación estándar de expresión de dicho gen en todo el conjunto de datos. De esta manera, se logra estandarizar los valores de expresión y que queden con media igual a 0 y desviación estándar igual a 1.

Finalmente, el conjunto ALS741 se utilizó para entrenar y testear todos los modelos, mientras que ALS301 se utilizó para la posterior validación independiente.

Microarray de ADN

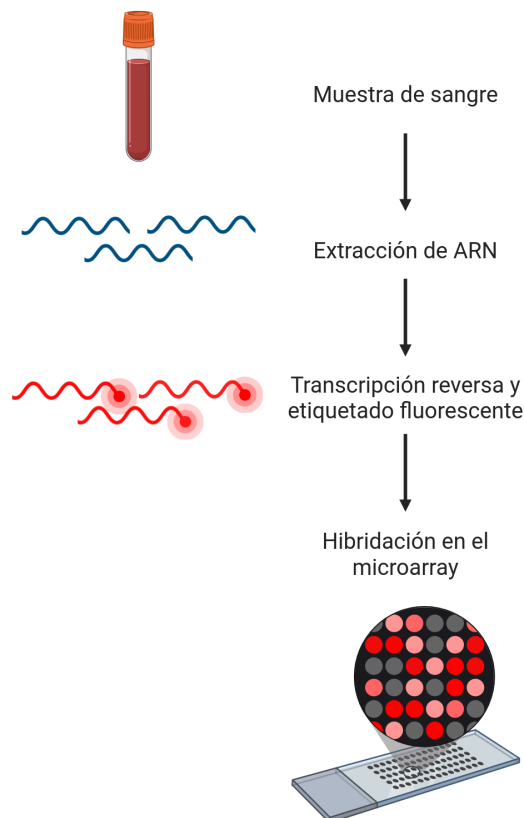


Figura 6. Esquema de los pasos fundamentales de la medición de expresión génica mediante *microarray* de ADN. Realizada con BioRender.

Se utilizaron datos de expresión génica como una aproximación indirecta de los niveles de proteína, asumiendo que, si bien la relación entre ambos no es siempre lineal ni directa, la abundancia del ARNm refleja tendencias generales en la actividad proteica.

3.1.2 Perfiles de expresión de pacientes de DFT

También se seleccionó un conjunto de datos de DFT (GEO *accession*: GSE140830) (35), que incluía perfiles de expresión génica también generados a partir de muestras de sangre y obtenidos utilizando la plataforma de expresión Illumina HumanHT-12 V4.0. Este conjunto contiene valores de intensidad de señal normalizada de 15.912 genes e incluía 251 pacientes con DFT y 268 controles sanos.

Los pacientes de DFT se encontraban divididos en cinco variantes clínicamente distintas: variante comportamental de la DFT (BvFTD), parálisis supranuclear progresiva (PSP), síndrome corticobasal (SCB), afasia progresiva primaria no fluida (NFvPPA, del inglés *Nonfluent variant Primary Progressive Aphasia*) y afasia progresiva primaria semántica (SvPPA, del inglés *Semantic variant Primary Progressive Aphasia*). Se agruparon las variantes en tres categorías clínicas más amplias, dependiendo de qué rasgos fueron mayoritariamente afectados en los pacientes: comportamental (BH, del inglés *behaviour*), que contenía todos los individuos con BvFTD; motora (MT), con individuos con PSP y CBS; y lenguaje (LG), con individuos con NFvPPA y SvPPA. (36).

Para estudiar los puntos en común entre las variantes de DFT directamente relacionadas con problemas en la motricidad y la ELA, se entrenaron y evaluaron modelos separados que clasificaron a los pacientes MT como muestras positivas y a los pacientes BH y LG como muestras negativas (denominados BH-MT y LG-MT, respectivamente). En la **Tabla 1** se muestra un resumen de todos los conjuntos de datos utilizados.

	ALS741		ALS301		DFT		
	<i>ELA</i>	<i>Control</i>	<i>ELA</i>	<i>Control</i>	<i>BH</i>	<i>MT</i>	<i>LG</i>
N	233	508	164	137	75	88	88
Género (% de mujeres)	38.6	45.3	41.5	42.3	43.8	54.4	53.8
Edad de inicio de enfermedad (IQR)	64 (56-71)		63 (55-71)				
Edad de toma de muestra (IQR)					63 (58-67)	70 (64-74)	68 (62-72)
Bulbar (%)	38.6		34.2				
Platforma	IHHT-12 V3.0		IHHT-12 V4.0				

Tabla 1. Resumen de los conjuntos de datos de ELA y DFT.

IHHT-12 = Matriz de expresión Illumina HumanHT-12. IQR = Rango intercuartílico

3.2 Construcción de redes PPI

Las proteínas no actúan de forma independiente, sino en una red de interacciones moleculares complejas. Por lo tanto, es importante identificar las interacciones entre proteínas. Las diferentes técnicas experimentales que se han desarrollado para medirlas varían considerablemente, sobre todo en cuanto a los datos que producen.

Cada base de datos de PPI presenta enfoques y criterios distintivos para definir una interacción, tales como experimentos requeridos y proteínas incluidas, entre otros. Adicionalmente presentan un nivel de curación diferente. Para realizar un análisis más robusto, se utilizaron tres redes de PPI (IntAct, BioGRID y HIPPIE) y se definieron los criterios para aceptar una interacción, como se detalla a continuación.

De cada red se filtraron los genes sin datos de expresión génica disponibles y, del subconjunto resultante, se retuvo únicamente el componente conectado de mayor tamaño, ya que GCNN no puede utilizar redes inconexas para la clasificación. Este procedimiento se aplicó de igual forma a los datos de DFT.

El número resultante de nodos y aristas de cada red de PPI se indica en la **Tabla 2**. El solapamiento de genes e interacciones entre redes se muestra en las **Figuras 7 y 8** para ELA y DFT, respectivamente.

	ELA		DFT	
	<i>Genes</i>	<i>Interacciones</i>	<i>Genes</i>	<i>Interacciones</i>
BioGRID	7.256	48.997	7.890	60.664
HIPPIE	9.037	68.298	9.140	77.246
IntAct	10.585	60.856	11.157	65.873

Tabla 2. Resumen del número de nodos (genes codificantes de proteínas) y aristas (interacciones proteína-proteína) en cada una de las redes de PPI seleccionadas utilizadas para entrenar y probar el modelo GCNN-GLRP. Los mismos conjuntos de genes se utilizaron como características de entrada en los clasificadores que no utilizan información de las interacciones.

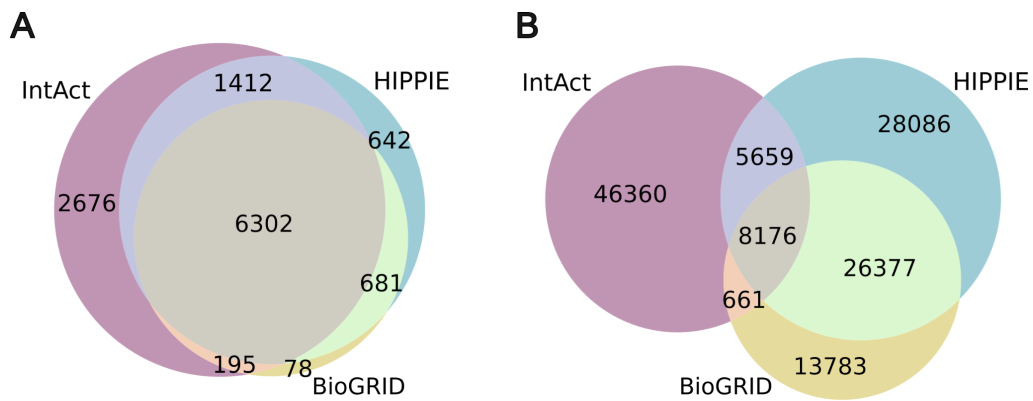


Figura 7. Diagramas de Venn con el número de **A.** nodos y **B.** aristas en las tres redes PPI utilizadas con el conjunto de datos de ELA.

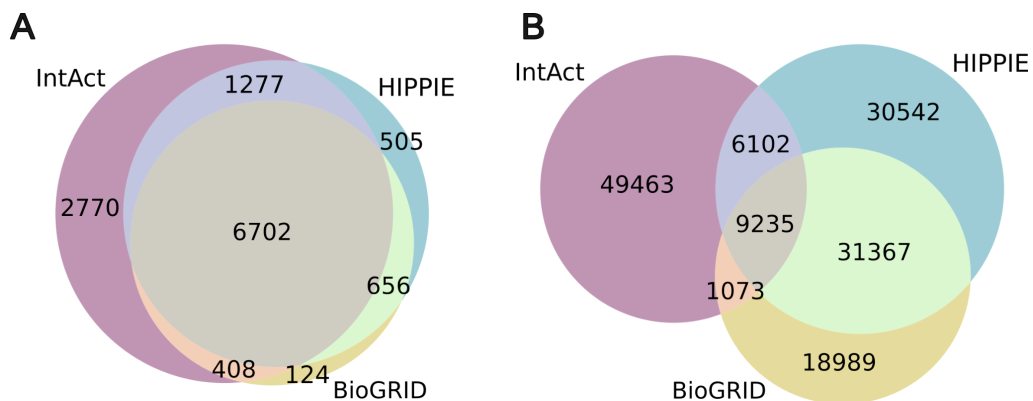


Figura 8. Diagramas de Venn con el número de **A.** nodos y **B.** aristas en las tres redes PPI utilizadas con el conjunto de datos de DFT.

Para los modelos que no incorporan información de redes de PPI para su clasificación, se utilizó el mismo conjunto de genes incluidos en las redes de PPI como datos de entrada para asegurar una comparación justa.

3.2.1 IntAct

IntAct (19) es un sistema de base de datos y herramientas gratuitas de código abierto que es miembro de IMEx (del inglés *International Molecular Exchange Consortium*), un proyecto internacional con el objetivo de recopilar de forma colaborativa datos sobre interacciones moleculares procedentes de literatura científica y de datos depositados directamente. IntAct hace hincapié en una curación exhaustiva que captura interacciones con suficiente detalle para respaldar la posterior presentación, agregación y análisis de los datos. En 2017, IMEx se convirtió en un recurso de datos básico de ELIXIR, siendo reconocido como parte de la infraestructura fundamental para las ciencias de la vida. IntAct es utilizada por todos los socios activos de IMEx (IntAct, DIP, UniProt, MINT, MatrixDB, UCLICS e IID) como plataforma común de curación y de difusión de datos.

Entre las herramientas disponibles en IntAct se hallan aquellas preparadas para la visualización y el análisis de datos. En diciembre de 2025, albergaba ~1,7 millones de interacciones binarias, meticulosamente seleccionadas por curadores. En este trabajo se utilizó su versión n° 248, seleccionando las interacciones definidas como "asociación física" o "interacción directa" con un puntaje MI (o *MI-score*) (37) superior a 0,4, considerado de confianza media a alta. Este puntaje evalúa el grado de confianza de que exista una interacción molecular. Para ello, considera tres factores: el método de detección de la interacción, el tipo de interacción y el número de publicaciones que la reportan, generando un puntaje normalizado entre 0 y 1.

3.2.2 BioGRID

BioGRID (del inglés *Biological General Repository for Interaction Datasets*) (21) es una base de datos de acceso abierto creado en 2006 con el objetivo específico de curar de forma exhaustiva todos los datos disponibles sobre interacciones biológicas generadas en la levadura *Saccharomyces cerevisiae*. En diciembre de 2025, contenía ~2,9 millones de interacciones proteicas y genéticas seleccionadas de los principales organismos modelo (incluyendo humanos) y otras especies menos estudiadas. Esta base de datos integra datos experimentales tanto de bajo como de alto rendimiento. También contiene modificaciones postraduccionales e interacciones entre fármacos u otras sustancias químicas y sus proteínas diana.

Debido a la gran extensión de la literatura biomédica humana, BioGRID ha adoptado en parte un enfoque centrado en los procesos biológicos y/o las enfermedades con el fin de profundizar en la curación de áreas críticas de la biología humana. Todas las interacciones proteicas y genéticas anotadas en BioGRID se derivan exclusivamente de la curación manual por parte de expertos de datos experimentales publicados en revistas revisadas por pares. BioGRID utiliza vocabularios estructurados personalizados para describir los datos de interacción, que incluyen 17 códigos de evidencia de PPI diferentes y 11 códigos de evidencia de interacción genética. De las PPI, el 90% de las cuales se derivan de métodos experimentales *in vivo*.

BioGRID además posee una herramienta de visualización de redes incorporada, que permite a los usuarios generar sus propias redes de interacción.

Se utilizó la versión v4.4.226, y de la misma se tomaron todas las interacciones físicas multi-validadas. BioGRID considera una interacción como multi-validada si existen dos publicaciones o fuentes experimentales diferentes que den cuenta de dicha relación.

3.2.3 HIPPIE

HIPPIE (del inglés *Human Integrated Protein-Protein Interaction rEference*) (20), lanzada en 2011, es una base de datos de redes PPI humanas que reúne interacciones altamente confiables y contexto-específicas, además de guiar su interpretación. Utiliza la interfaz PSICQUIC (38) para actualizar periódicamente el repositorio de PPIs, el cual contiene datos experimentales de IntAct, MINT, BioGRID, HPRD, DIP, BIND y MIPS.

En HIPPIE, para determinar la inclusión de una interacción en la base de datos, se asigna una puntuación de confianza normalizada. Ésta se calcula como la suma ponderada del recuento de estudios, la calidad y el recuento de técnicas, y la replicación entre especies.

HIPPIE ofrece también la posibilidad de visualizar resultados, lo cual está pensado para inspeccionar proteínas de interés en el interactoma humano (como aquellas asociadas a enfermedades, procesos biológicos, etc). Adicionalmente, la base de datos cuenta con estrategias bioinformáticas para anotar las PPI con información contextual y seleccionar subconjuntos significativos de toda la red.

En diciembre de 2025, HIPPIE contenía ~70.000 interacciones con puntuación de confianza. Para este trabajo se tomó la versión 2.3 y se conservaron todas aquellas con un valor de confianza superior a 0,73. Esto representa el 25% superior de las interacciones de esta base de datos, consideradas de alta confianza.

3.3 Modelos de *Machine Learning*

Se utilizó el modelo GCNN-GLRP (28) para la clasificación de pacientes con ELA (casos) e individuos sanos (controles), asignando datos de expresión génica a los nodos que representan los productos génicos (proteínas) dentro de una red PPI.

Adicionalmente, para comparar los resultados de la clasificación por este método, se entrenaron modelos de Regresión Logística (RL), *Random Forest* (RF) y

XGBoost. Estos clasificadores tratan cada gen como una variable independiente y no incorporan ninguna información de la red PPI.

El rendimiento se evaluó considerando distintas métricas, pero principalmente se consideró el área bajo la curva ROC (AUC-ROC, del inglés *Area Under the Receiver Operating Characteristic Curve*). Esta curva se calcula comparando la tasa de verdaderos positivos, también llamado ‘Sensibilidad’ o *recall*, con la de falsos positivos a diferentes umbrales. El área debajo de esta curva es una medida del desempeño de un modelo para realizar una clasificación binaria, distinguiendo entre casos positivos y negativos.

De un clasificador aleatorio se espera que clasifique azarosamente y por tanto obtendría un AUC-ROC igual a 0,5. En cambio, un predictor perfecto obtendría un valor de AUC-ROC de 1. En la **Figura 9 A** se muestra una matriz de confusión con todas las posibles predicciones, y en **B** una curva ROC esquemática mostrando el caso ideal, intermedio y aleatorio.

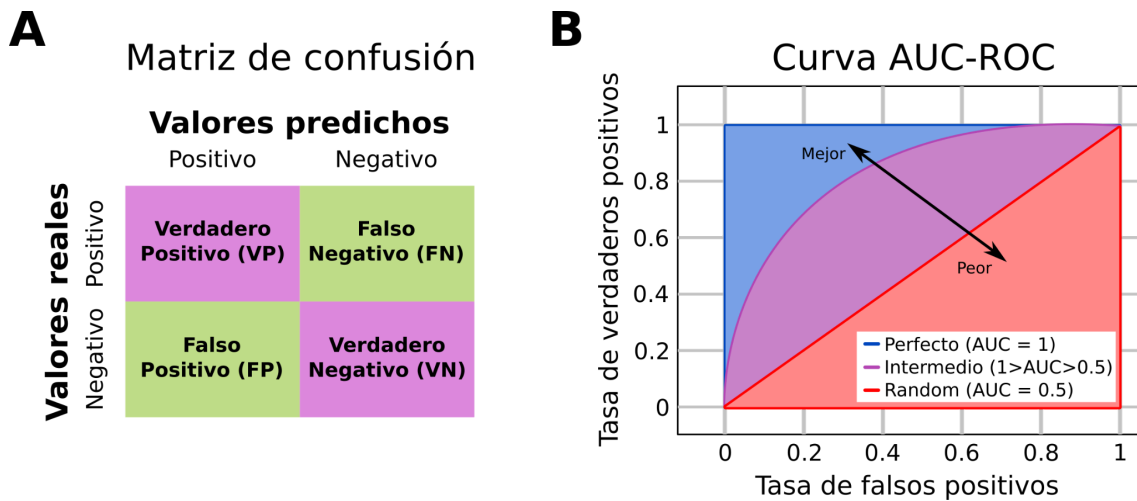


Figura 9 A. Matriz de confusión que resume los resultados de una clasificación binaria en términos de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). **B.** Curvas ROC que ilustran la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para tres escenarios: clasificador perfecto (AUC = 1), clasificador con desempeño intermedio ($1 > \text{AUC} > 0,5$) y clasificador aleatorio (AUC = 0,5).

Las otras métricas que se consideraron fueron la exactitud (*Accuracy*, en inglés), el *Recall* y el puntaje F1 (*F1-Score*). *Accuracy* da cuenta del porcentaje de muestras bien clasificadas:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

El *recall* indica la frecuencia con la que el modelo identifica correctamente los verdaderos positivos de todas las muestras positivas:

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

Y por último, el *F1-Score* es la media armónica de la precisión y el *recall*, representando simétricamente ambas métricas en una sola. Si se define la precisión como:

$$Precision = \frac{VP}{VP + FP} ,$$

podemos definir el *F1-Score* como:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Para minimizar el sesgo en la evaluación del desempeño de los modelos de ML, se empleó *10-fold cross validation* para el conjunto de datos de ELA y *5-fold* para el de DFT. Esta elección se basó en un análisis previo de rendimiento que indicó que dichas divisiones producían los mejores resultados para cada conjunto. El *k-fold cross validation* es una técnica que consiste en dividir el conjunto de datos en *k* subconjuntos de igual tamaño. En cada iteración, uno de los subconjuntos se reserva como conjunto de testeo mientras los *k-1* restantes se utilizan para el entrenamiento. Este proceso se repite *k* veces, de modo que cada subconjunto actúa una vez como conjunto de testeo, obteniendo así una evaluación más robusta y menos dependiente de una única partición de los datos. Por último, los modelos se ejecutaron 10 veces con cada red PPI y de cada métrica se calculó la mediana.

También se evaluó que la función de pérdida baje consistentemente. Ésta cuantifica la diferencia entre las predicciones dadas por los modelos y los valores reales, definiendo la señal que guía el proceso de optimización durante el entrenamiento mediante la minimización iterativa de dicho error.

3.3.1 GCNN-GLRP

GCNN es capaz de aprender los patrones de señal generados por las variables explicativas y su ubicación en una red, es decir, su contexto. No obstante, generalizar CNNs para que puedan procesar y analizar grafos no es trivial. Surge una limitación importante a la hora de realizar las tareas de convolución y *pooling*, ya que solo están definidas para redes regulares. Esto impide la definición de filtros localizados que sean eficientes para evaluar y aprender de dominios no regulares.

Por ende, resulta necesario diseñar filtros convolucionales locales que sean efectivos y tengan en cuenta las características propias de los grafos, y que el *pooling* sea capaz de agrupar nodos similares, reduciendo el tamaño del grafo pero manteniendo sus propiedades esenciales. Al mismo tiempo, otra condición relevante es que la complejidad computacional sea lineal con respecto a la dimensionalidad del grafo, para mantenerla baja y haciendo asequible el análisis.

Para abordar este desafío surge el denominado aprendizaje profundo geométrico (*Geometric Deep Learning*, en inglés), una modalidad que se enfoca en la generalización de los modelos de aprendizaje profundo a dominios no euclidianos. En particular, las GCNN (39,40) son algoritmos capaces de analizar y procesar patrones de señales estructurados en interacciones complejas. Las señales pueden ser cualquier función que transmita información sobre un fenómeno determinado. Un fenómeno asociado a cada uno de los nodos es una señal que puede integrarse a la red, conformando un ‘grafo de señales’ y permitiendo evaluar conjuntamente las conexiones y el fenómeno (**Figura 10**). En nuestro caso, para poder clasificar a los pacientes de ELA (de aquí en adelante denominados ‘casos’) y los controles, generamos grafos de señales para cada individuo considerando su expresión génica y una red de interacción proteína proteína (**Figura 11**). Las GCNN pueden realizar diversas tareas sobre estos grafos de señales, tales como la clasificación de nodos, la predicción de aristas o la clasificación del grafo mismo, como es nuestro caso.

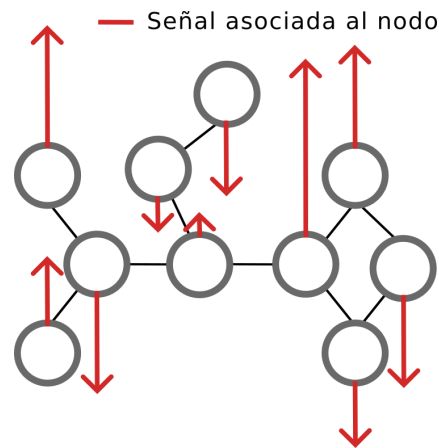


Figura 10. Esquema de un grafo de señales. Dependiendo el caso, las señales pueden tener valores positivos o negativos.

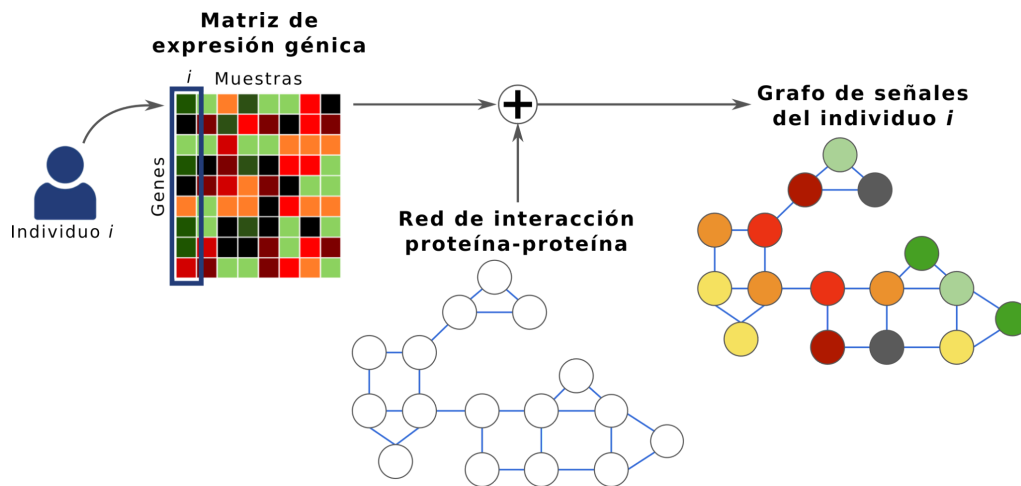


Figura 11. Esquema de la generación de un grafo de señal para un individuo i , el cual será utilizado para clasificar al individuo como caso o control. Los colores indican distintos niveles de expresión génica.

El procesamiento de señales en grafos (*Graph Signal Processing*, en inglés) no es trivial. Los filtros utilizados para estas tareas deben ser capaces de eliminar algunos componentes no deseados de la señal a estudiar y además el grafo debe estar representado como una matriz que pueda ser utilizada como dato de entrada para la convolución y *pooling*. Considerando este último punto, un grafo puede representarse mediante una matriz de adyacencia ($\mathbf{A}$): una matriz cuadrada donde cada elemento $\mathbf{A}_{mn}$ indica si existe conexión entre los nodos $\mathbf{m}$ y $\mathbf{n}$ (el elemento será 1 si existe conexión, o 0 si no la hay). También puede expresarse como una

matriz de pesos (**W**), donde el elemento indica algún tipo de peso de la conexión, en cuyo caso el elemento no será 1, sino el peso correspondiente. Un ejemplo es, en el caso de una red PPI, los niveles de evidencia que tiene la interacción entre dos proteínas.

Este tipo de matrices no pueden ser utilizadas en una red convolucional como tal, se deben transformar a un nivel que permita la formulación de operaciones fundamentales.

A partir de la matriz de pesos se puede definir la matriz de grados (**D**). Ésta es una matriz diagonal cuyos elementos D_{mm} corresponden a la suma de los pesos de todas las aristas incidentes en el nodo m:

$$D_{mm} = \sum_n W_{mn}$$

Con **W** y **D** se puede construir la matriz Laplaciana del grafo, definida como:

$$L = D - W$$

La matriz Laplaciana constituye un operador fundamental para el análisis espectral de grafos, ya que encapsula tanto la topología como la estructura de conectividad del grafo. Dado que se trata de una matriz simétrica y diagonalizable, puede factorizarse en términos de sus valores propios y vectores propios:

$$L = U\Lambda U^T$$

donde **U** es la matriz cuyas columnas corresponden a los vectores propios de la Laplaciana, definido como:

$$U = [u_0, u_1, \dots, u_{N-1}]$$

y Λ es una matriz diagonal que contiene los valores propios asociados ($\lambda_0, \lambda_1, \dots, \lambda_{N-1}$). Los vectores propios representan los modos de variación de las señales

sobre el grafo, y los valores propios indican qué tan suave o brusca es esa variación entre nodos vecinos.

Para analizar el contenido frecuencial de las señales, se utiliza la transformada de Fourier en grafos (*Graph Fourier Transform*, en inglés). Su forma análoga clásica sirve para descomponer una señal en la suma de componentes sinusoidales de distintas frecuencias. En ambas, la base espectral está dada por U , es decir, sus vectores propios. En este contexto, se define la Transformada de Fourier en grafos como la proyección de la señal sobre dicha base espectral:

$$\hat{x} = U^T x$$

La transformada inversa se obtiene mediante:

$$x = U \hat{x}$$

Siendo x la señal de interés, y $\hat{x}$ las componentes espectrales, representando éstas últimas la contribución de cada vector propio del grafo a la señal original.

Esta formulación permite definir la convolución en grafos de forma análoga a la de los dominios euclidianos. Con esta base teórica, las GCNN permiten extender el filtrado y el aprendizaje de señales a estructuras no euclidianas. En la **figura 12** se muestra la arquitectura del modelo GCNN utilizado en este trabajo.

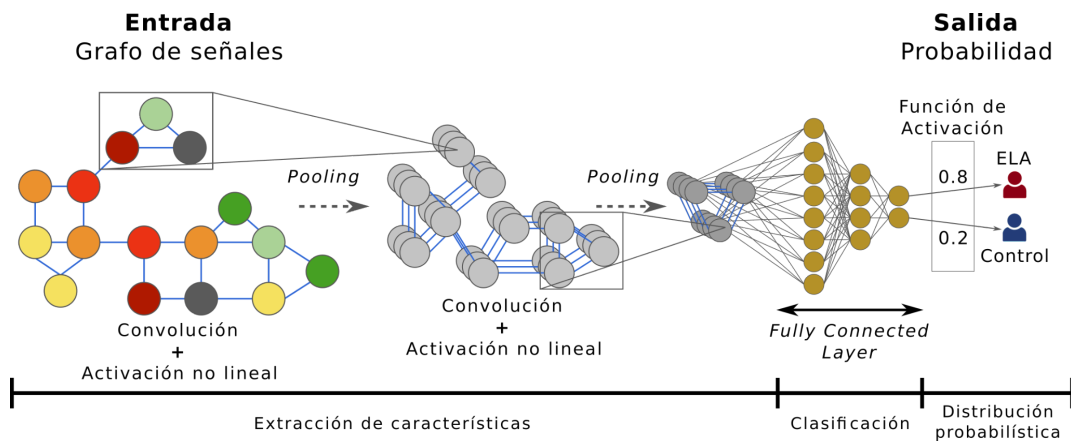


Figura 12. Esquema de la arquitectura de GCNN. Los datos de expresión génica de los pacientes se asignan a los nodos de la red PPI para generar el grafo de señales de entrada. Luego, la GCNN procesa los datos para finalmente predecir si el paciente es un caso o un control.

Para interpretar los resultados generados por la GCNN, se utilizó una herramienta de IA explicativa. Estos métodos buscan hacer interpretables las decisiones de clasificación de modelos complejos de ML. El método de LRP (31) opera mediante la redistribución retroactiva de la relevancia desde la capa de salida hacia la capa de entrada (**Figura 13**). Para cada neurona en una capa i , se cuantifica la contribución que cada neurona de la capa precedente ($i-1$) realizó a su activación. Este proceso se aplica de forma iterativa a través de todas las capas de la red, propagando la relevancia en sentido inverso al de la inferencia. La regla de propagación fundamental se define como:

$$R_j^{(i-1)} = \sum_k \frac{a_j \cdot w_{jk}}{\sum_{j'} a_{j'} \cdot w_{j'k}} R_k^{(i)}$$

donde $R_k^{(i)}$ es la relevancia de la neurona k en la capa i , a_j es la activación de la neurona j en la capa $i-1$, w_{jk} es el peso de la conexión entre ambas neuronas, y $R_j^{(i-1)}$ es la relevancia redistribuida hacia la neurona j . El denominador actúa como factor de normalización para garantizar la conservación de la relevancia total a lo largo de la propagación. Este proceso se repite hasta obtener una puntuación de relevancia asignada a cada elemento de la entrada original. En nuestro caso, utilizamos la técnica GLRP ya que se termina regenerando la red PPI original y cada uno de los nodos tendrá su respectivo valor de relevancia. Esto permite identificar qué características influyen de manera determinante en la predicción final del modelo.

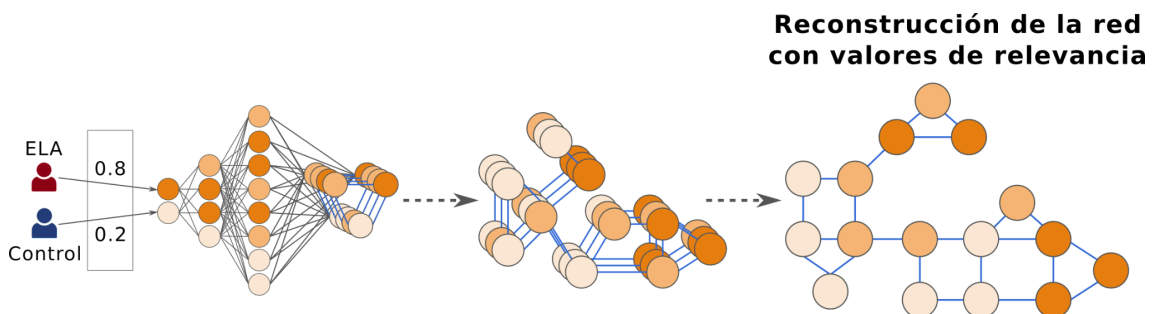


Figura 13. Esquema de GLRP, en la que los valores de relevancia son trasladados desde la salida hasta la entrada, devolviendo un valor de relevancia por nodo.

En el contexto de este trabajo, se aplicó el modelo GCNN-GLRP propuesto por (28) , utilizando como datos de entrada una matriz de expresión génica normalizada, una matriz de adyacencia representando todas las interacciones entre dos proteínas en la red PPI, y un vector binario que categoriza a los pacientes como "Caso" (1) o "Control" (0).

Para la optimización de hiperparámetros se realizó un barrido de parámetros (en inglés *parameter sweep*), corriendo todas las posibles combinaciones de los mismos buscando un rendimiento óptimo de los modelos. Se tuvieron en cuenta el siguiente listado de parámetros, estando entre paréntesis los valores que se pusieron a prueba:

- K-fold (10 o 5): Divide el dataset en K partes para entrenar y testear el modelo. Por ejemplo, de tomarse 10, el 90% del conjunto de datos se utiliza para entrenar y el 10% restante para testear.
- N° de *epochs*: Cantidad de veces que el modelo recorre el conjunto de datos completo durante el entrenamiento. Se comenzó con 50 y con los datos de ELA se llevó hasta 100, buscando la convergencia del modelo. En el caso de los datos de DFT, solo se necesitaron 40 *epochs*.
- *Pooling* (Max / Avg): Operación que reduce el tamaño de los mapas de características obtenidos luego de la convolución, tomando el valor máximo (Max) o el promedio (Avg) de una región.
- *Dropout* (0,5, 0,8 o 1): Proporción de neuronas activas aleatoriamente en cada *epoch* para reducir el sobreajuste. Éste se da cuando el modelo memoriza los datos de entrenamiento, incluyendo el ruido, en lugar de aprender patrones generales, logrando baja capacidad de generalizar. De este modo, nuevos valores suelen ser mal clasificados. En nuestro caso, si el *dropout* es igual a 0,8, el 20% de las neuronas se desactivan de forma aleatoria, forzando al modelo a no depender demasiado de ninguna neurona en particular
- Niveles (2 o 3): Número de capas convolucionales.

- Orden polinomial (2, 4 u 8): Grado del polinomio utilizado para aproximar el filtro convolucional al dominio espectral del grafo. Determina que cada nodo agrega información de hasta k nodos vecinos en su entorno local.
- *Learning rate* (0,1, 0,001 o 0,0001): Tamaño del paso con que el optimizador (encargado de minimizar la función de pérdida) ajusta los pesos en cada actualización durante el entrenamiento.

La combinación final utilizada para cada una de las redes en los conjuntos de datos de ELA y FTD se hallan en la **tabla 3**.

	ELA			BH-MT			LG-MT		
	Intact	BioGRID	HIPPIE	Intact	BioGRID	HIPPIE	Intact	BioGRID	HIPPIE
<i>K-fold</i>	10	10	10	5	5	5	5	5	5
<i>Epochs</i>	100	100	100	40	40	40	40	40	40
<i>Pooling</i>	Avg	Avg	Max	Avg	Max	Max	Max	Max	Max
<i>DO</i>	1	0,8	1	1	1	1	1	1	0,8
Niveles	2	2	2	3	2	2	2	2	2
O. P.	8	8	8	8	8	8	8	8	8
<i>L. R.</i>	0,001	0,001	0,001	0,001	0,001	0,001	0,0001	0,0001	0,0001

Tabla 3. Selección final de hiperparámetros de GCNN utilizados con cada conjunto de datos y red PPI luego del *parameter sweep*. DO = *Dropout*; O. P. = Orden polinomial; L.R. = *Learning rate*.

3.3.2 Regresión Logística

Uno de los modelos estadísticos más utilizados para realizar una clasificación binaria es el de regresión logística. En éste se ajustan los datos a una función logística (también conocida como sigmoidea), que tiene la siguiente forma:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Esta función permite obtener un valor de probabilidad entre 0 y 1, que representa la probabilidad de que una observación pertenezca a, por ejemplo, la clase positiva (1) (**Figura 14**).

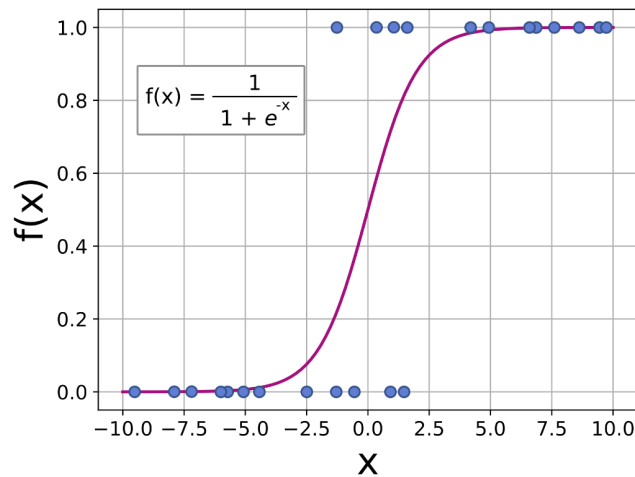


Figura 14. Gráfico de la función logística. Los puntos azules representan observaciones reales que fueron predichas como casos ($f(x)=1$), o controles ($f(x)=0$).

Para formular el modelo se considera un vector X de variables explicativas, definido como:

$$X = (x_1, x_2, \dots, x_p),$$

que en nuestro caso es la expresión génica perteneciente a un paciente. También se toma $Y \in \{0, 1\}$, la variable binaria que indicará a qué clase pertenece nuestro paciente (con o sin ELA, es decir, 1 o 0). La regresión logística modela entonces la probabilidad condicional como:

$$P(Y = 1 | X) = \pi(X).$$

Mediante la función logística antes mencionada queda como:

$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta^p X)}},$$

donde β_0 es la ordenada al origen y $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, es el vector de coeficientes del modelo.

Cada coeficiente β_j representa el cambio asociado a un incremento unitario en la variable x_j , manteniendo constantes las demás covariables. Es decir, el valor e^{β_j} cuantifica el factor multiplicativo de pertenecer a la clase positiva cuando x_j aumenta en 1.

En contextos de alta dimensionalidad (gran número de variables explicativas) o con multicolinealidad (variables explicativas muy correlacionadas), como es nuestro caso, la regresión logística es especialmente vulnerable al sobreajuste (*overfitting*, en inglés). Éste último ocurre cuando los modelos memorizan el ruido sin centrarse en los patrones generales, logrando así ajustarse de forma demasiado precisa a los datos y no siendo eficientes generalizando. Para evitar esto se pueden incorporar términos de regularización que buscan aumentar la capacidad de generalización. Las formas más utilizadas son:

- **L2 (Ridge)**: penaliza la suma de los cuadrados de los coeficientes, reduciendo su magnitud de forma uniforme sin eliminarlos por completo.
- **L1 (Lasso)**: penaliza la suma de los valores absolutos de los coeficientes, lo que tiende a llevar muchos de ellos a valer cero, produciendo que el modelo dependa solo de un subconjunto reducido de variables.
- **ElasticNet**: combinación ponderada de L2 y L1. Por defecto la contribución de cada uno es 0,5.

La regresión logística destaca por ser fácilmente interpretable, contar con gran solidez estadística y bajo costo computacional. Sin embargo, al asumir que las variables son independientes entre sí puede resultar algo restrictivo para sistemas complejos y no lineales, como los sistemas biológicos.

La regresión logística se implementó utilizando scikit-learn (v1.2.2). Se aplicó un máximo de iteraciones de 2000, y se utilizó L2 como regularizador, luego de dar mejores resultados por sobre L1 y Elastic Net.

3.3.3 Random Forest

Se trata de un modelo que se ha utilizado ampliamente en bioinformática, genómica y biología de sistemas debido a su alta capacidad para manejar grandes volúmenes de datos ruidosos (41). El principio fundamental de este modelo de aprendizaje supervisado consiste en construir un conjunto de modelos débiles (los árboles de decisión), altamente variables pero poco sesgados, y combinarlos para obtener un predictor robusto.

El árbol de decisión es una estructura jerárquica que divide recursivamente las variables explicativas en dos subgrupos mediante reglas del tipo $x_j \leq c$, siendo x_j la j -ésima variable y c un valor umbral aprendido a partir de los datos mismos. La elección de dicho corte se basa en maximizar una medida de pureza, la cual cuantifica en qué medida las observaciones que llegan a un nodo pertenecen predominantemente a una única clase. Finalizado este proceso, los nodos terminales (también llamados hojas) asignan una clase a cada subgrupo en función de la probabilidad estimada de pertenencia. La clasificación final de una nueva observación se determina mediante votación mayoritaria entre las hojas, asignando la clase más frecuente entre las hojas alcanzadas durante el recorrido del árbol (**Figura 15**).

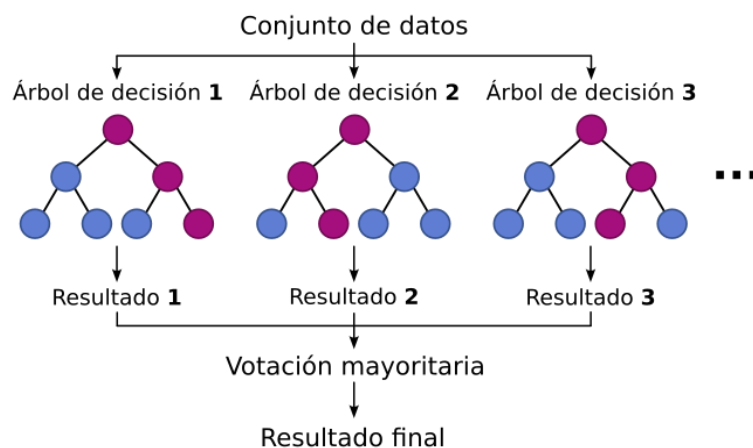


Figura 15. Esquema generalizado representando el proceso de clasificación realizado por *Random Forest*.

Si bien los árboles de decisión presentan diversas ventajas, al mismo tiempo son inherentemente inestables: pequeñas fluctuaciones en el conjunto de datos puede producir árboles significativamente diferentes. Para reducir la varianza se aplican dos mecanismos de aleatorización:

- *Bagging (Bootstrap Aggregating)*: Cada árbol no se entrena sobre el conjunto de datos total si no sobre una muestra *bootstrap* del conjunto de datos original, obtenida mediante muestreo con reemplazo.
- Selección aleatoria de variables: En cada nodo del árbol, la mejor partición se elige considerando únicamente un subconjunto aleatorio de variables explicativas, en lugar del conjunto completo.

Estos dos procesos disminuyen la correlación entre árboles, resultando crucial para la efectividad del ensamble de los mismos.

Para cuantificar la contribución de cada variable al modelo, se empleó la importancia basada en la reducción de impureza. El índice de pureza mide qué tan "mezcladas" están las clases en un nodo de un árbol de decisión. Un nodo es puro cuando todos sus ejemplos pertenecen a la misma clase. En cada nodo de los árboles de decisión, la división se selecciona de forma que maximice la disminución de, en nuestro caso, el índice de Gini. Éste mide la probabilidad de clasificar incorrectamente una variable elegida al azar si se asigna una clase según la distribución del nodo. Se define el índice de la forma siguiente:

$$Gini = 1 - \sum_{k=1}^K p_k^2$$

Siendo K el número de clases (2 en nuestro caso) y p_k la proporción de variables de la clase k en el nodo. Cuanto más se logre minimizar el índice, más puro es el nodo, significando entonces que si el índice es 0 nos encontramos con un nodo puro. El valor máximo que se puede obtener (si las clases están uniformemente distribuidas) es $1 - \frac{1}{K}$.

A partir de este índice se puede calcular la importancia “Gini” de una variable x_j . Ésta está definida como la suma de las reducciones de impureza producidas por todas las divisiones en las que dicha variable fue utilizada, a lo largo de todos los árboles del bosque. Valores elevados de la importancia Gini son indicativos de que la variable fue utilizada con frecuencia para dividir nodos, que la impureza se redujo considerablemente al hacerlo y que la variable contribuye en gran medida a la separación de clases. Nosotros utilizamos esta importancia para evaluar qué genes eran los más relevantes para la clasificación de pacientes.

Resulta ventajoso el uso de *Random Forest* dada su capacidad para modelar relaciones no lineales, su robustez frente al sobreajuste y su buen desempeño en contextos de alta dimensionalidad. No obstante, el modelo presenta ciertas limitaciones en términos de interpretabilidad global y puede mostrar sesgos en la estimación de la importancia de las variables.

Se utilizó nuevamente scikit-learn (v1.2.2) para generar el modelo de *Random Forest*, y el mismo se corrió con un número de árboles igual a 100.

3.3.4 XGBoost (eXtreme Gradient Boosting)

La técnica de potenciación del gradiente (*gradient boosting*, en inglés) busca ensamblar predictores débiles, construyendo modelos de forma escalonada y secuencial con el objetivo de aumentar la eficiencia de las clasificaciones. XGBoost es una implementación de ésta técnica con los árboles de decisión (42).

A diferencia del *Random Forest*, los árboles en XGBoost no se entrenan independientemente del resto, se construyen mediante un enfoque aditivo: se crean nuevos árboles de forma secuencial considerando el error de clasificadores anteriores (**Figura 16**). El algoritmo calcula los residuos de predicciones de cada árbol para determinar cuán alejadas estuvieron del modelo real. Los valores residuales expresan la diferencia entre los valores predichos y los reales.

Uno de los puntos críticos en este algoritmo es la definición de su función objetivo, la cual está conformada por un término que da cuenta de la pérdida, y otro que es el término de regularización. Para definirla primero debemos comprender que, dado un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$, la predicción de cada árbol de decisión estará dada por:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i),$$

siendo f_k un árbol de decisión del K total de árboles.

Por lo tanto, la función objetivo quedará definida como:

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Siendo $l(\cdot)$ la función de pérdida (por ejemplo, pérdida logística o cuadrática; en nuestro caso se utilizó la cuadrática) y $\Omega(f)$ el término de penalización, que se haya definido por la complejidad del árbol. De este modo se logra penalizar modelos excesivamente complejos y evitar el sobreajuste.

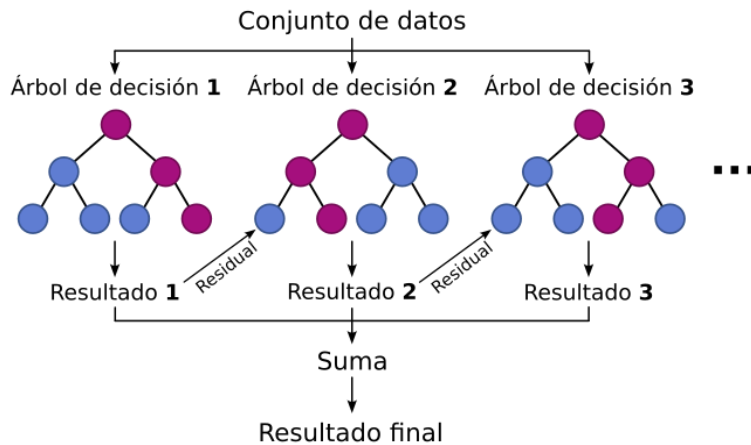


Figura 16. Esquema generalizado representando el proceso de clasificación realizado por XGBoost.

Además de esta estrategia, se incorporan diversos mecanismos para mantener el sobreajuste controlado, presentados como hiperparámetros que pueden ajustarse para lograr una mejor clasificación de los datos. Por último, al igual que con *Random Forest*, se puede tomar como medida de relevancia para las variables la importancia de Gini.

XGBoost ha sido ampliamente utilizado en investigaciones y aplicaciones industriales debido a su alto desempeño predictivo, su flexibilidad para modelar relaciones no lineales complejas y su capacidad para manejar grandes volúmenes de datos. No obstante, elevar la complejidad del modelo va en detrimento de su interpretabilidad.

XGBoost se implementó en este trabajo utilizando la biblioteca XGBoost para Python v1.7.6. El mejor rendimiento del modelo se dió con un número de árboles igual a 100 y una tasa de aprendizaje (*learning rate*, en inglés) igual a 0,01.

3.4 Cálculo del *ranking* de los genes según su importancia en la clasificación

Para obtener una estimación robusta de la importancia de los genes en la tarea de clasificación, se calculó un *ranking* ponderado para cada gen en las 10 corridas. En el caso de RL, RF y XGBoost, para cada uno de los 10 *folds*, los genes se rankearon según sus valores de importancia, asignando al gen más importante un valor de 1, al siguiente más importante un valor de 2, y así sucesivamente.

Sea $r_g^{(i,j)}$ la posición en el *ranking* del gen g en el j -ésimo *fold* de la i -ésima corrida. La posición media $\overline{r}_g$ para el gen g se calculó como:

$$\overline{r}_g = \frac{1}{10 \cdot 10} \sum_{i=1}^{10} \sum_{j=1}^{10} r_g^{(i,j)}$$

Para el modelo GCNN-GLRP, dado que los valores de relevancia se calcularon para cada gen en las m muestras (pacientes) en el subconjunto de testeo, la fórmula fue ligeramente adaptada. En este caso, $r_g^{(i,j)}$ denota la posición en el *ranking* del gen g en la j -ésima muestra de la i -ésima corrida. La posición media $\overline{r}_g$ para el gen g se calculó entonces como:

$$\overline{r}_g = \frac{1}{10 \cdot m} \sum_{i=1}^{10} \sum_{j=1}^m r_g^{(i,j)}$$

3.5 Expresión diferencial

Para el análisis de la expresión diferencial de genes en pacientes con ELA y controles sanos, se evaluaron los dos conjuntos de datos (ALS741 y ALS301) por separado. Para ello se utilizó el paquete Limma (43) (v 3.64.0) en R, y dado que identificar variaciones en las expresiones génicas determinadas con *microarray* no es trivial, se incluyó el método TREAT (44). Este último evalúa si la expresión diferencial de los genes supera un umbral predefinido de significancia biológica, permitiendo identificar si el cambio de expresión entre los dos grupos es lo suficientemente considerable como para definirlo biológicamente significativo.

Una vez realizado este análisis, se seleccionaron los genes con un p-valor ajustado menor a 0,05 y un *fold-change* absoluto de 0,1, ya que una vez utilizado TREAT, se debe considerar uno pequeño. El *fold-change* se encuentra definido como:

$$FC(x) = \frac{x}{x_0},$$

siendo x el valor observado y x_0 el valor de referencia.

3.6 Métricas de centralidad

Como parte de nuestro análisis, se propuso comparar la importancia de los genes relevantes identificados por GCNN y RL en las redes PPI. Con este fin, se calculó la centralidad de grado (*degree*, en inglés), de intermediación (*betweenness*, en inglés) y de cercanía (*closeness*, en inglés) para todos los nodos en cada una de las tres redes PPI utilizadas en el pipeline. En la **Figura 17** se muestran esquemas de los tres tipos de centralidad.

La centralidad de grado refleja cuántos vecinos directos tiene un nodo; la centralidad de intermediación mide la frecuencia con la que un nodo se encuentra en los caminos más cortos entre otros nodos; y la centralidad de cercanía indica cuán cerca está un nodo de todos los demás en la red, lo que indica su potencial para interactuar rápidamente con todo el sistema.

Se seleccionaron estas métricas porque se encuentran entre las medidas más relevantes para analizar la importancia de los nodos dentro de la estructura interna de una red. Permiten considerar no sólo su nivel de conectividad, sino también el grado en que actúan como intermediarios en la red y la rapidez con la que podrían propagar fenómenos que ocurran en ella.

Estos cálculos se realizaron utilizando NetworkX (versión 3.1) (45) en Python (versión 3.11.6).

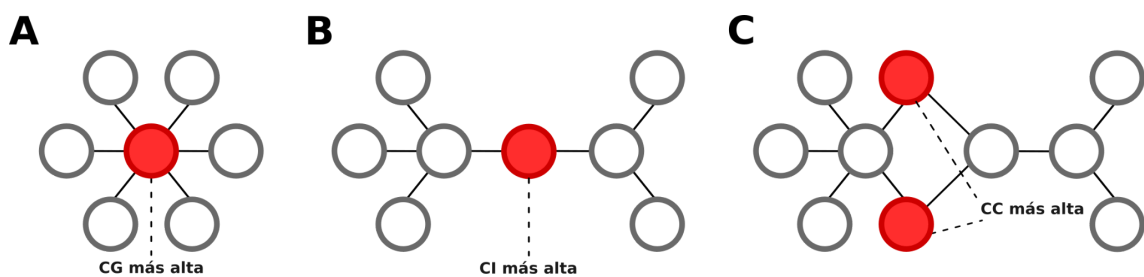


Figura 17. Esquemas representando redes en las cuales los nodos rojos son los que se hallan con mayor centralidad A. de grado (CG), B. de intermediación (CI) y C. de cercanía (CC).

3.7 Análisis de enriquecimiento de vías biológicas

Para analizar si procesos biológicos específicos estaban enriquecidos en los genes identificados como relevantes para la tarea de clasificación, se empleó la herramienta online Enrichr (46) y se tomaron los datos generados por la base de datos Reactome (47).

Para cada una de las redes y modelos, se seleccionaron los 50 genes más relevantes en cada corrida y se determinó cuáles genes se encuentran en ese listado en las 10 corridas. Éstos últimos se utilizan como dato de entrada para el análisis de enriquecimiento en vías biológicas. El conjunto de genes *background*, es decir, todos aquellos genes que podrían haber sido positivos, consistió en todos los genes de la red PPI utilizada durante el entrenamiento de los modelos.

Para limitar la búsqueda y centrarnos en resultados específicos, sólo se consideraron las vías con un P-valor ajustado $\leq 0,01$ y además, se mantuvieron sólo aquellas vías que contenían no más de 100 genes (48). Adicionalmente, se conservaron solamente las vías ‘hojas’ en la jerarquía (es decir, las más específicas), evitando así la inclusión de vías más amplias que incluyeran múltiples procesos (49).

Como paso final, se realizó una revisión bibliográfica exhaustiva con el fin de evaluar y contextualizar la relevancia biológica de las vías enriquecidas en relación con las enfermedades de interés (50).



Resultados



4.1 El Análisis de Componentes Principales no distingue a los pacientes con ELA de los controles sanos

Como una primera aproximación, se calcularon los componentes principales para ambos conjuntos de datos: ALS741 (utilizado para entrenar y testear los modelos) y ALS301 (utilizado para la validación independiente). El objetivo fue determinar si su variabilidad era suficiente para distinguir entre controles sanos y casos de ELA.

Como primer paso comparamos los dos conjuntos de datos. Al considerar los dos primeros componentes principales, se encontró que no están claramente separados, si bien las muestras de ALS301 están más agrupadas, gran parte de estas se solapan con las muestras de ALS741 (**Figura 18 A**).

Este comportamiento podría atribuirse a diferencias en las versiones del *beadchip* de expresión Illumina HumanHT-12 utilizado para medir la expresión génica o a variaciones en la forma en que se ejecutaron los protocolos en ambos casos.

En cuanto al estado de los pacientes, los dos primeros componentes principales no permiten distinguir entre personas con y sin ELA, ya que ambos grupos (en ambos conjuntos de datos) están sumamente mezclados (**Figura 18 B**).

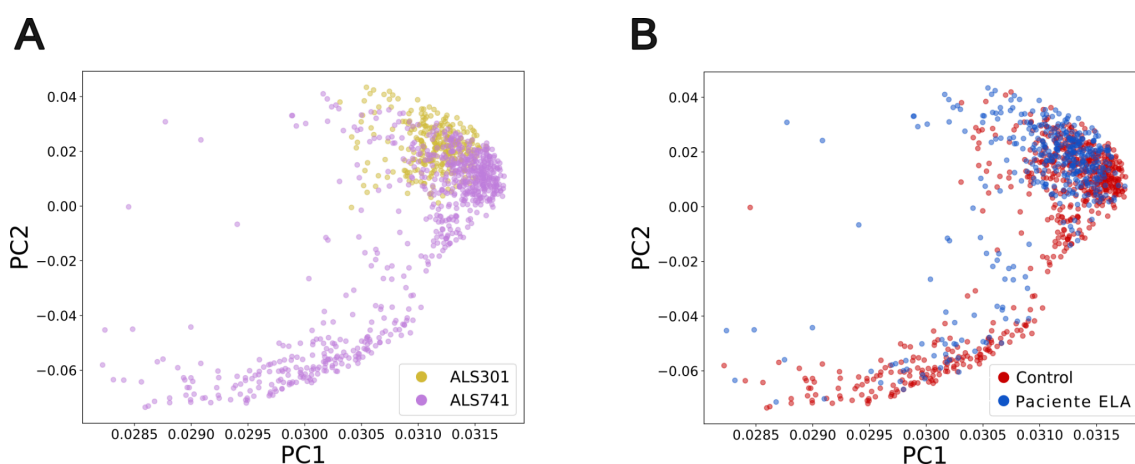


Figura 18. Gráficos de los dos primeros componentes principales de los datos de ELA. Cada punto representa a un paciente, coloreado según **A.** el conjunto de datos y **B.** el estatus del paciente.

4.2 Ciertos genes presentan expresión génica diferencial entre grupos

A continuación se analizó la expresión génica entre grupos. Se encontraron 831 genes expresados diferencialmente (DEGs, del inglés *Differentially Expressed Genes*) entre casos de ELA y controles sanos en el conjunto ALS741, siendo 544 *up*-regulados (expresión más elevada en casos de ELA respecto a los controles) y 287 *down*-regulados (expresión más baja) (**Figura 19 A**). El conjunto de datos ALS301, contiene 60 genes *up*-regulados y 32 *down*-regulados, lo que da un total de 92 DEGs (**Figura 19 B**). Las coincidencias entre los conjuntos de datos son de 12 y 38 genes *down* y *up*-regulados, respectivamente. (**Figura 20 A y B**, respectivamente).

Las diferencias en los números e identidades de los DEGs entre conjuntos podrían deberse a las mejoras en el rendimiento técnico y las capacidades de análisis de datos de la plataforma de *microarray* utilizada para ALS301 que, en comparación con la utilizada para ALS741, se trata de una versión más nueva y optimizada. Se ha estudiado que diferentes plataformas pueden devolver conjuntos de DEGs distintos, aunque biológicamente resulten complementarios (51).

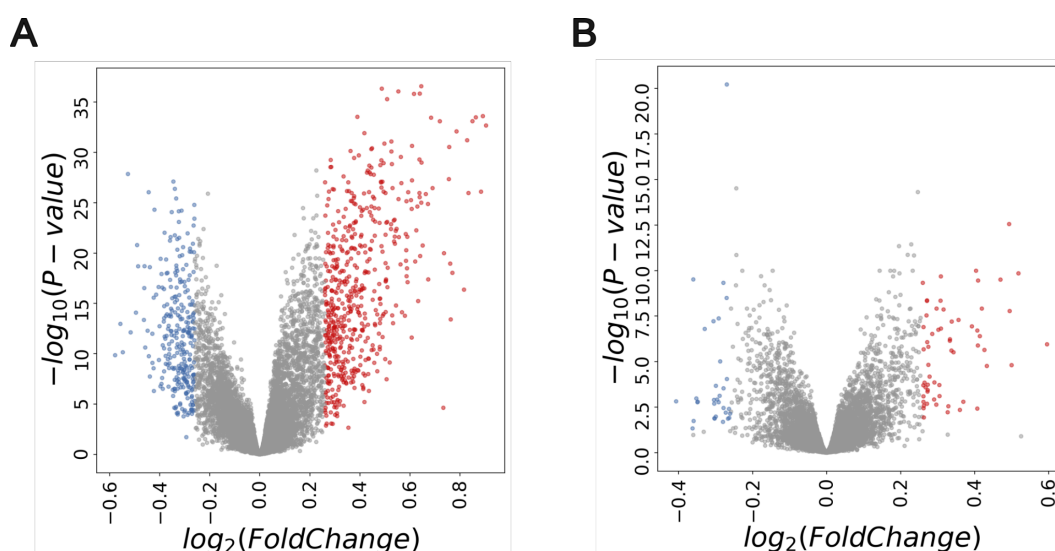


Figura 19. Volcano plot con los genes diferencialmente expresados entre individuos con y sin ELA. **A.** ELA741 **B.** ELA301. Ambas cohortes contienen el mismo número de genes. En azul se muestran los genes *down*-regulados y en rojo los *up*-regulados.

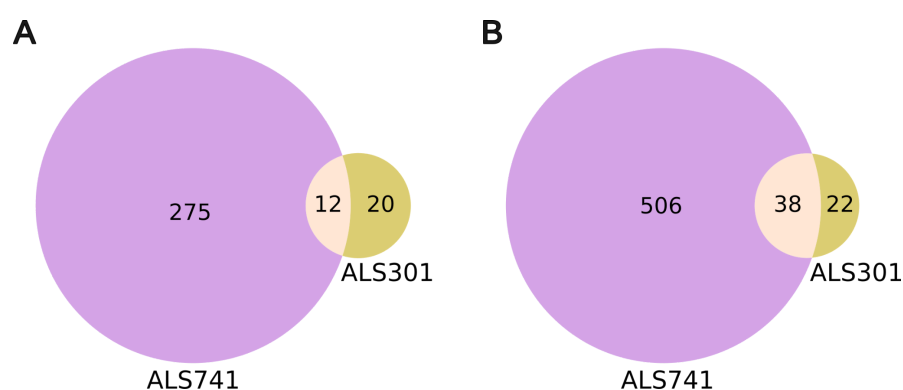


Figura 20. Diagrama de Venn mostrando el solapamiento de los genes diferencialmente expresados entre individuos con y sin ELA, tanto en ALS741 como en ALS301. En **A.** se muestran los *down*-regulados y en **B.** los *up*-regulados.

4.3 Los modelos GCNN y de Regresión Logística son clasificadores eficientes para casos y controles

Se intentó clasificar a los pacientes con un modelo de aprendizaje profundo (GCNN, *Graph Convolutional Neural Network*) y tres modelos más simples de ML (Regresión Logística, *Random Forest* y XGBoost). Los modelos más simples sólo utilizan los perfiles de expresión génica como dato de entrada, mientras que GCNN también utiliza una red PPI para la tarea de clasificación. Los modelos se entrenaron hasta llegar a convergencia, y el rendimiento se optimizó mediante el ajuste de hiperparámetros, tomando como punto de evaluación los valores de AUC-ROC.

Durante el testeo los modelos son capaces de discriminar entre casos de ELA y controles. Regresión Logística (RL) obtuvo el mayor rendimiento, a la par de GCNN, y se obtuvieron resultados similares con las tres redes PPI (**Tablas 4-6**).

En la **Figura 21** se puede ver que para GCNN la función de pérdida disminuye de forma consistente a lo largo de las iteraciones para las tres redes PPI. Ésto es indicativo de que no hubo indicios directos de sobreajuste. En la misma figura se pueden ver las curvas ROC y de *Accuracy* de los modelos creados también con las tres redes PPI.

La validación mostró nuevamente que RL y GCNN lograron una discriminación precisa en ambas cohortes de pacientes, siendo el primero el que obtuvo un rendimiento ligeramente superior. Por el contrario, RF y XGBoost presentaron un rendimiento inferior en la tarea de clasificación: si bien los valores de AUC resultaron predictivos, las restantes métricas evidenciaron un desempeño considerablemente más bajo.

	IntAct							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0,923	0,873	0,782	0,708	0,763	0,701	0,738	0,774
RL	0,951	0,893	0,828	0,783	0,815	0,721	0,786	0,945
RF	0,875	0,824	0,667	0,542	0,681	0,601	0,557	0,457
XGBoost	0,865	0,824	0,698	0,639	0,705	0,643	0,643	0,576

Tabla 4. Rendimiento de clasificación de pacientes con ELA con los modelos GCNN, RL, RF y XGBoost luego del testeo y la validación independiente, utilizando datos de IntAct como red PPI. En negrita se muestran los valores más altos comparando entre modelos.

	BioGRID							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0,924	0,867	0,787	0,708	0,752	0,694	0,760	0,884
RL	0,950	0,892	0,818	0,766	0,829	0,714	0,789	0,982
RF	0,871	0,824	0,649	0,532	0,675	0,615	0,594	0,518
XGBoost	0,844	0,824	0,684	0,609	0,675	0,605	0,557	0,463

Tabla 5. Rendimiento de clasificación de pacientes con ELA con los modelos GCNN, RL, RF y XGBoost luego del testeo y la validación independiente, utilizando datos de BioGRID como red PPI. En negrita se muestran los valores más altos comparando entre modelos.

	HIPPIE							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0,936	0,887	0,809	0,750	0,749	0,694	0,759	0,905
RL	0,944	0,892	0,819	0,783	0,823	0,718	0,786	0,945
RF	0,875	0,824	0,675	0,565	0,678	0,608	0,576	0,488
XGBoost	0,859	0,824	0,698	0,652	0,699	0,631	0,604	0,524

Tabla 6. Rendimiento de clasificación de pacientes con ELA con los modelos GCNN, RL, RF y XGBoost luego del testeo y la validación independiente, utilizando datos de HIPPIE como red PPI. En negrita se muestran los valores más altos comparando entre modelos.

En base a estos resultados, la incorporación de redes PPI no pareció tener un impacto significativo en el rendimiento predictivo, dado que RL, que no hace uso de dicha información, obtuvo resultados comparables en la tarea de clasificación.

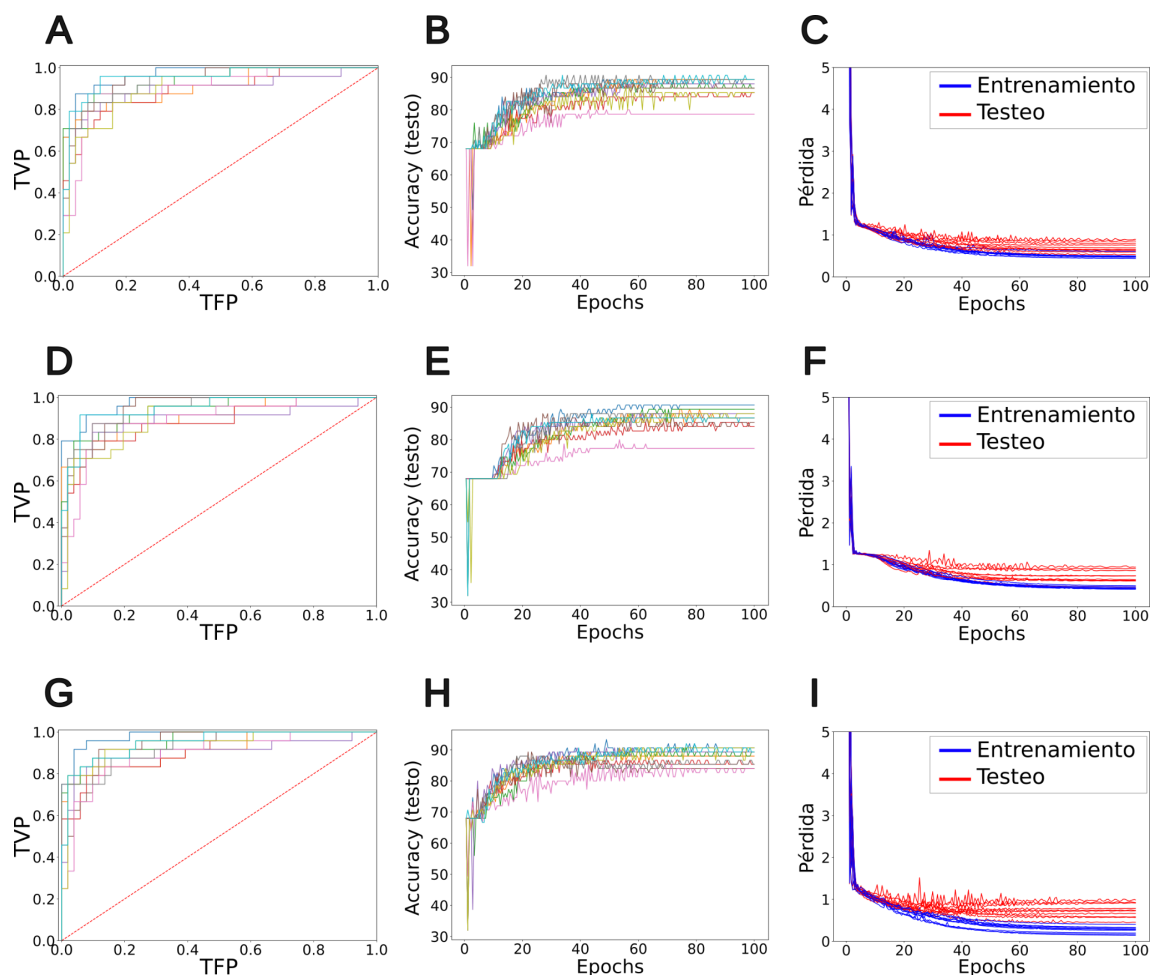


Figura 21. Rendimiento del modelo GCNN en diez corridas utilizando las tres redes PPI. Los paneles A-C corresponden a IntAct, D-F a BioGRID y G-I a HIPPIE. A, D y G muestran las curvas AUC-ROC en el testeo, B, E y H muestran las curvas de *accuracy* en el testeo y C, F e I las curvas de pérdida tanto en el entrenamiento como en el testeo.

4.4 Los genes predictivos más relevantes muestran una presencia estable tanto en GCNN-GLRP como en la Regresión Logística.

La obtención de modelos de ML capaces de clasificar correctamente pacientes con y sin ELA, nos permitió continuar con la identificación de qué variables (genes) fueron las más relevantes para la clasificación.

Para el modelo RL, la importancia de los genes se basó en los valores absolutos de los coeficientes obtenidos del modelo entrenado. Para RF y XGBoost, recuperamos los valores de importancia de Gini calculados durante el entrenamiento, y por último, para el modelo GCNN, adoptamos GLRP como se describe en (28) para calcular el valor de relevancia de las variables.

Utilizando los *ranking* ponderados calculados como se describe en la sección 3.4 de Materiales y Métodos, evaluamos cuán consistentes fueron los resultados a lo largo de las 10 corridas. La **figura 22** muestra el número de genes coincidentes en todas las posibles combinaciones sin repetición de las 10 corridas de cada modelo. Se consideran como punto base los 50 y 300 genes más relevantes. Se observa que las tres redes de PPI dan resultados similares. Se observa que tanto los modelos GCNN-GLRP como RL muestran resultados estables a lo largo de las 10 corridas. Por el contrario, RF y XGBoost muestran un solapamiento mínimo entre corridas cuando se consideran los 50 genes principales. Cuando se consideran los 300 genes principales, sus coincidencias aumentan, pero GCNN y RL siguen siendo más estables.

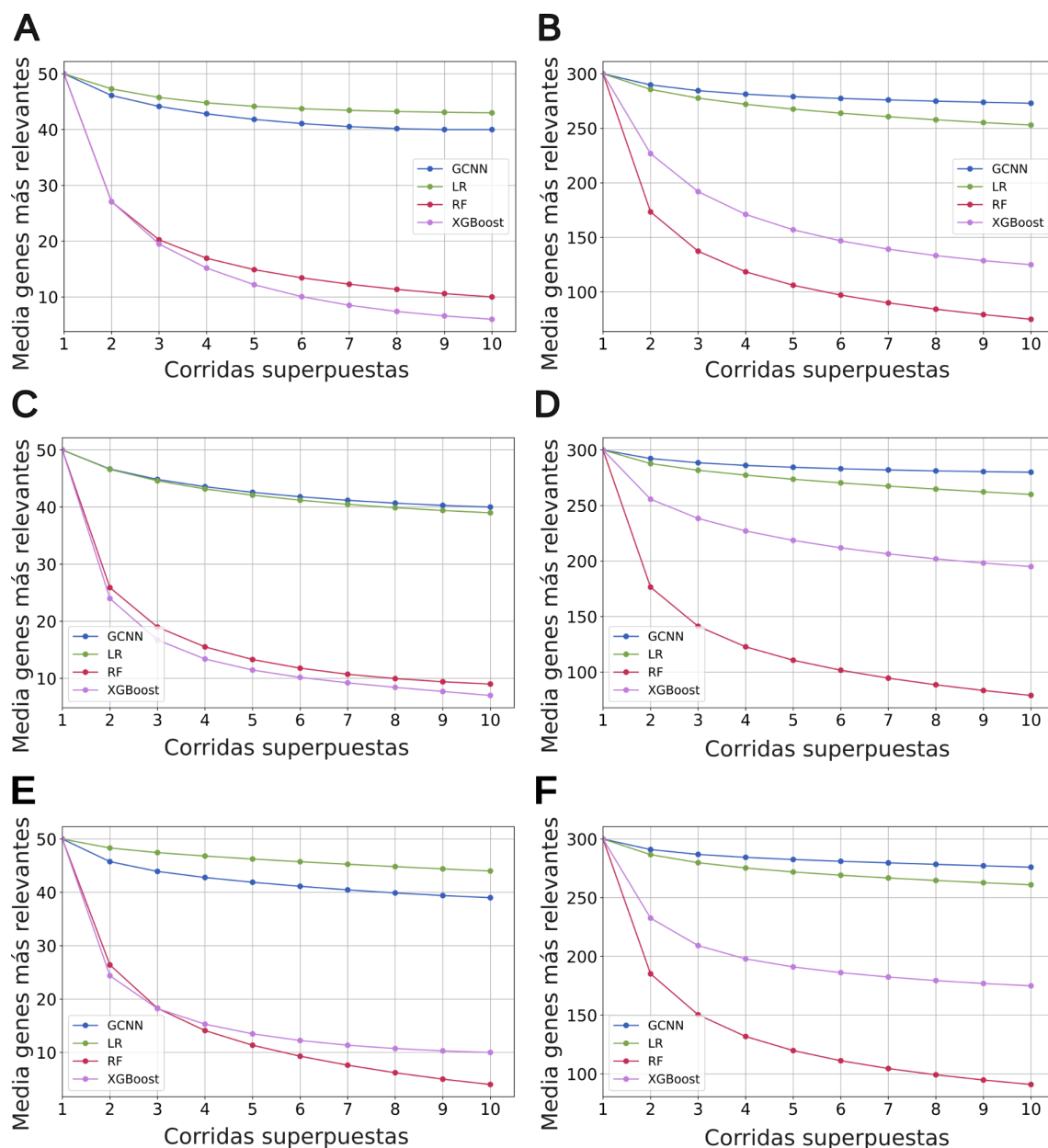


Figura 22. Número medio de genes relevantes comunes en las 10 corridas Número medio de genes más relevantes comunes en combinaciones sin repeticiones de 10 corridas utilizando las tres redes PPI: IntAct (A y B), HIPPIE (C y D) y BioGRID (E y F). En el gráfico, el eje x denota el número de corridas superpuestas, cuyo número de genes comunes se promedió con todas las combinaciones de corridas posibles. En A, C y E se consideraron como punto de partida los 50 genes más relevantes de cada corrida; en B, D y F los 300 más relevantes.

4.5 Variabilidad de los genes más relevantes entre modelos y redes PPI

A partir de este punto, nos centramos en analizar los genes que aparecen siempre entre los 50 más relevantes en las 10 corridas, de aquí en adelante denominados como los "genes más relevantes". Con cada una de las redes PPI, recuperamos aproximadamente 40 genes relevantes tanto con GCNN-GLRP como con RL. La **Tabla 7** muestra el número final en cada caso y en el material suplementario, **sección 10.2** se muestra el listado de genes. Sorprendentemente, considerando las tres redes PPI, el solapamiento de genes entre los dos modelos fue nulo, sugiriendo que, aunque ambos algoritmos clasifican las cohortes con éxito, se basan en información diferente para hacerlo.

	Superposición de los 50 genes más relevantes				Superposición de los 300 genes más relevantes			
	GCNN	RL	RF	XGBoost	GCNN	RL	RF	XGBoost
IntAct	40	43	10	6	273	253	75	125
BioGRID	39	44	4	10	276	261	91	175
HIPPIE	40	39	9	7	280	260	79	195

Tabla 7. Número de genes más relevantes comunes hallados en 10 corridas de los modelos GCNN, RL, RF, and XGBoost, usando IntAct, BioGRID, y HIPPIE como redes PPI. Se consideraron como punto de partida los 50 genes más relevantes de cada corrida; y luego los 300 más relevantes.

Por otro lado, también se determinó si los resultados obtenidos con las tres redes tienen genes relevantes en común. Se encontraron 15 genes relevantes compartidos entre las mismas cuando se corrió GCNN-GLRP, y 18 con RL (**Figura 23 A y B**). Esto representa aproximadamente el 38% de cada conjunto de genes relevantes obtenidos con cada red PPI.

Dado que las redes presentan diferencias sustanciales en los nodos y aristas que las componen, este número podría ser una indicación de que, en todos los casos, los modelos están recuperando información común. Es importante señalar que algunos genes identificados como relevantes al utilizar una red en particular, pueden no estar presentes en las otras, bien porque sus interacciones no alcanzan

el umbral de puntuación de confianza aplicado en este estudio y/o porque no forman parte del mayor componente conectado necesario para ejecutar el modelo GCNN.

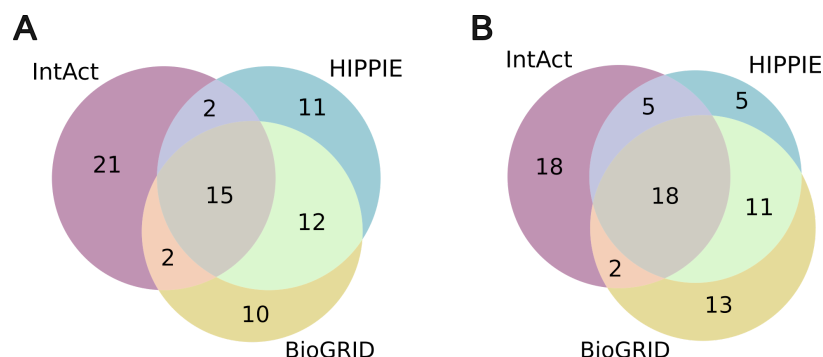


Figura 23. Diagrama de Venn que muestra el solapamiento de los genes más relevantes identificados utilizando los datos de las tres redes PPI en **A.** GCNN y **B.** RL.

También se analizó si el conjunto de genes relevantes hallados por GCNN y RL coinciden con los genes diferencialmente expresados. Hallamos que dentro de los relevantes según la RL, el porcentaje de DEGs es mayor que en el caso de los de GCNN (**Figura 24 A y B**). Esto podría deberse a que la RL solo tiene acceso a la expresión génica, por tanto es de esperar que priorice genes con expresiones diferentes entre pacientes sanos y enfermos. En el caso de GCNN este número es menor, indicando que no solo la expresión como tal sino también los patrones de señal y las conexiones de la red PPI son relevantes para determinar qué genes son los más importantes para clasificar las cohortes.

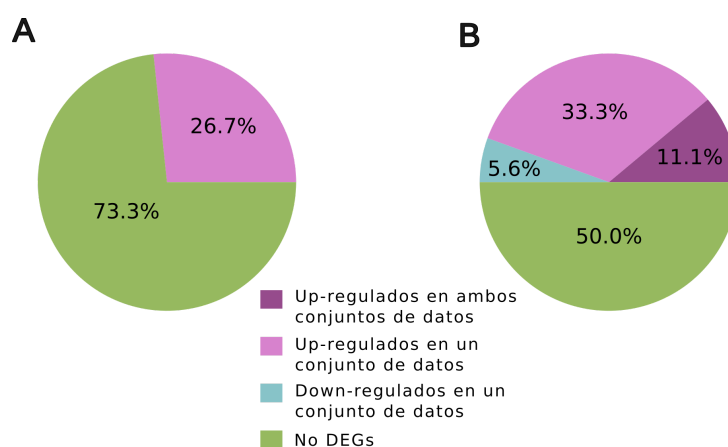


Figura 24. Gráficos de torta con el porcentaje de DEGs encontrados en los genes más relevantes comunes en las tres redes PPI obtenidos con **A.** GCNN y **B.** RL.

4.6 Los genes más relevantes obtenidos con GCNN muestran una mayor centralidad en todas las redes PPI en comparación con los de la Regresión Logística

Con el objetivo de evaluar la importancia y el nivel de conectividad de los genes más relevantes en las redes PPI, identificados tanto por GCNN como por RL, se calcularon las métricas de centralidad de grado, de intermediación y de cercanía. (Materiales y Métodos, sección 3.6). Estas métricas se seleccionaron para esta tarea ya que captan aspectos complementarios de la importancia topológica de un nodo. Como se muestra en la **Figura 25 A-C**, los genes identificados por GCNN muestran una mayor centralidad de grado, de intermediación y de cercanía en las tres redes, en comparación con los genes obtenidos con la RL.

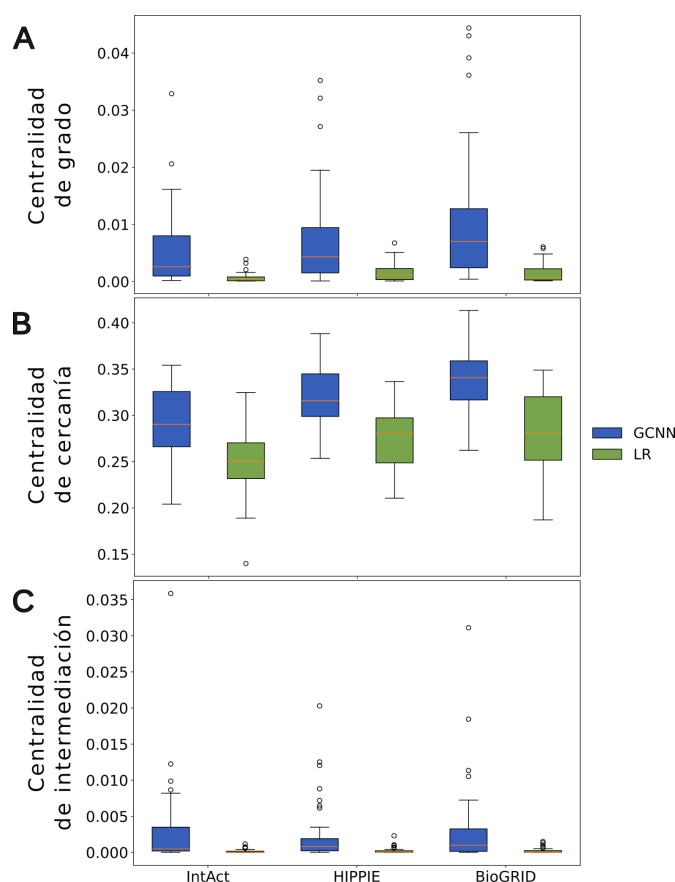


Figura 25. Puntuaciones de centralidad para las redes PPI más relevantes Puntuaciones de centralidad para los genes más relevantes encontrados con las redes IntAct, HIPPIE y BioGRID en los modelos GCNN y RL. **A.** Centralidad de grado **B.** Centralidad de proximidad **C.** Centralidad de distancia. Realizamos la prueba de Wilcoxon-Mann-Whitney entre modelos para cada PPI y puntuación de centralidad, siendo todos los resultados significativos (P-valor < 0,001).

Este patrón podría explicarse considerando que GCNN incorpora tanto los datos de expresión génica como la estructura topológica de las redes PPI durante el entrenamiento del modelo, mientras que RL toma cada gen como independiente de los demás.

4.7 Se identificaron vías enriquecidas en los genes más relevantes identificados por GCNN-GLRP

Uno de nuestros principales objetivos se basó en obtener información biológica relacionando los genes más relevantes con vías biológicas, buscando dilucidar procesos celulares que posiblemente estuvieran involucrados en la patología de la enfermedad. Calculamos entonces el enriquecimiento de vías biológicas en los genes más relevantes obtenidos con GCNN-GLRP y la RL, con cada una de las tres redes.

Se obtuvo que 60 vías están significativamente enriquecidas cuando se utilizan como entrada los 40 genes más relevantes de IntAct, 72 con HIPPIE (40 genes) y 68 con BioGRID (39 genes). Entre ellas, 48 vías se hallan en el listado de los tres casos (**Figura 26**). En la **sección 10.4** del material suplementario se ofrece información detallada sobre estas vías.

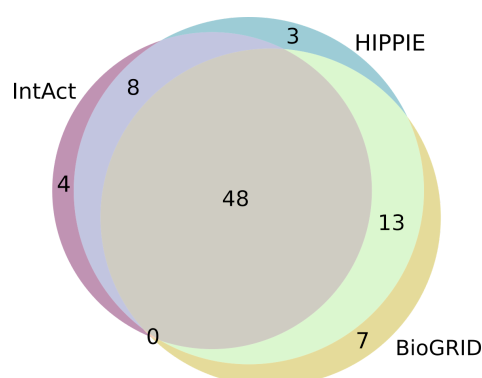


Figura 26. Diagrama de Venn mostrando el número de vías biológicas enriquecidas en genes relevantes obtenidos utilizando las tres redes PPI para clasificar con el modelo GCNN los datos de ELA.

Varias de estas vías están muy relacionadas a procesos biológicos que están alterados en pacientes con ELA. Entre ellas se incluyen la degradación del ARN mensajero mediada por mutación terminadora (en inglés *nonsense-mediated decay*) (16,52), tipos específicos de autofagia (53), ubiquitinación y desubiquitinación (54,55), mecanismos relacionados con el envejecimiento (56), neuroinflamación (57,58), metabolismo del glucógeno (59) y muerte celular (60,61).

Para nuestra sorpresa, al aplicar el mismo enfoque con los genes de RL, no se identificaron vías significativamente enriquecidas.

4.8 Se observan características compartidas entre la ELA y la DFT

La ELA y la DFT se describen a menudo como parte de un espectro de enfermedades, ya que una proporción considerable de pacientes presentan síntomas clínicos coincidentes, disfunción y agregación de TDP-43 y rasgos neuropatológicos distintivos, entre otros (12). Además, ambos trastornos comparten antecedentes genéticos comunes, con genes alterados como C9orf72, TARDBP, SQSTM1, VCP, FUS, TBK1 y CHCHD10 que desempeñan papeles críticos en su patogénesis (62).

Considerando esto, se analizó un nuevo conjunto de datos, esta vez de DFT. El objetivo se centró en explorar tanto posibles diferencias biológicas como puntos en común entre las dos enfermedades. El conjunto de datos seleccionado incluye pacientes diagnosticados con una de cinco variantes específicas de DFT. Los agrupamos en 3 categorías principales para su análisis, basándonos en si estaban afectadas principalmente el comportamiento (BH), las funciones motoras (MT) o las habilidades del lenguaje (LG).

Se corrieron los mismos 4 modelos con dos subconjuntos de datos de DFT como se explica en la sección 3.3 de Materiales y Métodos, para encontrar

particularidades de los pacientes afectados en sus habilidades motrices, que son los más relacionados con la ELA en cuanto a sintomatología. En las **Tablas 8-10** se muestra la mediana de los valores de AUC-ROC, *Accuracy*, F1-score y *Recall* luego del testeo de los modelos, después de 10 corridas utilizando las tres redes PPI: IntAct, BioGRID y HIPPIE.

	IntAct							
	BH-MT				LG-MT			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.696	0.545	0.706	1	0.620	0.542	0.487	0.472
RL	0.586	0.562	0.611	0.611	0.546	0.486	0.485	0.471
RF	0.637	0.606	0.683	0.722	0.533	0.543	0.529	0.500
XGBoost	0.563	0.562	0.600	0.611	0.525	0.514	0.500	0.515

Tabla 8. Rendimiento de clasificación entre pacientes con BH vs MT y entre pacientes con LG vs MT de los modelos GCNN, RL, RF y XGBoost durante el testeo, utilizando datos de IntAct como red PPI.

	BioGRID							
	BH-MT				LG-MT			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0,706	0,576	0,643	0,750	0,602	0,542	0,582	0,639
RL	0,589	0,576	0,603	0,611	0,523	0,500	0,493	0,500
RF	0,670	0,656	0,686	0,722	0,520	0,514	0,500	0,500
XGBoost	0,572	0,545	0,595	0,611	0,507	0,514	0,526	0,556

Tabla 9. Rendimiento de clasificación entre pacientes con BH vs MT y entre pacientes con LG vs MT de los modelos GCNN, RL, RF y XGBoost durante el testeo, utilizando datos de BioGRID como red PPI.

	HIPPIE							
	BH-MT				LG-MT			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0,691	0,652	0,725	0,806	0,554	0,500	0,660	0,972
RL	0,585	0,576	0,597	0,611	0,526	0,486	0,479	0,500
RF	0,642	0,606	0,667	0,706	0,499	0,514	0,485	0,471
XGBoost	0,615	0,585	0,632	0,657	0,523	0,514	0,526	0,556

Tabla 10. Rendimiento de clasificación entre pacientes con BH vs MT y entre pacientes con LG vs MT de los modelos GCNN, RL, RF y XGBoost durante el testeo, utilizando datos de HIPPIE como red PPI.

Se encontró que los modelos tienen un rendimiento predictivo pobre, bastante inferior al rendimiento obtenido en la clasificación del conjunto de datos de ELA y sus controles. Considerando el AUC, GCNN fue el modelo que mejor

funcionó en todos los casos. Como se muestra en las tablas, se obtuvieron resultados similares con las tres redes.

Al analizar la relevancia de los genes para la tarea de clasificación, encontramos que GCNN-GLRP fue de nuevo consistente con los datos DFT, como lo fue con los datos de ELA. No fue así el caso de la RL, que presentó resultados menos consistentes, aunque sí resultó más estable que RF y XGBoost. El número de genes más relevantes en cada caso se muestra en la **Tabla 11** y en el material suplementario **sección 10.3** se halla el listado de genes.

	Superposición de los 50 genes más relevantes							
	GCNN		RL		RF		XGBoost	
	BH-MT	LG-MT	BH-MT	LG-MT	BH-MT	LG-MT	BH-MT	LG-MT
IntAct	45	46	20	22	0	0	0	5
BioGRID	36	46	20	18	0	1	2	1
HIPPIE	38	44	23	21	0	0	16	14

Tabla 11. Número de genes más relevantes comunes hallados en 10 corridas de los modelos GCNN, RL, RF, and XGBoost, usando IntAct, BioGRID, y HIPPIE como redes PPI para clasificar los datos de BH-MT y LG-MT. Se consideraron como punto de partida los 50 genes más relevantes de cada corrida.

En cuanto al número de genes relevantes que coinciden utilizando las tres redes PPI en las 10 corridas de GCNN, se identifican de forma consistente 16 genes cuando el modelo se entrenó con el conjunto de datos BH-MT y 19 cuando se entrenó con el conjunto de datos LG-MT, como se muestra en la **Figura 27 A y B**, respectivamente.

A continuación se analizó el solapamiento de estos genes con los obtenidos con los datos de ELA. Observamos solapamiento entre los genes más relevantes obtenidos entre las tres clasificaciones (ELA, BH-MT y LG-MT), considerando todo gen común a las tres redes PPI. Cuatro genes fueron compartidos entre la ELA y los obtenidos con BH y MT, y seis con los genes obtenidos en la clasificación entre pacientes LG y MT. Los mismos se muestran en la **Tabla 12**. Notablemente, cuatro genes se encuentran representados de manera consistente en las tres

clasificaciones: Ubiquitina B (*UBB*), *Ubiquitin Like Modifier Activating Enzyme 3* (*UBA3*), Vimentina (*VIM*) y Beta-actina (*ACTB*).

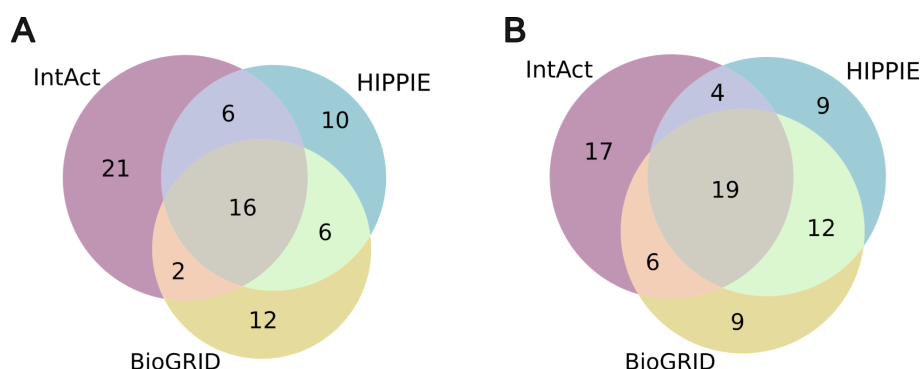


Figura 27. Diagrama de Venn que muestra el solapamiento de los genes más relevantes identificados utilizando los datos de las tres redes PPI cuando se clasificó con GCNN a **A.** BH-MT y **B.** LG-MT.

	BH-MT	LG-MT
UBB	x	x
UBA52	x	x
VIM	x	x
ACTB	x	x
YWHAB		x
PABPC1		x

Tabla 12. Genes más relevantes comunes entre la clasificación con GCNN de pacientes con ELA y ambas clasificaciones de DFT (BH-MT y LG-MT), identificados usando las redes IntAct, BioGRID y HIPPIE.

También repetimos el análisis de enriquecimiento de vías biológicas utilizando los genes más relevantes identificados a partir de las clasificaciones BH-MT y LG-MT con GCNN-GLRP. Al analizar el conjunto de datos BH-MT en las tres redes, se identificaron un total de 41 vías enriquecidas (Material suplementario, **sección 10.5**). En comparación, los resultados de LG-MT revelaron 112 vías enriquecidas en las mismas redes (**Figura 28 A y B**, respectivamente, y Material suplementario, **sección 10.6**).

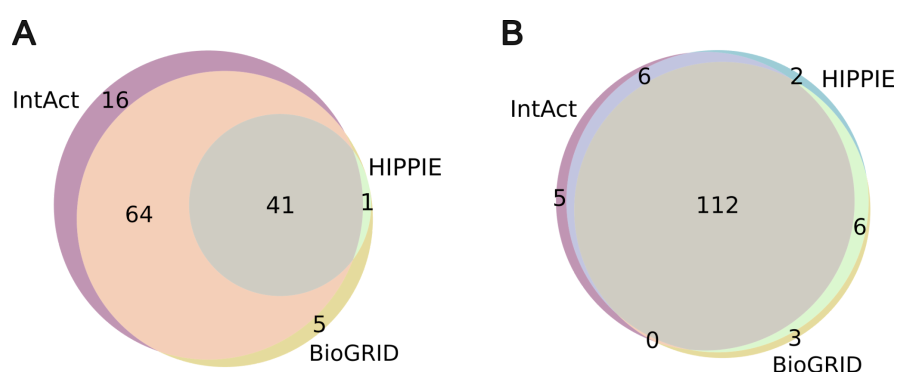


Figura 28. Diagrama de Venn mostrando el número de vías biológicas enriquecidas en genes relevantes obtenidos utilizando las tres redes PPI para clasificar con el modelo GCNN los datos de, **A.** BH-MT y **B.** LG-MT

Cabe destacar que las 41 vías enriquecidas encontradas con los datos BH-MT están incluidas en las 112 vías encontradas con los datos LG-MT, y 28 vías son comunes entre las obtenidas con las tres clasificaciones (ELA, BH-MT y LG-MT), como se muestra en la **figura 29**. Dado que la neurodegeneración es una característica central de ambas enfermedades, nos sorprendió positivamente observar un número considerable de vías compartidas, lo que sugiere un trasfondo biológico común. Algunas de ellas incluyen procesos como la autofagia selectiva, la deubiquitinación, mecanismos relacionados con el envejecimiento, la neuroinflamación, el metabolismo del glucógeno, la muerte celular y diferentes vías de señalización.

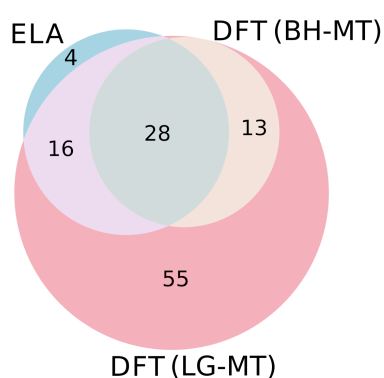


Figura 29. Diagrama de Venn mostrando el número de vías biológicas enriquecidas en genes relevantes obtenidos utilizando las tres redes PPI para clasificar con el modelo GCNN los datos de ELA, BH-MT y LG-MT.

Discusión

La ELA es una enfermedad compleja caracterizada por la pérdida progresiva de neuronas motoras. Aún carece de explicaciones etiológicas concluyentes y de biomarcadores específicos, por lo que el diagnóstico se basa principalmente en los síntomas clínicos. Como en muchos otros casos, años de investigación exhaustiva han llevado a la acumulación de una cantidad sustancial de información; sin embargo, la utilización eficaz de estos datos sigue siendo un reto.

En este estudio, clasificamos con éxito una cohorte independiente de pacientes con ELA y controles sanos, utilizando datos de expresión génica e incorporando a su vez información de redes PPI en un modelo de aprendizaje profundo y lo comparamos con métodos de *machine learning* más sencillos. La regresión logística nos permite una clasificación de forma más rápida, precisa y con menos recursos computacionales. GCNN-GLRP también logra una clasificación con un rendimiento similar y ambos métodos resultan más eficientes en comparación con *Random Forest* y XGBoost.

La principal diferencia observada entre los resultados de RL y GCNN-GLRP radica en los genes identificados como relevantes para la tarea definida. RL trata cada gen como independiente de los demás y, por lo tanto, sólo considera las diferencias en la expresión génica entre cohortes para su clasificación. Esto hace que se reconozcan como relevantes genes aislados. En cambio, GCNN-GLRP integra la expresión génica con los datos de redes de interacción entre proteínas. Por lo tanto, es razonable pensar que los genes identificados como relevantes por este método presentan correlación con procesos biológicos.

Esto se hizo más evidente al analizar la centralidad de los genes. Aquellos obtenidos por GCNN obtuvieron valores de centralidad más elevados en la red PPI, lo que indica que la conectividad no es trivial. También fue evidente en el análisis de enriquecimiento de vías: se encontraron numerosas vías significativamente enriquecidas en estos genes, mientras que no se identificó ninguna utilizando genes seleccionados por RL.

Varias de las vías enriquecidas identificadas utilizando los genes más relevantes de GCNN-GLRP, están estrechamente asociadas con procesos biológicos

bien estudiados en la ELA. Uno de ellos es el *nonsense-mediated decay*, un mecanismo de ‘vigilancia’ que se activa en presencia de ARNm anormales. Las mutaciones en TDP-43, presentes en aproximadamente el 90% de los casos de ELA, afectan su función normal y conducen a la acumulación de transcritos defectuosos que normalmente son el objetivo final de esta vía. TDP-43 también desempeña un rol fundamental en la regulación de la traducción: y la interrupción de la traducción local del ARNm en los axones se ha descrito como un factor decisivo en la progresión de la ELA (63). Asimismo, se observó de manera consistente un enriquecimiento significativo en diversas vías biológicas involucradas en la traducción y en la regulación de la estabilidad del ARNm dentro de nuestros resultados.

Otro rasgo distintivo de la enfermedad es la agregación de la forma aberrante, poli-ubiquitinada e hiperfosforilada de TDP-43 en el citoplasma. Esta proteína, que normalmente debería estar presente en el núcleo, se deslocaliza y su forma anormal y modificada postraduccionalmente sufre agregación. Este fenómeno está relacionado con la autofagia selectiva, como la agrefagia, la autofagia mediada por chaperonas y la microautofagia endosomal tardía (53). Estas vías se ven activadas cuando la acumulación de proteínas mal plegadas supera la capacidad de la célula para gestionarlas y requiere una eliminación de las mismas a gran escala. Las proteínas mal plegadas y agregadas se ubiquitinan, marcándose para su degradación (64). En particular, entre las vías enriquecidas aparecieron vías específicas directamente implicadas en la desubiquitinación. Este hallazgo es consistente con lo descrito en Dobson-Stone et. al (2020) donde una mutación con ganancia de función en la *CYLD lysine 63 deubiquitinase* (gen CYLD), descrita como un importante contribuyente en el desarrollo de la ELA y la DFT (54) en neuronas primarias de ratón, conduce a un aumento de la actividad deubiquitinasa, autofagia interrumpida, y deslocalización de TDP-43.

El estrés celular y el envejecimiento constituyen otros grandes procesos que subyacen a la fisiopatología de la ELA y otras enfermedades neurodegenerativas. Se ha descrito la presencia de estrés oxidativo en la enfermedad, niveles elevados de especies reactivas de oxígeno (ROS) y daño inducido por ROS (65). Entre las vías

enriquecidas identificamos la regulación de BACH1, un gen clave en el control de la producción de ROS. Otra vía afectada es el 'Fenotipo Secretor Asociado a la Senescencia' (SASP, del inglés *Senescence-Associated Secretory Phenotype*), en concordancia con el papel de la senescencia celular (un arresto irreversible del ciclo celular estrechamente vinculado al envejecimiento) en la ELA. Dado que la ELA es una enfermedad estrechamente relacionada con el avance de la edad, no es sorprendente que este proceso se haya descrito en muchos casos (66). Otro grupo de mecanismos relacionados con la edad, incluida la reparación del ADN (67), y varias vías asociadas, también se identificaron en nuestros resultados como enriquecidos.

Obtuvimos algunas vías enriquecidas en el conjunto de genes relevantes que están implicadas en la muerte celular programada. Dos de ellas ("p75NTR recluta complejos de señalización" y "NF- κ B se activa y da señales de supervivencia") desempeñan papeles directos en la supervivencia o muerte neuronal, dependiendo de su estado de activación (68–70). Esto es particularmente importante, dado que la ELA se caracteriza principalmente por la rápida degeneración de las neuronas motoras. Además, ambas vías están directamente relacionadas con la función inmunitaria y se encuentran reguladas por receptores tipo Toll (TLR) (56, 57). Esto es digno de mención, ya que varias vías intrínsecas a las cascadas TLR y la señalización de citoquinas también se identificaron como enriquecidas en nuestros resultados. Por ejemplo, la estimulación crónica de los TLR por agregados de proteínas provoca la activación proinflamatoria de la microglía y los astrocitos en el sistema nervioso central, lo que resulta en la liberación de citoquinas que exacerban el daño neuronal (71,72).

Por el contrario, ciertas citoquinas, como el factor de necrosis tumoral (TNF), pueden ejercer efectos beneficiosos al activar las vías de supervivencia en las neuronas motoras, contribuyendo así a la neuroprotección (73). Adicionalmente, los niveles de varios componentes del sistema inmune periférico, incluyendo linfocitos T, monocitos y citoquinas, experimentan cambios significativos a medida que la enfermedad progresa, lo que sugiere ajustes dinámicos en las respuestas inmunes que pueden modular la patología de la ELA (74).

El papel del sistema inmune en la ELA es complejo, e implica tanto mecanismos neuroprotectores como neurotóxicos que requerirían continuar la investigación para determinar su relación.

Por lo tanto, es importante señalar que nuestros resultados no especifican si estas vías biológicas están reguladas positiva o negativamente, pero sí muestran que los genes asociados y sus interacciones en la red PPI fueron fundamentales para distinguir a los pacientes con ELA de los controles sanos.

También identificamos puntos comunes entre la ELA y las variantes motoras de la DFT utilizando GCNN-GLRP, lo que no fue posible con otros modelos de ML. La mayoría de las vías biológicas enriquecidas entre los genes más relevantes en las tres clasificaciones son cruciales para la neurodegeneración y distintos mecanismos relacionados con el envejecimiento.

En particular, entre el subconjunto de genes que se identificaron como relevantes en los tres casos, encontramos UBB (ubiquitina B) y UBA52 (una proteína de fusión formada por ubiquitina y la proteína ribosomal eL40). Esto es coherente con la observación de que algunas proteínas relacionadas con la ELA y la DFT están muy marcadas para la degradación o presentan una actividad deubiquitinasa desregulada.

Otros tres genes relevantes compartidos (VIM, ACTB y YWHAB) están directamente asociados con el citoesqueleto. Se ha demostrado que la inestabilidad del citoesqueleto desempeña un papel clave en la supervivencia neuronal en las enfermedades neurodegenerativas. Por ejemplo, se sabe que las proteínas 14-3-3 (como YWHAB) interactúan con la proteína TAU, que, cuando produce agregados, provoca la degeneración de los microtúbulos en la enfermedad de Alzheimer. Se han descrito interacciones patológicas que implican a estas proteínas en varias enfermedades neurodegenerativas, incluida la ELA (75).

Por último, PABPC1 también surgió como un gen común entre la ELA y la DFT. Este gen desempeña un papel crítico en la traducción y estabilidad del ARNm y se

ha descrito que contribuye a ciertos mecanismos compensatorios en algunas formas de ELA y DFT (76).

Es importante destacar que la clasificación con los datos de DFT fue menos eficaz que aquella realizada con los pacientes con ELA y controles sanos. Si bien pudimos identificar algunos puntos en común, se requieren análisis adicionales para obtener resultados más sólidos y robustos a medida que se disponga de más datos, un importante limitante al momento de generar clasificadores.

Esto nos lleva a otras limitaciones de nuestra investigación. Aunque los datos de *microarrays* se utilizan ampliamente, se sabe que sus mediciones son susceptibles al ruido. Además, los datos se obtuvieron a partir de muestras de sangre total, que reflejan una señal media de todo el tejido. Esta señal no solo diluye firmas biológicas específicas, como las asociadas con la ELA, sino que también puede verse influenciada por otros procesos o enfermedades sistémicas. Como consecuencia, los análisis realizados de forma posterior, como el de vías enriquecidas, pueden verse afectados por este ruido inherente, lo que podría aumentar el número de resultados espurios.

No obstante, se ha demostrado que los transcriptomas de sangre total capturan información específica de tejidos menos accesibles, incluidas las neuronas (77). Por sobre todas las cosas, las mediciones en sangre cuentan con la invaluable ventaja de ser una muestra muy accesible, nada traumática ni intrusiva para el paciente.

El análisis de enriquecimiento de vías biológicas, al estar basado en métricas de sobrerrepresentación, también puede producir falsos positivos. Considerando que todos estos puntos son fundamentales dentro del paso a paso de nuestra metodología, aunque obtuvimos resultados estadísticamente significativos y se consideraron cortes de calidad estrictos, somos conscientes del impacto potencial de estas fuentes de ruido en nuestros resultados.

En cuanto a GCNN-GLRP, se trata de un algoritmo que requiere una red totalmente conectada y no dirigida como dato de entrada esencial. Adicionalmente,

cada PPI se encuentra condicionada por el umbral de calidad establecido de acuerdo con los criterios específicos empleados en la creación y curación de las bases de datos de donde fueron tomadas. Estos factores pueden haber llevado a la exclusión de genes que podrían haber sido importantes para la clasificación. Esto se hizo evidente, como ya se ha señalado, en el hecho de que algunos de los genes más relevantes identificados en una red no estaban presentes en las otras.

Cabe destacar que nuestros análisis arrojaron resultados valiosos, de gran significado biológico, con capacidad generadora de hipótesis y estadísticamente significativos, aun considerando las limitaciones tecnológicas y metodológicas inherentes al estudio.

Con este trabajo, destacamos el poder de integrar diversas fuentes de datos para obtener información biológicamente significativa. Si bien la tarea de clasificación es valiosa para identificar nuevos pacientes, también sirve como posible indicador para priorizar genes que pueden desempeñar un papel clave en la enfermedad.

Cabe destacar que los genes identificados como más relevantes por nuestros modelos no son aquellos tradicionalmente vinculados con la ELA o la DFT, lo que sugiere que las estrategias de integración podrían revelar patrones moleculares aún no estudiados exhaustivamente.

Creemos que estos genes, sus funciones biológicas y su contexto dentro de la red PPI y las vías en las que están enriquecidos pueden servir como valiosos puntos de partida para futuras investigaciones, lo que suma un valor adicional y único al trabajo.



Conclusiones



- La regresión logística fue más rápida, precisa y computacionalmente eficiente para distinguir pacientes con ELA respecto a controles sanos.
- GCNN-GLRP, comparable en rendimiento a la RL, identificó genes con una mayor conectividad biológica en redes PPI que los devueltos por RL, lo que sugiere que estas interrelaciones pueden aportar más información, ya que las proteínas no actúan de forma aislada, sino que existe una dependencia entre ellas en los procesos biológicos.
- Los genes recuperados por GCNN-GLRP permitieron identificar vías biológicas significativamente enriquecidas, vinculadas a procesos bien documentados en la ELA.
- Solo GCNN-GLRP permitió identificar genes compartidos entre ELA y variantes de DFT.
- Los datos de *microarrays* en sangre total introducen ruido y diluyen señales específicas, si bien captan información tejido-específica.
- Si bien los datos de DFT presentaron limitaciones en cuanto a su cantidad y calidad, lo que restringe la robustez de la clasificación, permitieron la generación de nuevas hipótesis.

Conclusión general

La integración de expresión génica con redes PPI en modelos de aprendizaje profundo revela patrones moleculares no identificados por métodos de aprendizaje automático tradicionales, abriendo nuevas posibles líneas de abordaje de patologías complejas como la ELA y la DFT.



Trabajos complementarios



7.1 Control de patrones de señal en las redes PPI durante la clasificación

Para verificar que la distinción de genes relevantes para la clasificación dependía de los patrones de señal distribuidos a lo largo de la red PPI, se llevó a cabo un control en el que se mantuvo la estructura de la red, pero se aleatorizó el orden de los genes en la matriz de expresión génica. De este modo, se generan nuevos patrones de señal que no son representativos del modelo bajo estudio. (Figura 30).

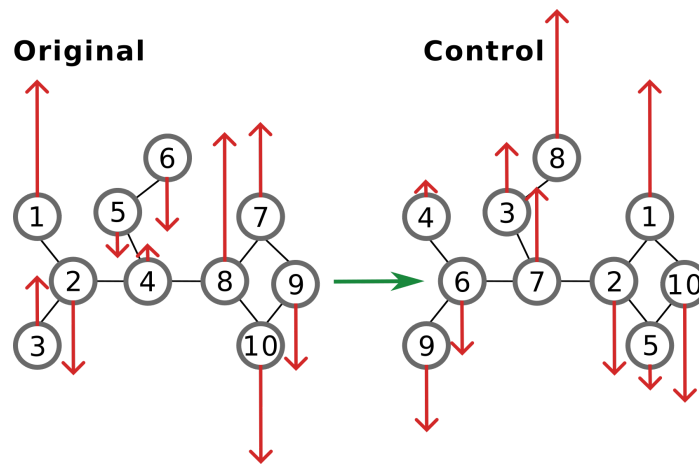


Figura 30. Diagrama que muestra cómo se encontraban ordenadas las señales de expresión génica de cada nodo de la red PPI en la corrida original (la señal coincide con su nodo) y en la corrida control (la señal no coincide con su nodo, el orden es aleatorio).

En primer lugar, se evaluó el rendimiento de GCNN con estos nuevos patrones. Como era de esperarse, el rendimiento de la clasificación se mantuvo en valores muy similares a los obtenidos originalmente, es decir, aquellos en los que la expresión génica sí correspondía con la ubicación en la red de la proteína codificada. Este comportamiento se observó de manera consistente en las tres redes PPI analizadas (Tablas 13-15). Dicho resultado podría atribuirse a que el modelo aprendió patrones de señal generados a partir de las nuevas ubicaciones asignadas aleatoriamente, las cuales se mantienen de forma coherente tanto en los casos de ELA como en los controles, independientemente de su relevancia biológica.

	IntAct							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
Original	0,923	0,873	0,782	0,708	0,763	0,701	0,738	0,774
Control	0.924	0.867	0.773	0.708	0.772	0.713	0.775	0.899

Tabla 13. Rendimiento de clasificación de pacientes con ELA con el modelo GCNN durante el testeo y la validación independiente, utilizando datos de IntAct como red PPI, comparando los valores originales con el control. En negrita se muestran los valores más altos.

	BioGRID							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
Original	0,924	0,867	0,787	0,708	0,752	0,694	0,760	0,884
Control	0.914	0.853	0.738	0.646	0.743	0.689	0.748	0.866

Tabla 14. Rendimiento de clasificación de pacientes con ELA con el modelo GCNN durante el testeo y la validación independiente, utilizando datos de BioGRID como red PPI, comparando los valores originales con el control. En negrita se muestran los valores más altos.

	HIPPIE							
	Testeo				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
Original	0,936	0,887	0,809	0,750	0,749	0,694	0,759	0,905
Control	0.932	0.887	0.809	0.750	0.771	0.719	0.776	0.905

Tabla 15. Rendimiento de clasificación de pacientes con ELA con el modelo GCNN durante el testeo y la validación independiente, utilizando datos de HIPPIE como red PPI, comparando los valores originales con el control. En negrita se muestran los valores más altos.

De mayor interés resultó analizar si los genes relevantes devueltos por GLRP tras la clasificación control coincidían con aquellos hallados cuando la expresión génica sí se encontraba correctamente ubicada en la red. Se encontró que, si bien la estabilidad fue muy similar (luego de 10 corridas se identificaban aproximadamente 40 genes comunes), los genes resultaron ser efectivamente distintos: con BioGRID se encontró únicamente un gen en común entre el modelo original y el control aleatorio, mientras que con IntAct e HIPPIE no se registró ningún gen relevante compartido entre ambas condiciones. Este hallazgo es de particular importancia, ya que la identificación de los mismos genes hubiera indicado que la estructura de la red no es determinante para evaluar su relevancia, y que, por ende, su incorporación no aportaría valor sustancial a la tarea de

clasificación, siendo otros los factores que determinarían la relevancia asignada a cada gen.

En conjunto, estos resultados permiten concluir que tanto la estructura de la red como los patrones de señal que emergen de ella desempeñan un papel fundamental en la determinación de los nodos relevantes para la clasificación.

7.2 Unión de redes de interacción proteína-proteína

Se utilizó una cuarta red PPI para el análisis realizado en esta tesis, estando la misma conformada por la unión de las redes de IntAct, BioGRID y HIPPIE (de aquí en adelante denominada '*Merge*'). La red *Merge* contiene 12.387 nodos y 139.695 aristas y se utilizó, al igual que las demás, en el protocolo explicado en Materiales y Métodos, sección xx.

El rendimiento de los cuatro clasificadores cuando se utilizó la red *Merge* fue semejante a los obtenidos con IntAct, BioGRID y HIPPIE por separado, con la regresión logística y GCNN dando las métricas más altas tanto en testeo como validación (**Tabla 16**).

Dado que, en el marco de la metodología adoptada, los análisis subsiguientes se fundamentaron en los resultados compartidos por las tres redes PPI, se optó por continuar trabajando con dichas redes en lugar de utilizar la red *Merge*.

	<i>Merge</i>							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.935	0.893	0.818	0.750	0.769	0.726	0.769	0.845
RL	0.949	0.893	0.831	0.783	0.824	0.724	0.789	0.951
RF	0.874	0.824	0.649	0.542	0.682	0.601	0.559	0.457
XGBoost	0.865	0.824	0.691	0.625	0.709	0.638	0.634	0.573

Tabla 16. Rendimiento de clasificación de pacientes con ELA con los modelos GCNN, RL, RF y XGBoost luego del testeo y la validación independiente, utilizando datos de la red *Merge* como red PPI. En negrita se muestran los valores más altos comparando entre modelos.

7.3 Organelas sin membrana y ELA

Las organelas sin membrana (*membraneless organelles*, MLOs, en inglés) son compartimentos formados a partir de la reunión y separación en fase líquido-líquido (LLPS, del inglés *Liquid-liquid Phase Separation*) de sus componentes, lo cuales suelen ser diversas proteínas y ARNs (78). Su presencia puede ser contexto específica, propia de tipos celulares determinados o ubicuas a todas las células del organismo (**Figura 31**, tomada de (79)). Sin una barrera física, se hallan realizando y regulando distintas funciones celulares. Es sabido que muchas de estas están intrínsecamente relacionadas a distintas enfermedades (80).

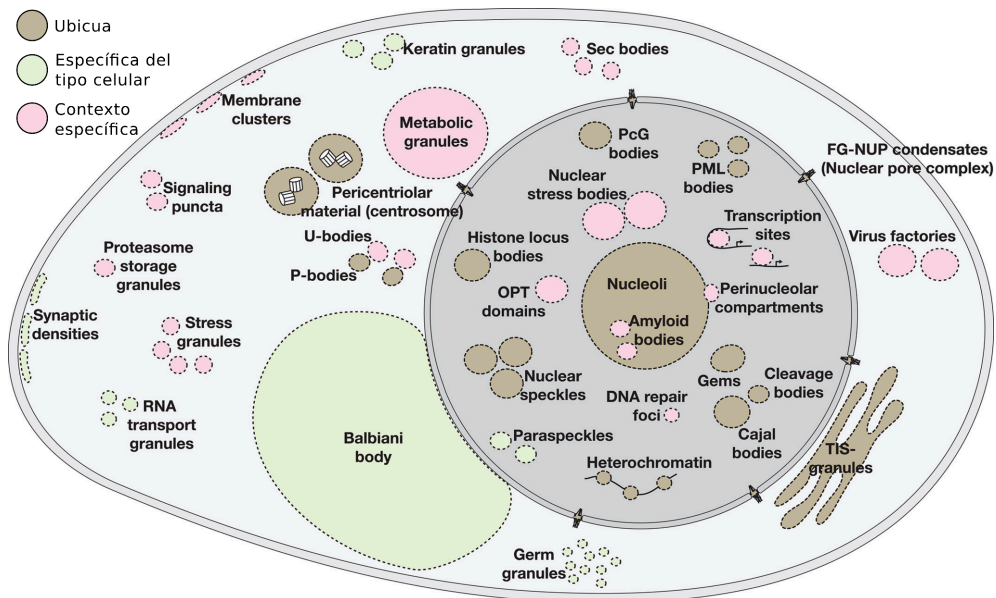


Figura 31. Diagrama de diferentes MLOs y sus ubicaciones dentro de la célula.

Se utilizó la base de datos PhaSeDis (81), para observar en qué afecciones se encontraban involucradas un mayor número de MLOs. Notamos que distintos tipos de cáncer y enfermedades neurodegenerativas eran las que tenían un mayor porcentaje de éstas (**Figura 32**), y considerando el segundo grupo, la ELA en particular es de las que tienen un número más elevado de MLOs asociadas a su patogénesis (**Figura 33**).

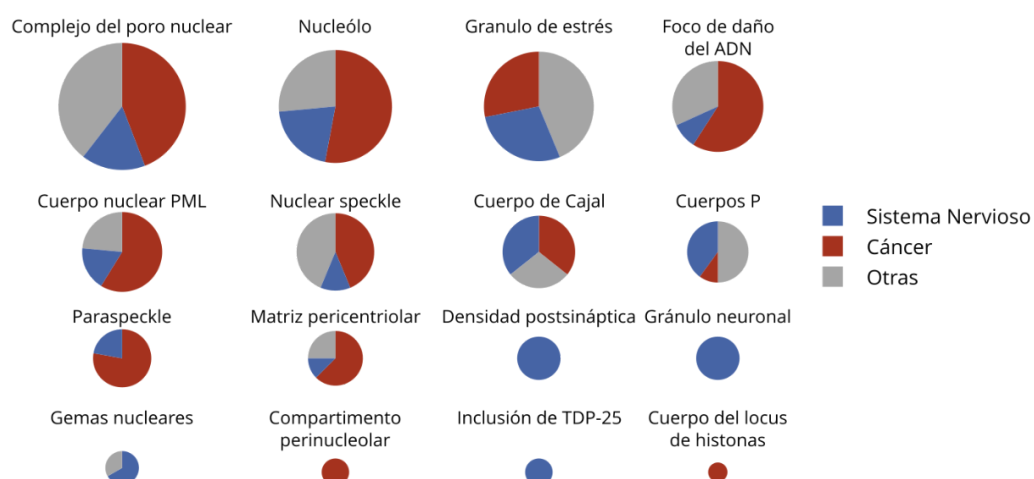


Figura 32. Gráficos de torta por MLO con la proporción de enfermedades neurodegenerativas, cáncer y otras patologías asociadas a estos condensados.

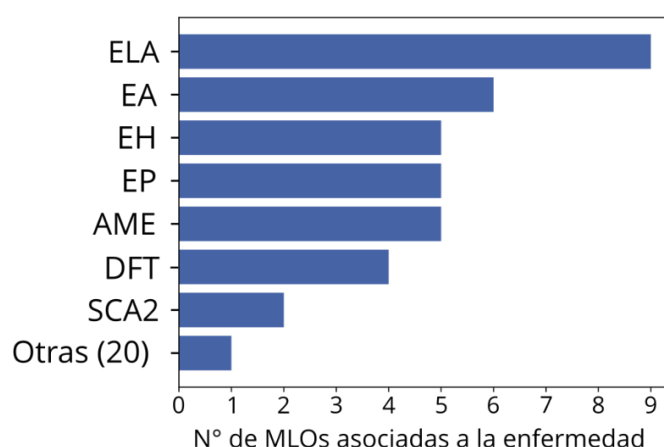


Figura 33. Gráfico de barras con el número de MLOs asociadas a distintas enfermedades neurodegenerativas. EA: Enfermedad de Alzheimer; EH: Enfermedad de Huntington; EP: Enfermedad de Parkinson; AME: Atrofia Muscular Espinal; SCA2: Ataxia Espinocerebelosa tipo 2.

Se consideró la posibilidad de que redes de MLOs que se saben involucradas en la enfermedad tal vez podrían contener información más específica y dar una mejor clasificación y posterior análisis de genes relevantes. Por lo tanto, se propuso realizar la misma clasificación de los pacientes con ELA y controles sanos, pero con la red PPI de las MLOs. Para definir las mismas se tomaron las proteínas presentes en cada MLO de la base de datos MLOsMetaDB (82), considerando aquellas tomadas de las bases de datos PhaSePro (83), PhaSepDB (84) y DrLLPS (85), ya

que los experimentos y la curación por las que fueron obtenidas son de mayor confianza. Las interacciones entre estas proteínas fueron nuevamente dadas por las redes IntAct, BioGRID y HIPPIE. El análisis finalmente se realizó solo con las redes del gránulo de estrés, *paraspeckle*, nucleolo y cuerpo P, ya que eran las únicas MLOs asociadas al ELA cuya red PPI totalmente conectada contenía más de 1000 proteínas. Esto fue decidido debido a que se necesita una red con un número considerable de nodos y aristas para que las operaciones de convolución y *pooling* puedan realizarse y así capturen información relevante.

Observando los rendimientos de los distintos modelos (Material suplementario, **sección 10.1**), pudimos notar que los resultados fueron más variables: los modelos que obtuvieron las mejores clasificaciones fueron levemente menos precisos que aquellos que utilizaron la red PPI completa. Además, se observó mayor variabilidad entre los resultados dependiendo de la red PPI usada y la MLO analizada.

7.4 AnswerALS, expresión génica obtenida mediante RNA-Seq

Se utilizó un tercer conjunto de datos de expresión génica para realizar la clasificación, esta vez provenientes del consorcio AnswerALS (86). Este programa se inició con muestras de células mononucleares de sangre periférica (PBMCs, del inglés *Peripheral Blood Mononuclear Cells*) provenientes de más de 1000 pacientes con ELA esporádico, a partir de las cuales se generaron líneas de neuronas motoras derivadas de células madre pluripotentes inducidas (iPSC-MNs, del inglés *induced pluripotent stem cell-derived motor neurons*). En la **Figura 34** se presenta un diagrama esquemático del proceso de obtención de esta línea. El consorcio aplicó sobre la misma distintas técnicas moleculares, bioquímicas y de imagen. El objetivo último es el de comprender la heterogeneidad de la ELA esporádica en estas neuronas derivadas de pacientes, intentando aproximarse a una biopsia. Los datos de genómica, epigenética, proteómica y transcriptómica obtenidos por el consorcio se encuentran accesibles a investigadores que presenten un proyecto de interés.

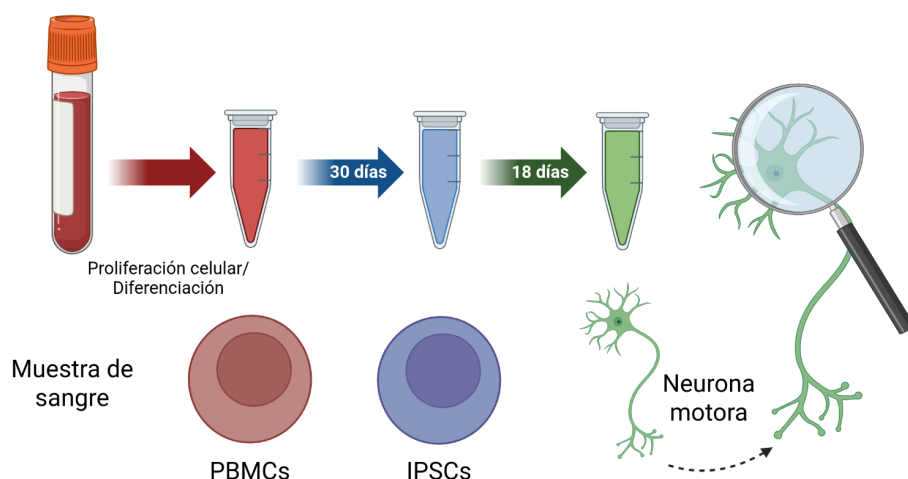


Figura 34. Diagrama de la obtención de la línea celular iPSC-MNs a partir de muestras de sangre. Realizada con BioRender.

Para la clasificación realizada en este trabajo se utilizaron nuevamente datos de transcriptómica, obtenidos mediante RNA-seq (del inglés *RNA sequencing*), una herramienta de secuenciación masiva que permite cuantificar el transcriptoma completo de una muestra biológica en un momento determinado (87). En esta técnica, a partir del ARN extraído, se generan millones de lecturas cortas (en inglés, *reads*) que se alinean a un genoma de referencia, permitiendo estimar el nivel de expresión de cada gen de forma simultánea y a escala global. A diferencia de los *microarrays*, que miden intensidad de señal luminosa como *proxy* de la expresión, RNA-seq cuantifica directamente la cantidad de transcripto presente en la muestra.

Se consideraron los datos en su forma más procesada, expresando los niveles de expresión como transcritos por millón (TPM), correspondientes a 389 pacientes de ELA y 72 controles sanos. Para definir el TPM del gen i , se define primero:

$$RPK_i = \frac{\text{reads del gen } i}{\text{longitud del gen } i \text{ (kb)}},$$

normalizando así la cantidad de transcripto por la longitud del gen, dado que genes más largos tienden a acumular un mayor número de *reads*. Finalmente, el TPM se define como:

$$TPM_i = \frac{RPK_i}{\sum_j RPK_j} \times 10^6$$

donde la sumatoria en el denominador recorre todos los genes j de la muestra, normalizando de este modo por la cantidad total de transcripto detectado.

Como paso final de preprocesamiento se los normalizó mediante transformación logarítmica.

En cuanto a las redes PPI, además de considerar nodos que tuvieran expresión génica en el conjunto de datos y el componente conectado de mayor tamaño, también nos aseguramos que todas las interacciones estuvieran presentes en las redes que se habían utilizado para clasificar los datos de *microarray*. Los pasos siguientes se realizaron de igual manera que con estos últimos. En la **Tabla 17** se muestra el número final de nodos y aristas que tuvieron las redes PPI en este caso.

El desbalance entre clases en este conjunto de datos era considerable, lo cual resultó en diversos intentos por lograr una correcta clasificación. En última instancia, se procedió realizando *subsampling*: Se tomaron de forma aleatoria 90 pacientes con ELA, para acercarnos a la cantidad de controles.

	Genes	Interacciones
BioGRID	7.249	48.951
HIPPIE	9.030	68.209
IntAct	10.461	57.545

Tabla 17. Resumen del número de nodos (genes codificadores de proteínas) y aristas (interacciones proteína-proteína) en cada una de las redes de PPI seleccionadas utilizadas para entrenar y probar el modelo GCNN-GLRP. Los mismos conjuntos de genes se utilizaron como *features* de entrada en los clasificadores que no utilizan información de las interacciones.

En este caso, el rendimiento de los modelos varió levemente dependiendo de qué red PPI fuera utilizada, pero en términos generales, con los datos de AnswerALS se obtuvieron resultados similares a los obtenidos con los datos de DFT. Los valores de AUC-ROC, *Accuracy*, *F1-Score* y *Recall* que se obtuvieron dieron

cuenta que, si bien los modelos fueron capaces de aprender más que un predictor *random*, no fueron extremadamente precisos en la clasificación (**Tabla 18**).

	IntAct				HIPPIE				BioGRID			
	AUC	Acc	F1	Recall	AUC	Acc	F1	Recall	AUC	Acc	F1	Recall
GCNN	0.604	0.647	0.714	0.889	0.639	0.588	0.681	0.889	0.583	0.529	0.692	1.0
RL	0.622	0.562	0.632	0.667	0.587	0.562	0.632	0.667	0.619	0.562	0.632	0.667
RF	0.562	0.562	0.636	0.778	0.579	0.562	0.636	0.778	0.578	0.562	0.632	0.667
XGBoost	0.492	0.500	0.577	0.667	0.508	0.500	0.563	0.556	0.508	0.529	0.588	0.667

Tabla 18. Rendimiento de clasificación de pacientes con y sin ELA de los modelos GCNN, RL, RF y XGBoost durante el testeo. Acc = *Accuracy*; F1 = *F1-Score*.



Publicaciones con referato



“Unveiling the Complexity of cis-Regulation Mechanisms in Kinases: A Comprehensive Analysis”

Alvaro M Navarro*, Macarena Alonso*, Elizabeth Martínez-Pérez, Tamas Lazar, Toby J Gibson, Javier A Iserte, Peter Tompa, Cristina Marino-Buslje.

Proteins. 2025 Mar;93(3):575-587. doi: 10.1002/prot.26751. Epub 2024 Oct 4.

*: contributed equally

Resumen

Los elementos reguladores *cis* (CREs, del inglés *cis-Regulatory Elements*) son regiones de las proteínas que modulan la actividad de las mismas mediante interacciones intramoleculares. Las quinasas, enzimas clave en numerosos procesos biológicos, suelen regularse a través de interacciones inhibitorias en *cis*. Este estudio profundiza en los mecanismos de regulación *cis* mediada por CREs en quinasas, combinando análisis estructural y de secuencia.

Para ello, recopilamos un extenso conjunto de datos de quinasas con CREs anotados, organizados en familias homólogas mediante alineamientos de secuencias múltiples. En todas las secuencias se mapearon sistemáticamente atributos moleculares clave: desorden y contenido de estructura secundaria, sitios activos y de unión a ATP, modificaciones postraduccionales y mutaciones asociadas a enfermedades. Asimismo, exploramos posibles cambios conformacionales entre los estados activo e inactivo de las quinasas y analizamos la presencia de éstas en orgánulos sin membrana, dilucidando sus funciones en dichos compartimentos.

Los CREs presentan un continuo estructural que abarca desde segmentos cortos desordenados hasta dominios completamente plegados. La diversidad de mecanismos mediante los cuales logran inhibir a la quinasa es notable: desde la interacción autoinhibitoria directa con el sitio activo, hasta la unión a sitios alternativos que inducen regulación alostérica. Estos distintos tipos de mecanismos inhibitorios se ilustran mediante sistemas representativos arquetípicos.

Si bien este trabajo ofrece un enfoque sistemático para comprender los CREs de las quinasas, se requieren investigaciones experimentales adicionales para desentrañar la complejidad propia de cada familia. Los conocimientos generados sientan las bases para futuros estudios orientados a descifrar los mecanismos moleculares de la desregulación de quinasas y a explorar posibles intervenciones terapéuticas (88).



RESEARCH ARTICLE

Unveiling the Complexity of *cis*-Regulation Mechanisms in Kinases: A Comprehensive Analysis

Alvaro M. Navarro¹ | Macarena Alonso¹ | Elizabeth Martínez-Pérez¹ | Tamas Lazar^{2,3} | Toby J. Gibson⁴ | Javier A. Iserte¹ | Peter Tompa^{2,3,5} | Cristina Marino-Buslje¹

¹Structural Bioinformatics Unit, Fundación Instituto Leloir, Buenos Aires, Argentina | ²VIB-VUB Center for Structural Biology, Flanders Institute for Biotechnology (VIB), Brussels, Belgium | ³Structural Biology Brussels, Department of Bioengineering, Vrije Universiteit Brussel, Brussels, Belgium | ⁴Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany | ⁵Research Centre for Natural Sciences, Hungarian Research Network, Institute of Enzymology, Budapest, Hungary

Correspondence: Cristina Marino-Buslje (cmb@leloir.org.ar)

Received: 26 April 2024 | **Revised:** 29 August 2024 | **Accepted:** 12 September 2024

Funding: This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas, <http://dx.doi.org/10.13039/501100002923>; European Commission, <http://dx.doi.org/10.13039/501100000780>; and Agencia Nacional de Promoción Científica y Tecnológica, <https://doi.org/10.13039/50110003074>. A.M.N. is a PhD fellow, E.M.P. is postdoctoral fellow, and J.I. and C.M.B. are researchers of National Scientific and Technical Research Council (CONICET)—Argentina; M.A. is a PhD fellow of National Agency for the Promotion of Research, Technological Development and Innovation (Agencia I + D + i)—Argentina. This work was partially funded by PICT-2018-01015. This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 778247 (IDPfun).

Keywords: auto-regulatory mechanism | *cis*-regulatory elements | conformational diversity | intrinsically disordered regions | kinases | liquid–liquid phase separation | mutations | protein autoregulation

ABSTRACT

Protein *cis*-regulatory elements (CREs) are regions that modulate the activity of a protein through intramolecular interactions. Kinases, pivotal enzymes in numerous biological processes, often undergo regulatory control via inhibitory interactions in *cis*. This study delves into the mechanisms of *cis* regulation in kinases mediated by CREs, employing a combined structural and sequence analysis. To accomplish this, we curated an extensive dataset of kinases featuring annotated CREs, organized into homolog families through multiple sequence alignments. Key molecular attributes, including disorder and secondary structure content, active and ATP-binding sites, post-translational modifications, and disease-associated mutations, were systematically mapped onto all sequences. Additionally, we explored the potential for conformational changes between active and inactive states. Finally, we explored the presence of these kinases within membraneless organelles and elucidated their functional roles therein. CREs display a continuum of structures, ranging from short disordered stretches to fully folded domains. The adaptability demonstrated by CREs in achieving the common goal of kinase inhibition spans from direct autoinhibitory interaction with the active site within the kinase domain, to CREs binding to an alternative site, inducing allosteric regulation revealing distinct types of inhibitory mechanisms, which we exemplify by archetypal representative systems. While this study provides a systematic approach to comprehend kinase CREs, further experimental investigations are imperative to unravel the complexity within distinct kinase families. The insights gleaned from this research lay the foundation for future studies aiming to decipher the molecular basis of kinase dysregulation, and explore potential therapeutic interventions.

Abbreviations: ABL1, tyrosine-protein kinase ABL1; AF, AlphaFold; CAMK, calcium/calmodulin-dependent protein kinase; CDK, cyclin-dependent kinase; CML, chronic myelogenous leukemia; CRE, *cis*-regulatory element; IDR, intrinsically disordered region; KD, kinase domain; MAPK, mitogen-activated protein kinase; MAP2K3, dual specificity mitogen-activated protein kinase kinase 3; MSA, multiple sequence alignment; PDB, Protein Data Bank; PML, promyelocytic leukemia body; PRKAA2, 5'-AMP-activated protein kinase catalytic subunit alpha-2; PSD, postsynaptic density; PTMs, post-translational modifications; RPS6KB1, ribosomal protein S6 kinase beta-1; RTK, receptor tyrosine kinase; SGs, stress granules; SH2, Src-homology-2; SH3, Src-homology-3; WNK, With-No-Lysine (K) kinase.

Alvaro M. Navarro and Macarena Alonso contributed equally to this study (co-first authors). Peter Tompa and Cristina Marino-Buslje contributed equally to this study.

© 2024 Wiley Periodicals LLC.

1 | Introduction

Kinases are ubiquitous enzymes that play essential signaling and regulatory roles in cells. They are central components in various cell signaling cascades, including pathways such as the mitogen-activated protein kinase (MAPK) pathway, regulating biological processes like growth, survival, differentiation and development [1, 2]. Kinases, such as cyclin-dependent kinases (CDKs), are also central to regulating the cell cycle, governing critical checkpoints that facilitate transitions between different phases [3].

As crucial intracellular signal transducers, kinases themselves must exhibit high responsiveness to upstream signals and regulatory cues. This responsiveness is intricately tied to their conformational flexibility, allowing for timely transitions between inactive and active states. Such transitions can be initiated either internally within the kinase (*cis*) or externally by another protein (*trans*). In the *trans* scenario, the interaction may entail another protomer of the same molecule or involve an entirely different protein. The former case can be exemplified by the transient dimerization of the kinase domains (KDs); for instance, members of the death-associated protein kinase (DAPK) family can generate face-to-face dimers that obstruct the catalytic sites of each protomer [4, 5]. An example of the second type of *trans* regulation is the B-Raf kinase, which is maintained in an auto-inhibited conformation through a dimer of 14-3-3 proteins binding to two phosphorylated residues in its C-terminus [6].

On the other hand, self-regulation represents a widespread regulatory mechanism facilitated through regions or domains within the same polypeptide chain, known as *cis*-regulatory elements (CREs) [7, 8]. In principle, the CRE can interact in a way to block the binding of the substrate or ATP, or distorts the catalytic site, making it nonfunctional [9]. They may also prevent post-translational modifications (PTMs), often resulting in an inactive state.

Intramolecular reorganizations, leading to the disengagement of the CRE and the KD, trigger the transition of the kinase back to its active state ([10]). A diverse array of CREs have been described, both in terms of sequence and structure. For instance, in the Src-family kinases CREs are fully folded domains, like the Src-homology-3 (SH3) and Src-homology-2 (SH2) domain. They bind to the opposite side of the activation loop, and stabilize an inactive conformation [11–14]. In certain receptor tyrosine kinases (RTKs), autoinhibitory interactions involve the juxtamembrane regions. Mutations or phosphorylations in this region can activate the kinase activity as is the case in some cancer types [14].

CREs may also encompass intrinsically disordered regions (IDRs). IDRs offer a multitude of interactions and conformational states [15] and a variety of autoinhibition mechanisms, including allosteric changes distal to the active site [8].

PTMs such as phosphorylation, acetylation, and ubiquitination play pivotal roles in regulating protein activity, stability, and localization. These modifications induce conformational changes or influence protein–protein interactions [12]. For instance, the

phosphorylation of Ser/Thr residues in the active site of protein kinases can enhance their catalytic activity [16].

Mutations in either the CRE or KD can interfere with activation cycles of the enzyme, often resulting in constitutive activation [17–20]. Moreover, mutations that release autoinhibition are likely to be a driver in cancer ([19, 21, 22]).

Kinase's spatial organization at the subcellular level is crucial for their signaling function and a proper strategy to achieve a high-fidelity signal response ([23]; L. [24]). Liquid–liquid phase separation (LLPS) is a process by which proteins and other biomolecules can separate into distinct liquid phases within a single cellular environment. LLPS is a reversible process that allows for the formation of dynamic, membrane-less compartments within cells, known as membraneless organelles (MLOs) [25–28]. Kinases exhibit the ability to assemble into enzymatic clusters, concentrating several of them, substrates, and cofactors locally in an MLO. One example is the kinase-control phase transition of MLOs in mitosis [29]. Kinases can be drivers, clients, and/or regulators. Drivers are essential proteins responsible for initiating the condensate formation [30–32], while clients, although not indispensable for the formation, can dynamically partition between the diluted phase and the condensate phase through transient interactions [33, 34]. Finally, regulators are proteins capable of influencing the formation and/or dissolution of condensates and are not necessarily within the condensate [35]. The precise regulation of protein activities, occurring at the right moment and location, is essential for normal growth, development, and homeostasis.

Despite its significance, our comprehension of *cis*-regulatory mechanisms remains limited, and a thorough investigation has yet to be conducted. Noteworthy is the scarcity of systematic inquiries into this intricate regulatory system in the context of kinases, which constitute the focal point of our ongoing research. In this study, we delve into the exploration of self-regulation mechanisms in kinase proteins that encompass *cis*-regulatory elements, with a particular focus on examining their conformational diversity.

2 | Results

2.1 | Dataset Description

In order to avoid redundancy, we clustered the original set of 280 kinases at 60% identity, ending up with a set of 250 clusters where one of the proteins is taken as the reference one. By applying the homology transfer protocol, we retrieved 8411 orthologous from 663 organisms (Figure S1 and Table S1). We then built a multiple sequence alignment (MSA) for each cluster (i.e., protein family) [36]. The distribution of the number of sequences across MSAs is shown in Figure 1A. Fifty-eight out of 250 MSAs have more than one experimentally solved structure in the Protein Data Bank (PDB), covering at least one residue of both the KD and the CRE, accounting for a total of 830 structures (Figure 1B). Forty-three MSAs have two or more structures (making up a total of 820 PDB structures), allowing the analysis of potential conformational changes within a family.

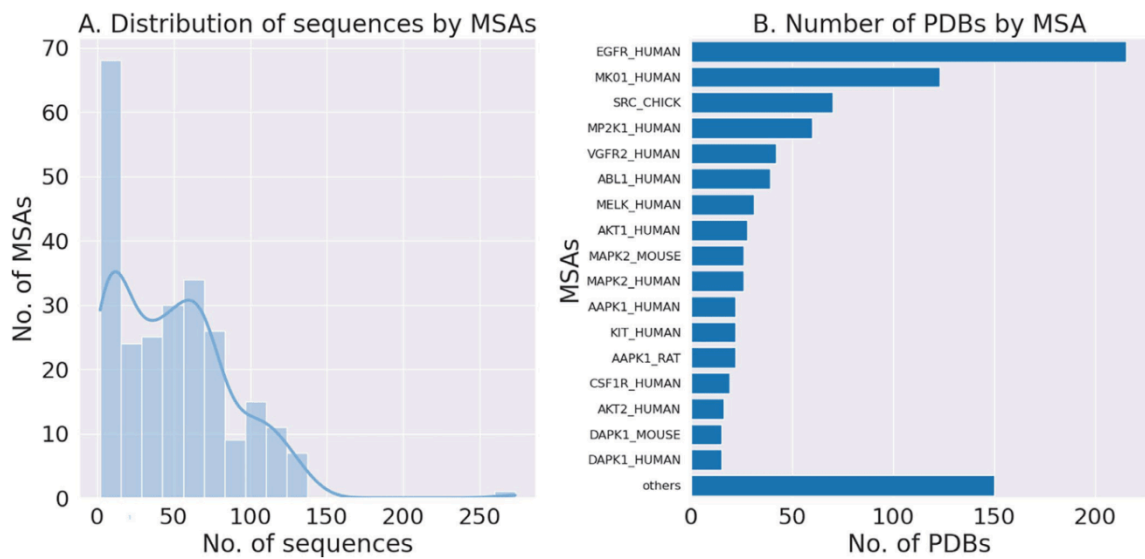


FIGURE 1 | MSAs description. (A) Distribution of 8411 sequences in the 250 MSAs. (B) Number of PDB structures across 58 MSAs. The protein name shown in the y-axis corresponds to the reference sequence in the MSA. Only MSAs with 10 or more sequences are shown and the rest grouped in the “others” category.

TABLE 1 | Kinases found associated with membraneless organelles in MLOsMetaDB.

MLO	CRE-containing kinases
Cajal body	RPS6KA1, FLT1, FLT3, ROR2, AKT3
Centrosome/Spindle pole body	CSNK1D, MAP3K11, PTK2, ROCK1, MAPK1, NTRK1, CSNK1E, MAPKAPK5, STK3, PRKAA1, MAP2K1, MAP2K2, PAK1, CAMK2B, ROCK2, PAK2, FGFR1, CAMK2G, PRKCB, BRAF, CHEK1
Nuclear speckle	MAP3K6, FGFR1, MAP3K4, HIPK2, MAP3K14, MAP4K1, RPS6KA1, KIT, MAP4K2
Nuclear stress body	MAPK1, MAP2K4
Nucleolus	TNK2, FGFR3, PDGFRB, CSNK1D, PDGFRA, FGFR2, NTRK1, CSF1R, DAPK2, CAMK4, MAP3K1, MAP2K1, EPHB2, MAP3K14, INSR, FGFR4, CDC42BPA, FGFR1, KIT, PRKAA1, PAK4, RPS6KA3, MUSK, MAP3K11, CAMK2G
P-body	PRKCA, PAK3, RPS6KA5, CAMKK2, PRKAA1, PRKCI, CSNK1D, CAMK4, PAK4, PTK2
PML nuclear body	ABL1, FGFR4, EGFR, MAP3K14, HIPK2, PAK3, INSR, PAK4, DAPK2, MAP3K1
Paraspeckle	MAP3K7
Postsynaptic density	PRKCG, CAMKK1, PAK1, CSNK1D, ROCK2, CAMK2G, MAP2K2, PRKCB, CAMK2B, CDC42BPA, CSNK1E, WNK1, MAP2K1, CAMKK2, PRKCA, ROCK1, PTK2, CAMK2D, MAPK1
Stress granule	ROCK1, MAPK1, RPS6KB1, PRKAA2, PRKCA, PAK4, RPS6KA4, ZAP70, MAP2K3, INSR, WNK1, PTK2, RPS6KA3

2.2 | Sequence and Structural Analysis

We observed a diverse range of CRE lengths spanning from as few as 6 to as many as 648 amino acids (Figure 2A). This result underscores the complexity of CREs and suggests the presence of both short stretches and folded domains. It is worth noting that the definition of CRE length can be imprecise due to the experimental methods used to determine it. It

is possible that the experimental removal of an entire protein domain caused activation, when in fact only a few residues were responsible.

The examination of the disorder content and predicted secondary structure elements for all kinases reveals a wide range of both disordered and structured CREs among the families, denoting again the potential occurrence of multiple binding modes between the

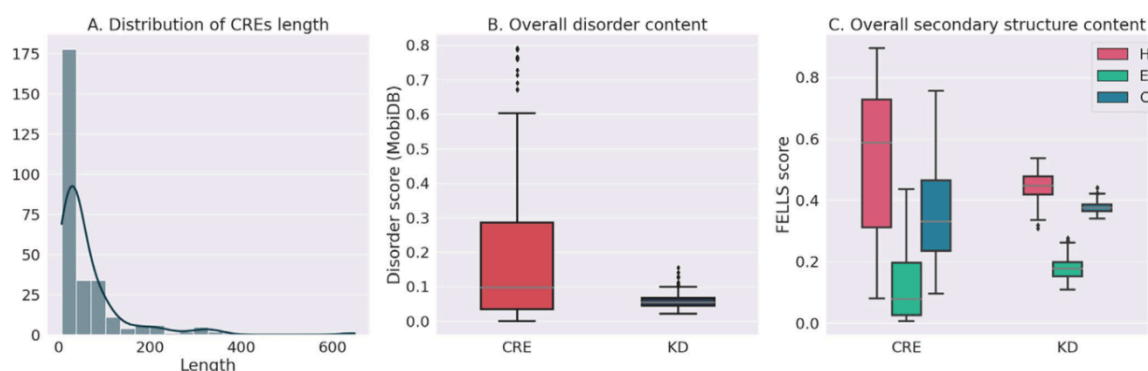


FIGURE 2 | CREs characteristics. (A) CREs sequence length distribution (286 CREs). min: six residues, max: 648 residues. (B) Distribution of MobiDB mean disorder score (from 0 to 1) in kinase MSAs. (C) FELLS score [37] means secondary structure propensity distribution (from 0 to 1) by MSA (H = Helix; E = β -strand and C = no secondary structure).

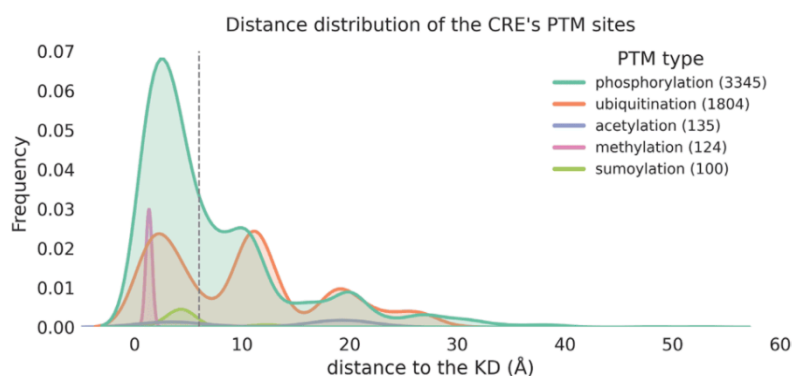


FIGURE 3 | Distribution of the CRE's PTMs distances to the KD. Normalized kernel density estimation of the interatomic distances from the PTMs located in the CRE to the KD. Each curve represents a type of PTM. Vertical dashed line indicates the contact threshold (6 Å). Normalization is considering all the PTMs together.

CRE and the KD (Figure 2B,C). The structural plasticity of CREs may confer advantages in interactions with the KD [38].

In order to elucidate possible regulatory mechanisms, we mapped the PTMs within the CRE in the available three-dimensional structures. PTMs are known for their crucial involvement in regulating kinase activity [39]; more specifically, most kinases need phosphorylation to be in an active state [40]. PTMs are also capable of controlling the secondary structure, binding, positioning and/or recognition of substrates [41].

As shown in Figure 3, 53% of the CRE residues prone to PTM acquisition are situated in close spatial proximity to the KD. There are substantial differences of locations by PTM type: methylation 98%, sumoylation 86%, phosphorylation 59%, acetylation 41% and ubiquitination 39% of the sites. Remarkably, ubiquitination displays a regular distribution pattern of around 10 Å cycles. This would align with the fact that there are proteins undergoing polyubiquitination (Y. [42, 43]) and the ubiquitin's conformation in multiple domains, such as UBB_HUMAN having three ubiquitin domains (Figure S2).

The above suggests that the accessibility of PTM sites and/or their modification could play a pivotal role in signal transduction by kinase. However, this analysis should be interpreted with care because the limited number of available structures and the type of proteins solved (the more studied) can bias the analysis.

Among PTMs, phosphorylation emerges as the most prevalent type in this set of proteins. Phosphorylations at crucial sites within the CRE intricately fine-tune the structural and functional properties of numerous protein kinases [41].

We also conducted an exhaustive structural analysis across 43 protein families, each featuring two or more experimentally solved 3D structures, which enabled us to compare various conformers to visualize conformational changes. Measuring the interatomic distances between CRE and KD residues, our investigation unveiled a spectrum of possibilities. The found variability strongly suggests that each kinase family employs their own mechanism to accomplish their autoinhibition.

Next we will deepen the structural analysis in four protein families representative of the observed mechanisms through the lens

of structures. This analysis is made with the current available information, more experimental determinations will be necessary to cover the entire structural space.

2.3 | FLT3 Family

The cell-surface receptor tyrosine-protein kinase FLT3, implicated in many important cellular processes such as growth, division and survival of cells, is activated when the specific ligand binds to it [44]. There are six structures available in the PDB, covering the CRE and KD regions. Superposing the six structures reveals their shared conformation, with most of the 38 CRE residues in contact with the KD (Figure 4A), thereby inhibiting its activity by a blockage of the catalytic site, as suggested by Griffith et al. [46]. Figure 4B shows the distance distribution of the CRE residues to the KD's active and ATP-binding sites. Figure 4D depicts the specific contacts between these two regions.

The structural superposition of the different structures shows that the CRE is mostly ordered (also in Figure S3A) and in direct contact with the KD's active site residues (Figure 4C), Figure 4D depicts the specific contacts between these two regions.

To investigate the potential for an alternative conformation, we also superposed the AlphaFold (AF) predicted model [47], which unsurprisingly has the same conformation. Sharing

conformations across six different PDBs solved in different circumstances provides strong evidence that the protein undergoes only minimal conformational changes. However, we cannot exclude the existence of alternative conformations, even integrating the AF model given its tendency to predict models in a more globular arrangement and its limited ability to capture structural conformers [19, 48] (Figure S4).

2.4 | ABL1 Family

The non-receptor tyrosine-protein kinase ABL1 is a proto-oncogene that plays an important role in several cell events such as cytoskeleton remodeling in response to extracellular stimuli, cell motility and adhesion, receptor endocytosis, autophagy, DNA damage response, and apoptosis. The CRE in this kinase (positions 61–233) is a tandem of an SH3 and an SH2 domains ([10]). These domains are located on the opposite side of the active site contacting the KD by few residues, thus probably inhibiting kinase activity through allosteric regulation ([10]) (Figure 5A). There are 39 available structures of the ABL1 family, which consistently exhibit the same spatial arrangement (Figure S5A), strongly suggesting a potentially favored conformation for maintaining the kinase in an autoinhibited state with minimal conformational changes. Again, the model predicted by AF presents an almost identical conformation to the ones retrieved from PDB (Figure S5B).

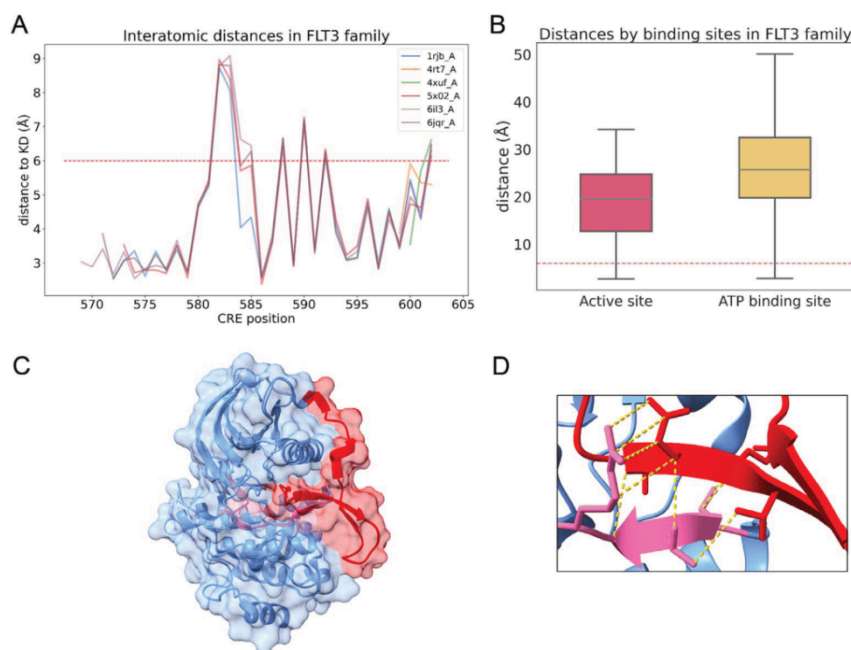


FIGURE 4 | FLT3 family structural analysis. (A) Interatomic distances between CRE and KD residues in FLT3 family of kinases, showing the CRE residues in the x-axis (CRE positions 565–602 in UniProt reference sequence coordinates, accession: P36888). The y-axis shows the shortest distance (between each pair of heavy atoms) of this residue of the CRE to any residue of the KD. Each curve is a PDB structure. The red dashed line indicates the contact distance threshold (6 Å). (B) Distance distribution of the CRE residues to the active and ATP binding sites. (C) Superposition of the PDB models. For the sake of simplicity only two models are shown (5X02 and 6JQR). The models are represented as ribbon and also showing a translucent surface. The KD is shown in light-blue, the CRE is colored in red and the active site in pink. (D) Contacts between the CRE and the active site are shown in yellow (contact distance ≤ 6 Å). Molecular figures were generated with ChimeraX [45].

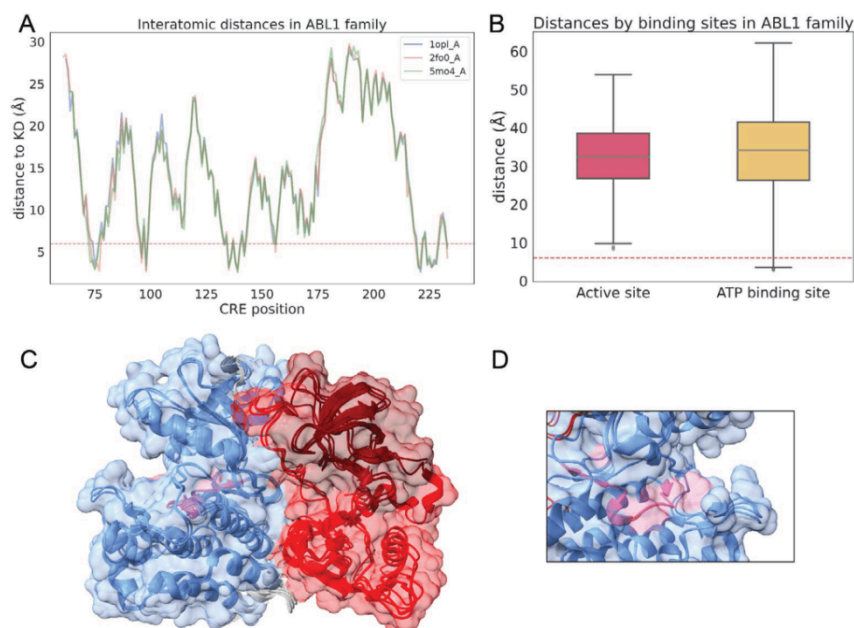


FIGURE 5 | ABL1 family structural analysis. (A) Interatomic distances within the ABL1 family, showing the CRE residues (x-axis, CRE positions 61–233 in UniProt reference sequence coordinates, accession: P00519) to the KD residues (y-axis). Out of the 35 PDB structures providing information on both CRE and KD, only three (1OPL, 2FO0 and 5MO4) comprehensively cover both regions. (B) Distance distribution of the CRE residues to the active and ATP binding sites of the KD. (C) Superposition of PDB structures depicted in (A). KD is shown in light blue and CRE in red, where SH2 domain is colored in dark red and SH3 domain in light red. (D) KD's active site is shown in pink. The image is the same as (C), zoomed and rotates 180°. All molecular figures were generated with ChimeraX [45].

The distribution of the interatomic distances between the KD's active and ATP-binding sites and the CRE residues show their position far from the contact threshold (Figure 5B). The CRE exhibits both disorder and order content, but still more disordered than the KD (disorder and secondary elements are shown in Figure S3B). The superposed PDB structures of the kinases in the autoinhibited state can be seen in Figure 5C and the KD's active site in more detail in Figure 5D.

2.5 | CAMK Families

In the Calcium/Calmodulin-dependent protein kinase (CAMK) family, the CRE is in contact with the KD in some of the conformers, denoting a specific condition. Typically, members of the CAMK family feature a N-terminal KD, followed by a CRE and a calmodulin (CaM)-binding domain that overlaps slightly with the CRE. These kinases are activated in response to the binding of calcium to CaM. Upon calcium/CaM binding, the kinase releases its autoinhibition by disrupting the interaction between the CRE and the KD. Subsequently, it auto-phosphorylates the CRE so that the kinase transitions to an active state, facilitating proper substrate binding [9, 49].

As an example, in the CAMK type 1 family (CAMK1), we were able to discriminate two distinctive conformations: in the first one (the “closed” conformation) the CRE is fully interacting with the KD, and in the second one (the “open” conformation) it is partially interacting with the KD, as shown in Figure 6A. The distance

distribution to the ATP-binding site is shown in Figure 6B. In this family, multiple inhibitory interactions occur when the CRE (structured as a helix–loop–helix segment) interacts with residues of the catalytic site. For the enzyme to change back to the active state, some residues of the CaM-binding domain have to protrude away from the catalytic site. Only then the interaction with CaM will be possible, promoting activation of the kinase [50].

In Figure 6C both conformations are overlapped, the “closed” one being entirely autoinhibitory, and the “open” one where the CRE is starting to extend, so that the kinase can be activated. A more detailed visualization of the contacts between the CRE and the active site can be seen in Figure 6D.

The CRE in CAMK1 is mostly structured and, since it is a helix–loop–helix segment, it is composed primarily of α -helices (Figure S3C).

In the CAMK type 2 subunit delta (CAMK2D) family, the CRE undergoes a conformational change, becoming structured upon binding in the closed conformers. The CRE interacts with the KD, forming an α -helix, while in the open conformation, the CRE is disordered and not interacting with the KD (Figure 7A,B). When activated, the CRE is bound to CaM in the open conformation. Figure 7C depicts the superposition of both close and open conformers, revealing discernible differences between them. Additionally, the figure also provides a detailed view of the spatial proximity between the CRE and the active site of the kinase in the closed conformation (Figure 7D).

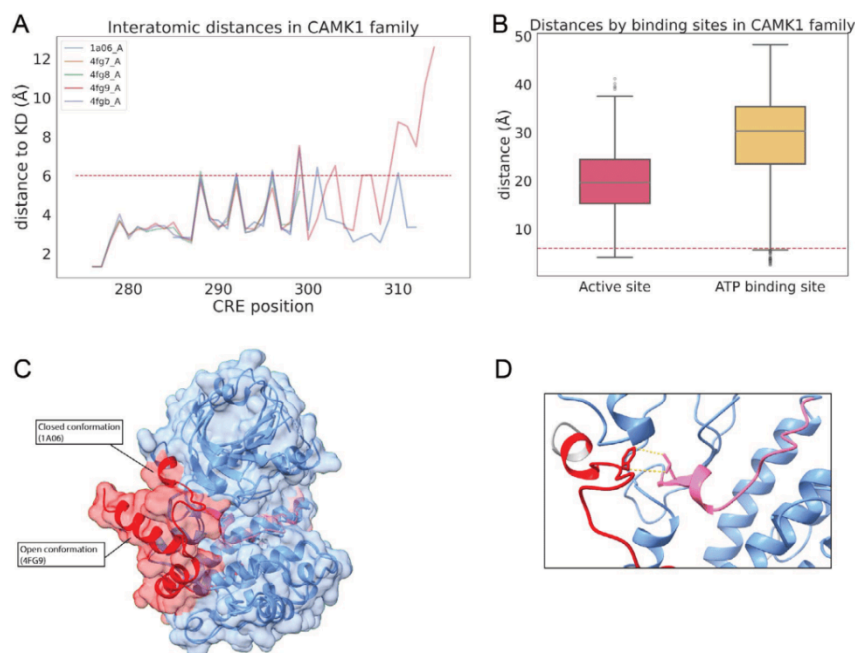


FIGURE 6 | CAMK1 family structural analysis. (A) Interatomic distances for the CAMK1 family, showing the CRE residues (x-axis, CRE positions 276–316 in UniProt reference sequence coordinates, accession: Q91YS8) to the KD residues (y-axis). (B) CRE distance distribution to the active and ATP binding sites of the KD. (C) Superposition of PDB structures, showing the close and open conformations (1A06, 4FG9 PDB accessions correspondingly). KD is shown in light blue and the CRE is colored in red. (D) Closed conformation (PDB 1A06), contacts indicated in dashed yellow lines. Contacts are shown in yellow dashed lines. Molecular figures were generated with ChimeraX [45].

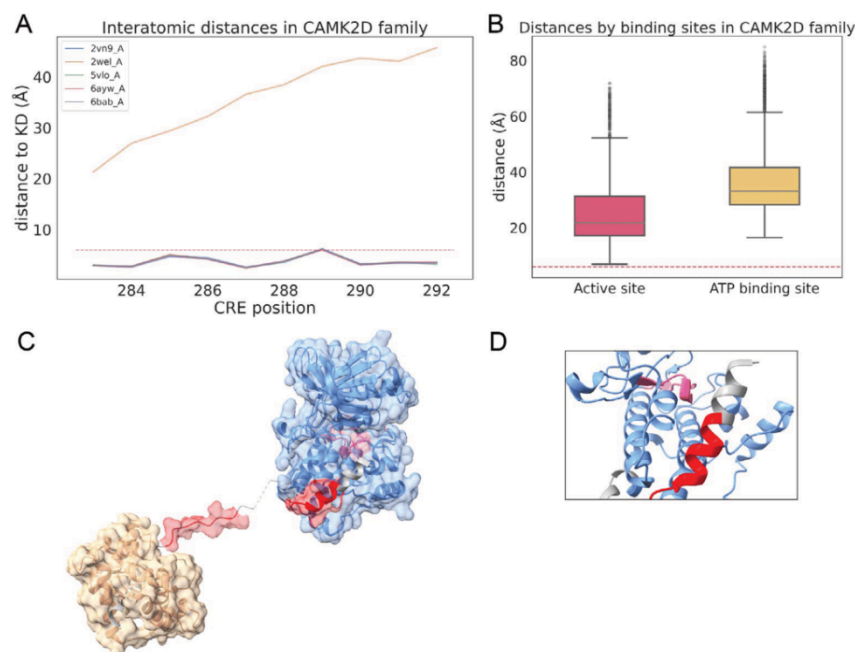


FIGURE 7 | CAMK2D family structural analysis. (A) Interatomic distances to the KD residues (y-axis) in the CAMK2D family, showing the CRE residues in the x-axis (CRE positions 283–292 in UniProt reference sequence coordinates, accession: Q13557). (B) Distance distribution of the CRE residues to the active and ATP binding sites of the KD. (C) Structural superposition of the closed and open conformations (PDB accessions 6BAB and 2WEL, respectively). (D) Zoom of the CRE forming an α -helix, situated at contact distance to the KD domain (PDB: 6BAB). Molecular figures were generated with ChimeraX [45].

We then explored how AF would resolve this structure, given the pronounced differences between the two PDB structures. As expected, the AF model is in the closed conformation, with the CRE structured and interacting with the KD, rendering a globular domain (Figure S6A).

2.6 | Mutations Analysis

Mutations in protein kinases exert a wide range of effects, influencing various aspects of their functionality, and thus are associated with diverse pathologies. Disruptions in the mechanisms of action of kinases can have profound and potentially life-threatening consequences.

Our analysis collated 927 (706 pathogenic and 221 benign) missense variants in our human kinases dataset (89 proteins), all of which were associated with disease mutations reported in ClinVar [51]. The distribution of the mutational rates in the CRE, ATP binding site, active site and the rest of the KD, is presented in Figure S7A. The CRE, ATP binding, and active sites exhibit significant differences ($p \leq 1.00.10^{-4}$) compared to the rest of the KD. All of the studied regions have more pathogenic than benign mutations, and, except for the active site, the differences in the mutation rates are significant (Table S2). However, it is worth noting that the active site shows the highest rate of pathogenic mutations. As an example, the kinases CSF1R and FGFR1 (0.62 and 0.54 mutation rates respectively, Figure S7A), are associated to severe disorders like leukoencephalopathy, a neurodegenerative disease [52, 53], and hypogonadotropic hypogonadism, a disorder characterized by low levels of serum gonadotropins leading to incomplete sexual maturation [54, 55]. Also, the number of mutations in the active site is 5-fold higher than expected (Chi-squared test, Table S3).

There is a subset of kinases with a high number of pathogenic mutations, as an example, BRAF has 132 mutations, all of them pathogenic. Also, there is a subset with kinases harboring all benign mutations, for example MAP2K3 (Figure S7B).

The most frequently associated disease terms with this set of proteins include various types of cancer, such as gastrointestinal stromal tumors, hereditary cancer-predisposing syndromes, and EGFR-related lung cancers. Additionally, terms related to diseases affecting signaling pathways, like RASopathies, constitute a group of syndromes caused by pathogenic mutations in components of the RAS/mitogen-activated protein kinase (RAS/MAPK) pathway [56].

2.7 | Kinases in Membraneless Organelles

Among the 91 human kinases harboring a CREs in our dataset, 67 were identified as being involved in LLPS and present in an MLO (MLOsMetaDB) [57]. Some of the most densely populated condensates with CRE-containing kinases include the nucleolus, centrosome/spindle pole body, and postsynaptic density (PSD) based on search in the.

Examples include FLT1 and FLT3 kinases in Cajal bodies, ABL1 and EGFR in Promyelocytic leukemia (PML) nuclear

bodies, MAPK1 and CAMK2D in PSDs, as well as PRKCA and WNK1 in stress granules (SGs), among others (Table 1). The numerous and diverse examples highlight the essential role of kinases in orchestrating precise spatiotemporal regulation in cells.

Kinases are found in cell type-specific condensates, such as PSDs in neurons, as well as in condition-dependent condensates like SGs during stress situations. Kinases have been shown to have an important role in both physiological and pathological MLOs. As an example, the phosphorylation of FUS by a kinase disrupts the intra- and intermolecular interactions that maintain pathological aggregation in cells [58].

Regarding their role within the MLO, Kinases are known as regulators in a plethora of processes, however, surprisingly in our dataset, most of them are found as clients (Figure 8). RPS6KB1 and PRKAA2 have a regulator role in SGs and CAMK2G in centrosomes and PSDs. On the other hand, the only LLPS driver in our dataset is ABL1 in PML body driver.

3 | Discussion

We conducted a comprehensive structural analysis of the auto-regulatory mechanisms across kinase families. We initially aimed to identify as many auto-regulatory kinases as possible to build a comprehensive dataset. Our search yielded 286 representative kinases from 250 kinase families, which allowed us to gather 8411 orthologous proteins for further analysis.

Our findings suggest that the self-regulatory mechanisms vary significantly among kinase families, even within closely related families (such as CAMK1 and CAMK2 families) the differences are substantial. This diversity arises from the wide spectrum of CREs present in these enzymes and the variety of possible ways these regions can bind and inhibit the kinase activity. The sizes of CREs span from a few residues to entire domains, and they can be either disordered or structured. Structured CREs can range from short α -helices and β -sheets to fully folded domains, such as the SH2 and SH3 domains of ABL1 kinase. Disordered CREs can provide advantages in the fine-tuning process of autoinhibition due to their ability to adopt multiple binding modes [8, 59, 60].

When multiple structures were experimentally solved for a particular protein family, we measured the interatomic distances between the CRE and the KD residues. The majority of kinases solved structures are in the auto-inhibited state, where the CRE is in direct contact with the KD. This observation suggests a preference among kinases to remain in a closed conformation, representing an inactive state, which contrasts with the fact that activation is needed for signal transduction. It is well known that the constitutive activation of kinases can lead to conditions linked to various diseases, including neurodegeneration, inflammation, autoimmunity, cancer, and cardiovascular diseases [17, 61–63], among others. While the observation is strongly supported due to the number of structures analyzed, caution should be exercised in concluding that the closed/inactive conformation is predominant due to the limited structural data currently available.

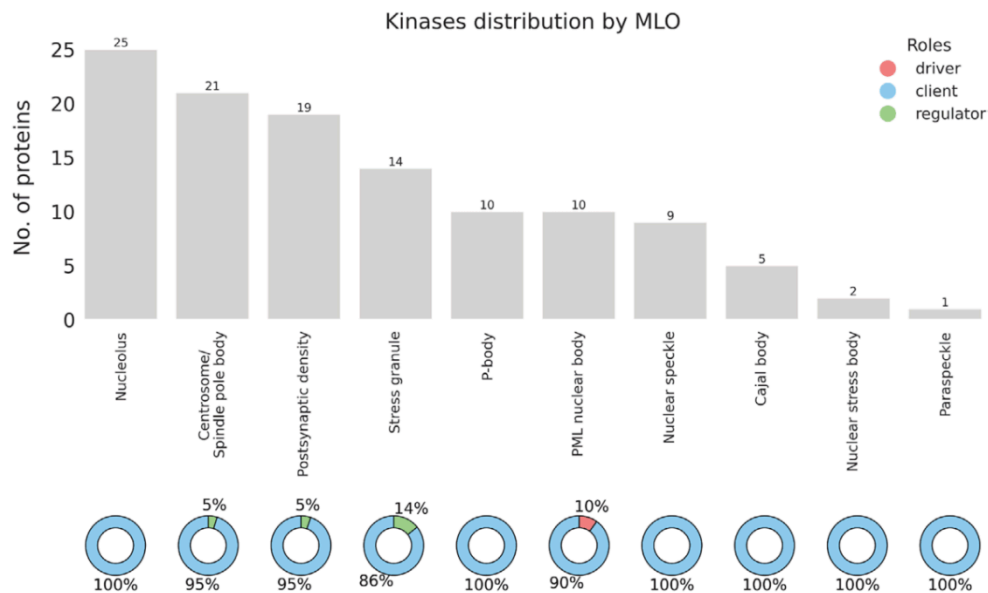


FIGURE 8 | Kinases distribution by MLO. Number of kinases found in each MLO and their biological role proportions given by MLOsMetaDB [57].

It is known that post-translational modifications have influence over kinase regulation. In our set of proteins, a substantial fraction of these modifications situated in the CRE are in contact distance to the KD. This proximity would cause the interactions to hide the modification sites and, therefore, modulate the kinase activity.

In the case of the CAMK family, the activation of its members depends on the concentration of Ca^{2+} /Calmodulin and the subsequent binding to their respective interaction sites. The CRE interacting with the active site in the CAMK family of kinases maintains the protein's inhibition while the binding of calmodulin leads to their activation. CAMK2D also requires phosphorylation to stay active. Members of the CAMK family have been associated with cancer when constitutively active [64, 65].

Mutations in kinases have a wide-ranging effects including differences in the stabilities between the active and inactive states [19], interference in the dimerization capability of some receptor kinases, for example, EGFR by mutations in the cysteine-rich domains [66], and constitutive activation [14], among others. As an example, mutations in the SH2 and SH3 domain in ABL1 interfere with the normal autoinhibition mechanism, facilitating the development of cancer [67, 68].

Mutations located in the CRE and in the activation loop in FLT3, can generate a constitutively activated protein and become pathogenic, for example, in acute myelogenous leukemia [44, 69–71].

Our analysis revealed that mutations occurring in the ATP binding, active, and CRE sites are significantly associated with pathogenicity and often appearing in higher frequencies. This correlation aligns with the intuitive understanding that disrupting these sites would likely perturb kinase activity and subsequently impact the signaling cascade they trigger.

Spatiotemporal confluence is key in many pathways and processes in which kinases are involved, hence condensates can act as catalysts for kinase activation recruiting the enzymes and concentrating their substrates [72]. As an example, WNK1 controls physiologic processes under stress conditions like hypertonicity, forming condensates capable of driving the flow of ions, helping to restore the normal cell volume [73].

It has been proved that the fusion BCR-ABL oncogenic protein not only accumulates in stress granules but also that this granular localization of the enzyme is essential for its leukemogenic activity [74]. Moreover, ABL1 together with MYC and ESR1 has been shown in aberrant proliferation signaling [75]. Another example of the importance of kinases in the LLPS process is DYRK3, a kinase that can drive the dissolution of SGs to release another kinase, mTORC1 [76] with all the implications that this event can cause.

Orti et al. conducted an extensive analysis of LLPS proteins, demonstrating that kinases typically serve as regulators of MLOs [77]. However, intriguingly, our investigation, limited to kinases that exhibit autoinhibition of their activity, reveals that they predominantly act as clients within MLOs, with the exception of ABL1, which is a driver. The underlying reasons for this divergent behavior among this subset of kinases, deviating from their conventional regulatory roles, warrant deeper examination and analysis.

Consequently, targeting the process of phase separation opens the possibility of having other control mechanisms that could constitute a novel therapeutic strategy in cancer [76].

Through this comprehensive analysis, we uncover several aspects oriented to shed light on the kinase self regulatory mechanism, such as disorder content, conformational diversity,

LLPS, PTMs, mutations, among others. We show that the *cis*-regulation might be accomplished by one or a combination of several factors.

While certain mechanisms share commonalities, the regulation varies distinctly and specifically across kinase families, necessitating a comprehensive, case-by-case analysis encompassing both their inactive and active states to elucidate crucial details of this paramount cellular regulatory mechanism.

The insights gleaned from this research contribute significantly to elucidating the *cis* regulation of kinases, laying essential groundwork for future studies aimed at unraveling the molecular mechanisms underlying their dysregulation and exploring potential therapeutic avenues.

4 | Materials and Methods

4.1 | Collection of Sequences Containing *cis*-Regulatory Elements

Kinase proteins were collected from Yeon [78] (a set of curated auto inhibited proteins) and from UniProtKB (release-2022_03), using the following keywords: “Autoinhibitory,” “Autoregulatory,” “AIS” (autoinhibitory segment), “Regulatory,” and “Auto-inhibitory” ending up with a set of 280 kinases from 34 organisms with a known CRE (Table S1), accounting for 286 CREs in this dataset. There are 90 human kinases in the dataset that we will use to study disease related mutations. We then clustered the sequences at 60% of identity to avoid redundancy and take one protein for each cluster as the reference one, which resulted in 250 clusters of kinases.

4.2 | Collection of Orthologous Sequences

We collected the orthologous sequences of each reference protein from OMaDB [79] and OrthoInspector [80], using a homology transfer pipeline, which generates good quality multiple sequence alignments (MSAs) [36]. We ended up with a total of 8411 sequences from 663 organisms distributed along 250 MSAs. The group of sequences obtained when querying the database with each reference protein, are considered a family for the purpose of this analysis.

4.3 | Structural Analysis

All kinase structures, covering at least one residue from the kinase domain (KD) and one residue from the CRE, were retrieved from the PDB [81]. Fifty-eight out of 250 MSAs contain at least one such structure in the PDB, accounting for a total of 830 structures. Forty-three MSAs contain two or more structures, such multiple instances enabling us to analyze conformational diversity within a family [82].

The KD and CRE mapping into the protein structure was done with the SIFT tool [83]. We mapped the residue numbering from UniProt to the PDB coordinates through sequence pairwise alignments using *in-house* scripts.

The distance calculation between CRE and KD amino acids was also done with *in-house* scripts. Two residues are considered to be in contact if any two heavy atoms from the CRE and the KD are closer than 6 Å. For the analysis we filtered the shortest distance among two heavy atoms of a given pair of CRE-KD residues.

4.4 | AlphaFold Models

AlphaFold (AF) models were downloaded from AlphaFoldDB [47] for single-chain cases. Complexes were predicted with ColabFold using the mmseq2 method and pdb100 templates [84].

4.5 | Disorder and Secondary Structure Content

Scores of structural disorder and secondary structure were retrieved from the MobiDB database (version 5.0) using MobiDB-lite and FESS tools, respectively [85].

4.6 | Molecular Features

We extracted relevant annotations within the kinase domain (KD), such as the ATP-binding site and active site, from the InterPro database [86]. Subsequently, we mapped these annotations onto the respective protein sequences. PTMs, including phosphorylation, ubiquitination, acetylation, and methylation, were retrieved from the PhosphoSitePlus database [87]. Additionally, annotations related to liquid-liquid phase separation for human kinases were acquired from MLOsmetaDB [57].

4.7 | Disease Mutations

Clinically relevant mutations in the human kinases were obtained from the ClinVar database (release 2023-12). Variants with “not provided” or “not specified” disease association were removed. For simplicity, variations annotated as *Likely benign* and *Benign/Likely benign* were integrated into the *Benign* category, and variations annotated as *Likely pathogenic*, *Pathogenic/Likely pathogenic*, *drug response* into the *Pathogenic* category. Only missense mutations were kept because more complex mutation types, such as insertions, deletions, indels or frameshifts could potentially alter the entire protein sequence. Mutational rates were calculated for each protein by dividing the number of mutations falling in a defined molecular feature (CRE, KD, active site and ATP-binding site) by their length.

Author Contributions

Alvaro M. Navarro: conceptualization, investigation, writing – original draft, methodology, validation, visualization, writing – review and editing, formal analysis, data curation. **Macarena Alonso:** writing – original draft, validation, visualization, data curation. **Elizabeth Martínez-Pérez:** methodology, formal analysis, supervision. **Tamas Lazar:** writing – review and editing, formal analysis.

Toby J. Gibson: conceptualization, investigation, methodology. **Javier A. Iserte:** methodology, validation, formal analysis. **Peter Tompa:** conceptualization, investigation, formal analysis. **Cristina Marino-Buslje:** conceptualization, funding acquisition, investigation, writing – original draft, writing – review and editing, formal analysis, project administration, supervision, resources.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data is available in the main manuscript, the [Supporting Information](#) and the Table S1 is in: <https://docs.google.com/spreadsheets/d/1acQjsZezDN3zHmjS0rGW704ZF5di7BNfIzCaq99XPx/edit?usp=sharing>.

Peer Review

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.26751>.

References

1. M. Qi and E. A. Elion, "MAP Kinase Pathways," *Journal of Cell Science* 118, no. Pt 16 (2005): 3569–3572.
2. M. Raman, W. Chen, and M. H. Cobb, "Differential Regulation and Properties of MAPKs," *Oncogene* 26, no. 22 (2007): 3100–3112.
3. L. Ding, J. Cao, W. Lin, et al., "The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer," *International Journal of Molecular Sciences* 21, no. 6 (2020): 1960, <https://doi.org/10.3390/ijms21061960>.
4. A. K. Patel, R. P. Yadav, V. Majava, I. Kursula, and P. Kursula, "Structure of the Dimeric Autoinhibited Conformation of DAPK2, a Pro-Apoptotic Protein Kinase," *Journal of Molecular Biology* 409, no. 3 (2011): 369–383.
5. B. Simon, A.-S. Huart, K. Temmerman, et al., "Death-Associated Protein Kinase Activity Is Regulated by Coupled Calcium/Calmodulin Binding to Two Distinct Sites," *Structure* 24, no. 6 (2016): 851–861.
6. E. Park, S. Rawson, K. Li, et al., "Architecture of Autoinhibited and Active BRAF-MEK1-14-3-3 Complexes," *Nature* 575, no. 7783 (2019): 545–550.
7. S. R. Hubbard, "Juxtamembrane Autoinhibition in Receptor Tyrosine Kinases," *Nature Reviews Molecular Cell Biology* 5, no. 6 (2004): 464–471.
8. T. Trudeau, R. Nassar, A. Cumberworth, E. T. C. Wong, G. Woollard, and J. Gsponer, "Structure and Intrinsic Disorder in Protein Autoinhibition," *Structure* 21, no. 3 (2013): 332–341.
9. M. T. Swilius and M. N. Waxham, "Ca(2+)/calmodulin-Dependent Protein Kinases," *Cellular and Molecular Life Sciences: CMLS* 65, no. 17 (2008): 2637–2657.
10. B. Nagar, O. Hantschel, M. A. Young, et al., "Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase," *Cell* 112, no. 6 (2003): 859–871.
11. T. J. Boggon and M. J. Eck, "Structure and Regulation of Src Family Kinases," *Oncogene* 23, no. 48 (2004): 7918–7927.
12. R. Reinhardt and T. A. Leonard, "A Critical Evaluation of Protein Kinase Regulation by Activation Loop Autophosphorylation," *Elife* 12, no. July (2023): e88210, <https://doi.org/10.7554/eLife.88210>.
13. R. Roskoski, "Src Protein-Tyrosine Kinase Structure and Regulation," *Biochemical and Biophysical Research Communications* 324, no. 4 (2004): 1155–1164.
14. K. A. Rygiel and J. M. Elkins, "Recent Advances in the Structural Biology of Tyrosine Kinases," *Current Opinion in Structural Biology* 82, no. October (2023): 102665.
15. P. Chakrabarti and D. Chakravarty, "Intrinsically Disordered Proteins/Regions and Insight Into Their Biomolecular Interactions," *Bio-physical Chemistry* 283 (2022): 106769.
16. N. Kannan, N. Haste, S. S. Taylor, and A. F. Neuwald, "The Hallmark of AGC Kinase Functional Divergence Is Its C-Terminal Tail, a cis-Acting Regulatory Module," *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 4 (2007): 1272–1277.
17. Z. Du and C. M. Lovly, "Mechanisms of Receptor Tyrosine Kinase Activation in Cancer," *Molecular Cancer* 17, no. 1 (2018): 58.
18. L. R. Klug, J. D. Kent, and M. C. Heinrich, "Structural and Clinical Consequences of Activation Loop Mutations in Class III Receptor Tyrosine Kinases," *Pharmacology & Therapeutics* 191 (2018): 123–134.
19. R. Nussinov, C.-J. Tsai, and H. Jang, "Autoinhibition Can Identify Rare Driver Mutations and Advise Pharmacology," *FASEB Journal* 34, no. 1 (2020): 16–29.
20. M. A. Pufall and B. J. Graves, "Autoinhibitory Domains: Modular Effectors of Cellular Regulation," *Annual Review of Cell and Developmental Biology* 18 (2002): 421–462.
21. S. C. Artim, J. M. Mendrola, and M. A. Lemmon, "Assessing the Range of Kinase Autoinhibition Mechanisms in the Insulin Receptor Family," *Biochemical Journal* 448, no. 2 (2012): 213–220.
22. A. Feichtner, V. Kugler, S. Schwaighofer, T. Nuener, J. Fleischmann, and E. Stefan, "Tracking Mutation and Drug-Driven Alterations of Oncokinase Conformations," *Memo* 15, no. 2 (2022): 137–142.
23. R. S. Gormal, R. Martinez-Marmol, A. J. Brooks, and F. A. Meunier, "Location, Location, Location: Protein Kinase Nanoclustering for Optimised Signalling Output," *Elife* 13, no. January (2024): e93902, <https://doi.org/10.7554/eLife.93902>.
24. L. Li, M. Herzog, S. Möbitz, and R. Winter, "Liquid Droplets of Protein LAF1 Provide a Vehicle to Regulate Storage of the Signaling Protein K-Ras4B and Its Transport to the Lipid Membrane," *Physical Chemistry Chemical Physics* 23, no. 9 (2021): 5370–5375.
25. S. Alberti, "The Wisdom of Crowds: Regulating Cell Function Through Condensed States of Living Matter," *Journal of Cell Science* 130, no. 17 (2017): 2789–2796.
26. S. Alberti, A. Gladfelter, and T. Mittag, "Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates," *Cell* 176, no. 3 (2019): 419–434.
27. S. Boeynaems, S. Alberti, N. L. Fawzi, et al., "Protein Phase Separation: A New Phase in Cell Biology," *Trends in Cell Biology* 28, no. 6 (2018): 420–435.
28. C. M. Fare, A. Villani, L. E. Drake, and J. Shorter, "Higher-Order Organization of Biomolecular Condensates," *Open Biology* 11, no. 6 (2021): 210137.
29. A. K. Rai, J. X. Chen, M. Selbach, and L. Pelkmans, "Kinase-Controlled Phase Transition of Membraneless Organelles in Mitosis," *Nature* 559, no. 7713 (2018): 211–216.
30. N. Farahi, T. Lazar, S. J. Wodak, P. Tompa, and R. Pancsa, "Integration of Data from Liquid-Liquid Phase Separation Databases Highlights Concentration and Dosage Sensitivity of LLPS Drivers," *International Journal of Molecular Sciences* 22, no. 6 (2021): 3017, <https://doi.org/10.3390/ijms22063017>.
31. T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, "Intrinsically Disordered Linkers Determine the Interplay Between Phase Separation and Gelation in Multivalent Proteins," *Elife* 6 (2017): e30294, <https://doi.org/10.7554/eLife.30294>.
32. R. Pancsa, W. Vranken, and B. Mészáros, "Computational Resources for Identifying and Describing Proteins Driving Liquid-Liquid

- Phase Separation," *Briefings in Bioinformatics* 22, no. 5 (2021): bbaa408, <https://doi.org/10.1093/bib/bbaa408>.
33. S. F. Banani, A. M. Rice, W. B. Peeples, et al., "Compositional Control of Phase-Separated Cellular Bodies," *Cell* 166, no. 3 (2016): 651–663.
 34. J. A. Ditlev, L. B. Case, and M. K. Rosen, "Who's In and Who's Out-Compositional Control of Biomolecular Condensates," *Journal of Molecular Biology* 430, no. 23 (2018): 4666–4684.
 35. D. Berchtold, N. Battich, and L. Pelkmans, "A Systems-Level Study Reveals Regulators of Membrane-Less Organelles in Human Cells," *Molecular Cell* 72, no. 6 (2018): 1035–1049.e5.
 36. E. Martínez-Pérez, M. Pajkos, C. E. T. Silvio, J. G. Toby, Z. Dosztanyi, and C. Marino-Buslje, "Pipeline for Transferring Annotations Between Proteins Beyond Globular Domains," *Protein Science* 32 (2023): e4655.
 37. D. Piovesan, I. Walsh, G. Minervini, and S. C. E. Tosatto, "FELLS: Fast Estimator of Latent Local Structure," *Bioinformatics* 33, no. 12 (2017): 1889–1891.
 38. M. Fuxreiter, "Classifying the Binding Modes of Disordered Proteins," *International Journal of Molecular Sciences* 21, no. 22 (2020): 8615, <https://doi.org/10.3390/ijms21228615>.
 39. P. Rellos, A. C. W. Pike, F. H. Niesen, et al., "Structure of the CaMKII δ /Calmodulin Complex Reveals the Molecular Mechanism of CaMKII Kinase Activation," *PLoS Biology* 8, no. 7 (2010): e1000426.
 40. F. Gesellchen, O. Bertinetti, and F. W. Herberg, "Analysis of Post-translational Modifications Exemplified Using Protein Kinase A," *Biochimica et Biophysica Acta* 1764, no. 12 (2006): 1788–1800.
 41. S. F. Steinberg, "Post-Translational Modifications at the ATP-Positioning G-Loop That Regulate Protein Kinase Activity," *Pharmacological Research* 135 (2018): 181–187.
 42. Y. Li, L. P. Wan, N.-N. Song, et al., "RNF220-Mediated K63-Linked Polyubiquitination Stabilizes Olig Proteins During Oligodendroglial Development and Myelination," *Science Advances* 10, no. 6 (2024): eadk3931.
 43. Y. Li, H. Wang, B. Sun, et al., "Smurf1 and Smurf2 Mediated Polyubiquitination and Degradation of RNF220 Suppresses Shh-Group Medulloblastoma," *Cell Death & Disease* 14, no. 8 (2023): 494.
 44. F. M. Abu-Duhier, A. C. Goodeve, G. A. Wilson, R. S. Care, I. R. Peake, and J. T. Reilly, "Identification of Novel FLT-3 Asp835 Mutations in Adult Acute Myeloid Leukaemia," *British Journal of Haematology* 113, no. 4 (2001): 983–988.
 45. E. F. Pettersen, T. D. Goddard, C. C. Huang, et al., "UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers," *Protein Science* 30, no. 1 (2021): 70–82.
 46. J. Griffith, J. Black, C. Faerman, et al., "The Structural Basis for Autoinhibition of FLT3 by the Juxtamembrane Domain," *Molecular Cell* 13, no. 2 (2004): 169–178.
 47. M. Varadi, S. Anyango, M. Deshpande, et al., "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space With High-Accuracy Models," *Nucleic Acids Research* 50, no. D1 (2022): D439–D444.
 48. D. V. Laurents, "AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function," *Frontiers in Molecular Biosciences* 9 (2022): 906437.
 49. O. W. Rokhlin, A. F. Taghiyev, K. Ulrich Bayer, et al., "Calcium/Calmodulin-Dependent Kinase II Plays an Important Role in Prostate Cancer Cell Survival," *Cancer Biology & Therapy* 6, no. 5 (2007): 732–742.
 50. J. Goldberg, A. C. Nairn, and J. Kuriyan, "Structural Basis for the Autoinhibition of Calcium/Calmodulin-Dependent Protein Kinase I," *Cell* 84, no. 6 (1996): 875–887.
 51. M. J. Landrum, S. Chitipiralla, G. R. Brown, et al., "ClinVar: Improvements to Accessing Data," *Nucleic Acids Research* 48, no. D1 (2020): D835–D844.
 52. T. Konno, M. Tada, M. Tada, et al., "Haploinsufficiency of CSF-1R and Clinicopathologic Characterization in Patients With HDLS," *Neurology* 82, no. 2 (2014): 139–148.
 53. C. Sundal and Z. K. Wszolek, "CSF1R-Related Adult-Onset Leukoencephalopathy With Axonal Spheroids and Pigmented Glia," in *GeneReviews*[®], eds. M. P. Adam, J. Feldman, G. M. Mirzaa, et al. (Seattle, WA: University of Washington, Seattle, 2012), <https://www.ncbi.nlm.nih.gov/books/NBK100239/>.
 54. M. Nie, Y. Bingqing, R. Chen, et al., "Novel Rare Variants in FGFR1 and Clinical Characteristics Analysis in a Series of Congenital Hypogonadotropic Hypogonadism Patients," *Clinical Endocrinology* 95, no. 1 (2021): 153–162.
 55. N. Xu, Y. Qin, R. H. Reindollar, S. P. T. Tho, P. G. McDonough, and L. C. Layman, "A Mutation in the Fibroblast Growth Factor Receptor 1 Gene Causes Fully Penetrant Normosmic Isolated Hypogonadotropic Hypogonadism," *Journal of Clinical Endocrinology and Metabolism* 92, no. 3 (2007): 1155–1158.
 56. K. A. Rauen, "Defining RASopathy," *Disease Models & Mechanisms* 15, no. 2 (2022): dmm049344, <https://doi.org/10.1242/dmm.049344>.
 57. F. Orti, M. L. Fernández, and C. Marino-Buslje, "MLOsMetaDB, a Meta-Database to Centralize the Information on Liquid-Liquid Phase Separation Proteins and Membraneless Organelles," *Protein Science* 33, no. 1 (2024): e4858.
 58. X. Tong, R. Tang, X. Jin, et al., "Liquid-Liquid Phase Separation in Tumor Biology," *Signal Transduction and Targeted Therapy* 7, no. 1 (2022): 221.
 59. M. Fenton, E. Gregory, and G. Daughdrill, "Protein Disorder and Autoinhibition: The Role of Multivalency and Effective Concentration," *Current Opinion in Structural Biology* 83 (2023): 102705.
 60. G. Gógl, A. P. Kornev, A. Reményi, and S. S. Taylor, "Disordered Protein Kinase Regions in Regulation of Kinase Domain Cores," *Trends in Biochemical Sciences* 44, no. 4 (2019): 300–311.
 61. C. L. Benn and L. A. Dawson, "Clinically Precedented Protein Kinases: Rationale for Their Use in Neurodegenerative Disease," *Frontiers in Aging Neuroscience* 12 (2020): 242.
 62. H. Patterson, R. Nibbs, I. McInnes, and S. Siebert, "Protein Kinase Inhibitors in the Treatment of Inflammatory and Autoimmune Diseases," *Clinical and Experimental Immunology* 176, no. 1 (2014): 1–10.
 63. S. Silnitsky, S. J. S. Rubin, M. Zerihun, and N. Qvit, "An Update on Protein Kinases as Therapeutic Targets-Part I: Protein Kinase C Activation and Its Role in Cancer and Cardiovascular Diseases," *International Journal of Molecular Sciences* 24, no. 24 (2023): 17600, <https://doi.org/10.3390/ijms242417600>.
 64. J. S. Brzozowski and K. A. Skelding, "The Multi-Functional Calcium/Calmodulin Stimulated Protein Kinase (CaMK) Family: Emerging Targets for Anti-Cancer Therapeutic Intervention," *Pharmaceuticals (Basel)* 12, no. 1 (2019): 8, <https://doi.org/10.3390/ph12010008>.
 65. Y.-Y. Wang, R. Zhao, and H. Zhe, "The Emerging Role of CaMKII in Cancer," *Oncotarget* 6, no. 14 (2015): 11725–11734.
 66. J. Cho, "Mechanistic Insights Into Differential Requirement of Receptor Dimerization for Oncogenic Activation of Mutant EGFR and Its Clinical Perspective," *BMB Reports* 53, no. 3 (2020): 133–141.
 67. D. Barilá and G. Superti-Furga, "An Intramolecular SH3-Domain Interaction Regulates c-Abl Activity," *Nature Genetics* 18, no. 3 (1998): 280–282.
 68. B. B. Brasher and R. A. Van Etten, "C-Abl Has High Intrinsic Tyrosine Kinase Activity That Is Stimulated by Mutation of the Src Homology 3 Domain and by Autophosphorylation at Two Distinct Regulatory Tyrosines," *Journal of Biological Chemistry* 275, no. 45 (2000): 35631–35637.
 69. S. Horiike, S. Yokota, M. Nakao, et al., "Tandem Duplications of the FLT3 Receptor Gene Are Associated With Leukemic Transformation of Myelodysplasia," *Leukemia* 11, no. 9 (1997): 1442–1446.

70. C. Thiede, C. Steudel, B. Mohr, et al., "Analysis of FLT3-Activating Mutations in 979 Patients With Acute Myelogenous Leukemia: Association With FAB Subtypes and Identification of Subgroups With Poor Prognosis," *Blood* 99, no. 12 (2002): 4326–4335.
71. Y. Yamamoto, H. Kiyoi, Y. Nakano, et al., "Activating Mutation of D835 Within the Activation Loop of FLT3 in Human Hematologic Malignancies," *Blood* 97, no. 8 (2001): 2434–2439.
72. T. P. López-Palacios and J. L. Andersen, "Kinase Regulation by Liquid-Liquid Phase Separation," *Trends in Cell Biology* 33, no. 8 (2023): 649–666.
73. C. R. Boyd-Shiwerski, D. J. Shiwerski, S. E. Griffiths, et al., "WNK Kinases Sense Molecular Crowding and Rescue Cell Volume via Phase Separation," *Cell* 185, no. 24 (2022): 4488–4506.e20.
74. S. Kashiwagi, Y. Fujioka, T. Kondo, et al., "Localization of BCR-ABL to Stress Granules Contributes to Its Oncogenic Function," *Cell Structure and Function* 44, no. 2 (2019): 195–204.
75. A. K. Chakravarty, D. J. McGrail, T. M. Lozanoski, et al., "Biomolecular Condensation: A New Phase in Cancer Research," *Cancer Discovery* 12, no. 9 (2022): 2031–2043.
76. Z. Liu, Z. Qin, Y. Liu, et al., "Liquid-Liquid Phase Separation: Roles and Implications in Future Cancer Treatment," *International Journal of Biological Sciences* 19, no. 13 (2023): 4139.
77. F. Orti, A. M. Navarro, A. Rabinovich, S. J. Wodak, and C. Marino-Buslje, "Insight Into Membraneless Organelles and Their Associated Proteins: Drivers, Clients and Regulators," *Computational and Structural Biotechnology Journal* 19 (2021): 3964–3977.
78. J. H. Yeon, "Systems-Wide Identification of cis-Regulatory Elements in Proteins," *Cell Systems* 2, no. 2 (2016): 89–100.
79. A. M. Altenhoff, C.-M. Train, K. J. Gilbert, et al., "OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More," *Nucleic Acids Research* 49, no. D1 (2020): D373–D379.
80. Y. Nevers, A. Kress, A. Defosset, et al., "OrthoInspector 3.0: Open Portal for Comparative Genomics," *Nucleic Acids Research* 47, no. D1 (2019): D411–D418.
81. H. M. Berman, J. Westbrook, Z. Feng, et al., "The Protein Data Bank," *Nucleic Acids Research* 28, no. 1 (2000): 235–242.
82. A. M. Monzon, D. J. Zea, C. Marino-Buslje, and G. Parisi, "Homology Modeling in a Dynamical World," *Protein Science* 26, no. 11 (2017): 2195–2206.
83. J. M. Dana, A. Gutmanas, N. Tyagi, et al., "SIFTS: Updated Structure Integration With Function, Taxonomy and Sequences Resource Allows 40-Fold Increase in Coverage of Structure-Based Annotations for Proteins," *Nucleic Acids Research* 47, no. D1 (2019): D482–D489.
84. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: Making Protein Folding Accessible to All," *Nature Methods* 19, no. 6 (2022): 679–682.
85. D. Piovesan, A. Del Conte, D. Clementel, et al., "MobiDB: 10 Years of Intrinsically Disordered Proteins," *Nucleic Acids Research* 51, no. D1 (2023): D438–D444.
86. T. Paysan-Lafosse, M. Blum, S. Chuguransky, et al., "InterPro in 2022," *Nucleic Acids Research* 51, no. D1 (2023): D418–D427.
87. P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek, "PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations," *Nucleic Acids Research* 43 (2015): D512–D520.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Bibliografía

1. Van Schependom J, D'haeseleer M. Advances in neurodegenerative diseases. *J Clin Med*. 2023 Feb 21;12(5).
2. Feigin VL, Vos T, Nichols E, Owolabi MO, Carroll WM, Dichgans M, et al. The global burden of neurological disorders: translating evidence into policy. *Lancet Neurol*. 2020 Mar;19(3):255–65.
3. Wijesekera LC, Leigh PN. Amyotrophic lateral sclerosis. *Orphanet J Rare Dis*. 2009 Feb 3;4:3.
4. del Aguila MA, Longstreth WT, McGuire V, Koepsell TD, van Belle G. Prognosis in amyotrophic lateral sclerosis: a population-based study. *Neurology*. 2003 Mar 11;60(5):813–9.
5. Talbot K. Motor neuron disease: the bare essentials. *Pract Neurol*. 2009 Oct;9(5):303–9.
6. Kiernan MC, Vucic S, Cheah BC, Turner MR, Eisen A, Hardiman O, et al. Amyotrophic lateral sclerosis. *Lancet*. 2011 Mar 12;377(9769):942–55.
7. Nguyen HP, Van Broeckhoven C, van der Zee J. ALS Genes in the Genomic Era and their Implications for FTD. *Trends Genet*. 2018 Jun;34(6):404–23.
8. van Rheenen W, van der Spek RAA, Bakker MK, van Vugt JJFA, Hop PJ, Zwamborn RAJ, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet*. 2021 Dec 6;53(12):1636–48.
9. Oskarsson B, Horton DK, Mitsumoto H. Potential environmental factors in amyotrophic lateral sclerosis. *Neurol Clin*. 2015 Nov;33(4):877–88.
10. Yu B, Pamphlett R. Environmental insults: critical triggers for amyotrophic lateral sclerosis. *Transl Neurodegener*. 2017 Jun 16;6:15.
11. Newell ME, Adhikari S, Halden RU. Systematic and state-of the science review of the role of environmental factors in Amyotrophic Lateral Sclerosis (ALS) or Lou Gehrig's Disease. *Sci Total Environ*. 2022 Apr 15;817:152504.
12. Burrell JR, Halliday GM, Kril JJ, Ittner LM, Götz J, Kiernan MC, et al. The frontotemporal dementia-motor neuron disease continuum. *Lancet*. 2016 Aug 27;388(10047):919–31.
13. Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*. 2006 Oct 6;314(5796):130–3.

14. Meneses A, Koga S, O’Leary J, Dickson DW, Bu G, Zhao N. TDP-43 Pathology in Alzheimer’s Disease. *Mol Neurodegener*. 2021 Dec 20;16(1):84.
15. Neumann M, Tolnay M, Mackenzie IRA. The molecular basis of frontotemporal dementia. *Expert Rev Mol Med*. 2009 Jul 29;11:e23.
16. Ling JP, Pletnikova O, Troncoso JC, Wong PC. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*. 2015 Aug 7;349(6248):650–5.
17. Mehta PR, Brown A-L, Ward ME, Fratta P. The era of cryptic exons: implications for ALS-FTD. *Mol Neurodegener*. 2023 Mar 15;18(1):16.
18. Seddighi S, Qi YA, Brown A-L, Wilkins OG, Bereda C, Belair C, et al. Mis-spliced transcripts generate de novo proteins in TDP-43-related ALS/FTD. *Sci Transl Med*. 2024 Feb 14;16(734):eadg7162.
19. Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D648–53.
20. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res*. 2017 Jan 4;45(D1):D408–14.
21. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021 Jan;30(1):187–200.
22. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D447–52.
23. Kotlyar M, Pastrello C, Ahmed Z, Chee J, Varyova Z, Jurisica I. IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D640–7.
24. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res*. 2009 Jan;37(Database issue):D767–72.
25. Chakraborty C, Bhattacharya M, Lee S-S, Wen Z-H, Lo Y-H. The changing scenario of drug discovery using AI to deep learning: Recent advancement,

- success stories, collaborations, and challenges. *Mol Ther Nucleic Acids*. 2024 Sep 10;35(3):102295.
26. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics*. 2022 Jul 25;16(1):26.
 27. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–324.
 28. Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, Auer F, et al. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med*. 2021 Mar 11;13(1):42.
 29. Yang Y, Tresp V, Wunderle M, Fasching PA. Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks. 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2018. p. 152–62.
 30. Parliament and C. of the European Union. General data protection regulation. [Internet]. 2016 [cited 2026 Mar 31]. Available from: <https://gdpr-info.eu/>
 31. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*. 2015 Jul 10;10(7):e0130140.
 32. Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol*. 2016;1418:93–110.
 33. van Rheenen W, Diekstra FP, Harschnitz O, Westeneng H-J, van Eijk KR, Saris CGJ, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS ONE*. 2018 Jun 25;13(6):e0198874.
 34. Swindell WR, Kruse CPS, List EO, Berryman DE, Kopchick JJ. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J Transl Med*. 2019 May 22;17(1):170.
 35. Nachun D, Ramos EM, Karydas A, Dokuru D, Gao F, Yang Z, et al. Systems-level analysis of peripheral blood gene expression in dementia patients reveals an innate immune response shared across multiple disorders. *BioRxiv*. 2019 Dec 13;
 36. Park HK, Chung SJ. New Perspective on Parkinsonism in Frontotemporal Lobar Degeneration. *J Mov Disord*. 2013 May 1;6(1):1–8.

37. Villaveces JM, Jiménez RC, Porras P, Del-Toro N, Duesbury M, Dumousseau M, et al. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. Database (Oxford). 2015 Feb 4;2015.
38. Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods. 2011 Jul;8(7):528–9.
39. Monti F, Boscaini D, Masci J, Rodolà E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model CNNs. arXiv. 2016;
40. Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. arXiv. 2016;
41. Breiman L. Random Forests. Springer Science and Business Media LLC. 2001;45:5–32.
42. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv. 2016;
43. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.
44. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009 Mar 15;25(6):765–71.
45. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. SciPy; 2008. p. 11–5.
46. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc. 2021 Mar;1(3):e90.
47. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The reactome pathway knowledgebase 2024. Nucleic Acids Res. 2024 Jan 5;52(D1):D672–8.
48. Karp PD, Midford PE, Caspi R, Khodursky A. Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. BMC Genomics. 2021 Mar 16;22(1):191.
49. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019 Feb;14(2):482–517.

50. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. *PLoS Comput Biol*. 2022 Aug 11;18(8):e1010348.
51. Pedotti P, 't Hoen PAC, Vreugdenhil E, Schenk GJ, Vossen RH, Ariyurek Y, et al. Can subtle changes in gene expression be consistently detected with different microarray platforms? *BMC Genomics*. 2008 Mar 10;9:124.
52. Kamelgarn M, Chen J, Kuang L, Jin H, Kasarskis EJ, Zhu H. ALS mutations of FUS suppress protein translation and disrupt the regulation of nonsense-mediated decay. *Proc Natl Acad Sci USA*. 2018 Dec 18;115(51):E11904–13.
53. Lamark T, Johansen T. Aggrephagy: selective disposal of protein aggregates by macroautophagy. *Int J Cell Biol*. 2012 Mar 22;2012:736905.
54. Dobson-Stone C, Hallupp M, Shahheydari H, Ragagnin AMG, Chatterton Z, Carew-Jones F, et al. CYLD is a causative gene for frontotemporal dementia - amyotrophic lateral sclerosis. *Brain*. 2020 Mar 1;143(3):783–99.
55. Sato Y, Terawaki S, Oikawa D, Shimizu K, Okina Y, Ito H, et al. Involvement of heterologous ubiquitination including linear ubiquitination in Alzheimer's disease and amyotrophic lateral sclerosis. *Front Mol Biosci*. 2023 Jan 16;10:1089213.
56. Tedeschi G, Trojsi F, Tessitore A, Corbo D, Sagnelli A, Paccone A, et al. Interaction between aging and neurodegeneration in amyotrophic lateral sclerosis. *Neurobiol Aging*. 2012 May;33(5):886–98.
57. Beers DR, Appel SH. Immune dysregulation in amyotrophic lateral sclerosis: mechanisms and emerging therapies. *Lancet Neurol*. 2019 Feb;18(2):211–20.
58. Chakraborty A, Diwan A. Biomarkers and molecular mechanisms of Amyotrophic Lateral Sclerosis. *AIMS Neurosci*. 2022 Nov 10;9(4):423–43.
59. Dodge JC, Treleaven CM, Fidler JA, Tamsett TJ, Bao C, Searles M, et al. Metabolic signatures of amyotrophic lateral sclerosis reveal insights into disease pathogenesis. *Proc Natl Acad Sci USA*. 2013 Jun 25;110(26):10812–7.
60. Guégan C, Przedborski S. Programmed cell death in amyotrophic lateral sclerosis. *J Clin Invest*. 2003 Jan 15;111(2):153–61.
61. Neel DV, Basu H, Gunner G, Chiu IM. Catching a killer: Mechanisms of programmed cell death and immune activation in Amyotrophic Lateral Sclerosis. *Immunol Rev*. 2022 Oct;311(1):130–50.

62. Abramzon YA, Fratta P, Traynor BJ, Chia R. The overlapping genetics of amyotrophic lateral sclerosis and frontotemporal dementia. *Front Neurosci.* 2020 Feb 5;14:42.
63. Pisciotanni A, Croci L, Lauria F, Marullo C, Savino E, Ambrosi A, et al. Neuronal models of TDP-43 proteinopathy display reduced axonal translation, increased oxidative stress, and defective exocytosis. *Front Cell Neurosci.* 2023 Nov 13;17:1253543.
64. Parakh S, Atkin JD. Protein folding alterations in amyotrophic lateral sclerosis. *Brain Res.* 2016 Oct 1;1648(Pt B):633–49.
65. Cunha-Oliveira T, Montezinho L, Mendes C, Firuzi O, Saso L, Oliveira PJ, et al. Oxidative stress in amyotrophic lateral sclerosis: pathophysiology and opportunities for pharmacological intervention. *Oxid Med Cell Longev.* 2020 Nov 15;2020:5021694.
66. Dashtimian AR, Darvishi FB, Arnold WD. Chronological and biological aging in amyotrophic lateral sclerosis and the potential of senolytic therapies. *Cells.* 2024 May 28;13(11).
67. Konopka A, Atkin JD. DNA damage, defective DNA repair, and neurodegeneration in amyotrophic lateral sclerosis. *Front Aging Neurosci.* 2022 Apr 27;14:786420.
68. Shu Y-H, Lu X-M, Wei J-X, Xiao L, Wang Y-T. Update on the role of p75NTR in neurological disorders: A novel therapeutic target. *Biomed Pharmacother.* 2015 Dec;76:17–23.
69. Mattson MP. NF-kappaB in the survival and plasticity of neurons. *Neurochem Res.* 2005;30(6–7):883–93.
70. Dresselhaus EC, Meffert MK. Cellular Specificity of NF-κB Function in the Nervous System. *Front Immunol.* 2019 May 9;10:1043.
71. Squillace S, Salvemini D. Toll-like receptor-mediated neuroinflammation: relevance for cognitive dysfunctions. *Trends Pharmacol Sci.* 2022 Sep;43(9):726–39.
72. Zhao W, Liu Z, Wu J, Liu A, Yan J. Potential targets of microglia in the treatment of neurodegenerative diseases: mechanism and therapeutic implications. *Neural Regen Res.* 2025 Mar 25;

73. Petitpain N, Devos D, Bagheri H, Rocher F, Gouraud A, Masmoudi K, et al. Is TNF inhibitor exposure a risk factor for amyotrophic lateral sclerosis? *Fundam Clin Pharmacol*. 2019 Dec;33(6):689–94.
74. McCombe PA, Lee JD, Woodruff TM, Henderson RD. The peripheral immune system and amyotrophic lateral sclerosis. *Front Neurol*. 2020 Apr 21;11:279.
75. Abdi G, Jain M, Patil N, Upadhyay B, Vyas N, Dwivedi M, et al. 14-3-3 proteins-a moonlight protein complex with therapeutic potential in neurological disorder: in-depth review with Alzheimer's disease. *Front Mol Biosci*. 2024 Feb 5;11:1286536.
76. Uozumi R, Mori K, Gotoh S, Miyamoto T, Kondo S, Yamashita T, et al. PABPC1 mediates degradation of C9orf72-FTLD/ALS GGGGCC repeat RNA. *iScience*. 2024 Mar 15;27(3):109303.
77. Basu M, Wang K, Ruppin E, Hannenhalli S. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci Adv*. 2021 Apr 2;7(14).
78. Mitrea DM, Kriwacki RW. Phase separation in biology; functional organization of a higher order. *Cell Commun Signal*. 2016 Jan 5;14:1.
79. Spann S, Tereshchenko M, Mastromarco GJ, Ihn SJ, Lee HO. Biomolecular condensates in neurodegeneration and cancer. *Traffic*. 2019 Dec;20(12):890–911.
80. Aguzzi A, Altmeyer M. Phase separation: linking cellular compartmentalization to disease. *Trends Cell Biol*. 2016 Jul;26(7):547–58.
81. Chen T, Tang G, Li T, Yanghong Z, Hou C, Du Z, et al. PhaSeDis: A Manually Curated Database of Phase Separation-disease Associations and Corresponding Small Molecules. *Genomics Proteomics Bioinformatics*. 2025 May 10;23(1).
82. Orti F, Fernández ML, Marino-Buslje C. MLOsMetaDB, a meta-database to centralize the information on Liquid-liquid phase separation proteins and Membraneless organelles. *BioRxiv*. 2023 Jul 25;
83. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D360–7.
84. You K, Huang Q, Yu C, Shen B, Sevilla C, Shi M, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D354–9.

85. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res.* 2020 Jan 8;48(D1):D288–95.
86. Baxi EG, Thompson T, Li J, Kaye JA, Lim RG, Wu J, et al. Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nat Neurosci.* 2022 Feb 3;25(2):226–37.
87. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009 Jan;10(1):57–63.
88. Navarro AM, Alonso M, Martínez-Pérez E, Lazar T, Gibson TJ, Iserte JA, et al. Unveiling the Complexity of cis-Regulation Mechanisms in Kinases: A Comprehensive Analysis. *Proteins.* 2025 Mar;93(3):575–87.



Material suplementario



El código utilizado para el desarrollo de esta tesis, así como las tablas suplementarias de vías biológicas enriquecidas (cuya información no se encuentra en su totalidad en el presente manuscrito por razones de espacio), están disponibles en el siguiente enlace:

https://github.com/malonso94/clasificacion_ELA_DFT

10.1 Tablas de rendimiento de la clasificación con redes PPI de MLOs

10.1.1 IntAct

	Gránulo de estrés							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.900	0.867	0.767	0.688	0.675	0.611	0.711	0.875
RL	0.895	0.838	0.739	0.739	0.752	0.706	0.748	0.796
RF	0.864	0.812	0.632	0.522	0.659	0.611	0.627	0.604
XGBoost	0.843	0.797	0.651	0.596	0.688	0.638	0.707	0.808

	Paraspeckle							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.828	0.800	0.625	0.521	0.620	0.538	0.457	0.357
RL	0.841	0.811	0.650	0.565	0.618	0.628	0.704	0.811
RF	0.845	0.797	0.611	0.500	0.570	0.548	0.565	0.543
XGBoost	0.791	0.772	0.585	0.522	0.551	0.538	0.552	0.524

	Nucleolo							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.905	0.860	0.753	0.667	0.707	0.621	0.732	0.951
RL	0.897	0.851	0.744	0.696	0.698	0.635	0.741	0.970
RF	0.870	0.811	0.632	0.511	0.652	0.595	0.586	0.524
XGBoost	0.823	0.784	0.622	0.565	0.632	0.595	0.638	0.652

	Cuerpo P							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.847	0.813	0.674	0.583	0.631	0.563	0.498	0.396
RL	0.854	0.811	0.681	0.596	0.695	0.558	0.444	0.323
RF	0.822	0.797	0.595	0.478	0.598	0.568	0.545	0.476
XGBoost	0.768	0.784	0.585	0.522	0.586	0.551	0.563	0.530

10.1.2 BioGRID

	Gránulo de estrés							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.911	0.847	0.729	0.667	0.720	0.661	0.750	0.927
RL	0.907	0.865	0.762	0.739	0.754	0.698	0.759	0.878
RF	0.868	0.824	0.650	0.522	0.672	0.625	0.646	0.634
XGBoost	0.865	0.824	0.674	0.609	0.706	0.645	0.696	0.762

	Paraspeckle							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.877	0.833	0.706	0.667	0.693	0.638	0.725	0.872
RL	0.892	0.838	0.723	0.696	0.695	0.615	0.729	0.951
RF	0.865	0.811	0.641	0.522	0.638	0.591	0.598	0.555
XGBoost	0.817	0.784	0.612	0.553	0.608	0.578	0.621	0.628

	Nucleolo							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.894	0.840	0.729	0.667	0.710	0.659	0.746	0.924
RL	0.908	0.851	0.756	0.708	0.725	0.581	0.722	1
RF	0.867	0.824	0.649	0.522	0.676	0.621	0.641	0.622
XGBoost	0.839	0.797	0.637	0.583	0.625	0.591	0.640	0.652

	Cuerpo P							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.863	0.807	0.696	0.625	0.671	0.630	0.695	0.780
RL	0.853	0.824	0.690	0.652	0.685	0.638	0.625	0.555
RF	0.840	0.797	0.597	0.478	0.625	0.578	0.568	0.506
XGBoost	0.776	0.770	0.579	0.522	0.607	0.571	0.610	0.607

10.1.3 HIPPIE

	Gránulo de estrés							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.906	0.840	0.727	0.667	0.696	0.630	0.735	0.948
RL	0.905	0.865	0.773	0.739	0.771	0.711	0.767	0.869
RF	0.870	0.824	0.649	0.522	0.671	0.621	0.652	0.665
XGBoost	0.872	0.824	0.682	0.625	0.695	0.636	0.689	0.738

	Paraspeckle							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.860	0.813	0.682	0.583	0.659	0.616	0.693	0.796
RL	0.868	0.824	0.708	0.652	0.716	0.661	0.740	0.893
RF	0.860	0.811	0.632	0.511	0.628	0.581	0.574	0.524
XGBoost	0.805	0.784	0.605	0.522	0.605	0.581	0.615	0.616

	Nucleolo							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.909	0.847	0.729	0.646	0.674	0.625	0.733	0.963
RL	0.918	0.865	0.771	0.739	0.726	0.611	0.734	0.988
RF	0.863	0.824	0.649	0.532	0.673	0.616	0.628	0.598
XGBoost	0.835	0.797	0.650	0.583	0.644	0.608	0.655	0.692

	Cuerpo P							
	Test				Validación			
	AUC	Acc	F1-Score	Recall	AUC	Acc	F1-Score	Recall
GCNN	0.859	0.807	0.674	0.625	0.652	0.618	0.664	0.701
RL	0.848	0.811	0.682	0.625	0.693	0.618	0.576	0.473
RF	0.849	0.797	0.602	0.478	0.626	0.581	0.554	0.482
XGBoost	0.780	0.770	0.582	0.478	0.618	0.578	0.594	0.567

10.2 Lista de genes más relevantes para la clasificación de pacientes de ELA con GCNN y RL

Genes más relevantes luego de 10 clasificaciones de pacientes con ELA					
GCNN			Regresión Logística		
IntAct	BioGRID	HIPPIE	IntAct	BioGRID	HIPPIE
ACTB	ACTB	ACTB	ABCA1	ABCA1	ABCA1
ALPP	AHR	AHR	ACTA2	ACTA2	ACTA2
BANP	CCND3	ARHGDIB	AIF1	ACTG1	ARGLU1
BASP1	CDC37	CCND3	AKR1C3	ARGLU1	BRI3
CCND3	CEBPB	CCNI	APOBEC3A	ATXN10	CAMK1G
CCNI	COX4I1	CD81	AQP10	B2M	CAT
CD74	CSNK2B	CDC37	ARG1	CA4	CKAP4
CDC37	DAZAP2	CEBPB	ARGLU1	CAT	COX5B
CEBPB	EEF2	COR01A	CA4	CBX4	CREB1
DAZAP2	FLOT2	COX4I1	CAMK1G	CDK6	DHRS3
EMP3	GAPDH	DAZAP2	CAT	CKAP4	ECH1
GRN	GNAS	FCGRT	CATSPER2	CLDND1	FIS1
HBB	HDAC1	FLOT2	CD93	COX5B	GMFG
HGS	HSPA8	GAPDH	CKAP4	CREB1	GTF2B
HSP90AB1	IFITM1	GNAI2	CLEC12A	ECH1	HSD17B7
IL18	LAMP1	GNAS	GTF2B	FIS1	LPCAT2
ITGB2	LYN	HSPA8	HDC	GTF2B	LRG1
LAIR1	MYH9	ITM2B	IL2RB	IRF5	LRWD1
LASP1	MYL6	LAMP1	KLRB1	LRP3	MEGF10
LITAF	NFKBIA	LITAF	LPCAT2	LRWD1	MRPL33
LYN	PABPC1	LYN	LRG1	MBTD1	MRPS18C
MYL6	PCBP1	MYL6	LRWD1	MEGF10	NQO2
NKG7	PFN1	NFKBIA	LYPD2	MRPL33	NUDT18
PABPC1	PSAP	PABPC1	MMP9	MRPS18C	PDK4
PCBP1	RAC2	PSAP	MRPL33	NAMPT	PNN
PDE4C	RHOA	RHOA	MT1A	NCALD	PPP3R1
RPLP1	RPL19	RHOG	NQO2	PNN	QPCT
RPS25	RPLP1	RNASET2	NUDT18	PPP3R1	RHOC
RPS3	RPS16	RPS14	PGLYRP1	QPCT	ROPN1L
S100A4	RPS3	RPS16	PNN	RHOC	RPIA
S100A9	RPS6	RPS3	PPP3R1	ROPN1L	SAR1A
SNCA	S100A4	S100A4	QPCT	RPIA	SLC20A1
TMBIM6	TUBA1B	SLC25A39	RPIA	RPS29	SRP14
TSP0	UBA52	TMSB10	SRP14	SAR1A	SRPK1
TYROBP	UBB	UBA52	SRPK1	SLC20A1	TPST1
UBA52	VIM	UBB	TESC	SRP14	TPT1
UBB	WBP2	VIM	TMEM216	SRPK1	TRAM1
VIM	YWHAB	WBP2	TNFSF14	TESC	TUBB2A
WBP2	YWHAH	YWHAH	TPST1	TPST1	TXN
YWHAB		YWHAH	TPT1	TPT1	
			TUBB2A	TRAM1	

			TXN	TUBB2A	
			VNN2	TXN	
				ZYX	

10.3 Lista de genes más relevantes para la clasificación de pacientes BH-MT y LG-MT con GCNN y RL

Genes más relevantes luego de 10 clasificaciones de pacientes con subtipos de DFT					
GCNN					
BH-MT			LG-MT		
IntAct	BioGRID	HIPPIE	IntAct	BioGRID	HIPPIE
ACTB	ACTB	ACTB	ACTB	ACTB	ACTB
ACTG1	ACTG1	ACTG1	ACTG1	ACTG1	ACTG1
B2M	ALOX5AP	ADGRE5	B2M	CDC37	B2M
BANP	CLUAP1	ARHGDIB	BANP	CEBPB	CLUAP1
CCNDBP1	COX4I1	B2M	CCNDBP1	CLUAP1	COX4I1
CDC37	CREB1	CLUAP1	CDC37	COX4I1	CREB1
CLUAP1	EEF1A1	CREB1	CLUAP1	CREB1	DAZAP2
CREB1	F2R	DAZAP2	CREB1	EEF1A1	EEF1A1
DAZAP2	FAM177A1	EEF1A1	DAZAP2	F2R	F2R
EEF1A1	FTL	F2R	EEF1A1	FAM177A1	FAM177A1
FAM177A1	GGA1	FAM177A1	EMP3	FTL	FARSB
FTL	GJC1	FARSB	FAM177A1	GGA1	FTL
GABPB2	HSPA8	FTL	FTL	HSPA8	GGA1
GGA1	KLF2	GABPB2	GABPB2	MSH3	GJC1
HBB	MSH3	GGA1	GGA1	ORC6	ITM2B
ITGB2	MSRB1	GJC1	HBB	PABPC1	LITAF
LASP1	MYL6	GNAS	ITGB2	PCBP1	MSH3
LITAF	NACA	ITM2B	LASP1	PCLAF	OAZ1
MSH3	OAZ1	LITAF	LITAF	PDCD7	ORC6
ORC6	PCLAF	MSH3	LYN	PFN1	PABPC1
PABPC1	PDCD7	OAZ1	MSH3	RACK1	PFN1
PRR13	PFN1	ORC6	ORC6	RHOA	PIP4K2A
RAC2	RACK1	PFN1	PABPC1	ROCK2	POTEF
RACK1	RNASET2	PIP4K2A	PRR13	RPL18	RHOA
ROCK2	ROCK2	POTEF	RAC2	RPL18A	ROCK2
RPLP2	RPS6	RHOA	RACK1	RPL19	RPL18
RPS12	SQSTM1	ROCK2	ROCK2	RPL6	RPL18A
RPS19	TPT1	RPL18A	RPLP1	RPLP2	RPS11

RPS25	TUBA1A	S100A4	RPLP2	RPS11	RPS14
RPS3	TXNIP	SHCBP1	RPS12	RPS12	RPS16
S100A4	UBA52	TPT1	RPS19	RPS16	RPS19
S100A9	UBB	TUBA1A	RPS25	RPS19	RPS2
SAT1	UBC	TXNIP	S100A4	RPS3	RPS27A
SNCA	VIM	UBA52	S100A9	RPS6	RPS6
TCAF1	XPNPEP3	UBB	SNCA	S100A4	SHCBP1
TMSB4X	YWHAZ	VIM	TCAF1	SHCBP1	SQSTM1
TPT1		XPNPEP3	TMSB4X	SQSTM1	TPT1
TXNIP		YWHAZ	TPT1	TUBA1A	TUBA1A
UBA52			TXNIP	TXNIP	TXNIP
UBB			UBA52	UBA52	UBA52
UBC			UBB	UBB	UBB
VIM			UBC	UBC	VIM
WAS			VIM	VIM	YWHAB
YWHAB			WAS	YWHAB	YWHAZ
YWHAZ			YWHAB	YWHAB	
			YWHAZ	YWHAZ	

10.4 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con ELA

Nombre de vía biológica	Identificador Reactome	P-valor Aj. IntAct	P-valor Aj. BioGRID	P-valor Aj. HIPPIE
Translesion Synthesis by REV1	R-HSA-110312	0,007	0,010	0,006
Downregulation of ERBB4 Signaling	R-HSA-1253288	0,003	0,004	0,003
Spry Regulation of FGF Signaling	R-HSA-1295596	0,006	0,009	0,006
L13a-mediated Translational Silencing of Ceruloplasmin Expression	R-HSA-156827	0,000	0,000	0,000
Peptide Chain Elongation	R-HSA-156902	0,001	0,000	0,001
TICAM1, RIP1-mediated IKK Complex Recruitment	R-HSA-168927	0,007	0,008	0,006
SRP-dependent Cotranslational Protein Targeting to Membrane	R-HSA-1799339	0,001	0,000	0,002
NRIF Signals Cell Death From the Nucleus	R-HSA-205043	0,006	0,008	0,006
p75NTR Recruits Signalling Complexes	R-HSA-209543	0,005	0,007	0,004
NF-kB Is Activated and Signals Survival	R-HSA-209560	0,005	0,000	0,000
TGF-beta Receptor Signaling in EMT (Epithelial to Mesenchymal Transition)	R-HSA-2173791	0,007	0,001	0,000
Selenocysteine Synthesis	R-HSA-2408557	0,001	0,000	0,001
Senescence-Associated Secretory Phenotype (SASP)	R-HSA-2559582	0,004	0,007	0,004
FCERI Mediated NF-kB Activation	R-HSA-2871837	0,006	0,001	0,001

Glycogen Synthesis	R-HSA-3322077	0,003	0,006	0,004
Activated TAK1 Mediates P38 MAPK Activation	R-HSA-450302	0,007	0,010	0,008
AUF1 (hnRNP D0) Binds and Destabilizes mRNA	R-HSA-450408	0,002	0,000	0,000
Regulation of FZD by Ubiquitination	R-HSA-4641263	0,009	0,010	0,008
Josephin Domain DUBs	R-HSA-5689877	0,004	0,005	0,003
Metalloprotease DUBs	R-HSA-5689901	0,005	0,007	0,004
DNA Damage Recognition in GG-NER	R-HSA-5696394	0,002	0,004	0,002
Regulation of TP53 Activity Through Methylation	R-HSA-6804760	0,007	0,010	0,007
Translation Initiation Complex Formation	R-HSA-72649	0,002	0,000	0,000
Formation of a Pool of Free 40S Subunits	R-HSA-72689	0,001	0,000	0,001
GTP Hydrolysis and Joining of the 60S Ribosomal Subunit	R-HSA-72706	0,001	0,000	0,002
Eukaryotic Translation Termination	R-HSA-72764	0,001	0,000	0,001
PTK6 Regulates RTKs and Their Effectors AKT1 and DOK1	R-HSA-8849469	0,002	0,003	0,002
VLDLR Internalisation and Degradation	R-HSA-8866427	0,006	0,010	0,007
Regulation of PTEN Localization	R-HSA-8948747	0,002	0,004	0,002
TICAM1,TRAF6-dependent Induction of TAK1 Complex	R-HSA-9014325	0,002	0,003	0,002
Regulation of Signaling by CBL	R-HSA-912631	0,001	0,001	0,001
IRAK1 Recruits IKK Complex	R-HSA-937039	0,005	0,006	0,004
IRAK2 Mediated Activation of TAK1 Complex	R-HSA-937042	0,002	0,003	0,002
TRAF6-mediated Induction of TAK1 Complex Within TLR4 Complex	R-HSA-937072	0,005	0,005	0,004
Chaperone Mediated Autophagy	R-HSA-9613829	0,000	0,000	0,000
Late Endosomal Microautophagy	R-HSA-9615710	0,000	0,000	0,000
Response of EIF2AK4 (GCN2) to Amino Acid Deficiency	R-HSA-9633012	0,000	0,000	0,000
Alpha-protein Kinase 1 Signaling Pathway	R-HSA-9645460	0,002	0,003	0,002
Aggrephagy	R-HSA-9646399	0,001	0,000	0,001
RAS Processing	R-HSA-9648002	0,007	0,007	0,008
Pexophagy	R-HSA-9664873	0,003	0,004	0,003
Signaling by CSF1 (M-CSF) in Myeloid Cells	R-HSA-9680350	0,001	0,002	0,001
Inactivation of CSF3 (G-CSF) Signaling	R-HSA-9705462	0,001	0,001	0,001
Negative Regulation of FLT3	R-HSA-9706369	0,006	0,006	0,004
Regulation of BACH1 Activity	R-HSA-9708530	0,003	0,004	0,003
Regulation of NF-kappa B Signaling	R-HSA-9758274	0,007	0,009	0,006
Nonsense Mediated Decay (NMD) Independent of the Exon Junction Complex (EJC)	R-HSA-975956	0,000	0,000	0,000
Nonsense Mediated Decay (NMD) Enhanced by the Exon Junction Complex (EJC)	R-HSA-975957	0,000	0,000	0,000

10.5 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con subtipos de DFT (BH vs MT)

Nombre de vía biológica	Identificador Reactome	P-valor Aj. IntAct	P-valor Aj. BioGRID	P-valor Aj. HIPPIE
Translesion Synthesis by REV1	R-HSA-110312	0,000	0,000	0,009
ER-Phagosome Pathway	R-HSA-1236974	0,000	0,004	0,009
Downregulation of ERBB4 Signaling	R-HSA-1253288	0,000	0,000	0,005
Spry Regulation of FGF Signaling	R-HSA-1295596	0,000	0,000	0,010
Translocation of SLC2A4 (GLUT4) to the Plasma Membrane	R-HSA-1445148	0,000	0,001	0,003
TICAM1, RIP1-mediated IKK Complex Recruitment	R-HSA-168927	0,000	0,000	0,009
Formation of Annular Gap Junctions	R-HSA-196025	0,002	0,002	0,006
NRIF Signals Cell Death From the Nucleus	R-HSA-205043	0,000	0,000	0,009
p75NTR Recruits Signalling Complexes	R-HSA-209543	0,000	0,000	0,007
NF-kB Is Activated and Signals Survival	R-HSA-209560	0,000	0,000	0,007
TGF-beta Receptor Signaling in EMT (Epithelial to Mesenchymal Transition)	R-HSA-2173791	0,000	0,000	0,001
Glycogen Synthesis	R-HSA-3322077	0,000	0,000	0,006
Deactivation of the Beta-Catenin Transactivating Complex	R-HSA-3769402	0,000	0,000	0,005
EPHB-mediated Forward Signaling	R-HSA-3928662	0,001	0,002	0,001
Interaction Between L1 and Ankyrins	R-HSA-445095	0,005	0,002	0,009
Regulation of FZD by Ubiquitination	R-HSA-4641263	0,000	0,000	0,008
RHO GTPases Activate IQGAPs	R-HSA-5626467	0,007	0,001	0,004
Translesion Synthesis by POLK	R-HSA-5655862	0,000	0,000	0,009
Translesion Synthesis by POLI	R-HSA-5656121	0,000	0,000	0,010
Josephin Domain DUBs	R-HSA-5689877	0,000	0,000	0,006
Ovarian Tumor Domain Proteases	R-HSA-5689896	0,001	0,001	0,005
Metalloprotease DUBs	R-HSA-5689901	0,000	0,000	0,009
DNA Damage Recognition in GG-NER	R-HSA-5696394	0,000	0,000	0,006
Orc1 Removal From Chromatin	R-HSA-68949	0,000	0,005	0,009
PTK6 Regulates RTKs and Their Effectors AKT1 and DOK1	R-HSA-8849469	0,000	0,000	0,005
VLDLR Internalisation and Degradation	R-HSA-8866427	0,000	0,000	0,009
Regulation of PTEN Localization	R-HSA-8948747	0,000	0,000	0,005
TICAM1,TRAF6-dependent Induction of TAK1 Complex	R-HSA-9014325	0,000	0,000	0,006
IRAK1 Recruits IKK Complex	R-HSA-937039	0,000	0,000	0,007
IRAK2 Mediated Activation of TAK1 Complex	R-HSA-937042	0,000	0,000	0,006
TRAF6-mediated Induction of TAK1 Complex Within TLR4 Complex	R-HSA-937072	0,000	0,000	0,009
Negative Regulation of NOTCH4 Signaling	R-HSA-9604323	0,000	0,000	0,007
Chaperone Mediated Autophagy	R-HSA-9613829	0,000	0,000	0,000
Late Endosomal Microautophagy	R-HSA-9615710	0,000	0,000	0,004
Alpha-protein Kinase 1 Signaling Pathway	R-HSA-9645460	0,000	0,000	0,006
Aggrephagy	R-HSA-9646399	0,000	0,000	0,000
Pexophagy	R-HSA-9664873	0,000	0,000	0,006
Negative Regulation of FLT3	R-HSA-9706369	0,000	0,000	0,007
Regulation of BACH1 Activity	R-HSA-9708530	0,000	0,000	0,006
Regulation of NF-kappa B Signaling	R-HSA-9758274	0,000	0,000	0,010
Amyloid Fiber Formation	R-HSA-977225	0,000	0,000	0,000

10.6 Lista de vías biológicas enriquecidas en genes relevantes para la clasificación de pacientes con subtipos de DFT (LG vs MT)

Nombre de vía biológica	Identificador Reactome	P-valor Aj. IntAct	P-valor Aj. BioGRID	P-valor Aj. HIPPIE
Translesion Synthesis by REV1	R-HSA-110312	0,000	0,000	0,000
Recognition of DNA Damage by PCNA-containing Replication Complex	R-HSA-110314	0,001	0,002	0,001
Translesion Synthesis by POLH	R-HSA-110320	0,000	0,001	0,000
Activation of NF-kappaB in B Cells	R-HSA-1169091	0,003	0,007	0,003
Oxygen-dependent Proline Hydroxylation of Hypoxia-inducible Factor Alpha	R-HSA-1234176	0,002	0,006	0,003
ER-Phagosome Pathway	R-HSA-1236974	0,000	0,008	0,001
Downregulation of ERBB4 Signaling	R-HSA-1253288	0,000	0,000	0,000
Spry Regulation of FGF Signaling	R-HSA-1295596	0,000	0,000	0,000
Translocation of SLC2A4 (GLUT4) to the Plasma Membrane	R-HSA-1445148	0,000	0,000	0,000
Peptide Chain Elongation	R-HSA-156902	0,000	0,000	0,000
TICAM1, RIP1-mediated IKK Complex Recruitment	R-HSA-168927	0,000	0,000	0,000
SCF-beta-TrCP Mediated Degradation of Emi1	R-HSA-174113	0,002	0,004	0,002
APC-Cdc20 Mediated Degradation of Nek2A	R-HSA-179409	0,001	0,001	0,001
EGFR Downregulation	R-HSA-182971	0,001	0,001	0,001
Formation of Annular Gap Junctions	R-HSA-196025	0,002	0,003	0,002
NRIF Signals Cell Death From the Nucleus	R-HSA-205043	0,000	0,000	0,000
p75NTR Recruits Signalling Complexes	R-HSA-209543	0,000	0,000	0,000
NF-kB Is Activated and Signals Survival	R-HSA-209560	0,000	0,000	0,000
Regulation of Activated PAK-2p34 by Proteasome Mediated Degradation	R-HSA-211733	0,001	0,003	0,002
NOTCH1 Intracellular Domain Regulates Transcription	R-HSA-2122947	0,002	0,004	0,002
Activated NOTCH1 Transmits Signal to the Nucleus	R-HSA-2122948	0,001	0,001	0,001
Downregulation of TGF-beta Receptor Signaling	R-HSA-2173788	0,001	0,001	0,001
TGF-beta Receptor Signaling in EMT (Epithelial to Mesenchymal Transition)	R-HSA-2173791	0,000	0,000	0,000
Selenocysteine Synthesis	R-HSA-2408557	0,000	0,000	0,000
Oxidative Stress Induced Senescence	R-HSA-2559580	0,004	0,009	0,005
Senescence-Associated Secretory Phenotype (SASP)	R-HSA-2559582	0,002	0,001	0,003
Oncogene Induced Senescence	R-HSA-2559585	0,001	0,002	0,001
FCERI Mediated NF-kB Activation	R-HSA-2871837	0,001	0,008	0,005
NOTCH2 Activation and Transmission of Signal to the Nucleus	R-HSA-2979096	0,000	0,001	0,000
Glycogen Synthesis	R-HSA-3322077	0,000	0,000	0,000
Autodegradation of the E3 Ubiquitin Ligase COP1	R-HSA-349425	0,001	0,003	0,002
Deactivation of the Beta-Catenin Transactivating Complex	R-HSA-3769402	0,000	0,000	0,000
Rap1 Signalling	R-HSA-392517	0,003	0,004	0,003

EPHB-mediated Forward Signaling	R-HSA-3928662	0,000	0,000	0,000
EPHA-mediated Growth Cone Collapse	R-HSA-3928663	0,006	0,009	0,008
Interaction Between L1 and Ankyrins	R-HSA-445095	0,005	0,003	0,003
Activated TAK1 Mediates P38 MAPK Activation	R-HSA-450302	0,000	0,001	0,000
AUF1 (hnRNP D0) Binds and Destabilizes mRNA	R-HSA-450408	0,000	0,000	0,000
Asymmetric Localization of PCP Proteins	R-HSA-4608870	0,002	0,005	0,003
Degradation of AXIN	R-HSA-4641257	0,002	0,004	0,002
Degradation of DVL	R-HSA-4641258	0,002	0,005	0,002
Regulation of FZD by Ubiquitination	R-HSA-4641263	0,000	0,000	0,000
PINK1-PRKN Mediated Mitophagy	R-HSA-5205685	0,001	0,000	0,000
Regulation of TNFR1 Signaling	R-HSA-5357905	0,000	0,000	0,002
TNFR1-induced NF-kappa-B Signaling Pathway	R-HSA-5357956	0,000	0,000	0,001
Dectin-1 Mediated Noncanonical NF-kB Signaling	R-HSA-5607761	0,002	0,005	0,003
Degradation of GLI1 by the Proteasome	R-HSA-5610780	0,002	0,005	0,003
Degradation of GLI2 by the Proteasome	R-HSA-5610783	0,002	0,005	0,003
GLI3 Is Processed to GLI3R by the Proteasome	R-HSA-5610785	0,002	0,005	0,003
RHO GTPases Activate IQGAPs	R-HSA-5626467	0,007	0,001	0,001
Translesion Synthesis by POLK	R-HSA-5655862	0,000	0,000	0,000
Translesion Synthesis by POLI	R-HSA-5656121	0,000	0,001	0,000
Termination of Translesion DNA Synthesis	R-HSA-5656169	0,001	0,000	0,001
Regulation of RAS by GAPs	R-HSA-5658442	0,002	0,006	0,003
MAP2K and MAPK Activation	R-HSA-5674135	0,001	0,002	0,001
NIK-->noncanonical NF-kB Signaling	R-HSA-5676590	0,002	0,005	0,003
UCH Proteinases	R-HSA-5689603	0,001	0,002	0,001
Josephin Domain DUBs	R-HSA-5689877	0,000	0,000	0,000
Ovarian Tumor Domain Proteases	R-HSA-5689896	0,001	0,000	0,000
Metalloprotease DUBs	R-HSA-5689901	0,000	0,000	0,000
Processing of DNA Double-Strand Break Ends	R-HSA-5693607	0,003	0,007	0,004
DNA Damage Recognition in GG-NER	R-HSA-5696394	0,000	0,000	0,000
Formation of Incision Complex in GG-NER	R-HSA-5696395	0,002	0,003	0,002
Gap-filling DNA Repair Synthesis and Ligation in GG-NER	R-HSA-5696397	0,000	0,001	0,001
Dual Incision in GG-NER	R-HSA-5696400	0,001	0,003	0,002
Formation of TC-NER Pre-Incision Complex	R-HSA-6781823	0,002	0,006	0,003
Dual Incision in TC-NER	R-HSA-6782135	0,004	0,008	0,005
Gap-filling DNA Repair Synthesis and Ligation in TC-NER	R-HSA-6782210	0,004	0,008	0,005
Regulation of TP53 Degradation	R-HSA-6804757	0,001	0,003	0,002
Regulation of TP53 Activity Through Methylation	R-HSA-6804760	0,000	0,001	0,000
Orc1 Removal From Chromatin	R-HSA-68949	0,000	0,001	0,001
CDK-mediated Phosphorylation and Removal of Cdc6	R-HSA-69017	0,003	0,008	0,005
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	R-HSA-69601	0,001	0,003	0,002
Translation Initiation Complex Formation	R-HSA-72649	0,000	0,000	0,000
Formation of a Pool of Free 40S Subunits	R-HSA-72689	0,000	0,000	0,000
Formation of the Ternary Complex, and Subsequently, the 43S Complex	R-HSA-72695	0,003	0,000	0,000
Ribosomal Scanning and Start Codon Recognition	R-HSA-72702	0,003	0,000	0,000

Eukaryotic Translation Termination	R-HSA-72764	0,000	0,000	0,000
Ubiquitin-dependent Degradation of Cyclin D	R-HSA-75815	0,001	0,003	0,002
PTK6 Regulates RTKs and Their Effectors AKT1 and DOK1	R-HSA-8849469	0,000	0,000	0,000
FBXL7 Down-Regulates AURKA During Mitotic Entry and in Early Mitosis	R-HSA-8854050	0,002	0,004	0,002
VLDLR Internalisation and Degradation	R-HSA-8866427	0,000	0,000	0,000
E3 Ubiquitin Ligases Ubiquitinate Target Proteins	R-HSA-8866654	0,002	0,003	0,002
RUNX1 Regulates Transcription of Genes Involved in Differentiation of HSCs	R-HSA-8939236	0,003	0,007	0,003
Regulation of RUNX2 Expression and Activity	R-HSA-8939902	0,003	0,007	0,004
Regulation of RUNX3 Expression and Activity	R-HSA-8941858	0,002	0,004	0,002
Regulation of PTEN Localization	R-HSA-8948747	0,000	0,000	0,000
Regulation of PTEN Stability and Activity	R-HSA-8948751	0,003	0,007	0,004
ER Quality Control Compartment (ERQC)	R-HSA-901032	0,000	0,001	0,000
TICAM1,TRAF6-dependent Induction of TAK1 Complex	R-HSA-9014325	0,000	0,000	0,000
Regulation of Signaling by CBL	R-HSA-912631	0,000	0,001	0,000
Endosomal Sorting Complex Required For Transport (ESCRT)	R-HSA-917729	0,001	0,002	0,001
IRAK1 Recruits IKK Complex	R-HSA-937039	0,000	0,000	0,000
IKK Complex Recruitment Mediated by RIP1	R-HSA-937041	0,000	0,000	0,000
IRAK2 Mediated Activation of TAK1 Complex	R-HSA-937042	0,000	0,000	0,000
TRAF6-mediated Induction of TAK1 Complex Within TLR4 Complex	R-HSA-937072	0,000	0,000	0,000
Negative Regulation of NOTCH4 Signaling	R-HSA-9604323	0,000	0,001	0,000
Chaperone Mediated Autophagy	R-HSA-9613829	0,000	0,000	0,000
Regulation of Localization of FOXO Transcription Factors	R-HSA-9614399	0,002	0,003	0,002
Late Endosomal Microautophagy	R-HSA-9615710	0,000	0,000	0,000
Response of EIF2AK4 (GCN2) to Amino Acid Deficiency	R-HSA-9633012	0,000	0,000	0,000
Alpha-protein Kinase 1 Signaling Pathway	R-HSA-9645460	0,000	0,000	0,000
Aggrephagy	R-HSA-9646399	0,000	0,000	0,000
RAS Processing	R-HSA-9648002	0,000	0,000	0,000
Pexophagy	R-HSA-9664873	0,000	0,000	0,000
Signaling by CSF1 (M-CSF) in Myeloid Cells	R-HSA-9680350	0,000	0,001	0,001
Inactivation of CSF3 (G-CSF) Signaling	R-HSA-9705462	0,000	0,001	0,001
Negative Regulation of FLT3	R-HSA-9706369	0,000	0,000	0,000
Regulation of BACH1 Activity	R-HSA-9708530	0,000	0,000	0,000
Regulation of NF-kappa B Signaling	R-HSA-9758274	0,000	0,000	0,000
Nonsense Mediated Decay (NMD) Independent of the Exon Junction Complex (EJC)	R-HSA-975956	0,000	0,000	0,000
Amyloid Fiber Formation	R-HSA-977225	0,000	0,000	0,000