

NIFPTML: Aprendizaje Automático por Teoría de Perturbaciones con Fusión de Información de Redes Biomoleculares en Química Médica, Cromosómica, y Nanoinformática

Autora: Viviana Fernanda Quevedo Tumaili

Tesis doctoral UDC / 2022

Director: Prof. Dr. Alejandro Pazos Sierra

Co-Director: Prof. Dr. Humberto González Díaz

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones



UNIVERSIDADE DA CORUÑA



Prof. Dr. D. Alejandro Pazos Sierra, Catedrático del Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Informática, CITIC-Centro de Investigación de Tecnologías de la Información y las Comunicaciones, Universidade da Coruña, Instituto de Investigación Biomédica de a Coruña (INIBIC), A Coruña, Miembro fundador de IKERDATA.

Prof. Dr. D. Humberto González Díaz, Prof. IKERBASQUE del Departamento de Química Orgánica e Inorgánica, Facultad de Ciencia y Tecnología, Universidad del País Vasco, UPV/EHU; IKERBASQUE, Fundación Vasca para la Ciencia; y, BIOFISIKA: Centro Vasco de Biofísica, CSIC-UPV/EHU, Leioa, Miembro fundador de IKERDATA.

HACEN CONSTAR QUE:

La memoria de investigación **“NIFPTML: APRENDIZAJE AUTOMÁTICO POR TEORÍA DE PERTURBACIONES CON FUSIÓN DE INFORMACIÓN DE REDES BIOMOLECULARES EN QUÍMICA MÉDICA, CROMOSÓMICA, Y NANOINFORMÁTICA”** ha sido realizada por **Dña. VIVIANA FERNANDA QUEVEDO TUMAILLI**, bajo nuestra dirección en el Programa de doctorado en TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES, y constituye la Tesis que presenta para optar al Grado de Doctora de la Universidade da Coruña.

A Coruña, 21 de marzo de 2022

Prof. Dr. Alejandro Pazos Sierra
Director de Tesis

Prof. Dr. Humberto González Díaz
Co-Director de Tesis

A mi esposo y a mis hijos,

a mi padres

a mis hermanos y querida familia

AGRADECIMIENTOS

En primer lugar, deseo expresar mi agradecimiento a los directores de esta tesis los Profesores Dr. Alejandro Pazos Sierra y Dr. Humberto González Díaz, por toda la ayuda y dedicación que han brindado a este trabajo. Asimismo, quiero agradecer a sus colaboradores de la Universidade da Coruña, especialmente a los grupos de Redes de Neuronas Artificiales y Sistemas Adaptativos – Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR), Instituto de Investigación Biomédica de A Coruña (INIBIC) y al Centro de Investigación en Tecnologías de la Información (CITIC). También agradecer a sus colaboradores de la Universidad del País Vasco, especialmente al Departamento de Química Orgánica e Inorgánica. Agradecer también a la Universidad Estatal Amazónica por la beca de estudios doctorales otorgada, así como también a los compañeros docentes/administrativos que estuvieron pendientes en este proceso.

Deseo expresar también todo mi agradecimiento a mis padres, a mis hermanos, a mi esposo y a mis hijos por su apoyo incondicional y su confianza a lo largo de este trayecto.

A mis demás familiares, amigos y a todas aquellas personas que han estado presentes y han dedicado parte de su tiempo al desarrollo de este trabajo, por su colaboración y paciencia.

¡Muchas gracias a todos!

RESUMO

A teoría das redes complexas permite estudar sistemas biomoleculares. Dado que os grafos poden representar redes, nunha rede de proteínas, por exemplo, os nodos son os aminoácidos e os eixes son as secuencias e/ou interaccións/proximidades espaciais entre os aminoácidos. Para cuantificar a estrutura destes sistemas utilízanse parámetros/índices numéricos extraídos destas Invariantes de Rede (NI). Estes parámetros pódense correlacionar con propiedades biolóxicas mediante técnicas de Aprendizaxe Automática (ML), o que permite atopar modelos predictivos. Ademais, é necesario utilizar técnicas de Fusión de Información (IF) de diversas fontes para obter un conxunto de datos enriquecido. Os operadores da Teoría da Perturbación (PT) procesan a información cuantificando perturbacións/desviacións en variables estruturais con respecto aos valores esperados para diferentes subconxuntos de variables categóricas. Se propoñen utilizar a estratexia NIFPTML combinando as fases mencionadas anteriormente (NI+IF+PT+ML) de un xeito innovador, necesario para estudar problemas que impliquen un ou máis destes sistemas ao mesmo tempo. Aplícanse NIFPTML a varios problemas complexos con diferentes sistemas (fármacos, proteínas, xenes, cromosomas, nanopartículas). Defínense por primeira vez redes complexas GOIN (Gene Orientation Inversion Network) e os seus parámetros numéricos. Isto permitiu exemplificar o uso de NIFPTML en problemas que implican cromosomas, entrando directamente na aplicación de ML na nova área coñecida como Chromosomal.

RESUMEN

La teoría de redes complejas permite estudiar sistemas biomoleculares. Dado que los grafos pueden representar redes, en una red de proteína, como ejemplo, los nodos son los aminoácidos y los ejes son las secuencias y/o interacciones/proximidades espaciales entre los aminoácidos. Para cuantificar la estructura de estos sistemas se usan parámetros/índices numéricos extraídos de estas Invariantes de Redes (NI). Estos parámetros pueden ser correlacionados con propiedades biológicas mediante técnicas de Aprendizaje Automático (ML), permitiendo encontrar modelos predictivos. Adicionalmente, es necesario utilizar técnicas de Fusión de Información (IF) de diversas fuentes para obtener un conjunto de datos enriquecido. Los operadores de la Teoría de Perturbación (PT) procesan la información cuantificando las perturbaciones/desviaciones en las variables estructurales con respecto a valores esperados para diferentes subconjuntos de variables categóricas. Se propone usar la estrategia NIFPTML combinando las fases mencionadas anteriormente (NI+IF+PT+ML) de una manera innovadora necesaria para estudiar problemas que involucran uno o más de estos sistemas a la vez. Se aplican NIFPTML a varios problemas complejos con distintos sistemas (fármacos, proteínas, genes, cromosomas, nanopartículas). Se definen por primera vez las redes complejas GOIN (Gen Orientation Inversion Network) y sus parámetros numéricos. Esto permitió ejemplificar el uso de NIFPTML en problemas que involucran cromosomas, incursionando directamente en la aplicación de ML en la nueva área conocida como Cromosómica.

ABSTRACT

Complex network theory allows the study of biomolecular systems. Since graphs can represent networks, in a protein network, as an example, the nodes are the amino acids and the axes are the sequences and/or spatial interactions/proximities between the amino acids. Numerical parameters/indexes extracted from these Networks Invariants (NI) are used to quantify the structure of these systems. These parameters can be correlated with biological properties using Machine Learning (ML) techniques, allowing predictive models to be found. Additionally, it is necessary to use Information Fusion (IF) techniques from various sources to obtain an enriched dataset. Perturbation Theory (PT) operators process the information by quantifying the perturbations/deviations in the structural variables with respect to expected values for different subsets of categorical variables. It is proposed to use the NIFPTML strategy combining the above-mentioned phases (NI+IF+PT+PT+ML) in an innovative way necessary to study problems involving one or more of these systems at a time. NIFPTML is applied to several complex problems with different systems (drugs, proteins, genes, chromosomes, nanoparticles). GOIN (Gen Orientation Inversion Network) complex networks and their numerical parameters are defined for the first time. This allowed to exemplify the use of NIFPTML in problems involving chromosomes, making a direct incursion into the application of ML in the new area known as Chromosomics.

ÍNDICE DE CONTENIDOS

AGRADECIMIENTOS	7
RESUMO.....	9
RESUMEN	11
ABSTRACT	13
ÍNDICE DE CONTENIDOS	15
ÍNDICE DE FIGURAS	17
ÍNDICE DE TABLAS	19
LISTA DE ABREVIATURAS.....	20
1. <i>Introducción</i>	25
1.1. Objetivos	31
1.2. Hipótesis de trabajo	35
2. <i>Fundamentos teóricos</i>	39
2.1. Base de datos disponibles de distintos sistemas biomoleculares	39
2.1.1. ChEMBL.....	39
2.1.2. UniProt.....	39
2.1.3. NCBI	40
2.2. Representación de los datos de sistemas biomoleculares	40
2.2.1. Sistemas complejos.....	40
2.2.2. Red Compleja de un sistema biomolecular	41
2.2.3. Grafo	42
2.2.4. Grafo estrella	43
2.2.5. Grafo de secuencia	44
2.2.6. Grafos moleculares	45
2.2.7. Red Neuronal Recurrente	45
2.3. Caracterización numérica de sistemas biomoleculares.....	46
2.3.1. Descriptores moleculares	46
2.3.2. Índice de Shannon	47
2.3.3. Cadenas de Markov	48
2.4. Software de computación avanzada para el cálculo de parámetros numéricos de sistemas biomoleculares.....	48

2.4.1.	Dragon	48
2.4.2.	CentiBiN	49
2.4.3.	MARCH-INSIDE	50
2.4.4.	S2Snet	51
2.5.	Métodos de IA/ML para análisis de parámetros numéricos de sistemas biomoleculares	53
2.5.1.	LDA/GDA	53
2.5.2.	Arbol de Clasificación.....	53
2.5.3.	Máquinas de Vectores de Soporte.....	54
2.5.4.	Redes de Neuronas Artificiales ó “Artificial Neural Networks” (ANN).....	54
2.5.5.	Modelo de Aprendizaje Automático con Teoría de la Perturbación	55
2.5.6.	Fusión de la Información para el Modelo de Aprendizaje Automático con Teoría de la Perturbación.....	55
2.6.	Software para entrenar modelos AI/ML.....	56
2.6.1.	Statistics.....	56
2.6.2.	Weka	56
2.7.	Criterios de calidad de métodos de AI/ML de clasificación para sistemas biomoleculares	57
2.7.1.	Sensibilidad	57
2.7.2.	Especificidad	57
2.7.3.	Análisis del Área bajo la Curva Característica Operativa del Receptor	57
2.7.4.	Entrenamiento y Validación.....	58
3.	<i>Trabajo Experimental</i>	61
3.1.	Capítulo 1. Modelos PTML de Compuestos Anti-leishmania.....	61
3.2.	Capítulo 2. Modelos ML en Redes GOINs de Compuestos AntiPlasmodium.....	81
3.3.	Capítulo 3. Modelos forestales aleatorios para sistemas de liberación de Nanopartículas decoradas con fármacos antimaláricos	102
3.4.	Capítulo 4. Mapeo IFPTML de gráficos de fármacos con red estructural de proteínas y cromosomas frente a información de ensayos preclínicos para el descubrimiento de compuestos antipalúdicos	119
4.	<i>Conclusiones</i>	147
5.	<i>Futuros desarrollos</i>	149
6.	<i>Bibliografía</i>	151

ÍNDICE DE FIGURAS

Figura 1. Ejemplo de una pequeña red con 8 nodos y 10 enlaces	41
Figura 2. Ejemplo de los tipos de redes presente en este documento. Una visualización de la estructura de la red GOIN del Cromosoma I de <i>P. Falciparum</i> con un total de 157 nodos (genes) y 289 enlaces (inversión genética).....	42
Figura 3. Ejemplo matriz de adyacencia y su grafo. Una visualización de la matriz A de 157 nodos que representan a los genes del Cromosoma I de <i>P. Falciparum</i> y 289 aristas.....	43
Figura 4. El grafo estrella S_6 (sin tomar en cuenta el nodo interno) o S_7 (con nodo interno) ..	43
Figura 5. Grafo de Estrella (Recurrencia) vs. Grafo de Secuencia dirigido y no-dirigido	44
Figura 6. Generación de representación de la estructura molecular de la fórmula $C_{13}H_{18}N_4O_3$ (lado izquierdo) en el grafo molecular con hidrógenos suprimidos (lado derecho).....	45
Figura 7. Representación básica y compacta de una RNN.....	46
Figura 8. Representación básica y compacta de una RNN de una secuencia finita	46
Figura 9. Interfaz gráfica de la aplicación de escritorio Dragon 6.0	49
Figura 10. Interfaz gráfica de la versión online E-Dragon 1.0	49
Figura 11. Interfaz de la aplicación CentiBIN.....	50
Figura 12. Interfaz gráfica de la aplicación MARCH-INSIDE	51
Figura 13. Interfaz gráfica de la aplicación S2Snet.....	52
Figura 14. Interfaz gráfica de la aplicación Statistics.....	56
Figura 15. Guía para la interpretación de AUROC	58
Figura 16. Fármacos antileishmanicos con motivos heterocíclicos y miltefosina	63
Figura 17. C-10 sustituidas 5,10-dihydropyrrolo[1,2- <i>b</i>]isoquinolines 2 y 3 y examinadas contra <i>L. amazonensis</i> y <i>L. donovani</i>	66
Figura 18. Flujo de trabajo del estudio de las GOIN del conjunto de datos del Proteoma Plasmodium.	83
Figura 19. Ilustración de patrones de inversión génica con diferentes valores de corte.	85

Figura 20. Construcción de matrices de adyacencia a partir de datos genéticos de cada cromosoma.	85
Figura 21. Ilustración del GOIN y S-GOIN para el cromosoma I en la interfaz del software CentiBiN.....	91
Figura 22. Promedio de proximidad de S-GOINs frente a R-GOINs de 14 cromosomas.	92
Figura 23. Grado de S-GOINs frente a R-GOINs de 14 cromosomas.	94
Figura 24. Ilustración de las comunidades del cromosoma II.	95
Figura 25. Resultados del análisis de la curva ROC para el modelo lineal.	100
Figura 26. Flujo de trabajo para el desarrollo de Modelos PTML.	104
Figura 27. Diagrama de caja de los valores AUC de los clasificadores ML (CV de 10 veces).	113
Figura 28. Selección de características mediante ExtraTrees: disminución de la impureza promedio por característica, en orden original y en orden descendente; las barras rosas representan las características reintroducidas tras el filtrado inicial.	115
Figura 29. Impacto de la característica en la variable de salida para el mejor modelo basado en los valores medios de SHAP.	117
Figura 30. Flujo de trabajo general de los pasos dados en este trabajo.	121
Figura 31. Variables pre-procesadas frente a post-procesadas.	123
Figura 32. Árbol de decisión del modelo IFPTML-CTUS.....	125
Figura 33. Árbol de decisión del modelo IFPTML-CTLC.....	127
Figura 34. Un ejemplo del modelo IFPTML-CTLC.	132
Figura 35. Desarrollo del modelo IFPTML y proceso de IF.....	135
Figura 36. Ilustración de diferentes representaciones para representar sistemas moleculares múltiples.	140

ÍNDICE DE TABLAS

Tabla 1. Efectos leishmanicidas y citotóxicos IC ₅₀ de los derivados de pirroloisoquinolina (expresados en μM) en el ensayo “ <i>in vitro</i> ” de promastigotes.	67
Tabla 2. Efectos leishmanicidas y citotóxicos IC ₅₀ de los derivados de pirroloisoquinoline (expresados en μM) en el ensayo “ <i>in vitro</i> ” de promastigotes.	68
Tabla 3. Resultados del modelo PTML-LDA	73
Tabla 4. Predicción PTML del valor medio de probabilidad $p(\text{IC}_{50}(\mu\text{M}) < 0.01)_{\text{avg}}$ para las pirroloisoquinolinas 2ad, 2bb, 3h, and 3i contra > 20 especies diferentes de <i>Leishmania</i>	74
Tabla 5. Definición de parámetros más relevantes utilizados para describir redes complejas.	87
Tabla 6. Modelos GOIN / S-GOIN vs. R-GOIN (ejemplos seleccionados).....	93
Tabla 7. Proteínas en las comunidades más pequeñas.....	96
Tabla 8. Resultados del análisis discriminante.....	98
Tabla 9. Principales genes codificando proteínas con altas probabilidades RIFIN según modelo	99
Tabla 10. Condiciones de ensayo ChEMBL (ejemplos seleccionados)	106
Tabla 12. Condiciones de ensayo de las nanopartículas decoradas (ejemplos seleccionados)	108
Tabla 13. Parámetros de actividad de la Nanopartícula (c ₀).....	109
Tabla 14. Área bajo las características operativas del receptor (AUC) para los modelos de clasificación de referencia.	113
Tabla 15. Resultado del modelo IFPTML-GDA.....	124
Tabla 16. Valores seleccionados de promedios de condiciones múltiples para diferentes combinaciones de condiciones de ensayo.	125
Tabla 17. Coeficientes del modelo IFPTML-CTUS.	126
Tabla 18. Coeficientes del modelo IFPTML-CTLG.	127
Tabla 19. Comparación de modelos con diferentes algoritmos.....	129
Tabla 20. Funciones más relevantes utilizadas en la etapa de pre-procesamiento de datos...	138
Tabla 21. Particiones y niveles (valores únicos) tomados por las variables de entrada categóricas (no ordenadas).	141
Tabla 22. Variables de entrada de los modelos IFPTML desarrollados.....	144

LISTA DE ABREVIATURAS

ANN: Analysis Neuronal Network, Redes Neuronales Artificiales

AUROC: Analisis of the Area Under Receiver Operating Characteristic Curve, Análisis del Área bajo la Curva Característica Operativa del Receptor

CC-BY: Creative Commons-Attribution, Creative Commons-Atribución

CentiBiN: Centralities in Biological Networks, Centralidades en las Redes Biológicas

CTLC: Classification Tree with Linear Combinations, Árbol de Clasificación con Combinaciones Lineales

CT: Classification Tree, Árbol de Clasificación

CTUS: Classification Tree with Univariate Split, Árbol de Clasificación con División Univariante

DDNP: Drug-Decorated Nanoparticles, Nanopartículas Decoradas con Fármacos

DNA: Deoxyribonucleic Acid, Ácido desoxirribonucleico

FN: False Negative, Falsos Negativos

FNR: False Negative Rate, Tasa de Falsos Negativos

FP: False Positive, Falsos Positivos

FPR: False Positive Rate, Tasa de Falsos Positivos

GDA: General Discriminant Analysis, Análisis General Discriminante

GDV: Genome Data Viewer, Visor de Datos del Genoma

GOIN: Gene Orientation Inversion Networks, Redes de Orientación de Inversión en los Genes

IF: Information Fusion, Fusión de la Información

IFPTML: Information Fusion Perturbation Theory Machine Learning, Aprendizaje Automático con Teoría de la Perturbación y Fusión de la Información

MARCH-INSIDE: MARKovian CHEmicals IN SILico Design, Química Marcoviana en Silo Diseño

MMA: Multi-condition combined Mobiles Average, Medias Móviles combinadas Multicondición

LDA: Linear Discriminant Analysis, Análisis Discriminante Lineal

LNN: Linear Neuronal Network, Red Neuronal Lineal

ML: Machine Learning, Aprendizaje Automático

NA: Network Analysis, Análisis de Redes

NI: Networks Invariants, Invariantes de Redes

NCBI: National Center for Biotechnology Information, Centro Nacional de Información Biotecnológica

NIFPTML: Network Information Fusion Perturbation Theory Machine Learning, Aprendizaje Automático con Teoría de la Perturbación y Fusión de la Información de Redes Complejas.

TNR: True Negative Rate, Tasa de Verdaderos Negativos

TN: True Negative, Verdaderos Negativos

PIN: Protein Interactions, Interacciones entre proteínas

PT: Perturbation Theory, Teoría de la Perturbación

PTML: Perturbation Theory Machine Learning, Aprendizaje Automático con Teoría de la Perturbación

QSAR: Quantitative Structure-Activity Relationship, Relación Cuantitativa Estructura-Actividad

R-GOIN: Random- Gene Orientation Inversion Networks, GOIN Aleatoria

RNA: Ribonucleic Acid, Ácido Ribonucleico

S2Snet: Sequence to Star Network, Red de Secuencia a Estrella

S-GOIN: Spatial- Gene Orientation Inversion Networks, GOIN Espacial

SMILES: Simplified Molecular Input Line Entry Specification, Especificación Simplificada de la línea de entrada Molecular

Sp: Specificity, Especificidad

Sn: Sensitivity, Sensibilidad

SV: Sequence Viewer, Visor de Secuencia

SVM: Support-Vector Machines, Máquinas de Vectores de Soporte

TI: Topology Index, Índice Topológico

TP: True Positive, Verdaderos Positivos

TPR: True Positive Rate, Tasa de Verdaderos Positivos

UniProt: Universal Protein Resource, Recurso Proteico Universal

WHO: World Health Organization, Organización Mundial de la Salud

1. Introducción

1. Introducción

La teoría de redes complejas permite el estudio de sistemas biomoleculares (fármacos, proteínas, genes, redes metabólicas, *etc.*). Las redes pueden ser representadas como grafos a través de conjuntos de nodos y ejes. Un ejemplo es el grafo molecular donde los nodos y ejes corresponden a los átomos y enlaces químicos de la molécula de un fármaco respectivamente. Otro ejemplo es la red de una proteína donde los nodos son aminoácidos y los ejes la secuencia y/o interacción/proximidad espacial entre los aminoácidos. En principio, la aproximación de redes complejas pudiera usarse para estudiar también sistemas como los cromosomas, nanopartículas, poblaciones, *etc.* Los recientes avances en la tecnología de secuenciación del ADN están permitiendo un rápido aumento del número de genomas secuenciados. No obstante, muchas preguntas básicas de la biología del genoma siguen aún sin respuesta, debido a que los datos de las secuencias por sí solos no permiten comprender cómo se organiza el genoma en cromosomas, la posición e interacción de esos cromosomas en la célula y cómo cambian los cromosomas y sus interacciones entre sí en respuesta a estímulos ambientales o a lo largo del tiempo. No obstante, hasta donde se conoce, las redes complejas no han sido usadas para representar/estudiar cromosomas como un todo. Esto abriría las puertas a estudios con técnicas y procedimientos de Computación avanzada e Inteligencia Artificial, como es el Aprendizaje Automático o Machine Learning (AI/ML) en la nueva área de la Cromosómica. Esta área se ocupa de la plasticidad de los cromosomas en relación con la posición tridimensional de los genes, que afectan a la función celular de forma específica para el desarrollo y los tejidos durante el ciclo celular. La cromosómica incluye la investigación de cambios en la arquitectura de los cromosomas mediados por la modificación de la cromatina de los cromosomas pudiendo influir en las funciones y la duración de la vida de las células, los tejidos, los órganos y los individuos (Claussen, 2005).

Para cuantificar la estructura de estos sistemas biomoleculares se usan parámetros numéricos extraídos de estas redes que cubren múltiples parámetros de actividad biológica. Los parámetros de las redes pueden ser estructurales y condiciones experimentales de los ensayos de todos los subsistemas involucrados.

Estos parámetros o índices numéricos invariantes de redes ó Networks Invariants (NI) pueden ser correlacionados con las propiedades biológicas de dichos sistemas mediante técnicas de AI y, o, ML. Además, en muchos problemas de interés se hace necesario fusionar información sobre varios de estos sistemas a la vez. Las técnicas para la Fusión de Información ó “Information Fusion” (IF) de diversas fuentes permiten obtener un conjunto de datos

enriquecido. Los operadores de la Teoría de Perturbación ó “Perturbation Theory” (PT) permiten cuantificar las perturbaciones/desviaciones en las variables estructurales con respecto a los valores esperados para diferentes subconjuntos de variables categóricas. Por último, los métodos IA/ML permiten encontrar modelos predictivos para las propiedades biológicas de los sistemas (fármacos, proteínas, *etc.*). La combinación de todas estas fases se puede conceptualizar como una estrategia NIFPTML. En efecto NIFPTML combina todas las fases mencionadas anteriormente (NI + IF + PT + AI/ML). La primera fase, usa parámetros numéricos de invariantes de redes (NI, invariants networks) para cuantificar la estructura de los sistemas, la fase IF fusiona los datos de múltiples sistemas provenientes de distintas fuentes, la fase PT procesa la información, y la fase AI/ML encuentra el modelo predictivo. En la tesis se aplica la estrategia NIFPTML a varios problemas complejos con distintos sistemas (fármacos, proteínas, genes, cromosomas, nanopartículas).

Por tanto, en esta tesis se plantea como primer objetivo proponer e ilustrar el uso de una innovadora estrategia NIFPTML para estudiar problemas que involucran a uno o más de uno de estos sistemas a la vez. Como algunos de los sistemas estudiados involucran cromosomas en esta tesis se definen por primera vez las redes complejas GOIN (Gen Orientation Inversion Network) y sus parámetros numéricos. Otro de los objetivos de la tesis es introducir redes complejas para el estudio de Cromosomas. Esto ha permitido ejemplificar el uso de la estrategia NIFPTML en problemas que involucran cromosomas. Lo cual lleva a adentrarse directamente en la aplicación de AI/ML en la nueva área conocida como Cromosómica. Finalmente, otro de los objetivos propuestos ha sido la colaboración con otros grupos de investigación en la corroboración experimental de los resultados predichos por los modelos NIFPTML desarrollados. Los distintos trabajos presentados en esta tesis, usando la estrategia NIFPTML incluyendo en algunos casos estudios con redes GOINs, se resumen a continuación.

En el primer trabajo se usó la estrategia NIFPTML para estudiar fármacos anti-leishmania. Se ha demostrado que los algoritmos de NIFPTML son útiles para modelar grandes (>145.000 casos) conjuntos de datos ChEMBL de ensayos preclínicos anti-leishmania. En este trabajo no se usaron redes complejas de proteínas por lo que se omite la fase NI. Es posible utilizar el modelo PTML desarrollado para reducir los costes de los ensayos prediciendo la probabilidad con la que un compuesto de consulta de esta u otra serie de compuestos alcanza un nivel deseado para múltiples parámetros (IC_{50} , K_i , *etc.*) frente a diferentes especies de *Leishmania* y proteínas diana, con altos valores de especificidad (>98%) y sensibilidad (>90%), tanto en las series de entrenamiento como de validación. Los estudios predictivos sirvieron de

complemento a estudios experimentales llevados a cabo por otros investigadores del grupo. En estos estudios se sintetizaron y ensayaron compuestos químicos de tipo pirrolo[1,2-b]isoquinolina. La evaluación de la actividad leishmanicida de las pirrolo[1,2-b]isoquinolina “*in vitro*” contra la leishmaniasis visceral (*L. donovani*) y cutánea (*L. amazonensis*) reveló que casi todos los compuestos mostraron una citotoxicidad muy baja, $CC_{50} > 100 \mu\text{g/mL}$ en células J774 (dosis más alta probada). Esta es una característica importante, ya que la toxicidad de los fármacos es una de las principales limitaciones de la quimioterapia actual para la leishmaniasis. En general, las pirroloisoquinolinas sustituidas por 10-arilmetilo mostraron la mejor actividad contra *L. amazonensis* en ensayos “*in vitro*” de promastigotes. En particular, **2ad** ($IC_{50} = 3,30 \mu\text{M}$, $SI > 77,01$) y **2bb** ($IC_{50} = 3,93 \mu\text{M}$, $SI > 58,77$) fueron aproximadamente 10 veces más potentes y selectivas que el fármaco de referencia (miltefosina). Por otro lado, **2ae** fue el compuesto más activo en los ensayos “*in vitro*” con amastigotes ($IC_{50} = 33,59 \mu\text{M}$, $SI > 8,93$). Estos resultados experimentales estaban en concordancia en líneas generales con las predicciones hechas por los modelos NIFPTML.

En el segundo trabajo se construyó una red compleja de la distribución espacial y orientación de los genes en el cromosoma por primera vez llamada GOINs. Se pudo observar que la orientación de los genes no sigue un patrón aleatorio en los cromosomas de *P. falciparum*. También se concluyó mediante modelo PTML que existe una relación entre la orientación del gen y la función biológica de la proteína RIFIN en el proteoma de *P. falciparum*.

En el tercer trabajo se usó la estrategia IFPTML resultando útil para clasificar los fármacos en función de su constante a muchas nanopartículas diferentes y su capacidad para actuar contra el *Plasmodium*. El mejor modelo de clasificación se ha obtenido utilizando Random Forest (RF) con sólo 27 características seleccionadas de fármacos y nanopartículas en todas las condiciones experimentales consideradas. El rendimiento del modelo RF demostró el poder de la fusión de información de las características experimentales de los fármacos y las nanopartículas para la predicción de la probabilidad, relacionada con la actividad antimalárica de la nanopartícula-fármaco/compuesto.

En el cuarto trabajo, se desarrolló el método NIFPTML utilizando todas las fases propuestas en la estrategia llevando a cabo la fusión de datos de diferentes fuentes, aplicando los operadores de perturbación y permitiendo la predicción computacional de nuevos compuestos antimaláricos para diferentes sistemas biomoleculares.

El algoritmo IFPTML tuvo éxito a la hora de tener en cuenta tanto la información numérica (parámetros estructurales) como la información categórica (múltiples condiciones experimentales) de los tres conjuntos de datos. Las medidas de entropía de Shannon Sh_k (variables numéricas) fueron útiles para cuantificar la información sobre la estructura de los fármacos, las secuencias de proteínas, las secuencias de genes y los cromosomas. Además, las Medias Móviles combinadas Multicondición ó Multi-condition combined Mobiles Average (MMA) de diferentes particiones de variables categóricas del conjunto de datos ChEMBL fueron útiles para codificar múltiples condiciones experimentales de ensayos preclínicos e información sobre proteínas, genes y cromosomas diana. El modelo IFPTML-CTLC es el más complejo en términos de número de variables de entrada, número de LCs y número de reglas de división. Sin embargo, el modelo IFPTML-CTLC mostró un mejor rendimiento que el IFPTML-GDA e incluye más información biológicamente relevante que el modelo IFPTML-CTUS. Este modelo podría convertirse en una herramienta útil para la optimización de ensayos preclínicos de nuevos compuestos antimaláricos teniendo en cuenta la estructura del fármaco, la especie de *Plasmodium*, la secuencia de la proteína diana y otros múltiples parámetros.

1.1. Objetivos

1.1. Objetivos

1. Definir las redes complejas GOINs en forma de redes de tipo secuencia-recurrencia que incluyan información sobre la disposición (locus) y orientación de genes en el cromosoma.
2. Verificar la factibilidad de usar al *P. Falciparum* como caso de estudio para construir las GOINs para todo el genoma de un organismo tipo y la posibilidad del uso de herramientas propias de computación avanzada adaptadas a este contexto.
3. Estudiar, mediante técnicas y procedimientos de computación avanzada y AI, la distribución de las redes tipo GOINs construidas en comparación con modelos de redes aleatorias con vistas a detectar una posible relevancia biológica del fenómeno.
4. Estudiar, con técnicas de computación avanzada y AI la formación de comunidades y su posible utilidad en la detección de clusters de genes/proteínas con relevancia biológica dentro de las redes tipo GOINs construidas.
5. En estudios de ML y otras técnicas de AI, calcular centralidades de nodos para genes en redes tipo GOINs y verificar su utilidad como variables de entrada para predecir función biológica de los genes.
6. Definir la matriz de Markov asociada a una red tipo GOIN y calcular sus índices de información de Shannon a partir de la matriz de Markov para diferentes sistemas biomoleculares relacionados con el *P. Falciparum* incluyendo: grafos moleculares de fármacos, redes de secuencia-recurrencia de genes y proteínas, y, redes de tipo GOIN para obtener índices de estos sistemas en escalas comparables.
7. Desarrollar nuevos modelos NIFPTML para problemas biológicos complejos que involucren varios de los sistemas biomoleculares anteriores a la vez usando los índices de Shannon de estos sistemas definidos en el objetivo anterior.
8. Validar los modelos NIFPTML desarrollado, mediante la comprobación experimental de los resultados obtenidos.

1.2. Hipótesis de trabajo

1.2. Hipótesis de trabajo

En problemas que involucren sistemas biomoleculares complejos compuestos por varios sub-sistemas (fármacos, proteínas, genes, cromosomas, *etc.*) se pueden utilizar representaciones existentes o definir nuevas representaciones de estos sistemas como redes complejas de los sub-sistemas del problema y cuantificar su estructura con parámetros numéricos extraídos de dichas redes que podrían correlacionarse con propiedades biológicas de dichos sistemas usando una estrategia que combine las fases NI+IF+PT+ML. En este sentido es especialmente prometedor los estudios de cromosómica dedicados a la definición de redes complejas, cálculo de índices numéricos, y desarrollo de modelos predictivos “*ad hoc*” para sistemas que involucren cromosomas, utilizando para ello de manera completamente innovadora herramientas, técnicas y procedimientos de computación avanzada e AI/ML en la nueva área de la Cromosómica.

2. Fundamentos teóricos

2. Fundamentos teóricos

2.1. Base de datos disponibles de distintos sistemas biomoleculares

2.1.1. ChEMBL

ChEMBL es una base de datos abierta de sistemas biomoleculares. Los sistemas incluidos son moléculas bioactivas candidatas a fármacos. También suele incluir información de los sistemas biomoleculares diana de los fármacos como son proteínas y genes, pero no reporta información sobre su estructura. Esta base de datos es actualizada manualmente con propiedades de tipo farmacéuticas. Es accesible a través de una interfaz web en <https://www.ebi.ac.uk/chembl>. Permite buscar compuestos probados, ensayos realizados que evalúan la bioactividad, e información adicional sobre el objeto de estos ensayos como: la proteína, ácidos nucleicos, fracciones subcelulares, líneas celulares, tejidos, organismos, estructuras bidimensionales, propiedades moleculares calculadas como LogP y peso molecular. Hay una amplia información adicional, gracias al intercambio con otras bases de datos como PubChem, BioAssay y BindingDB. Además, se registra un número de acceso principal de UniProt y nombres e identificadores a las dianas de organismos de la taxonomía del NCBI (National Center for Biotechnology Information) para poder ampliar aún más la información. Con un conjunto de datos específico extraído de ChEMBL se puede llevar a cabo el entrenamiento de modelos ML para la predicción de dianas y la identificación de herramientas químicas para una diana en concreto, entre otras aplicaciones (Gaulton *et al.*, 2012; Mayr *et al.*, 2018; Mendez *et al.*, 2019).

2.1.2. UniProt

El Recurso Proteico Universal ó Universal Protein Resource (UniProt), es una base de datos abierta de secuencias de proteínas de todas las ramas de la vida de alta calidad con información funcional biológica del sistema biomolecular. Es accesible a través de una interfaz web en <https://www.uniprot.org/> bajo una licencia CC-BY (Creative Commons-Attribution) facilitando que los datos puedan ser reutilizados por otros usuarios. UniProt tiene cuatro componentes principales para usos distintos: la base de conocimiento, los grupos de referencia, el archivo y la base de datos de secuencias metagenómicas y ambientales. Los conjuntos de datos específicos se pueden utilizar para realizar investigaciones biológicas y biomédicas. UniProt mantiene referencias cruzadas con muchas bases de datos como ChEMBL, GenBank, GeneID, NCBI con el objetivo de ampliar aún más su información (UniProt, 2021).

2.1.3. NCBI

El Centro Nacional de Información Biotecnológica ó National Center for Biotechnology Information (NCBI), es una base de datos de biología molecular disponible de manera gratuita. Es accesible a través de una interfaz web en <https://www.ncbi.nlm.nih.gov/>. Contiene información relacionada con secuencias genómicas en la herramienta GenBank y artículos científicos relacionados con biotecnología, biomedicina, genómica entre otros. Tiene herramientas para la visualización y el análisis de las secuencias de ADN (Ácido desoxirribonucleico ó DNA, “DeoxyriboNucleic Acid”), ARN (Ácido Ribonucleico ó RNA, “Ribonucleic Acid”), o sobre las proteínas acompañados por los modelos de genes y otros datos relacionados con la secuencia. Una de las herramientas del NCBI es el Visor de Datos del Genoma (GDV, “Genome Data Viewer”) compuesto de un Visor de Secuencia (SV, Sequence Viewer) mostrando datos de secuencia, de rastreo y de información para realizar búsquedas en un ensamblaje completo del genoma acotando correctamente el cromosoma, secuencia, región o gen específico. Es accesible a través de una interfaz web en <https://www.ncbi.nlm.nih.gov/genome/gdv/>. El GDV reemplazó al NCBI Map Viewer que permitía la visualización del genoma completo. Entre los organismos que incluye GDV están las levaduras y los protistas unicelulares, las plantas, los hongos, los artrópodos, el ganado y todos los organismos habituales (Rangwala *et al.*, 2021; Wolfsberg, 2011).

2.2. Representación de los datos de sistemas biomoleculares

2.2.1. Sistemas complejos

No existe una sola definición para lo que se conoce como sistema complejo biomolecular. De manera generalizada, se podría decir que un sistema complejo es un conjunto de elementos en interacción que cuando uno de los elementos del sistema es modificado todos los demás elementos se ven afectados y, por lo tanto, el sistema complejo cambia. Para que el comportamiento de un sistema complejo sea descrito adecuadamente, es importante conocer bien no solamente sus elementos y las interacciones entre ellos, sino también los valores instantáneos y los cambios dinámicos que pueda tener cada elemento. Este comportamiento es muy difícil de modelar debido a las relaciones, competencias y otras interacciones entre sus elementos o entre un sistema determinado y su entorno. Hoy en día, este término aparece en muchas y muy diversas áreas: economía, sociología, física, química, biología molecular, matemáticas, *etc.* Los sistemas complejos pueden tener distintas propiedades producto de las interacciones, tales como: conexión, retroalimentación, adaptación, emergencia, orden

espontáneo, retroalimentación y no linealidad, entre otras. Los elementos que interactúan en un sistema complejo forman una red compleja, siendo normalmente representado como un grafo de vértices conectados por aristas (Bar-Yam, 2002; Newman, 2003).

2.2.2. Red Compleja de un sistema biomolecular

Una red compleja básicamente tiene dos elementos principales: nodos y enlaces (Figura 1). El conjunto de nodos representa a los elementos y los enlaces representan a las interacciones de un sistema complejo biomolecular. Hoy en día, este término está presente en diferentes niveles de organización de la materia, como las redes biomoleculares, biológicas, tecnológicas y sociales. Por ejemplo, la World Wide Web puede ser modelada como una red de información, donde los elementos serán las páginas informativas y los enlaces entre estos elementos serán los hipervínculos. Aplicando en otro nivel, el cromosoma puede ser modelado como una red GOIN, donde los elementos son los genes y los enlaces entre estos genes es la inversión genética (Figura 2). Las redes complejas pueden tener distintas propiedades, tales como el grado y la centralidad de intermediación. El grado representa el número de enlaces que conectan a un nodo. Los nodos con una centralidad alta sirven como puentes entre diferentes partes de un nodo (Proulx *et al.*, 2005; Mashaghi *et al.*, 2004; Newman, 2003; Quevedo *et al.*, 2018).

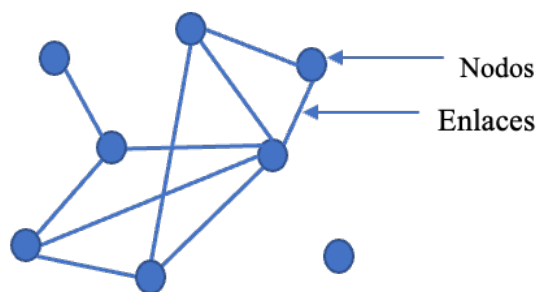


Figura 1. Ejemplo de una pequeña red con 8 nodos y 10 enlaces

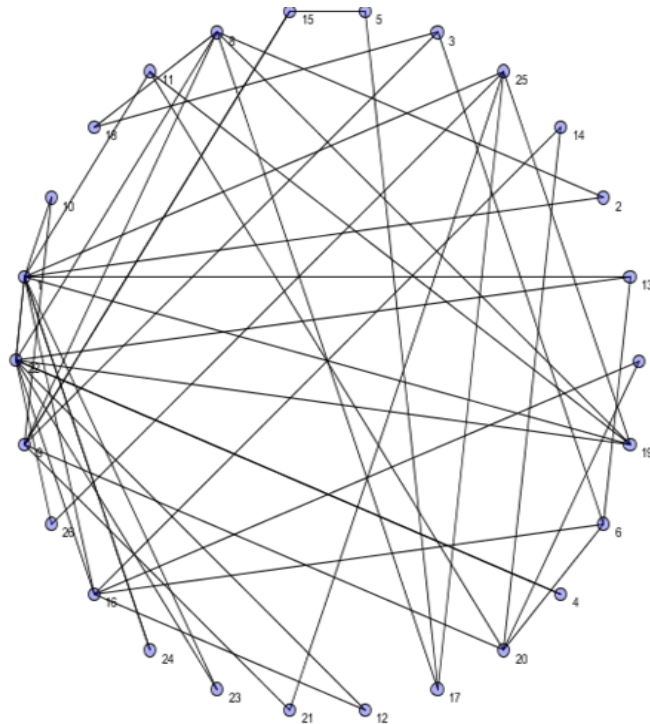


Figura 2. Ejemplo de los tipos de redes presente en este documento. Una visualización de la estructura de la red GOIN del Cromosoma I de *P. Falciparum* con un total de 157 nodos (genes) y 289 enlaces (inversión genética).

2.2.3. Grafo

Una red compleja puede estar representada a través de un grafo. Los elementos de un grafo son los vértices o llamados nodos (V) y las aristas o llamados arcos (E). Un grafo es un par (V, E) y se representa a través de la siguiente fórmula $G = (V, E)$, donde V es el conjunto no vacío de elementos y E es un subconjunto de pares $\{x, y\}$ ordenados o no-ordenados de vértices donde $x, y \in V$. Cuando el grafo G es dirigido, el subconjunto de pares ordenados (x, y) muestran la existencia de una arista desde el vértice x hacia el vértice y , no necesariamente tiene la arista inversa desde el vértice y hacia x . Cuando el grafo G es no dirigido, muestran aristas bidireccionales, debido a que los pares no-ordenados $\{x, y\}$ e $\{y, x\}$ de vértices adyacentes son equivalentes. Se puede representar el par (V, E) mediante la matriz de adyacencia A con N filas y N columnas, donde N es el número de vértices del grafo, por lo que la matriz es cuadrada (Figura 3). Se establece una ordenación entre los elementos de V para asignar filas y columnas de la matriz A a los vértices; de manera general, se puede tomar $V = \{v_1, v_2, \dots, v_N\}$ y la ordenación $v_1 < v_2 < \dots < v_N$. De acuerdo con esta ordenación, la fila i -ésima de la matriz A es asignada al vértice v_i para indicar las aristas con origen v_i y destino v_j mediante el elemento $a_{ij} = 1$ si existe una arista desde v_i hasta v_j , y 0 en caso contrario (Bender *et al.*, 2010; Claude, 1958; Biggs *et al.*, 1986).

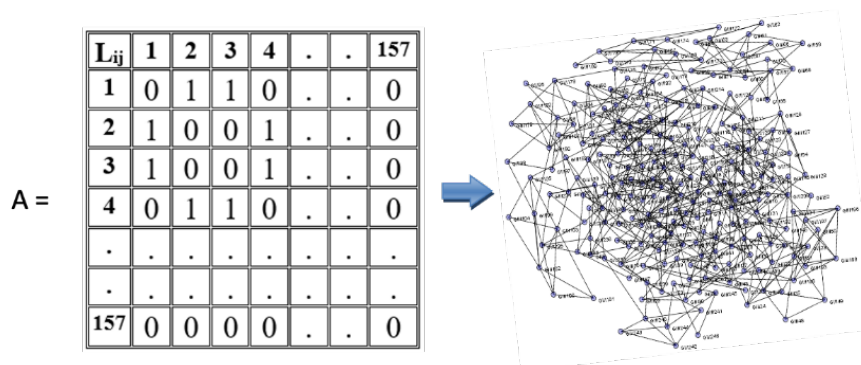


Figura 3. Ejemplo matriz de adyacencia y su grafo. Una visualización de la matriz A de 157 nodos que representan a los genes del Cromosoma I de *P. Falciparum* y 289 aristas.

2.2.4. Grafo estrella

Un grafo estrella está representada por S_k , es un grafo bipartito completo $k_{1,k}$, un árbol con un nodo interno y k hojas, pero sin nodos internos y k+1 hojas cuando k es menor o igual a 1. En la Figura 4 el número de vértices es $k=6+1$, con un solo nodo interno y 6 hojas, además tiene 6 aristas y su diámetro es igual a 2. Son grafos conectados, donde un solo nodo tiene grado mayor que uno (Fertin *et al.*, 2004; Faudree *et al.*, 1997).

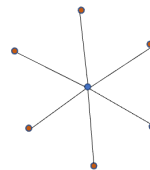


Figura 4. El grafo estrella S_6 (sin tomar en cuenta el nodo interno) o S_7 (con nodo interno)

El grafo estrella es un caso especial de árboles con N vértices donde un vértice tiene N-1 grados de libertad y los vértices restantes tienen un solo grado de libertad (Harary, 1969). En el caso de las proteínas, cada una de las 20 posibles ramas ("rayos") de la estrella contiene el mismo tipo de aminoácido y el centro de la estrella es un vértice no aminoácido. Una secuencia primaria de proteínas puede representarse mediante diferentes formas de grafos, que pueden asociarse a distintas matrices de distancia. El mejor método para construir un grafo estrella estándar es el siguiente: cada aminoácido/vértice contiene la posición en la secuencia original y las ramas se etiquetan en el orden alfabético del código de aminoácidos de 3 letras (Randic *et al.*, 2007). El grafo incrustado contiene la conectividad inicial de la secuencia en la cadena de proteínas. La Figura 5 lado izquierdo, presenta los grafos en estrella no incrustados de dos ejemplos (ATGGAATATC y ATGGCAATAT) utilizando el orden alfabético del código de nucleótidos de una letra. Hay un nodo central con grado mayor a dos, todos los demás nodos tienen grado igual a 2 excepto los últimos nodos mas alejados del nodo central con grado 1.

2.2.5. Grafo de secuencia

Los grafos de secuencia, también llamado grafos de alineación, tiene 2 nodos terminales de grado 1 y los demás nodos tienen grado 2. En la Figura 5 lado derecho se puede observar los grafos de secuencia, es un grafo en el que los vértices representan los aminoácidos y las aristas representan la secuencia de los genes de un cromosoma (Quevedo-Tumaili *et al.*, 2021).

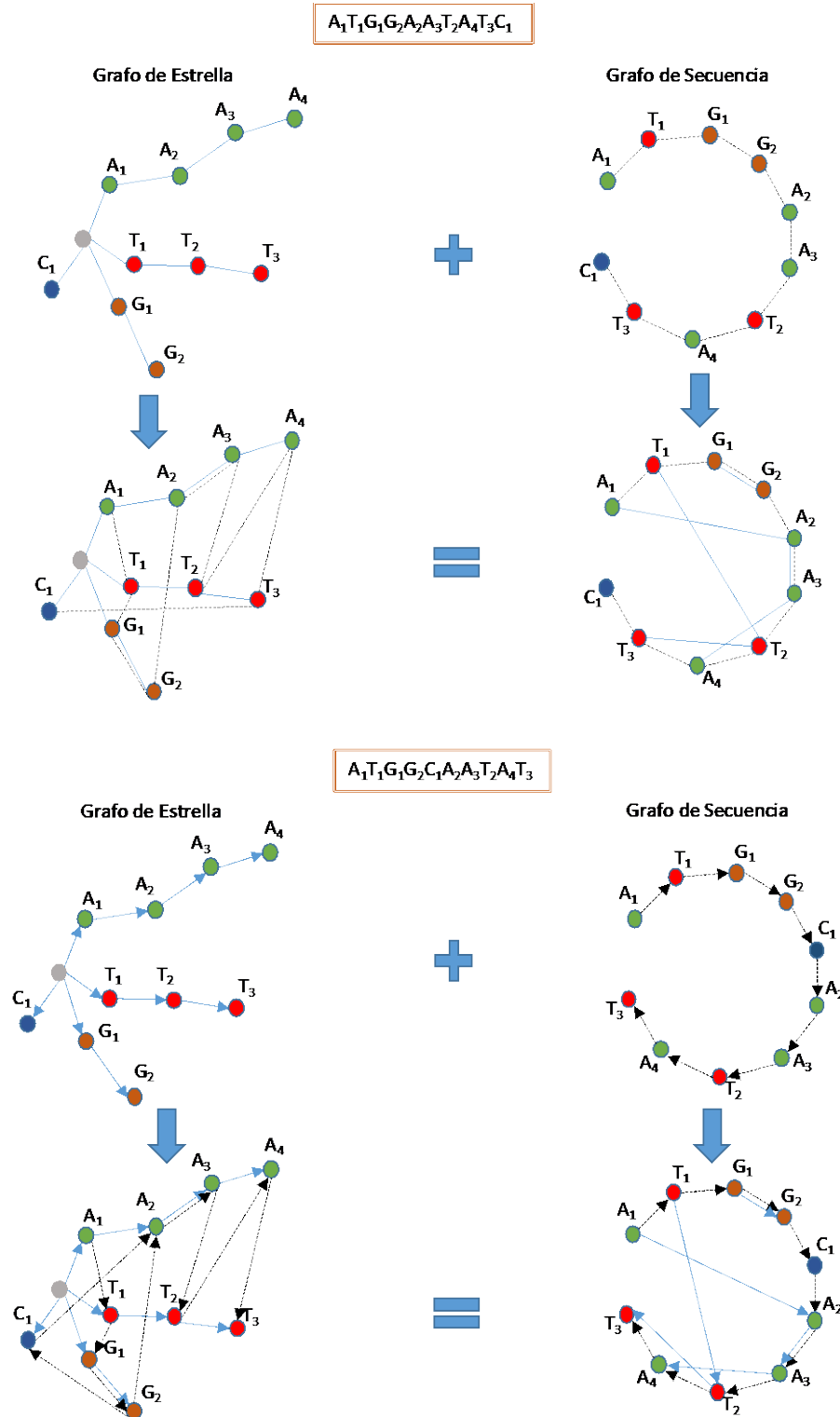


Figura 5. Grafo de Estrella (Recurrencia) vs. Grafo de Secuencia dirigido y no-dirigido

2.2.6. Grafos moleculares

En términos de teoría de grafos, un grafo molecular o llamado también grafo químico es una representación de la fórmula estructural de un compuesto químico. Un grafo molecular es un grafo donde los vértices representan a los átomos del compuesto químico y las aristas a los enlaces químicos (Figura 6). Sus vértices están etiquetados con los tipos de átomos correspondientes y las aristas con los tipos de enlaces (Minkin, 1999). Para fines particulares se pueden ignorar cualquiera de los etiquetados. Un grafo molecular sin hidrógeno o con hidrógeno suprimido es el grafo molecular con los vértices de hidrógeno suprimidos. En algunos casos, como por ejemplo para calcular los índices topológicos, un grafo molecular es un grafo conectado y no dirigido que admite una correspondencia uno a uno con la fórmula estructural de un compuesto químico en el que los vértices del grafo corresponden a los átomos de la molécula y las aristas del grafo corresponden a los enlaces químicos entre estos átomos (King, 1983).

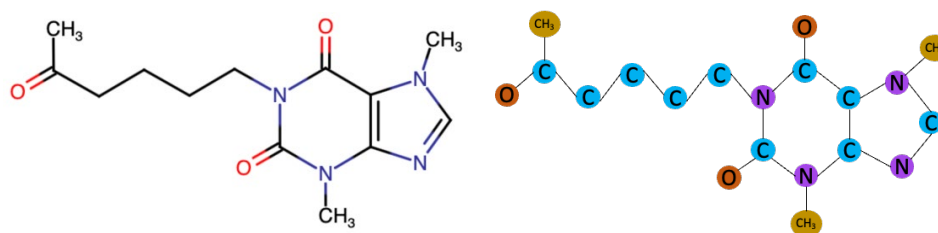


Figura 6. Generación de representación de la estructura molecular de la fórmula $C_{13}H_{18}N_4O_3$ (lado izquierdo) en el grafo molecular con hidrógenos suprimidos (lado derecho)

2.2.7. Red Neuronal Recurrente

Las redes de neuronas artificiales recurrentes ó “Recurrent Neural Networks” (RNN), es una familia de redes de neuronas artificiales, donde las conexiones entre los nodos forman un grafo dirigido, o no dirigido, a lo largo de una secuencia temporal permitiendo un comportamiento dinámico temporal. Teniendo en cuenta que secuencia temporal no tiene que significar literalmente tiempo, puede significar también la posición de un objeto en una secuencia. En este tipo de redes la salida de un nodo depende de las salidas de los nodos anteriores de la red. El reconocimiento de escritura a mano o del habla son ejemplos donde se aplicaron este tipo de redes (Dupond, 2019; Graves *et al.*, 2009; Sak *et al.*, 2014). La representación básica y compacta es la siguiente (Figura 7):

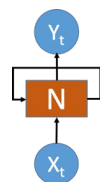


Figura 7. Representación básica y compacta de una RNN

La representación básica de una red recurrente de una secuencia finita para predecir el siguiente elemento de la secuencia en función únicamente de los N pasos atrás, siendo posible entrenar y tener resultados, se puede observar en la Figura 8:

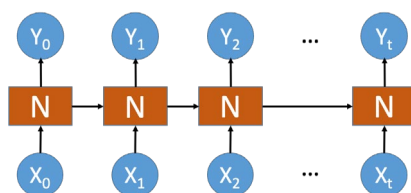


Figura 8. Representación básica y compacta de una RNN de una secuencia finita

2.3. Caracterización numérica de sistemas biomoleculares

2.3.1. Descriptores moleculares

Los descriptores moleculares son parámetros numéricos que representan la información química codificada contenida en la molécula que permite realizar estudios matemáticos sobre las moléculas. Este término desempeña un papel importante en las ciencias farmacéuticas, la química, la política de protección del medio ambiente y las investigaciones sanitarias.

Todeschini y Consonni definen al "descriptor molecular" como el resultado final de un procedimiento lógico y matemático que transforma la información química codificada en una representación simbólica de una molécula, en un número útil o en el resultado de algún experimento estandarizado (Todeschini & Consonni, 2000).

Según esta definición, los descriptores moleculares se dividen en dos categorías principales: a) las mediciones experimentales y b) los teóricos. Como un ejemplo de la primera categoría se puede mencionar: el log P , el momento dipolar, la refractividad molar, la polarizabilidad y, en general, las propiedades fisicoquímicas aditivas. La segunda categoría que se derivan de una representación simbólica de la molécula y que pueden clasificarse a su vez según los distintos tipos de representación molecular. Las principales clases de descriptores moleculares teóricos son:

- a) 0D: son descriptores constitucionales, descriptores de recuento
- b) 1D: es la lista de fragmentos estructurales, huellas dactilares

- c) 2D: son características topológicas ó Índices Topológicos ó Topological Indices (ITs), teoría de grafos
- d) 3D: basados en descriptores de información espacial como de tamaño, superficie y volumen
- e) 4D: es la combinación de los anteriores añadiendo características cuánticas

Algunas propiedades de los descriptores moleculares son: la invariancia, que permite obtener el mismo valor numérico de forma independiente a características específicas de la representación molecular y la degeneración, en la cual dos moléculas distintas no pueden tener el mismo descriptor molecular (Mauri *et al.*, 2017).

2.3.2. Índice de Shannon

El parámetro numérico llamado índice de Shannon ó entropía de Shannon fue propuesto por Claude Shannon en 1948 para cuantificar la entropía en cadenas de texto con relación al contenido de información. Es decir, cuantas más letras haya, y cuanto más cercanas sean sus abundancias proporcionales en la cadena de interés, más difícil será predecir correctamente qué letra será la siguiente en la cadena.

El índice de Shannon pondera la incertidumbre (entropía) asociada a esta predicción. Se puede calcular de la siguiente manera:

$$Sh = - \sum_{i=1}^R p_i \ln p_i$$

Donde, p_i es la proporción de caracteres que pertenecen al i -ésimo tipo de letra en la cadena de interés. En ecología, p_i suele ser la proporción de individuos que pertenecen a la i -ésima especie en el conjunto de datos de interés. Entonces, la entropía de Shannon cuantifica la incertidumbre en la predicción de la identidad de la especie de un individuo que se toma al azar del conjunto de datos. En los sistemas moleculares, son útiles para cuantificar la información sobre las condiciones estructurales y experimentales de ensayo de todos los sistemas implicados (fármacos, proteínas, redes de genes, *etc.*).

La ecuación tiene el logaritmo natural, pero al tratarse de calcular el índice de Shannon puede elegirse libremente. El propio Claude Shannon popularizó las bases logarítmicas 2, 10 y e en las aplicaciones. La base 2 corresponde a la unidad de medida denominado dígitos binarios (bits), la base 10 a los dígitos decimales (decits) y la base e a los dígitos naturales (nats).

Es necesario comparar los valores del índice de Shannon en iguales bases logarítmicas, caso contrario se puede convertir la base a a la base b con la multiplicación por \log_{ba} (Shannon, 1948).

El índice de Shannon (Sh) está relacionado con la media geométrica ponderada de las abundancias proporcionales de los tipos. (Tuomisto, 2010).

2.3.3. Cadenas de Markov

Es un modelo estocástico que describe una secuencia de eventos posibles en la que la probabilidad de cada evento depende solo del estado alcanzado en el evento anterior. Las cadenas de Markov son parámetros numéricos usados para calcular los índices de información de Shannon de diferentes sistemas, incluyendo simulaciones relevantes en Epidemiología de transmisión de la enfermedad (Riera-Fernandez *et al.*, 2012). Permiten también cuantificar la información estructural/funcional relevante de los sistemas, como por ejemplo la estructura química de los fármacos, las secuencias de proteína, los genes y los cromosomas (Munteanu *et al.*, 2008).

2.4. Software de computación avanzada para el cálculo de parámetros numéricos de sistemas biomoleculares

2.4.1. Dragon

DRAGON es una aplicación desarrollada por el Grupo de investigación de Quimiometría y QSAR (Relación Cuantitativa Estructura-Actividad, en idioma inglés es Quantitative Structure-Activity Relationship) de Milán (<http://michem.disat.unimib.it/chm/>) que permite calcular descriptores moleculares con el fin de evaluar las relaciones moleculares estructura-actividad o estructura-propiedad, y también, permite analizar la similitud y cribado de alto rendimiento de bases de datos de moléculas. Puede calcular desde lo más simple, como el tipo de átomos, hasta los descriptores topológicos y geométricos. La aplicación funciona en Windows y Linux, con una interfaz gráfica y una línea de comandos. Permite calcular 1600 descriptores moleculares agrupados en 20 bloques lógicos.

DRAGON fue lanzado en 1997 y se siguen obteniendo las actualizaciones y las inclusiones de nuevos descriptores moleculares para avanzar en la investigación en QSAR (Todeschini, Consonni, Mauri, y Pavan, 2005). Algunas propiedades moleculares han sido calculadas mediante esta aplicación tales como ΔLOGP , TPSA y NRV. Las dos primeras propiedades se refieren al coeficiente de partición n-octanol/agua y al área de superficie polar topológica

respectivamente, estas dos propiedades permiten cuantificar la estructura molecular del fármaco mediante una suma ponderada de fragmentos moleculares en las moléculas. Mientras que NRV es el número de violaciones a la regla de V, lo que cuantifica la semejanza/similitud de un compuesto con respecto a los fármacos conocidos basándose en los enlaces de hidrógeno, en el peso molecular, *etc.* DRAGON crea formatos tales como, SMILES (Simplified Molecular Input Line Entry Specification, Especificación Simplificada de la línea de entrada Molecular), CML, MDL, HyperChem, Sybyl, MacroModel. La Figura 9 es una imagen de la interfaz gráfica de la aplicación de escritorio Dragon 6.0 y la Figura 10 es una imagen de la interfaz gráfica de la versión online gratuita de E-DRAGON 1.0.

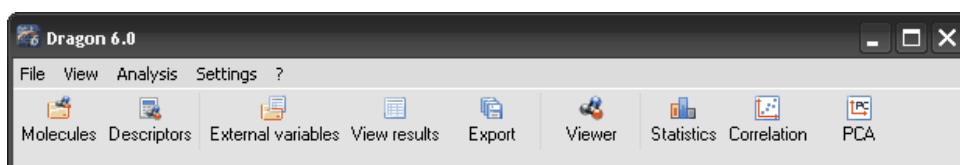


Figura 9. Interfaz gráfica de la aplicación de escritorio Dragon 6.0

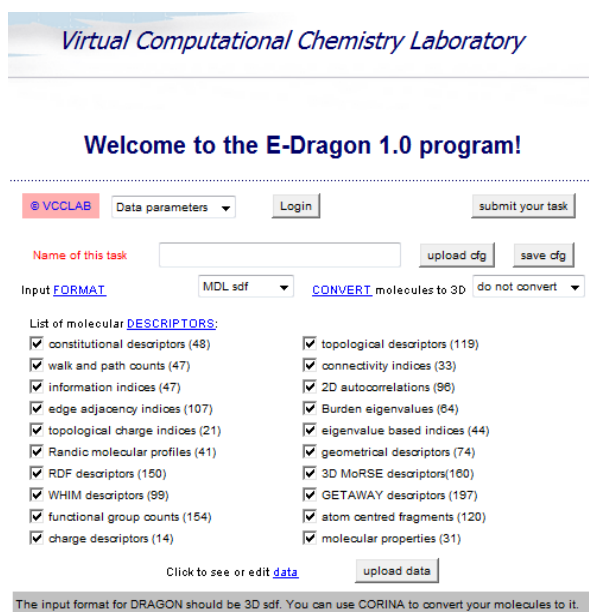


Figura 10. Interfaz gráfica de la versión online E-Dragon 1.0

2.4.2. CentiBiN

La aplicación gratuita de cálculo de Centralidades en las Redes Biológicas ó “Centralities in Biological Networks” (CentiBiN), es utilizada para el cálculo y la exploración de centralidades en redes biológicas incluyendo redes de sistemas biomoleculares, como las redes de interacción proteína-proteína. Está disponible en <http://centibin.ipk-gatersleben.de/> en

dos versiones diferentes: una aplicación Java Web Start y una aplicación Windows instalable que incluye una guía de usuario y varios conjuntos de datos de ejemplo. Puede calcular 17 centralidades diferentes para redes dirigidas o no dirigidas, desde medidas locales, es decir, medidas que sólo consideran la vecindad directa de un elemento de la red, hasta medidas globales. Esta aplicación soporta la exploración de la distribución de centralidad mediante la visualización de los elementos centrales dentro de la red y proporciona varios mecanismos de diseño para la generación automática de representaciones gráficas de una red. CentiBiN soporta diferentes formatos de entrada, especialmente para redes biológicas, y la exportación de las centralidades calculadas a otras aplicaciones. CentiBiN ayuda a los investigadores de biología de sistemas a identificar los elementos cruciales de las redes biológicas (Junker et al., 2006), tal como muestra el interfaz de la aplicación en la Figura 11.

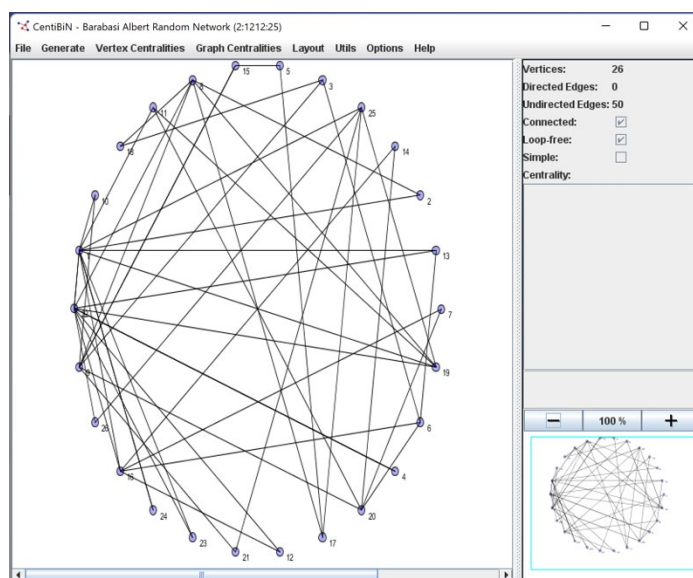


Figura 11. Interfaz de la aplicación CentiBiN

2.4.3. MARCH-INSIDE

La aplicación de Invariantes químicas de Markov para la simulación y el diseño de redes ó Markov Chemical Invariants for Networks Simulation and Design (MARCH-INSIDE), está desarrollada por González-Díaz et al., y permite el cálculo de parámetros numéricos de sistemas biomoleculares basados en cadenas de Markov. Dichos parámetros pueden ser usados por algoritmos AI/ML para encontrar Relaciones Cuantitativas Estructura-Actividad conocida como Quantitative Structure-Activity Relationship (QSAR). MARCH-INSIDE utiliza la teoría de las cadenas de Markov para generar parámetros que describen numéricamente la estructura química de los fármacos y las dianas farmacológicas, permite calcular las entropías de

información de Shannon de los fármacos y de las dianas farmacológicas (Zhao *et al.*, 2019). También se ha usado esta aplicación para introducir los códigos denominados Simplified Molecular Input Line Entry Specification, o Especificación Simplificada de la Línea de Entrada Molecular (SMILES) de los compuestos. MARCH-INSIDE, cuya interfaz gráfica se puede observar en la Figura 12, utiliza grafos para representar las partes del subsistema (átomos) y las relaciones (enlaces químicos) entre ellas en la estructura del subsistema. La aplicación asocia una matriz de adyacencia de átomos a los respectivos grafos para llevar a cabo una representación numérica del sistema. La matriz de adyacencia se transforma en matriz de Markov y de esta se calculan las potencias naturales de orden k^{th} . La aplicación utiliza las ecuaciones de Chapman-Kolmogórov para calcular las probabilidades absolutas de cada nodo de un subsistema determinado (Munteanu *et al.*, 2008; Zhao *et al.*, 2019).

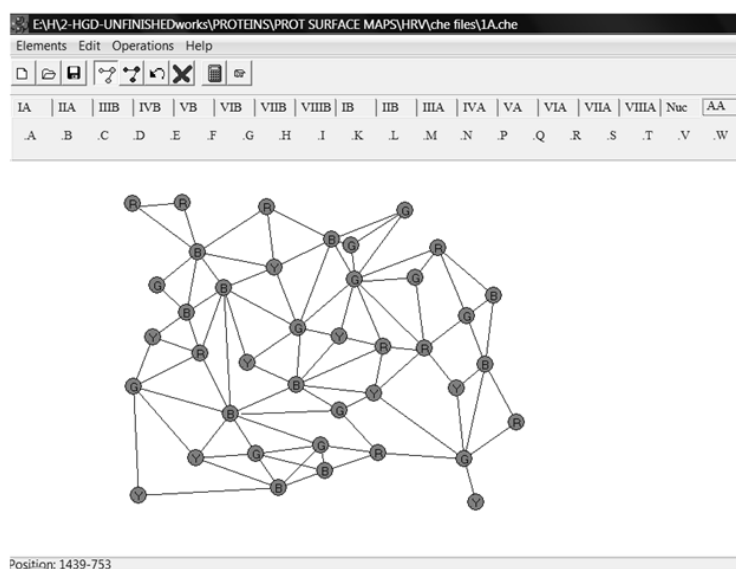


Figura 12. Interfaz gráfica de la aplicación MARCH-INSIDE

2.4.4. S2Snet

Red de Secuencia a Estrella ó “Sequence to Star Network” (S2SNet) es un software de gran utilidad para la representación y el cálculo de parámetros numéricos de sistemas biomoleculares. S2SNet es una aplicación gratuita de Python que utiliza wxPython para la interfaz gráfica de usuario y Graphviz como back-end de trazado. Se encuentra disponible en <https://github.com/muntisa/S2SNet>. Esta aplicación transforma secuencias de caracteres en índices topológicos (TIs) de redes complejas de tipo estrella (Star Network, SN) (Munteanu *et al.*, 2008). Con estos índices se pueden realizar diversos análisis estadísticos o crear modelos QSAR. Las cadenas de aminoácidos de las proteínas y los ácidos nucleicos son algunos

ejemplos de secuencias. Se puede utilizar S2SNet para estudiar distintos sistemas, desde sistemas de átomos simples en pequeñas moléculas anti-cancerígenas, hasta sistemas complejos de redes metabólicas, sociales, computacionales o sistemas biológicos. Algunos ejemplos de secuencias son las cadenas de aminoácidos de las proteínas, las cadenas de ADN/ARN, los espectros de masas o los electroencefalogramas (EEG). S2SNet puede calcular los valores del índice de información de Shannon de la Proteína, del Gen y el Cromosoma sobre la secuencia y la recurrencia de diferentes aminoácidos en las proteínas, nucleótidos en los genes, y genes en los cromosomas.

S2SNet, (como puede verse en la Figura 13) utiliza un gráfico para representar las partes del subsistema (nodos) y las relaciones (enlace) entre ellas en la estructura del subsistema de proteínas, genes y cromosomas. Las partes de los subsistemas son átomos, aminoácidos, bases de nucleótidos o genes. Los enlaces entre ellos son enlaces químicos, enlaces peptídicos, secuencia de genes o posición de genes según el sistema. S2SNet asocia una matriz de adyacencia de nodos $A(\text{Subsistema}_s)$ a los respectivos grafos para llevar a cabo una representación numérica del sistema. A continuación, transforma la matriz de adyacencia de cada subsistema $A(\text{Subsistema}_s)$ en una matriz de Markov $\Pi_1(\text{Subsistema}_s)$. Después, calcula las potencias naturales de orden k^{th} para cada matriz $\Pi_1(\text{Subsistema}_s)$. Por último, utiliza las ecuaciones de Chapman-Kolmogórov para calcular las probabilidades absolutas ${}^a p(n/s)_k$ de cada nodo de un subsistema determinado (n/s) (Munteanu *et al.*, 2008; Zhao *et al.*, 2019). Con estas probabilidades, S2Snet realiza el cálculo de los diferentes valores de Shannon de la proteína, gen y cromosoma.

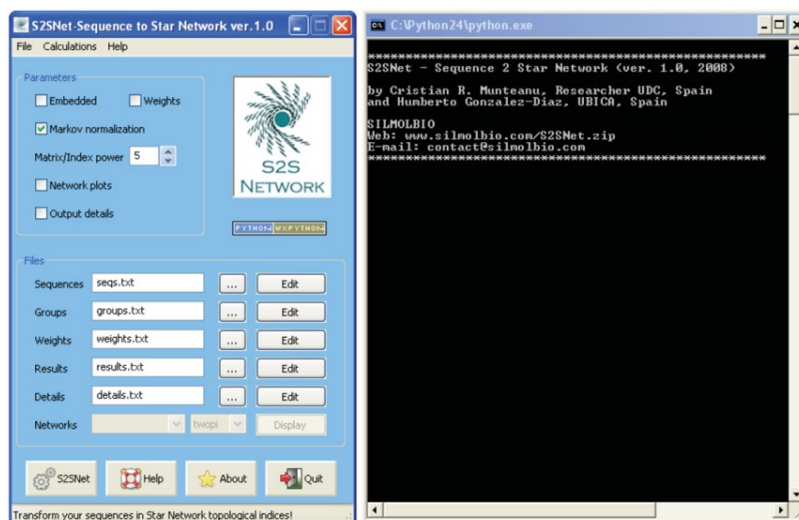


Figura 13. Interfaz gráfica de la aplicación S2Snet

2.5. Métodos de IA/ML para análisis de parámetros numéricos de sistemas biomoleculares

2.5.1. LDA/GDA

El Análisis Discriminante Lineal ó “Linear Discriminant Analysis” (LDA) es un método IA/ML de gran utilidad para el estudio de sistemas biomoleculares. Es una técnica fundamental de análisis de datos propuesta por R. Fisher para discriminar entre diferentes tipos de flores (Fisher, 1936). Es una técnica estadística que trata con dos tipos de problemas: discriminación descriptiva, que describe si dos o más poblaciones son diferentes entre sí; y clasificación de sentido estricto; en la que, dados dos o mas poblaciones, se averigua a cuál de las poblaciones pertenece un objeto (Abraira & Pérez, 1996; Gómez & Martínez, 2001). LDA requiere la obtención de combinaciones lineales o no lineales de características independientes que discriminarán entre grupos definidos “*a priori*”, de manera que los errores que clasifican mal deben ser mínimos. Este método se aplica en la predicción de quiebra, reconocimiento facial, marketing, estudios biomédicos y ciencias de la tierra, entre otras. Un caso general del LDA es el método de Análisis General Discriminante ó “General Discriminant Analysis” (GDA) que es un método para el análisis discriminante no lineal utilizando el operador de la función kernel. El método GDA proporciona un “mapeo” de los vectores de entrada en un espacio de características de alta dimensión. El objetivo de LDA/GDA es encontrar una proyección de las características en un espacio de menor dimensión maximizando la relación entre la dispersión de un grupo y la dispersión de otro.

2.5.2. Arbol de Clasificación

El árbol de clasificación ó “Classification Tree” (CT), es una herramienta muy usada en ML para sistemas biomoleculares que sirve de apoyo para tomar decisiones, permite reconocer, diferenciar y comprender los objetos e ideas utilizando un modelo de decisiones en forma de árbol y sus posibles consecuencias, incluyendo los costos de los recursos y la utilidad (Brinton, 1939). CT tiene tres tipos de nodos: los nodos de decisión que generalmente están representados por cuadrados, los nodos de probabilidad representados por círculos y los nodos finales por triángulos. Es una técnica simple de entender e interpretar y se puede combinar con otras técnicas de toma de decisión (Quevedo-Tumaili *et al.*, 2021).

2.5.3. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte ó “Support-Vector Machines” (SVM), es uno de los modelos de ML más usados, versátil y potente en sistemas biomoleculares, capaz de realizar clasificaciones lineales o no lineales, regresiones e incluso detección de valores atípicos. Funciona bien con la clasificación de conjuntos de datos complejos de tamaño pequeño o mediano (Patle & Chouhan, 2013). Son aplicaciones de SVM: la categorización de textos, el reconocimiento de caracteres escritos a mano, la clasificación de imágenes, el análisis de biosecuencias, etc.

Las SVM realizan la clasificación construyendo un hiperplano de N dimensiones y tiene cuatro conceptos básicos (Furey *et al.*, 2000; Cristianini & Shawe-Taylor, 2000):

- 1) El hiperplano de separación separa las diferentes muestras del conjunto de datos.
- 2) El hiperplano de margen máximo adopta la máxima distancia de cualquiera de los perfiles de expresión dados.
- 3) El margen suave permite manejar y lidiar con los errores en los datos sin afectar el resultado final.
- 4) La función del núcleo permite realizar una clasificación bidimensional de un conjunto de datos originalmente unidimensionales. Las SVM permiten una clasificación casi perfecta en el conjunto de datos, puede tratar con más de dos variables predictoras y separar los puntos con curvas no lineales.

2.5.4. Redes de Neuronas Artificiales ó “Artificial Neural Networks” (ANN)

Las ANN, tienen un comportamiento excelente clasificando patrones. Nacen a partir de los trabajos realizados en el ámbito de la Cibernética a principios de los años 40, en concreto en 1943 con la presentación de la “neurona formal”, por parte de Warren MacCulloch y Walter Pitts, en su artículo “*A Logical Calculus of Ideas Inmanent in Nervous Activity*” que ha servido para inspirar en la década de los 50 las propuestas de modelos conexionistas y han crecido exponencialmente en los últimos años. Están formadas por elementos que proceden de forma similar a las funciones más elementales de la neurona biológica. Por su potencial, se aplican en varios campos como la ingeniería, la física y la biología, entre otros muchos. Los modelos ANN pueden aprender bajo supervisión o son de tipo autoorganizado. El desarrollo del hardware para desarrollar modelos ANN es muy limitado aunque la arquitectura paralela sea atractiva para el procesamiento paralelo (Khaparde *et al.*, 1991). Ultimamente, la utilización de tarjetas gráficas, usadas de forma primaria para la gestión y procesamiento de imágenes, en el procesamiento

paralelo de la información está aportando nuevas posibilidades, sobre todo en el ámbito del “Deep Learning”.

2.5.5. Modelo de Aprendizaje Automático con Teoría de la Perturbación

El Modelo de Aprendizaje Automático con Teoría de la Perturbación ó “Perturbation Theory Machine Learning” (PTML), como su nombre indica, es un modelo que combina ideas de la Teoría de la Perturbación (PT) con desarrollos en Machine Learning (ML). Puede ser construido de forma aditivo lineal hasta obtener uno complejo, que parte de un valor conocido de la actividad biológica esperada para un grupo de compuestos y añade el efecto de las perturbaciones de la estructura y, o, condiciones de ensayo del nuevo caso con respecto a la referencia. El modelo PTML utiliza los operadores de la PT para cuantificar la desviación (perturbaciones) en las variables continuas (parámetros estructurales, tiempo, concentración, *etc.*) con respecto a la información funcional codificada por las variables categóricas (condiciones experimentales) (Nocedo-Mena *et al.*, 2019). Los modelos PTML permiten predecir diferentes parámetros de actividad biológica y, o, toxicidad (K_i , IC_{50} , LD_{50} , K_m , % de inhibición, *etc.*) para la interacción de diferentes compuestos con diferentes dianas biológicas (proteínas, tejidos, líneas celulares, organismos patógenos, *etc.*). La utilización de los modelos PTML reducen los costos (Ferreira *et al.*, 2018; Díaz-Alarcía *et al.*, 2019; Santana *et al.*, 2020).

2.5.6. Fusión de la Información para el Modelo de Aprendizaje Automático con Teoría de la Perturbación

La Fusión de la Información para el Modelo de Aprendizaje Automático con Teoría de la Perturbación ó “Information Fusion Perturbation Theory Machine Learning” (IFPTML) se ha utilizado en la nanotecnología, química médica, *etc.* Permite modelizar un gran conjunto de datos que puede incluir características de Big Data. El modelo IFPTML combina técnicas de Fusión de Información (IF) de datos de diversas fuentes, incluso heterogéneas, con modelos de PTML. Pueden incluir datos sobre la secuencia de proteínas del GenBank, redes metabólicas, nanopartículas, o incluso información sobre datos de epidemiología en los condados de Estados Unidos, *etc.* (Gonzalez-Diaz *et al.*, 2013; Santana *et al.*, 2019; Nocedo-Mena *et al.*, 2019).

2.6. Software para entrenar modelos AI/ML

2.6.1. Statistics

Statistics, cuya interfaz gráfica puede verse en la Figura 14, es un paquete de software propietario desarrollado originalmente por StatSoft a mediados de la década de 1980. A partir del 15 de mayo de 2017 el software es mantenido por TIBCO. Su página oficial es www.tibco.com/data-science-and-streaming. Permite el análisis avanzado de datos, gestión de datos, minería de datos para la extracción y descubrimiento de patrones, ML, estadísticas, análisis de texto y procedimientos de visualización de datos para la representación gráfica de datos. Incluye además modelos predictivos, de clasificación, agrupamiento y técnicas exploratorias. Puede construir gráficos analíticos y exploratorios con gráficos de 2 y 3 dimensiones (Hill y Lewicki, 2007).

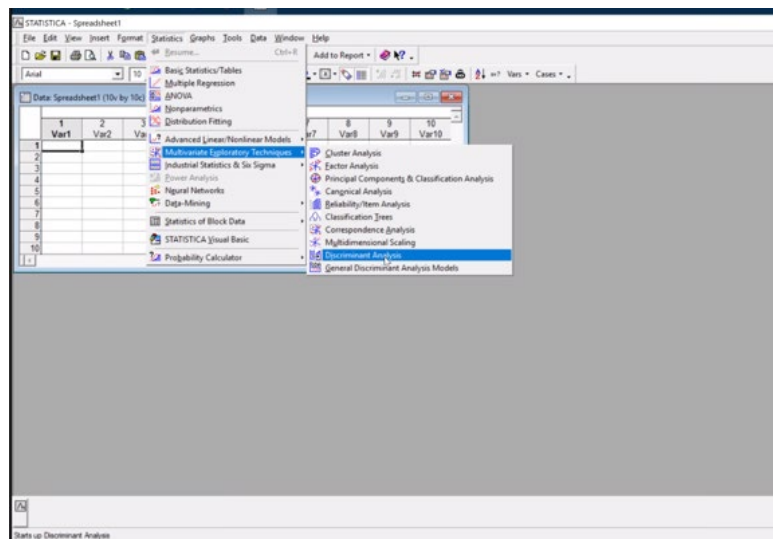


Figura 14. Interfaz gráfica de la aplicación Statistics

2.6.2. Weka

El software libre de entorno Waikato para el análisis del conocimiento ó “Waikato Environment for Knowledge Analysis” (Weka), es un software con licencia GNU desarrollado por la Universidad de Waikato de Nueva Zelanda e implementado en Java. Está disponible en www.cs.waikato.ac.nz/ml/weka. El software tiene interfaces gráficas de usuario para un fácil acceso a estas funciones (Witten *et al.*, 2011). Contiene una colección de módulos de visualización y algoritmos para el análisis de datos y la modelización predictiva. Permite probar y comparar rápidamente diferentes tipos de ML en nuevos conjuntos de datos (Hall *et al.*, 2009).

En la minería de datos incluye varias tareas como el preprocesamiento de datos, clasificación, regresión, análisis de conglomerados, visualización y selección de variables.

2.7. Criterios de calidad de métodos de AI/ML de clasificación para sistemas biomoleculares

2.7.1. Sensibilidad

Sensibilidad, es un método utilizado para representar la Tasa de Verdaderos Positivos ó “True Positive Rate” (TPR) y se calcula de la siguiente manera:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

Donde, TP (True Positive) son Verdaderos Positivos, un resultado de prueba que indica correctamente la presencia de una condición o característica; P (condition Positive) es la condición Positiva que representa el número de casos positivos reales en los datos, FN (False Negative) son los Falsos Negativos, un resultado de prueba que indica erróneamente que una condición o atributo particular está ausente. Por último, FNR (False Negative Rate) es la Tasa de Falsos Negativos (Parikh *et al.*, 2008; Altman & Bland, 1994).

2.7.2. Especificidad

Especificidad, es un método utilizado para representar la Tasa de Verdaderos Negativos ó True Negative Rate (TNR) y se calcula de la siguiente manera:

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

Donde, TN (True Negative) son los Verdaderos Negativos, un resultado de prueba que indica correctamente la ausencia de una condición o característica; N (condition Negative) es la condición Negativa que representa el número de casos negativos reales en los datos, FP (False Positive) son los Falsos Positivos, un resultado de prueba que indica erróneamente que esta presente una condición o característica particular. Por último, FPR (False Positive Rate) es la Tasa de Falsos Positivos (Parikh *et al.*, 2008; Altman & Bland, 1994).

2.7.3. Análisis del Área bajo la Curva Característica Operativa del Receptor

El Análisis del Área bajo la Curva Característica Operativa del Receptor ó “Analysis of the Area Under Receiver Operating Characteristic Curve” (AUROC) es un parámetro numérico

utilizado con los clasificadores binarios para la representación gráfica de la sensibilidad frente a la especificidad. AUROC traza la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR). La TPR mide hasta qué punto una prueba diagnóstica es capaz de clasificar las instancias positivas correctamente, entre todos los casos positivos disponibles durante la prueba. La FPR es la proporción de instancias negativas que se clasifican de manera incorrecta como positivas y es igual a $1 - \text{TNR}$. Un espacio ROC está definido por FPR en el eje “x” y VPR en el eje “y”, representa los intercambios entre verdaderos positivos (en principio, beneficios) y falsos positivos (en principio, costes). El gráfico AUROC también es conocido como la representación de sensibilidad frente a $(1 - \text{TNR})$ dado que TPR es equivalente a sensibilidad y FPR es igual a $1 - \text{especificidad}$. Cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio AUROC. Según se observa en la Figura 15, la mejor predicción en el espacio AUROC se ubicaría en un punto en la esquina superior lado izquierdo o coordenada $(0, 1)$ con un 100% de sensibilidad y un 100% de especificidad. La diagonal divide el espacio AUROC. Los puntos por encima de la diagonal representan buenos resultados de clasificación (mejor que el azar), los puntos por debajo de la línea representan resultados pobres (peor que al azar). A manera de guía para interpretar la curva AUROC se establecen los siguientes intervalos para los puntos de AUAUC: $[0.5]$: Como lanzar una moneda, $[0.5, 0.6)$: Test malo, $[0.6, 0.75)$: Test regular, $[0.75, 0.9)$: Test bueno, $[0.9, 0.97)$: Test muy bueno y $[0.97, 1)$: Test excelente (Fawcett, 2004; Swets, 1996).

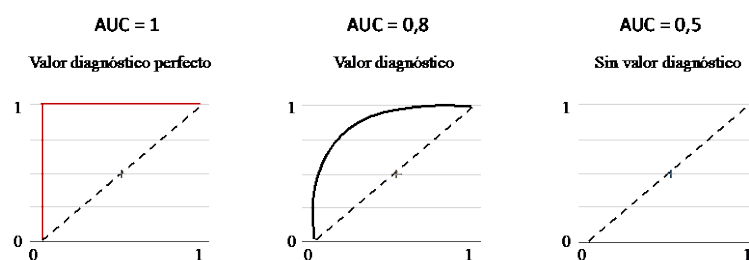


Figura 15. Guía para la interpretación de AUROC

2.7.4. Entrenamiento y Validación

El entrenamiento es el proceso mediante el cual una muestra de datos es utilizada para ajustar el modelo. La validación es el proceso mediante el cual una muestra de datos es utilizada para proporcionar una evaluación imparcial de un ajuste del modelo en el conjunto de datos de entrenamiento mientras se ajustan los hiperparámetros del modelo. La evaluación se vuelve más sesgada a medida que la habilidad del conjunto de datos de validación se incorpora a la

configuración del modelo. Cuando se entrena el modelo, se divide el conjunto de datos en dos subconjuntos de datos: de entrenamiento y de validación. Se selecciona de manera aleatoria, estratificada y representativamente de los casos de entrenamiento/validación. Generalmente se selecciona uno de cada cuatro casos para formar la serie de entrenamiento, es decir, un 75% de los casos y una serie de validación con un 25% de los casos (Cabrera-Andrade, 2020; Santana, 2020).

3. Trabajo Experimental

3.1. Capítulo 1. Modelos PTML de Compuestos Anti-leishmania

Artículo 1

Ref. Barbolla I, Hernández-Suárez L, Quevedo-Tumaili V., Nocedo-Mena D, Arrasate S, Dea-Ayuela MA, González-Díaz H, Sotomayor N, Lete E. (2021). Palladium-mediated synthesis and biological evaluation of C-10b substituted Dihydropyrrolo[1,2-b] isoquinolines as antileishmanial agents. *Europesn Journal of Medicinal Chemistry*. 220:113458. doi: 10.1016/j.ejmech.2021.113458. Epub 2021 Apr 16. PMID: 33901901

3.1.1. Introducción

Leishmaniasis es una enfermedad parasitaria olvidada, endémica en unos 100 países, cuya morbilidad y mortalidad aumentan cada día. La enfermedad está causada por patógenos protozoarios del género *Leishmania* que son transmitidos por los flebótomos. La *Leishmania* spp existe en dos formas morfológicamente distintas: una forma móvil flagelada (promastigotes) y una forma intracelular no flagelada (amastigotas). Existen cuatro formas principales de la enfermedad: la leishmaniasis visceral (VL, también conocida como kala-azar); la leishmaniasis dérmica post-kala-azar (PKDL); la leishmaniasis cutánea (CL); y la leishmaniasis mucocutánea (MCL). Mientras que la CL es la forma más común de la enfermedad, la VL es la más grave y puede ser mortal si no se trata (WHO, 2014). Los pacientes inmunodeprimidos relacionados con la co-infección por el VIH o el trasplante de órganos sólidos son propensos a la infección por *Leishmania*, que también puede favorecer el desarrollo del cáncer (Schwing *et al.*, 2019; Akuffo *et al.*, 2018).

Los tratamientos actuales contra la *Leishmania* se ven obstaculizados por la toxicidad de los fármacos, el elevado costo, la necesidad de administración parenteral, el aumento de las tasas de fracaso del tratamiento y la aparición de especies o cepas de *Leishmania* multirresistentes (MDR). El tratamiento no sólo depende de la especie etiológica y del tipo de infección (CL, VL, PKDL o MCL), sino también de la ubicación geográfica donde se adquirió la enfermedad (Sundar & Chakravarty, 2013), (Jones, 2018). Además, hay pocas dianas moleculares bien validadas en *Leishmania*, pero se desconocen las dianas moleculares de las moléculas clínicas actuales. La VL se trata con diferentes terapias multidrogas, diferentes combinaciones de antimoniales pentavalentes, paromomicina, anfotericina B liposomal y

miltefosina. Su uso está además limitado por las toxicidades asociadas que ponen en peligro la vida, como las arritmias cardíacas, la prolongación del intervalo QT (QTc), los latidos prematuros ventriculares, la taquicardia o la fibrilación en los antimoniales pentavalentes o la nefrotoxicidad, la hipopotasemia y la miocarditis en el tratamiento con anfotericina B (Sundar & Singh, 2018). Por una parte, la toxicidad y el estrecho margen terapéutico es una de las principales limitaciones de los compuestos utilizados actualmente contra la leishmaniasis, por lo que es importante buscar nuevas alternativas terapéuticas que presenten una toxicidad reducida. Además, estos fármacos pueden tener importantes efectos secundarios, por ejemplo, la miltefosina puede provocar defectos de nacimiento si se toma en los tres meses previos al embarazo (Dorlo, 2012). Por otra parte, el tratamiento de la CL debe decidirse en función de las lesiones clínicas, la especie etiológica, la diferente respuesta a los fármacos de las especies y cepas, y la posibilidad de que se convierta en una leishmaniasis mucosa (Sundar & Chakravarty, 2013; Bilbao-Ramos *et al.*, 2017). Pero los tratamientos probados de la CL son escasos, siendo los antimoniales pentavalentes, la paromomicina, la pentamidina y el derivado tri-azólico fluconazol los fármacos más eficaces (Minodier & Parola, 2007) (Figura 16).

Entre los últimos avances en este campo destacan las nuevas series químicas de fármacos antileishmánicos basados en andamios heterocíclicos de nitrógeno, como la pirazolopirimidina GSK3186899/DDD853651 (Wyllie *et al.*, 2018; Thomas *et al.*, 2019) y los derivados del benzabenzoxazol GNF6702 y LXE408 (Khare *et al.*, 2016; Nagle *et al.*, 2020) (Figura 16). Además, la sitamaquina, una 8-aminoquinolina, es un fármaco candidato para el tratamiento de la VL por vía oral, aunque los ensayos clínicos de fase II señalan algunos efectos adversos, como la metahemoglobinemia y la nefrotoxicidad, que tienen que ser considerados para una decisión de desarrollo posterior (Loiseau *et al.*, 2011) (Figura 16). Más recientemente, las 4-aminoestirilquinolinas, los quinolona-metronidazoles y los derivados basados en la ferrocenilquinolina también han mostrado perfiles antileishmania prometedores (Staderini *et al.*, 2019; Upadhyay *et al.*, 2019; Bhat *et al.*, 2020; Mukherjee *et al.*, 2020). En la actualidad, la identificación de nuevos fármacos contra la leishmania eficaces y seguros es crucial para avanzar en la obtención de nuevos compuestos líderes y en el control de la enfermedad (Hendrickx *et al.*, 2019).

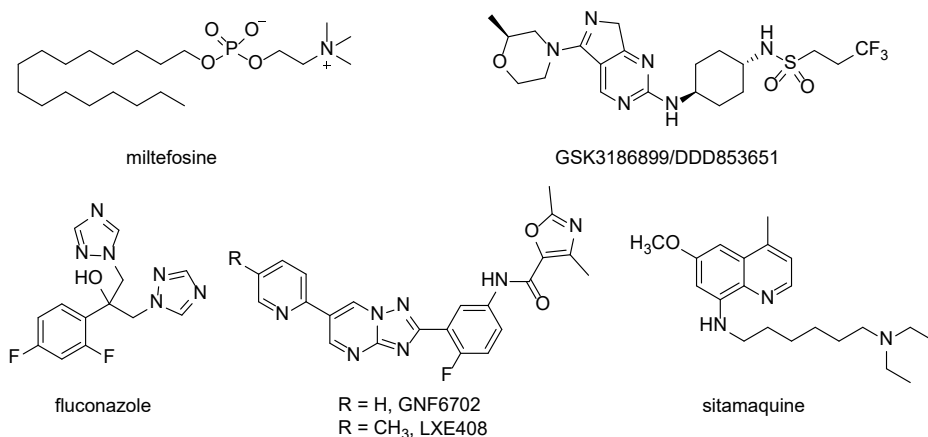


Figura 16. Fármacos antileishmanicos con motivos heterocíclicos y miltefosina

Nuestro grupo tiene experiencia en el desarrollo de metodologías sintéticas para la preparación de diferentes heterociclos benzo(hetero)fusionados de seis miembros (Martínez *et al.*, 2009; Lage *et al.*, 2009; Coya *et al.*, 2014; Ortíz de Elguea *et al.*, 2015; Carral-Menoyo *et al.*, 2020) en particular la quinolina, la isoquinolina y sus homólogos dihidro, cuyos núcleos estructurales se encuentran entre muchos productos naturales y farmacéuticos biológicamente activos (Evidente & Kornienko, 2009; Pereira *et al.*, 2015; Cortés *et al.*, 2015; He *et al.*, 2015; Nair *et al.*, 2016; Zhan *et al.*, 2017). En particular, los alcaloides del tipo Lycorane de las Amaryllidaceae pueden presentar actividad biológica contra enfermedades tropicales causadas por parásitos protozoarios (Osorio *et al.*, 2008; Prabhu *et al.*, 2015; Nair & Staden, 2019). Por otro lado, Bringmann y Moll han informado de que la presencia de un anillo aromático unido al átomo de nitrógeno del núcleo de la isoquinolina es crucial para la actividad leishmanicida de estos heterociclos (Ponte-Sucre *et al.*, 2009; Bringmann *et al.*, 2010). Por lo tanto, se considera que el cambio del grupo arilo por un anillo heteroaromático, el pirrol, pero fusionado al núcleo de isoquinolina por el lado b, daría lugar al núcleo de pirrolo[1,2-b]isoquinolina, combinando así características estructurales de ambos tipos de heterociclos. Recientemente se ha informado de la síntesis de pirrolo [1,2-b] isoquinolines 2 sustituidas (Barbolla *et al.*, 2019) Heck/Suzuki (Esquema 1) (Biemolt & Ruijter, 2018; Ping *et al.*, 2019; Barbolla *et al.*, 2019; Coya *et al.*, 2015; Blázquez-Barbadillo *et al.*, 2016). Por lo tanto, se decide explorar la actividad biológica de estos compuestos.

En este contexto, los modelos quimiinformáticos pueden ser útiles para llevar a cabo un cribado previo “in silico” computacional de alto rendimiento de grandes bibliotecas de compuestos. Estos estudios permiten priorizar algunas familias de compuestos (potenciales

compuestos líderes) en los ensayos farmacológicos con el fin de reducir el tiempo y los costes (recursos materiales y humanos) del proceso de descubrimiento de fármacos. De este modo, se evitan los ensayos de nuevos compuestos mediante pruebas de ensayo y error. Sin embargo, hasta donde se tiene conocimiento, no hay informes de modelos computacionales para compuestos antileishmaniales que incluyan datos para múltiples especies y tipos de ensayos.

Los principales escollos de los modelos clásicos de Cheminformatics son la imposibilidad de predecir de forma simultánea múltiples parámetros de actividad biológica de los fármacos contra diferentes proteínas diana, líneas celulares, organismos de ensayo, *etc.* No son capaces de realizar una clasificación multietiqueta y multisalida de nuevos compuestos.

Como se comenta en la parte introductoria, se ha introducido el algoritmo IFPTML para resolver problemas similares (múltiples estructuras frente a múltiples especies y condiciones de ensayo) en el proceso de descubrimiento de fármacos (Gonzalez-Díaz *et al.*, 2013; Ortega-Tenezaca *et al.*, 2020). Los modelos PTML parten de un valor conocido (referencia) de la actividad biológica esperada para un grupo de compuestos o propiedad y añaden el efecto de las perturbaciones (desviaciones) de la estructura y, o, las condiciones de ensayo del nuevo caso (compuesto de consulta) con respecto a la referencia. El enfoque PTML utiliza operadores PT (PTO) como desviaciones, medias móviles, *etc.* para cuantificar el efecto de estas desviaciones o perturbaciones sobre la actividad biológica final.

Los modelos PTML más sencillos son modelos aditivos lineales, pero se pueden construir modelos PTML más complejos y generales. Los modelos PTML se han utilizado con éxito para predecir diferentes parámetros de actividad biológica y/o toxicidad (K_i , IC_{50} , LD_{50} , K_m , % de inhibición, *etc.*) para la interacción de diferentes compuestos con diferentes dianas biológicas (proteínas, tejidos, líneas celulares, organismos patógenos, *etc.*) (Ferreira *et al.*, 2018; Díaz-Alarcía *et al.*, 2019; Nocedo-Mena *et al.*, 2019; Santana *et al.*, 2020). Por lo tanto, los modelos PTML son útiles para seleccionar compuestos de la serie que se enviarán a los ensayos farmacológicos.

Se describe aquí la evaluación biológica “*in vitro*” contra dos especies de *Leishmania*, *L. amazonensis* y *L. donovani*, causantes de CL y VL, respectivamente de una serie de andamios de pirrolo[1,2-b] isoquinoline. Además, se ha desarrollado el primer modelo PTML de propósito general para predecir la actividad antileishmanial de estas series de pirroloisoquinolines en diferentes ensayos biológicos contra diferentes especies de *Leishmania*.

3.1.2. Ensayos antileishmanios

Los nuevos derivados de 5,10-dihidropirrolo[1,2-b]isoquinoline sintetizados **3**, junto con los derivados (hetero)arilmetilos sustituidos C-10 obtenidos previamente **2** (Barbolla *et al.*, 2019), fueron ensayados frente a dos especies de *Leishmania*, *L. amazonensis* y *L. donovani*, causantes de CL y VL, respectivamente (Figura 17). Se llevaron a cabo ensayos de susceptibilidad “*in vitro*” de los promastigotes y ensayos de susceptibilidad “*in vitro*” de los amastigotes intracelulares, así como ensayos de citotoxicidad en la línea celular de macrófagos J774, una línea de macrófagos utilizada para probar la citotoxicidad de los fármacos “*in vitro*” antes de los ensayos en animales (véase la sección experimental). La miltefosina fue el fármaco de referencia, ya que puede utilizarse para el tratamiento de diferentes formas de la enfermedad. El cribado inicial en los ensayos de promastigotes “*in vitro*” reveló que algunas 5,10-dihidropirrolo[1,2-b]isoquinolinas se comparan favorablemente con la miltefosina en términos de actividad y selectividad contra *L. amazonensis*. En general, la mejor actividad se encontró para los derivados **2aa**, **2ab**, **2ad**, **2ae**, **2ag**, **2ah**, **2bb** y **2db** sustituidos por 10-arilmetilo. En particular, los compuestos más activos y selectivos **2ad** ($IC_{50} = 3,30 \pm 2,80 \mu M$, SI > 77,01) y **2bb** ($IC_{50} = 3,93 \pm 0,23 \mu M$, SI > 58,77) fueron aproximadamente 10 veces más potentes y selectivos que el fármaco de referencia (miltefosina), seguido del compuesto **2db** ($IC_{50} = 8,00 \pm 0,28 \mu M$, SI > 34,4). Por el contrario, la presencia de un grupo cianometilo en C-10 dio lugar a compuestos con una actividad antileishmanial más débil contra *L. amazonensis* (Tabla 1, entradas 12-19). Sólo los compuestos **3h** y **3i**, que tienen un grupo O-bencil en la posición 7 del núcleo de la pirroloisoquinolina, tienen una actividad contra *L. amazonensis* comparable a la de la miltefosina (Tabla 1, entradas 18-19 vs. entrada 20). Por lo tanto, la presencia de un grupo bencílico en el núcleo de pirroloisoquinolina parece ser crucial para la actividad antileishmaniana.

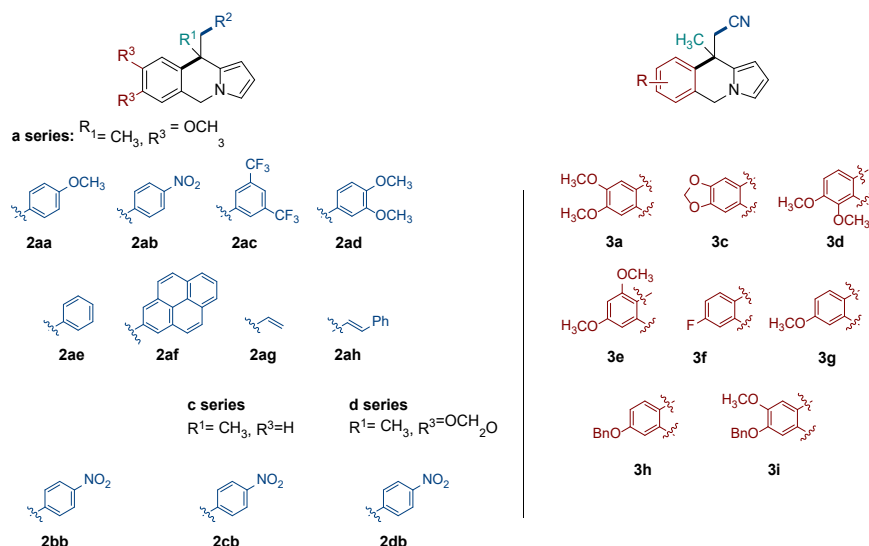


Figura 17. C-10 sustituidas 5,10-dihydropyrrolo[1,2-*b*]isoquinolines 2 y 3 y examinadas contra *L. amazonensis* y *L. donovani*.

El patrón de sustitución aromática del núcleo de pirroloisoquinoline de esta serie **2** también juega un papel importante, teniendo los sustituyentes donadores de electrones un impacto positivo en la actividad anti-leishmania contra *L. amazonensis*. Así, el compuesto **2cb**, con un anillo aromático no sustituido, es 4-47 veces menos potente que otros compuestos de la serie **2** (Tabla 1, entrada 10 vs. entradas 1 y 11). Además, como se muestra en la Tabla 1, la sustitución del grupo metilo en la posición 10 del sistema de anillos de pirroloisoquinoline por un grupo trifluorometilo dio lugar a compuestos más activos, siendo el **2bb** más de 4 veces más activo contra *L. amazonensis* que el **2ab** (Tabla 1, entrada 2 vs. entrada 9). En cuanto al patrón de sustitución de la fracción bencílica, la presencia de dos grupos trifluorometilos fue perjudicial para la actividad y la selectividad, siendo **2ac** el derivado menos activo y más tóxico de ambas series (Tabla 1, entrada 3). Sin embargo, hay que señalar que la introducción de este grupo CF₃ en C-10 tiene un efecto menor (Tabla 1, entrada 2 vs. entrada 9).

Por el contrario, todas las pirroloisoquinolinas ensayadas son notablemente menos activas que la miltefosina para el tratamiento de *L. donovani*, siendo de nuevo el compuesto **2bb** uno de los más activos y selectivos de ambas series con $\text{IC}_{50} = 16,41 \pm 4,90 \mu\text{M}$ y $\text{SI} > 14,09$. Además, y muy notablemente, casi todas las pirroloisoquinolinas son mucho menos tóxicas que el fármaco de referencia con valores de concentración del compuesto que produce una reducción del 50% de la viabilidad celular (Concentración Citotóxica, CC_{50}) en el rango 195-416 μM en células J774. De hecho, el índice de selectividad ($\text{SI} = \text{CC}_{50}/\text{IC}_{50}$) es mayor para casi todas las pirroloisoquinolinas que para la miltefosina, cuyo SI es sólo 4,43.

También hay que señalar que, aunque se han notificado mejores valores de IC₅₀ para distintos tipos de moléculas contra diferentes especies de *Leishmania* (Ortalli *et al.*, 2020; Thomas *et al.*, 2020), suelen tener valores de CC₅₀ más bajos que la mayoría de nuestras pirroloisoquinolinas, que no son tóxicas a la dosis más alta ensayada (CC₅₀> 100 µg/mL). Por lo tanto, se trata de un resultado muy interesante, teniendo en cuenta que la quimioterapia disponible actualmente para la leishmaniasis se ve obstaculizada por la toxicidad y la resistencia a los fármacos.

Tabla 1. Efectos leishmanicidas y citotóxicos IC₅₀ de los derivados de pirroloisoquinolina (expresados en µM) en el ensayo “*in vitro*” de promastigotes.

Entrada	Compuesto	<i>L. amazonensis</i>		<i>L. donovani</i>		Macrófagos
		IC ₅₀ ± SD (µM) ^a	SI ^b	IC ₅₀ ± SD (µM) ^a	SI ^b	J774 CC ₅₀ ± SD (µM) ^c
1	2aa	26.41±2.20	>10.40	47.40±1.79	>5.80	≥ 275.13 ^e
2	2ab	18.00±0.79	>14.68	79.93±4.36	>3.31	≥ 264.24 ^e
3	2ac	39.62±2.13	1.18	111.34±10.8	0.42	46.80±4.00
4	2ad	3.30±2.80	>77.01	38.30±3.36	>6.63	≥ 254.14 ^e
5	2ae	17.09±0.60	>17.54	30.29±2.01	>10.09	≥ 299.91 ^e
6	2af	36.71±0.44	>5.95	86.01±1.60	>2.54	≥ 218.54 ^e
7	2ag	23.29±1.41	>15.15	124.43±14.70	>2.84	≥ 352.89 ^e
8	2ah	12.24±0.56	>23.00	33.66±2.06	>8.27	≥ 278.20 ^e
9	2bb	3.93±0.23	>58.77	16.41±4.90	>14.09	≥ 231.30 ^e
10	2cb	157.35±6.28	>2.00	159.40±26.20	>1.97	≥ 314.09 ^e
11	2db	8.00±0.28	>34.49	30.35±3.92	>9.09	≥ 275.95 ^e
12	3a	309.20±41.10	>1.14	226.21±11.60	>1.57	≥ 354.18 ^e
13	3c	77.35±1.13	>4.85	120.12±12.01	>3.13	≥ 375.51 ^e
14	3d	171.07±6.38	>2.07	nd ^d	nd ^d	≥ 354.18 ^e
15	3e	93.85±6.02	>3.77	nd ^d	nd ^d	≥ 354.20 ^e
16	3f	96.55±1.66	>4.31	192.19±7.66	>2.17	≥ 416.18 ^e
17	3g	123.25±10.70	>3.21	359.99±34.44	>1.10	≥ 396.32 ^e
18	3h	15.83±0.06	14.04	60.25±0.34	3.69	222.30±42.70
19	3i	24.82±0.28	7.86	34.37±5.30	5.68	195.30±27.30

20 miltefosine 30.70±0.98 4.43 0.15±0.02 906.00 135.90±10.30

^a IC₅₀: Concentración del compuesto que produjo una reducción del 50% de los parásitos; SD: Desviación Estándar. ^b SI: Índice de Selectividad, $SI = CC_{50}/IC_{50}$. ^c CC₅₀: Concentración del compuesto que produjo una reducción del 50% de la viabilidad celular en las células de cultivo tratadas con respecto a las no tratadas. ^d nd: no determinado. ^e Los valores de CC₅₀, expresados en μM , corresponden a 100 $\mu\text{g/mL}$, que fue la dosis más alta ensayada.

A continuación, los compuestos más activos de los ensayos con promastigotes se examinaron contra los amastigotes de *L. amazonensis* y *L. donovani* (Tabla 2). Todas las pirroloisoquinolinas probadas fueron menos activas que la miltefosina cuando se probaron contra los amastigotes de *L. donovani*. Sin embargo, estos compuestos tienen valores de IC₅₀ en el rango 33,59-77,12 μM , que son similares o incluso mejores que la miltefosina (IC₅₀ = 47,60 ± 7,04 μM) en el ensayo contra *L. amazonensis*. Todos ellos tienen valores de SI entre 3,58 y 8,93, que son de 2 a 4 veces superiores a la miltefosina (SI = 2,85). En este caso, la pirroloisoquinoline **2ae** mostró la mejor actividad con IC₅₀ = 33,59 ± 2,64 μM y mayor selectividad con SI > 8,93.

Tabla 2. Efectos leishmanicidas y citotóxicos IC₅₀ de los derivados de pirroloisoquinoline (expresados en μM) en el ensayo “*in vitro*” de promastigotes.

Entrada	Compuesto	<i>L. amazonensis</i>		<i>L. donovani</i>		Macrófagos
		IC ₅₀ ± SD (μM) ^a	SI ^b	IC ₅₀ ± SD (μM) ^a	SI ^b	J774 CC ₅₀ ± SD (μM) ^c
1	2aa	51.56±12.00	>5.34	45.00±7.51	>6.12	≥ 275.13 ^e
2	2ab	56.12±13.50	>4.70	29.62±6.87	>8.92	≥ 264.24 ^e
3	2ad	60.79±5.74	>4.18	46.30±0.33	>5.49	≥ 254.14 ^e
4	2ae	33.59±2.64	>8.93	55.03±3.96	>5.45	≥ 299.91 ^e
5	2ag	43.54±9.53	>8.10	255.88±42.10	>1.38	≥ 352.89 ^e
6	2ah	69.26±6.96	>4.02	16.74±0.14	>16.61	≥ 278.20 ^e
7	2bb	56.72±4.58	>4.08	22.17±4.16	>10.43	≥ 231.30 ^e
8	2db	77.12±19.10	>3.58	68.65±9.63	>4.02	≥ 275.95 ^e
9	3h	nd ^d	-	44.75±8.53	4.97	222.30±42.70

10	miltefosine	47.60±7.04	2.85	0.37±0.05	369.30	135.90±10.30
----	-------------	------------	------	-----------	--------	--------------

^a IC_{50} : Concentración del compuesto que produjo una reducción del 50% de los parásitos; SD: Desviación estándar.

^b SI: Índice de Selectividad, SI $1/4 CC_{50}/IC_{50}$.

^c CC_{50} : Concentración del compuesto que produjo una reducción del 50% de la viabilidad celular en las células de cultivo tratadas con respecto a las no tratadas.

^d nd: no determinado.

^e Los valores de CC_{50} , expresados en μM , corresponden a 100 mg/mL, que fueron las dosis más altas ensayadas.

Se puede observar que la serie de dihidroloisoquinoline sustituida por 10-arilmetilo parece ser más selectiva y presenta una actividad similar o incluso mejor que el fármaco de referencia para el tratamiento de los amastigotes de *L. amazonensis*. El resultado es prometedor porque en el presente modelo de ensayo experimental los amastigotes están dentro de las células de macrófagos humanos. Por lo tanto, parece que estas pirroloisoquinolines podrían ser capaces de atravesar las barreras del huésped y del parásito para ejercer su actividad sin dañar la membrana del huésped (membrana del macrófago). En primer lugar, tienen que atravesar la membrana de las células del huésped (macrófagos), después tienen que atravesar la membrana de la vacuola parasitófora ó Parasitophorous Vacuole Membrane (PVM). Por último, si la actividad no es sobre la PVM directamente, los compuestos deben alcanzar/atrasar la membrana del parásito para alcanzar su objetivo molecular en la membrana o dentro del parásito. El PVM impide la acidificación del medio por parte de los lisosomas de la célula huésped para destruir un parásito invasor. El PVM lo forma el parásito utilizando partes de la membrana de la célula huésped. El PVM rodea al parásito intracelular, creando una burbuja separada de membrana plasmática llena de citoplasma dentro de la célula huésped (Kemp *et al.*, 2013).

3.1.3. Modelo computacional

Como se ha descrito anteriormente, se han medido los valores experimentales de $IC_{50}(\mu M)$ para las series de pirroloisoquinolines **2** y **3** a partir de los ensayos contra dos especies de *Leishmania*, *L. amazonensis* y *L. donovani*, en dos estadios diferentes (S) de desarrollo del parásito: amastigotes (A) y promastigotes (P). Se ha podido observar una cierta tendencia en el comportamiento de estas pirroloisoquinolines frente a los promastigotes de las dos especies diferentes de *Leishmania* estudiadas. De hecho, se ha encontrado un coeficiente de regresión de $R = 0,67$ para la IC_{50} de los derivados de pirroloisoquinoline en *L. amazonensis* frente a *L.*

donovani. Sin embargo, aún no se tienen pistas sobre las posibles proteínas diana de estos compuestos. Además, hay muchas otras especies de *Leishmania* (incluidas las cepas MDR) que no han podido ser ensayadas y que pueden presentar diferentes susceptibilidades a los mismos fármacos.

Para poner los resultados obtenidos en contexto, se descargó y exploró un conjunto de datos ChEMBL de >145.000 ensayos preclínicos de compuestos putativos anti-leishmania, como se detalla en la Sección Experimental y en el Material Suplementario. Los experimentos incluyen hasta 10 condiciones diferentes de ensayo $c_j = [c_0, c_1, c_2, \dots, c_{10}]$, que se dividieron en dos particiones o subconjuntos de condiciones experimentales c_I y c_{II} . Las condiciones principales c_I son las que caracterizan el experimento biológico *per se* (parámetro medido, proteína objetivo, estadio del parásito, organismo del ensayo, *etc.*). El conjunto de datos incluye valores para $n(c_0) > 50$ parámetros biológicos diferentes medidos para compuestos frente a al menos una de $n(c_1) = 32$ proteínas diana, $n(c_2) = 29$ líneas celulares, $n(c_3) = 40$ organismos de ensayo (no todos son parásitos), y $n(c_4) = 37$ especies o cepas de parásitos de *Leishmania*, *etc.* En lugar de los resultados de todos los experimentos posibles, el conjunto de datos incluye $n(c_I) > 240$ combinaciones de estas condiciones principales c_I (experimentos diferentes) para numerosos compuestos. Además, el conjunto de datos incluye $n(c_{II}) = 80$ combinaciones de condiciones secundarias c_{II} relacionadas con la naturaleza biológica y, o, la exactitud de los datos (tipo de diana, mapeo de la diana, exactitud del valor, *etc.*). Este estudio reveló que más de 70 ensayos con 20 especies/cepas de *Leishmania* son los más utilizados. Desgraciadamente, incluso limitando el análisis a estos ensayos puede ser costoso en términos de recursos y tiempo. Probar una serie corta de sólo 10 compuestos requeriría $n_{\text{assay}} = 10 \cdot 70 = 700$ ensayos experimentales. Por lo tanto, se ha decidido realizar un estudio computacional preliminar de la susceptibilidad de otras especies a esta serie de compuestos. Como se ha dicho anteriormente, no hay informes sobre un modelo computacional para realizar dicho estudio. Por lo tanto, primero se ha desarrollado un nuevo modelo PTML para llevar a cabo este tipo de predicciones. La ecuación del mejor modelo PTML encontrado es la siguiente:

$$\begin{aligned}
 f(v_{ij})_{\text{calc}} = & 59.918599 \cdot f(v_{ij})_{\text{ref}} + 1.631537 \cdot \Delta D_1(c_I) & (1) \\
 & + 0.041494 \cdot \Delta D_2(c_I) - 2.675709 \cdot \Delta D_3(c_I) \\
 & - 1.562187 \cdot \Delta D_1(c_{II}) - 0.041886 \cdot \Delta D_2(c_{II}) \\
 & + 2.649511 \cdot \Delta D_3(c_{II}) - 25.182671 \\
 n = & 109389 \quad \chi^2 = 135169.7 \quad p < 0.05
 \end{aligned}$$

La idea de este modelo PTML es comenzar utilizando como entrada una función de referencia $f(v_{ij})_{ref}$, que se obtiene a partir de los ensayos experimentales de un conjunto de compuestos de referencia realizados bajo condiciones específicas c_j . Esta función cuantifica la probabilidad previa esperada de dar un resultado positivo en los ensayos específicos para un compuesto seleccionado al azar. A continuación, se añadieron a la $f(v_{ij})_{ref}$ los valores de los operadores PT (PTO) con la forma general $\Delta D_k(c_j)$. Estos PTOs cuantifican la desviación (Δ) de los descriptores moleculares D_k (estructura) de nuestro compuesto de consulta con respecto al grupo de compuestos de referencia (véase la Sección Experimental). Los descriptores D_k son $D_1 = \Delta \text{LOGP}$, el coeficiente de partición n-octanol/agua, $D_2 = \text{TPSA}$, el Área de Superficie Polar Topológica, y $D_3 = \text{NRV}$, el Número de Violaciones a la Regla de V (regla de Lipinski o de Pfizer) fueron utilizados para identificar cada compuesto en la ecuación. Los descriptores ΔLOGP y TPSA cuantifican la estructura molecular del fármaco mediante una suma ponderada de diferentes fragmentos moleculares en las moléculas. La regla NRV cuantifica la semejanza/similitud de un compuesto con respecto a los fármacos conocidos basándose en el peso molecular, los enlaces de hidrógeno, *etc.* (Todeschini & Consonni, 2000). Los valores de D_k se extrajeron del conjunto de datos ChEMBL y/o se calcularon con el software DRAGON para los nuevos compuestos. Finalmente, la salida del modelo $f(v_{ij})_{calc}$ es una función de puntuación utilizada para calcular la probabilidad de actividad $p(f(v_{ij})_{pred} = 1)$ de los diferentes compuestos. Para entrenar este modelo, se ha seleccionado al azar una gran serie de entrenamiento de $n = 109.389$ ensayos preclínicos descargados de la base de datos ChEMBL. Los valores de Especificidad (Sp) y Sensibilidad (Sn) están en el rango $\approx 90-98\%$ para las series de entrenamiento (ver Material Suplementario), que son valores excelentes para este tipo de modelos de clasificación ML. Además, el nivel $p < 0,05$ para la prueba de Chi-cuadrado con $\chi^2 = 135,169,7$ apunta a una discriminación estadísticamente significativa entre compuestos activos ($f(v_{ij})_{obs} = 1$) y no activos ($f(v_{ij})_{obs} = 0$) en todos estos ensayos. Este modelo puede utilizarse para predecir nuevos compuestos no incluidos en las series de entrenamiento. En primer lugar, los valores de $f(v_{ij})_{ref}$ y PTO (que contienen D_k de fármaco y $\langle D_k(c_j) \rangle$ de ensayo) se sustituyeron en la ecuación para calcular la función de salida $f(v_{ij})_{calc}$. A continuación, los valores de $f(v_{ij})_{calc}$ se transformaron en probabilidades posteriores de éxito $p(f(v_{ij})_{pred} = 1)$ para cada compuesto en diferentes ensayos utilizando una función sigmoidea. Una vez calculados los valores de las probabilidades, se pudieron clasificar los compuestos. Así, aquellos resultados en el rango de probabilidad $p(f(v_{ij})_{pred} = 1) > 0,5$ se consideran interesantes para el ensayo $f(v_{ij})_{pred} = 1$. A continuación, se probó el presente modelo PTML con una serie de validación muy amplia, obteniendo valores de Sp y Sn también en el rango $\approx 90-98\%$ para la serie de

validación externa. En conclusión, este sencillo pero potente modelo PTML predice muy bien (exactitud general = 97,8%) un gran conjunto de datos (entrenamiento + validación) de ensayos preclínicos de actividad antileishmanial ($n_{\text{assay}} > 145.000$) que incluyen 96.800 compuestos únicos.

3.1.4. Estudio predictivo

Como se ha mencionado anteriormente, algunos compuestos de nuestra serie muestran valores de IC_{50} interesantes y pueden considerarse activos ($f(v_{ij}) = 1$) utilizando el corte de $IC_{50} = 10 \mu\text{M}$ en el rango de los estudios experimentales más típicos. A continuación, se ha decidido utilizar este nuevo modelo para realizar una predicción computacional de los resultados de las 19 pirroloisquinolinas de nuestra serie frente a diferentes especies de *Leishmania* (>20) en más de 160 ensayos preclínicos diferentes. Este modelo PTML es capaz de predecir la probabilidad posterior $p(f(v_{ij})_{\text{pred}} = 1)$ de obtener el nivel deseado para más de 50 propiedades biológicas diferentes (IC_{50} , K_i , K_m , etc.). En este estudio preliminar, se seleccionó el IC_{50} como propiedad única, ya que fue la primera propiedad que se midió experimentalmente en las primeras etapas del cribado, utilizando un valor de corte muy bajo de $IC_{50} = 0,01 \mu\text{M}$, en un esfuerzo por reducir el número de falsos positivos desde los primeros pasos del cribado. En consecuencia, el modelo se utilizó para calcular los valores de probabilidad $p(IC_{50}(\mu\text{M}) < 0,01)_{\text{calc}}$ para los que un compuesto mostraría una $IC_{50}(\mu\text{M}) < 0,01$ en diferentes ensayos. El número total de cálculos para los ensayos “*in vitro*” que no especifican la proteína objetivo fue $n_{\text{calc1}} = n_{\text{cmpd}} \cdot n_{\text{assay}} = 19 \cdot 162 = 3078$. Los resultados completos de este estudio predictivo se recopilan en la Tabla S9 del Material Suplementario que pueden encontrarse en línea en <https://doi.org/10.1016/j.ejmech.2021.113458>.

También se predijeron los valores de $p(IC_{50}(\mu\text{M}) < 0,01)_{\text{calc}}$ para ensayos con dianas proteicas conocidas, siendo el número de cálculos $n_{\text{calc2}} = n_{\text{cmpd}} \cdot n_{\text{prot}} = 19 \cdot 32 = 608$ para los 19 compuestos frente a 32 proteínas diana de diferentes especies. Esto hace un total de $n_{\text{calc}} = n_{\text{calc1}} + n_{\text{calc2}} = 3686$ cálculos para nuestros 19 compuestos en diferentes ensayos. En aras de la simplicidad, sólo se han cambiado las condiciones c_I (ensayo *per se*) utilizando un subconjunto fijo de condiciones c_{II} para los 3686 cálculos. En consecuencia, el número total de valores predichos de $p(IC_{50}(\mu\text{M}) < 0,01)_{\text{calc}}$ fue de 3686. Para resumir los resultados y extraer conclusiones, se ha calculado el valor medio de la probabilidad $p(IC_{50}(\mu\text{M}) < 0,01)_{\text{avg}}$, que es la probabilidad media de que se prediga que el compuesto tiene una $IC_{50}(\mu\text{M}) < 0,01$ en múltiples ensayos frente a la misma especie. El modelo PTML predice un comportamiento

coherente para esta serie de compuestos como serie homóloga. Esto significa que se predice que las especies resistentes/susceptibles a un compuesto son resistentes a la acción de casi todos los compuestos de la serie completa. El modelo no detecta diferencias globales significativas en el comportamiento de ambas subseries (subserie **2aa-2db** vs. subserie **3a-3i**) de compuestos, lo que podría ser coherente con el hecho de que ambas subseries de compuestos tienen el núcleo de pirroloisoquinoline. El valor medio de la probabilidad $p(\text{IC}_{50}(\mu\text{M}) < 0,01)_{\text{avg}}$ para casi todos los compuestos de la serie están en el rango 0,1-0,4, $p(\text{IC}_{50}(\mu\text{M}) < 0,01) < 0,5$, para muchas especies de *Leishmania*, lo que está de acuerdo con los hallazgos experimentales.

La Tabla 3 muestra los resultados seleccionados del estudio predictivo frente a diferentes especies para los dos compuestos principales de cada subserie (**2ad**, **2bb**, **3h** y **3i**) probados experimentalmente en este trabajo (véase el Material Suplementario para los detalles completos). Curiosamente, el modelo PTML predice valores de $p(\text{IC}_{50}(\mu\text{M}) < 0,01)_{\text{avg}} > 0,8$ para algunas especies/cepas no probadas previamente, a pesar del exigente valor umbral computacional utilizado en el estudio computacional. Véanse, por ejemplo, los valores de $p(\text{IC}_{50}(\mu\text{M}) < 0,01)_{\text{avg}}$ de los cuatro compuestos más exitosos para *L. braziliensis cepa M2904* (*L. brm.*), y *L. major cepa Friendlin* (*L. maf.*), y *L. mexicana* (*L. mex.*), que se muestran en la Tabla 3. En cuanto a la posible proteína diana, el modelo predice valores de $p(\text{IC}_{50}(\mu\text{M}) < 0,01)_{\text{avg}} > 0,7$ para ambas subseries de pirroloisoquinolines frente a algunas proteínas. El estudio señala la Tripanotión reductasa de *L. donovani* (P39050) (Taylor *et al.*, 1994; Chan *et al.*, 1998), la _L-isoaspartato(_D-aspartato) O-metil-transferasa (P22061) de *L. donovani* (Ingrosso *et al.*, 1989) y la Ornitina decarboxilasa de *L. donovani* (P27116) (Hanson *et al.*, 1992) como dianas plausibles a ensayar, aunque también se recogen más candidatos en la Tabla 3. Es interesante que algunas de estas proteínas son enzimas con derivados de aminoácidos (Tripanotión, _D-aspartato y Ornitina) como sustratos. Por lo tanto, este estudio computacional abre la puerta a un ensayo posterior de estos compuestos o sus derivados frente a otras especies de *Leishmania* y a ensayos específicos frente a probables dianas proteicas.

Tabla 3. Resultados del modelo PTML-LDA

Conjunto de datos	Modelo PTML		Clases observadas	Clases predichas	
	Estadístico	(%)		$f(v_{ij})_{\text{pred}} = 0$	$f(v_{ij})_{\text{pred}} = 1$
Entrenamiento	Sp(%)	98.3	$f(v_{ij})_{\text{obs}} = 0$	100919	1759
Serie	Sn(%)	90.6	$f(v_{ij})_{\text{obs}} = 1$	631	6080
Validación	Sp(%)	98.3	$f(v_{ij})_{\text{obs}} = 0$	33619	593

Serie	Sn(%)	90.8	$f(v_{ij})_{obs} = 1$	207	2043
-------	-------	------	-----------------------	-----	------

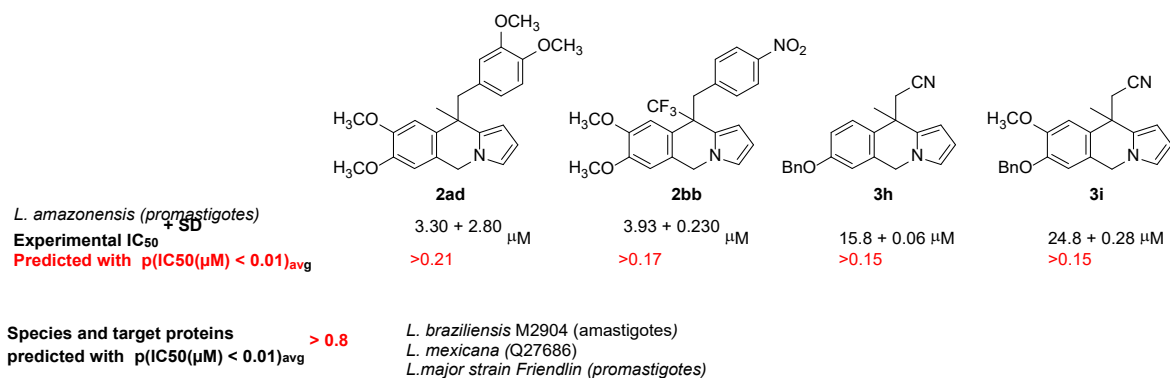


Tabla 4. Predicción PTML del valor medio de probabilidad $p(\text{IC}_{50}(\mu\text{M}) < 0.01)_{avg}$ para las pirroloisoquinolinas **2ad**, **2bb**, **3h**, and **3i** contra > 20 especies diferentes de *Leishmania*

Leish.	S ^b	Target	Compuesto				Leish.	S ^b	Target	Compuesto			
Especies ^a		Proteína	2ad	2bb	3h	3i	Especies ^a		Proteína	2ad	2bb	3h	3i
L. act.	A	-	0.02	0.01	0.01	0.01	L. maj.	-	P37268	0.41	0.34	0.30	0.32
L. act.	P	-	0.02	0.01	0.01	0.01	L. maj.	-	Q01782	0.58	0.51	0.47	0.49
L. ama.	A	-	0.10	0.08	0.06	0.07	L. maj.	-	Q0GKD7	0.43	0.36	0.32	0.34
L. ama.	P	-	0.21	0.17	0.15	0.15	L. maj.	-	Q4Q5S8	0.04	0.03	0.03	0.03
L. ama.	-	O96394	0.64	0.58	0.53	0.55	L. maj.	-	Q4Q5W4	0.78	0.73	0.69	0.71
L. ari.	-	-	0.21	0.17	0.15	0.16	L. maj.	-	Q4QBL1	0.52	0.46	0.41	0.43
L. bra.	A	-	0.12	0.09	0.08	0.09	L. maj.	-	Q4QE15	0.02	0.02	0.01	0.02
L. bra.	P	-	0.20	0.16	0.14	0.15	L. maj.	-	Q6S996	0.17	0.13	0.11	0.12
L. brm.	A	-	0.93	0.91	0.89	0.90	L. maj.	-	Q9LM02	0.04	0.03	0.02	0.03
L. brm.	P	-	0.11	0.09	0.07	0.08	L. maj.	P	-	0.06	0.05	0.04	0.05
L. cha.	A	-	0.18	0.15	0.13	0.14	L. maj.	P	Q01782	0.41	0.35	0.31	0.33
L. cha.	P	-	0.25	0.21	0.18	0.19	L. maf.	P	-	0.94	0.92	0.90	0.91
L. don.	A	-	0.30	0.27	0.25	0.25	L. mex.	A	-	0.30	0.26	0.24	0.25
L. don.	-	P39050	0.83	0.79	0.76	0.77	L. mex.	-	P04406	0.73	0.67	0.63	0.65
L. don.	-	Q95WR6	0.16	0.13	0.11	0.12	L. mex.	-	P36400	0.35	0.29	0.26	0.27
L. don.	-	Q95Z89	0.29	0.24	0.21	0.22	L. mex.	-	Q01558	0.60	0.54	0.50	0.51

L. don.	-	Q9NJG8	0.07	0.05	0.04	0.05	L. mex.	-	Q27686	0.96	0.95	0.95	0.95
L. don.	P	-	0.25	0.23	0.22	0.23	L. mex.	-	Q4U254	0.41	0.35	0.31	0.32
L. don.	P	P22061	1.00	1.00	1.00	1.00	L. mex.	-	Q9U5N6	0.34	0.28	0.25	0.26
L. don.	P	P27116	1.00	0.99	0.99	0.99	L. mex.	P	-	0.27	0.22	0.20	0.21
L. dod.	-	-	0.21	0.17	0.15	0.16	L. mex.	P	P11166	0.09	0.07	0.06	0.06
L. enr.	P	-	0.13	0.10	0.09	0.09	L. mem.	-	Q27686	0.29	0.24	0.21	0.22
L. gar.	-	-	0.21	0.17	0.15	0.16	L. mem.	P	-	0.02	0.01	0.01	0.01
L. guy.	A	-	0.08	0.06	0.05	0.06	L. mev.	-	-	0.21	0.17	0.15	0.16
L. guy.	P	-	0.28	0.23	0.20	0.21	L. pan.	A	-	0.10	0.08	0.07	0.07
L. inf.	A	-	0.32	0.28	0.26	0.27	L. pan.	P	-	0.42	0.38	0.35	0.36
L. inf.	-	Q8I6E4	0.28	0.23	0.20	0.21	L. per.	P	-	0.23	0.19	0.16	0.17
L. inf.	P	-	0.20	0.17	0.15	0.15	L. pif.	A	-	0.11	0.09	0.07	0.08
L. maj.	A	-	0.23	0.20	0.18	0.18	L. pif.	P	-	0.21	0.17	0.14	0.15
L. maj.	-	O15826	0.76	0.70	0.67	0.68	L. pro.	P	-	0.11	0.09	0.07	0.08
L. maj.	-	O96526	0.07	0.06	0.05	0.05	L. tar.	-	-	0.10	0.08	0.07	0.07
L. maj.	-	P00374	0.21	0.17	0.15	0.16	L. tro.	P	-	0.18	0.14	0.12	0.13
L. maj.	-	P07382	0.22	0.18	0.16	0.16	L. tur.	-	-	0.08	0.06	0.05	0.05

^a *Leishmania species*: *L. aethiopica* = *L. aet.*, *L. amazonensis* = *L. ama.*, *L. aristidesi* = *L. ari.*, *L. braziliensis* = *L. bra.*, *L. braziliensis M2904* = *L. brm*, *L. chagasi* = *L. cha.*, *L. donovani* = *L. don.*, *L. donovani donovani* = *L. dod.*, *L. enriettii* = *L. enr.*, *L. garnhami* = *L. gar*, *L. guyanensis* = *L. guy*, *L. infantum* = *L. inf*, *L. major* = *L. maj*, *L. Mexicana* = *L. mex*, *L. mexicana mexicana* = *L. mem*, *L. mexicana venezuelensis* = *L. mev*, *L. panamensis* = *L. pan.*, *L. peruviana* = *L. per*, *L. pifanoi* = *L. pif*, *L. tarentolae* = *L. tar*, *L. tropica* = *L. tro*, *L. turanica* = *L. tur.*, *L. major strain Friendlin* = *L. maf*. ^b S = Stage: A = Amastigotes, P = Promastigotes.

3.1.5. Conclusiones

En conclusión, la evaluación de la actividad leishmanicida de las pirrolo[1,2-b]isoquinolina “*in vitro*” contra la leishmaniasis visceral (*L. donovani*) y cutánea (*L. amazonensis*) reveló que casi todos los compuestos mostraron una citotoxicidad muy baja, CC₅₀> 100 µg/mL en células J774 (dosis más alta probada). Esta es una característica importante, ya que la toxicidad de los fármacos es una de las principales limitaciones de la

quimioterapia actual para la leishmaniasis. En general, las pirroloisoquinolinas sustituidas por 10-arilmetilo mostraron la mejor actividad contra *L. amazonensis* en ensayos “*in vitro*” de promastigotes. Por un lado, **2ad** ($IC_{50} = 3,30 \mu M$, $SI > 77,01$) y **2bb** ($IC_{50} = 3,93 \mu M$, $SI > 58,77$) fueron aproximadamente 10 veces más potentes y selectivas que el fármaco de referencia (miltefosina). Por otro lado, **2ae** fue el compuesto más activo en los ensayos “*in vitro*” con amastigotes ($IC_{50} = 33,59 \mu M$, $SI > 8,93$). Además, se ha demostrado que los algoritmos de ML de la Teoría de la Perturbación (PTML) son útiles para modelar grandes (>145.000 casos) conjuntos de datos ChEMBL de ensayos preclínicos anti-leishmania. Es posible utilizar el modelo PTML desarrollado para reducir los costes de los ensayos prediciendo la probabilidad con la que un compuesto de consulta de esta u otra serie de compuestos alcanza un nivel deseado para múltiples parámetros (IC_{50} , K_i , *etc.*) frente a diferentes especies de *Leishmania* y proteínas diana, con altos valores de especificidad (>98%) y sensibilidad (>90%) tanto en las series de entrenamiento como de validación.

3.1.5.1. Detalles del ensayo biológico para las determinaciones de IC_{50} y CC_{50} contra *L. donovani* y *L. amazonensis*

Parásitos y procedimiento de cultivo. Se utilizaron las siguientes especies de *Leishmania*: *L. donovani* (MHOM/IN/80/DD8) que fue purificada (ATCC, EE.UU.) y *L. amazonensis* (MHOM/Br/79/Maria) que fue amablemente proporcionada por el Prof. Alfredo Toraño (Instituto de Salud Carlos III, Madrid). Los protastigotes se cultivaron en medio para insectos de Schneider complementado con 10% de suero fetal bovino inactivado por calor (FBS) y 1000 U/L de penicilina más 100 mg/L de estreptomycin en frascos de cultivo de 25 mL a 26 °C.

Ensayo de susceptibilidad de promastigotes “*in vitro*”. El ensayo se llevó a cabo como se ha descrito anteriormente [35]. En resumen, se cultivaron promastigotes en fase logarítmica ($2,5 \cdot 10^5$ parásitos/pocillo) en placas de plástico de 96 pocillos. Las soluciones madre de los compuestos probados se solubilizaron a 50 mg/mL en Dimetil Sulfoxido (DMSO). Se realizaron diluciones seriadas 1:2 de los compuestos de prueba en medio de cultivo fresco (100, 50, 25, 12,5, 6,25, 3,12, 1,56 y 0,78 $\mu g/mL$) hasta un volumen final de 200 μL . También se incluyó el control de crecimiento y la señal de ruido. Las concentraciones finales de disolvente (DMSO) nunca superaron el 0,5% (v/v), lo que garantizó que no se produjera ningún efecto sobre la proliferación o la morfología de los parásitos. Después de 48 h a 26 °C, se añadieron 20 μL de una solución de resazurina 2,5 mM a cada pocillo y las placas se volvieron

a colocar en la incubadora durante otras 3 h. Se determinaron las unidades de fluorescencia relativa (RFU) (longitud de onda de excitación-emisión de 535 nm - 590 nm) en un fluorómetro (Infinite 200 Tecan i- Control). La inhibición del crecimiento (%) se calculó mediante $100 - [(RFU \text{ pozos tratados} - RFU \text{ señal-ruido}) / (RFU \text{ no tratados} - RFU \text{ señal-ruido}) \times 100]$. Todos los ensayos se realizaron por triplicado. La miltefosina (Sigma-Merck, Madrid, España) se utilizó como fármaco de referencia y se evaluó en las mismas condiciones. La eficacia de cada compuesto se estimó calculando el IC_{50} (concentración del compuesto que produjo una reducción del 50% de los parásitos) mediante un análisis probit multinomial incorporado en el software SPSS v21.0. El índice de selectividad (SI) se calculó como la relación entre la citotoxicidad (CC_{50}) y la actividad contra los parásitos (IC_{50}).

Ensayo de susceptibilidad de amastigotes intracelulares “*in vitro*”. El ensayo se llevó a cabo como se ha descrito previamente [36]. Brevemente, se sembraron 5×10^4 macrófagos J774 y promastigotes estacionarios en una proporción de 1:5 en cada pozo de una placa de microtitulación, se suspendieron en 200 μ L de medio de cultivo y se incubaron durante 24 horas a 33 °C en una cámara con 5% de CO_2 . Después de esta primera incubación, se aumentó la temperatura hasta 37 °C durante otras 24 horas. A continuación, las células se lavaron varias veces en medio de cultivo mediante centrifugación a 1.500g durante 5 minutos para eliminar los promastigotes libres no internalizados. Por último, se sustituyó el sobrenadante por 200 μ L/pocillo de medio de cultivo que contenía diluciones seriadas de 2 veces de los compuestos de prueba, como en el ensayo de promastigotes. También se incluyó el control de crecimiento y la señal de ruido. Tras la incubación durante 48 horas a 37 °C, 5% de CO_2 , el medio de cultivo se sustituyó por 200 μ L/pocillo de la solución de lisis (RPMI-1640 con 0,048% de HEPES y 0,01% de SDS) y se incubó a temperatura ambiente durante 20 minutos. A continuación, se centrifugaron las placas a 3.500 g durante 5 minutos y se sustituyó la solución de lisis por 200 μ L/pocillo de medio para insectos de Schneider. Las placas de cultivo se incubaron a 26 °C durante otros 4 días para permitir la transformación de los amastigotes viables en promastigotes y su proliferación. Después, se añadieron 20 μ L/pocillo de resazurina 2,5 mM y se incubaron durante otras 3 horas. Finalmente, se midió la emisión de fluorescencia y se estimó el IC_{50} como se ha descrito anteriormente. Todos los ensayos se realizaron por triplicado. La miltefosina (Sigma-Merck, Madrid, España) se utilizó como fármaco de referencia y se evaluó en las mismas condiciones. El IC_{50} y el SI se calcularon como en la sección anterior.

Ensayo de citotoxicidad en macrófagos. El ensayo se llevó a cabo como se ha descrito previamente [37]. Se sembraron líneas celulares de macrófagos J774 (5×10^4 células/pozo) en microplacas de fondo plano de 96 pozos con 100 μL de medio RPMI 1640. Se dejó que las células se adhirieran durante 24 h a 37 °C, 5% de CO_2 , y se sustituyó el medio por diferentes concentraciones de los compuestos en 200 μL de medio y se expusieron durante otras 24 horas. También se incluyeron controles de crecimiento y de señalización.

Después, se añadió un volumen de 20 μL de la solución de resazurina de 2,5 mM, y las placas se devolvieron a la incubadora durante otras 3 horas para evaluar la viabilidad celular.

La reducción de la resazurina se determinó por fluorometría como en el ensayo de promastigotes. Cada concentración se ensayó tres veces. El efecto citotóxico de los compuestos se definió como la reducción del 50% de la viabilidad celular de las células del cultivo tratado con respecto al cultivo no tratado (CC_{50}) y se calculó mediante un análisis probit multinomial incorporado en el software SPSS v21.0.

3.1.5.2. Métodos computacionales

Se ha desarrollado un modelo PTML para la predicción de la probabilidad $p(f(v_{ij})_{\text{obs}} = 1)$ con la que los valores experimentales v_{ij} de la actividad biológica el fármaco probado en las condiciones del ensayo alcanza un nivel deseado $f(v_{ij})_{\text{obs}} = 1$ o no $f(v_{ij})_{\text{obs}} = 0$. El flujo de trabajo general para el PTML implica los siguientes pasos generales: (1) carga de datos, (2) preprocesamiento de datos, (3) entrenamiento y validación del modelo, y (4) aplicación del modelo para el estudio predictivo. La etapa (2) incluye las siguientes operaciones (2a) cálculo de la función objetivo, (2b) cálculo de la función de referencia, y (2c) cálculo de los Operadores de Perturbación-Teoría (PTOs). Los diferentes pasos se explican en la sección SI.

Descarga y preprocesamiento de los datos del modelo PTML. En primer lugar, se descargó de ChEMBL un conjunto de datos de 145.851 ensayos preclínicos de compuestos putativos anti-leishmania. Todos los valores experimentales de las propiedades biológicas v_{ij} se transformaron en una función objetivo categórica $f(v_{ij})_{\text{obs}}$, con el fin de minimizar los problemas de precisión. La estructura de los fármacos se codificó mediante un vector de descriptores moleculares D_{ki} para cada fármaco. Los elementos de este vector son los valores numéricos de los diferentes descriptores moleculares D_{ki} de cada fármaco. Como se indicó en la sección de resultados, los descriptores $D_1 = \Delta\text{LOGP}$, $D_2 = \text{PSA}$ y $D_3 = \text{NRV}$ se utilizaron para identificar el compuesto en la ecuación. Los valores de D_k se extrajeron del conjunto de datos ChEMBL

y, o, se calcularon con el software DRAGON en el caso de los nuevos compuestos. A continuación, se codificaron las condiciones del ensayo mediante un vector de condiciones c_j (ver detalles en la siguiente sección). Debido al elevado número de parámetros biológicos con diferentes escalas y niveles de error, se discretizaron mediante la función booleana $f(v_{ij})_{obs}$ obtenida para desarrollar un modelo de clasificación. A continuación, se calcularon los Operadores PT (PTOs) para comprimir en una sola variable la información tanto de la estructura del fármaco (vectores D_k) como del ensayo biológico realizado (vectores c_j). A continuación, los PTO se definieron mediante 10 condiciones de contorno diferentes, etiquetas o variables discretas de cada ensayo $c_1, c_2, c_3, \dots, c_{10}$. Los PTO tienen la siguiente forma $\Delta D_k(c_j) = D_k - \langle D_k(c_j) \rangle$. Estos operadores cuantifican la desviación (Δ) de los descriptores moleculares de un determinado compuesto D_k (estructura) con respecto a los valores esperados $\langle D_k(c_j) \rangle$ de estos descriptores para todos los compuestos ensayados previamente en las mismas condiciones c_j .

Modelo lineal PTML. La técnica de modelización PTML es útil para buscar modelos predictivos para conjuntos de datos complejos con múltiples características de Big Data [38,19a]. Se puede predecir la probabilidad posterior $p(f(v_{ij}) = 1)$ con la que el compuesto de consulta alcanza el nivel deseado de actividad $f(v_{ij})_{obs} = 1$ en el ensayo preclínico con un subconjunto específico de condiciones de ensayo seleccionadas de entre múltiples condiciones de ensayo $c_j = (c_0, c_1, c_2, \dots, c_{jmax})$. Como entrada, se utilizan OTPs similares a los operadores MA de Box-Jenkins [39,19a]. Los modelos lineales PTML tienen la siguiente forma:

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{k=1, j=0}^{k_{max}, j_{max}} a_{kj} \cdot \Delta D_k(c_j) \quad (2)$$

El resultado del modelo es un parámetro adimensional $f(v_{ij})_{calc}$, que puede transformarse fácilmente en las probabilidades deseadas $p(f(v_{ij}) = 1)$ mediante una función sigmoidea. La primera variable de entrada es la función de referencia $f(v_{ij})_{ref}$. Es la probabilidad esperada “*a priori*” con la que un compuesto de referencia aleatorio puede tener el nivel de actividad deseado para el parámetro biológico c_0 . Los PTO se han calculado para todos los descriptores D_k con respecto a dos tipos de condiciones experimentales c_I y c_{II} (subconjuntos) de variables categóricas. Se han creado para codificar por separado la información sobre las condiciones experimentales de los ensayos preclínicos (c_I) o sobre la naturaleza y la calidad de los datos (c_{II}). La primera partición c_I , o subconjunto de condiciones, incluye condiciones relacionadas con el sistema de ensayo *per se*, la proteína diana, la línea celular, el organismo, *etc.*, mientras que las condiciones/etiquetas del ensayo recogidas en la segunda partición c_{II} están relacionadas con la naturaleza y la calidad de los datos. Se han incluido numerosos ejemplos de valores

numéricos de c_0 , $f(v_{ij})_{ref}$, y $\langle D_k(c_1) \rangle$ y $\langle D_k(c_j) \rangle$ Para muchas condiciones de ensayo en las Tablas S6, S7 y S8 del Material Suplementario. Los datos complementarios de este artículo pueden encontrarse en línea en <https://doi.org/10.1016/j.ejmech.2021.113458>.

Predicción del modelo PTML de los nuevos resultados. Se llevó a cabo la predicción computacional de los resultados de 19 compuestos de la serie propia frente a >20 especies de *Leishmania* en más de 70 ensayos preclínicos diferentes utilizando el nuevo modelo PTML. Se sustituyeron en la ecuación los valores de la $f(v_{ij})_{ref}$ y los PTO ($\Delta D_k(c_j) = D_k - \langle D_k(c_j) \rangle$) para cada compuesto en diferentes ensayos. Los valores $\langle D_k(c_j) \rangle$ son los valores medios de los descriptores D_k para todos los compuestos ensayados previamente en las mismas condiciones c_j . En consecuencia, pueden utilizarse indirectamente para definir las condiciones específicas de ensayo que se pretenden predecir [17]. Tras sustituir los valores D_k para cada compuesto y los valores de $f(v_{ij})_{ref}$ y $\langle D_k(c_j) \rangle$ para cada ensayo, se obtuvieron los valores de la función de salida $f(v_{ij})_{calc}$. A continuación, estos valores se transformaron en probabilidades utilizando una función sigmoidea (ver detalles en el SI). Así, una vez calculados los valores posteriores de las probabilidades, se pueden clasificar los compuestos. Aquellos compuestos en el rango de probabilidad $p(f(v_{ij})_{pred} = 1) > 0,5$ se consideran interesantes para el ensayo $f(v_{ij})_{pred} = 1$.

3.2. Capítulo 2. Modelos ML en Redes GOINs de Compuestos AntiPlasmodium

Artículo 2

Ref. Quevedo-Tumaili, V. F., Ortega-Tenezaca, B., & Gonzalez-Diaz, H. (2018). Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *Journal Proteome Research*. 17(3). 1258–1268. doi: 10.1021/acs.jproteome.7b00861.

3.2.1. Introducción

La distribución espacial (Hurst *et al.*, 2002) de los genes en los cromosomas parece no ser aleatoria. Por ejemplo, sólo el 10% de los genes se transcriben a partir de promotores bidireccionales en humanos y muchos más se organizan en grandes agrupaciones.

Curiosamente, los genes vecinos son frecuentemente coexpresados, pero raramente relacionados funcionalmente. *Kustatscher et al.*, encontraron que la coexpresión de pares de genes bidireccionales y los genes cercanos en general es amortiguado a nivel proteico (*Kustatscher et al.*, 2017). Según estos autores, agrupar genes humanos a lo largo de la secuencia del genoma es biológicamente relevante para reducir el ruido de expresión. En cualquier caso, *Kustatscher et al.*, ha planteado preguntas intrigantes: ¿Por qué un gen no relacionado funcionalmente está agrupado en el genoma? ¿Cómo puede la célula tolerar su coexpresión? Preguntaron estos autores en un trabajo anterior (*Kustatscher et al.*, 2017). Sería interesante agregar aquí más preguntas, relacionadas con las inversiones de la orientación de los genes: ¿La orientación del gen (inversión) sigue un patrón aleatorio? ¿Es relevante para la actividad biológica de alguna manera? Además, el Análisis de Redes (NA, Network Analysis) se ha expandido a diferentes niveles de organización (Newman, 2003). De hecho, NA puede ayudarnos a estudiar la distribución a largo plazo de muchos patrones estructurales diferentes, características de conectividad y motivos regulatorios en diversos sistemas complejos. A nivel molecular y biológico, se puede usar NA para estudiar las interacciones del fármaco objetivo (Lin *et al.*, 2015), la estructura de estos *targets* (mapas de contacto de proteínas), las Interacciones entre las Proteínas (PIN, Protein Interactions) (Estrada, 2006), a las vías metabólicas, la corteza cerebral o los ecosistemas. A una escala mayor se puede usar NA que estudiará grandes redes sociales como Internet, redes financieras (Sharma *et al.*, 2017; Garcia-Bernardo *et al.*, 2017; Duenas, 2017) o Corte Suprema de EE. UU. (Fowler & Jeon, 2008; Fowler *et al.*, 2007). Inferir las relaciones de estructura y, o, propiedad en redes complejas a partir de datos observables son significativas en muchas áreas de la ciencia (Yang *et al.*, 2017).

Se pueden usar las herramientas ML en NA para realizar la inferencia de red (Sander *et al.*, 2017). Por ejemplo, Ghanat *et al.*, utilizaron modelos ML para reconstruir una red de cáncer (Ghanat *et al.*, 2017). Una forma de realizar esta tarea es mediante el cálculo de un parámetro de tipo denominado índices topológicos (TIs) (para redes completas) y, o, vértice Centralidades (para nodos). Estos índices son útiles para caracterizar patrones numéricos de conectividad entre nodos o actores en una red (representado como un gráfico matemático). En particular, el nodo o los valores de centralidad de vértice son un atributo estructural, estrictamente dependiente de las conexiones del nodo (red de nodo ubicación). El concepto de centralidad fue introducido por Bavelas en 1948 (Bavelas, 1948). En este sentido, para definir la centralidad se debe encontrar un parámetro que cuantifique la contribución de un nodo a la red, desde su ubicación en él. A continuación, estos índices se pueden usar como variables de entrada para algoritmos ML como el GDA, ANN, *etc.* De esta manera, se pueden encajar Modelos cuantitativos capaces de predecir las propiedades de las redes que dependen de su estructura. La combinación de NA y ML puede ser útil para estudiar la interrelación entre la estructura y las propiedades de muchos tipos de redes que incluyen, por ejemplo, Proteomas, Corteza cerebral, Epidemiológicas, Redes sociales, *etc.* (Herrera-Ibata *et al.*, 2015; Duardo-Sanchez *et al.*, 2014; Gonzalez-Diaz *et al.*, 2014; Vazquez-Prieto *et al.*, 2014; Gonzalez-Diaz *et al.*, 2013; Gonzalez-Diaz & Riera-Fernandez, 2012; Riera-Fernandez *et al.*, 2012; Duardo-Sanchez, 2011; Duardo-Sanchez, 2010; Gonzalez-Diaz *et al.*, 2008). En este trabajo, se define por primera vez un nuevo tipo de redes complejas para estudiar características en la microestructura de los cromosomas. Se acuña esta nueva clase como la Red de Inversiones de la Orientación de los Genes (GOINs). Esto implica analizar al mismo tiempo la distribución espacial de los genes en el cromosoma y la aparición de inversiones en la orientación de los genes a lo largo del cromosoma. Al hacerlo, se selecciona como caso de estudio el parásito *P. falciparum*. La malaria es un importante problema de salud pública, considerada una enfermedad tropical desatendida, en muchas partes del mundo, especialmente en África, incrementado por el desarrollo de fármacos y los problemas de resistencia (Arie, 2017). Este organismo tiene 14 cromosomas y 5365 proteínas en el proteoma; codificado por un número idéntico de genes que codifican proteínas (Lasonder *et al.*, 2002). Muchas investigaciones se han realizado sobre diferentes familias de proteínas en este parásito (Ranjan *et al.*, 2011; Nirmalan *et al.*, 2007; Sam-Yellowe, 2004). Sin embargo, muchas de estas proteínas son proteínas hipotéticas cuya función permanece desconocida hasta ahora. En primer lugar, se construye la matriz de adyacencia (A_k) y la respectiva GOIN para cada uno de los 14 cromosomas. A continuación, se calculan las centralidades del nodo (C_i) para todos los genes en cada red. Se realiza un estudio

comparativo de la distribución y las propiedades topológicas de estas redes con modelos de redes aleatorias. Por último, se utilizan las centralidades de nodo como entrada a los algoritmos ML para predecir ejemplos específicos de funciones biológicas sin depender de la información genética. La Figura 18 ilustra el flujo de trabajo general utilizado en este trabajo para desarrollar el nuevo modelo.

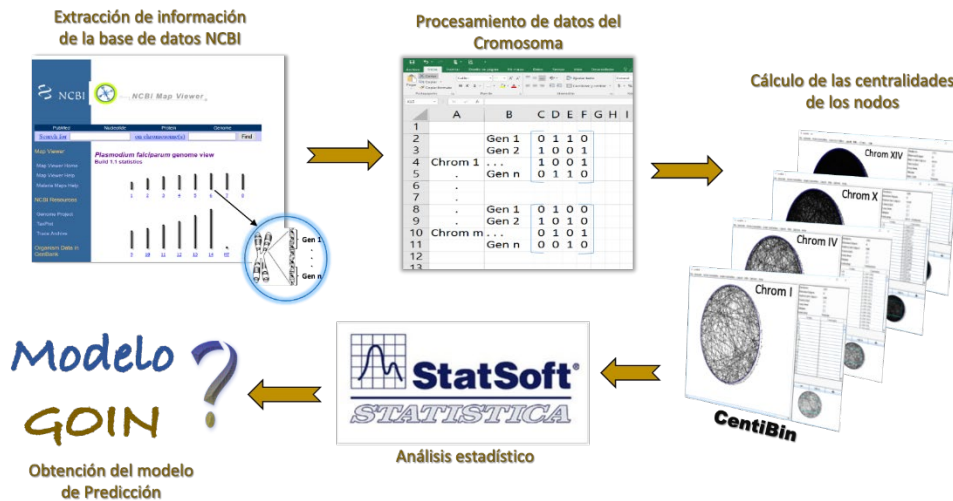


Figura 18. Flujo de trabajo del estudio de las GOIN del conjunto de datos del Proteoma Plasmodium.

3.2.2. Materiales y métodos

3.2.2.1. Conjunto de datos sobre el Genoma y Proteoma de *P. falciparum* en Mapviewer.

Se descarga toda la información sobre el gen (genoma) y las proteínas *P. falciparum* de la base de datos Mapviewer (<https://www.ncbi.nlm.nih.gov/projects/mapview/>) (Wolfsberg, 2011), (Wolfsberg, 2010), (Wolfsberg, 2007). El genoma de *P. falciparum* reportado en Mapviewer está organizado en 14 cromosomas. Cada cromosoma contiene un cierto número de genes totales. A su vez, la base de datos registra las coordenadas dentro del cromosoma para cada gen (posición inicial y final), utilizadas para obtener las secuencias.

Además, la base de datos informa el símbolo, positivo (+) o negativo (-) para determinar la Orientación (O_{ik}) del gen i ($Gene_{ik}$) a lo largo del cromosoma k_{th} , $O_{ik} = 1$ (positivo) u $O_{ik} = -1$ (negativo). También se obtuvo la Posición (P_{ik}) de cada gen en el cromosoma y una descripción de la función biológica de cada proteína en el proteoma del parásito. En la Tabla 5 del archivo S1 de la Información de Apoyo, se detallan las características de los 5365 genes y proteínas del proteoma de *P. falciparum*.

3.2.2.2. Redes de Inversión de Orientación de los Genes (GOINs). Se han construido dos tipos de grafos GOIN. El primero tipo es el grafo GOIN con información sobre la orientación de los genes solamente. El segundo tipo son los grafos S-GOIN que incluye información sobre la orientación de los genes y la distribución espacial. Se ha realizado los siguientes pasos. En primer lugar, se enumeran los valores de O_{ik} y P_{ik} de cada $Gene_{ik}$ en un archivo de Excel.

Después de eso, se han etiquetado a los nodos con números n_i ($i = 1, 2, \dots, n_{max}$) que representan los genes del cromosoma k . Estas etiquetas siguen el orden establecido por Mapviewer, es decir, el número 1 representa $Gene_{1k}$ y la posición está representado por $P_{1k} = 1$. Finalmente, se han utilizado dos enfoques diferentes en GOIN y S-GOIN para calcular la existencia ($L_{ij} = 1$) o no ($L_{ij} = 0$) de enlaces entre nodos en la red. Para los grafos GOIN, interconectamos los nodos de acuerdo con el patrón de inversión de la orientación del gen de sus vecinos. Entonces, usamos la siguiente función:

$$L_{ij} = Si(Y(O_{ik} * O_{jk} = -1; abs(P_{ik} - P_{jk}) < valor_de_corte); 1; 0) \quad (1)$$

Dónde, O_{ik} , O_{jk} son la orientación y P_{ik} , P_{jk} son la posición de $Gene_{ik}$ y $Gene_{jk}$, $0 < i, j \leq n_k$, n_k es el número de nodos (genes) del cromosoma k . Se establece como valor de corte (Ver Figura 19) el mínimo valor que garantiza que el grado de Centralidad fue $C_{deg}(Gene_i, Chr_k) \geq 1$ para todos los nodos de la red. Significa que no hay nodos aislados.

Evaluando la función, $L_{ij} = 1$ si $O_{ik} \neq O_{jk}$ (inversión de orientación) y $|P_{ik} - P_{jk}| < \text{valor de corte}$ y $L_{ij} = 0$ de lo contrario. Se pueden organizar los valores de L_{ij} en adyacencia cuadrada $n \times n$ matrices (A_k). Estas matrices se pueden visualizar como representaciones gráficas del GOIN. En el archivo S2 de la Información de Apoyo, se publica la información en formato .net para los 14 GOINs; uno para cada cromosoma. La Figura 20 muestra el proceso utilizado en este trabajo para construir las matrices.

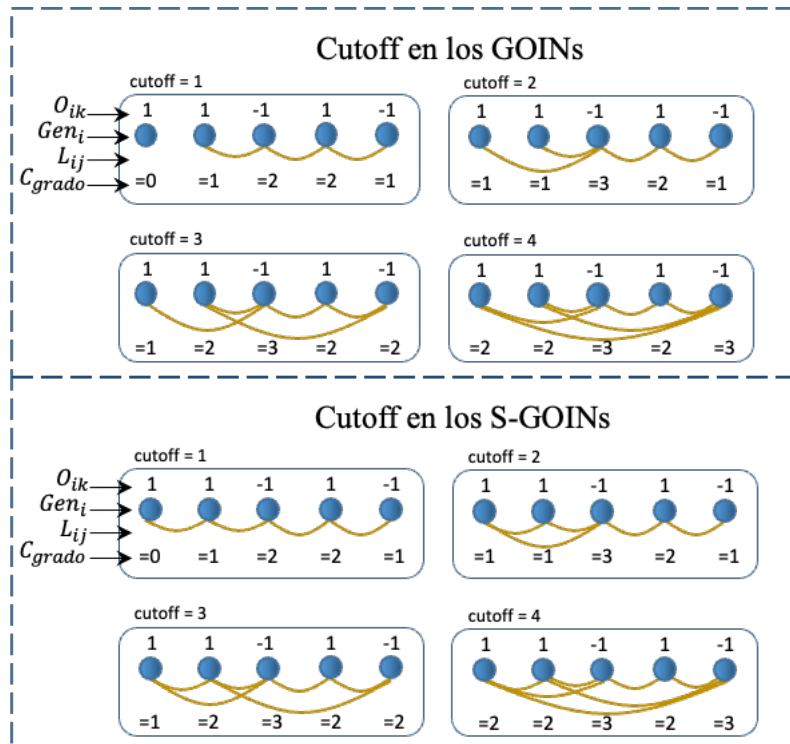


Figura 19. Ilustración de patrones de inversión génica con diferentes valores de corte.

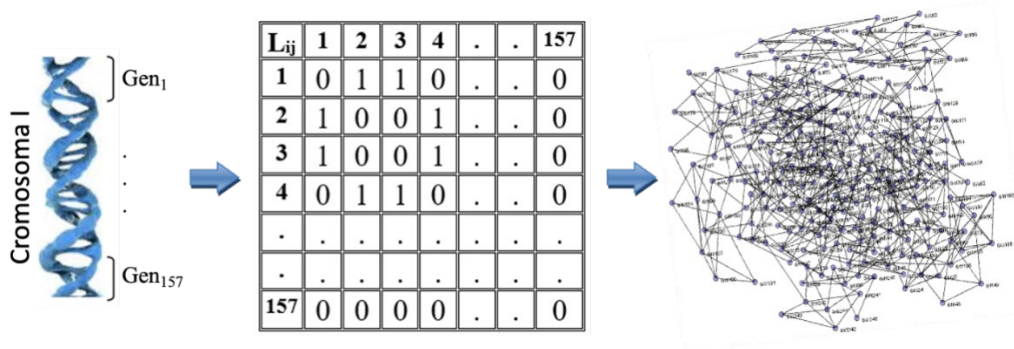


Figura 20. Construcción de matrices de adyacencia a partir de datos genéticos de cada cromosoma.

Para el segundo tipo de grafos, se comienza con la GOIN y se agrega la adyacencia espacial entre los genes a lo largo del cromosoma. Se denomina a esta segunda red espacial GOIN (S-GOIN). Al hacerlo, se usa la siguiente función anidada para determinar la existencia de enlaces ($L_{ij} = 1$) caso contrario ($L_{ij} = 0$):

$$L_{ij} = Si(O(abs(P_{ik} - P_{jk}); 1; Y(O_{ik} * O_{jk} = -1; abs(P_{ik} - P_{jk}) < valor_de_corte)); 1; 0) \quad (2)$$

La condición $|P_{ik} - P_{jk}| = 1$ garantiza que hay un enlace $L_{ij} = 1$ entre pares de genes $Gene_{ik}$ y $Gene_{jk}$ con adyacencia espacial en el cromosoma k . En el archivo S3 de la Información de Apoyo, se lanzan los archivos en formato .net para los 14 S-GOIN, uno para cada cromosoma. Como resultado, se generan 28 matrices, 14 para GOIN y otras 14 para S-GOIN. Además, se construyen 14 modelos de redes aleatorias (R) llamadas R-GOIN para cada uno de los S-GOIN, con fines comparativos. Estos 14 R-GOINs fueron construidos con el mismo número de nodos usando los modelos de red aleatoria Erdos-Renyi con el software CentiBiN versión 1.4.3 (Junker *et al.*, 2006). Se calculó la probabilidad del enlace de cada GOIN, S-GOIN y R-GOIN utilizando la siguiente función:

$$p_k = \frac{L_k}{L_{max}} = L_k / \binom{n_k}{2} = \frac{2 \times L_k}{n_k^2 - n_k} \quad (3)$$

Donde, $p(L_k)$ es la probabilidad de formación de enlaces L_{ij} entre dos nodos. L_k es el número de enlaces entre dos nodos, y n_k es el número de nodos en la red del cromosoma k^{th} . En el archivo S4 de la Información de Apoyo, se lanza la información en formato .net para los 14 R-GOINs. Los valores de L_k y $p(L_k)$ de la S-GOIN se usaron para construir el R-GOIN en el software CentiBin de acuerdo al modelo de redes aleatorias Erdos-Renyi (Junker, 2006). El archivo S5 de la Información de Apoyo muestra los valores de las matrices para todas las GOINs.

3.2.2.3. Centralidades de las redes complejas. En el caso particular de las redes de genes, las centralidades de los nodos pueden jugar un papel muy importante. La centralidad de un nodo, en teoría de grafos y redes complejas, se refiere a los parámetros que, de alguna manera, miden la importancia relativa del nodo dentro de la red. El valor de la centralidad de un nodo es útil, por ejemplo, para detectar vecinos relevantes.

En la Tabla 5, se resumen algunas Centralidades de nodo calculadas por CentiBiN, software utilizado en este trabajo. CentiBiN soporta centralidades de nodo para redes no dirigidas. Calcula cinco tipos diferentes de centralidades (C_1), que van desde medidas locales. Esas medidas solo consideran la vecindad directa de un elemento de red a las medidas globales (Junker *et al.*, 2006). El archivo S5 de la Información de Apoyo muestra los valores de las centralidades y los promedios calculados.

Tabla 5. Definición de parámetros más relevantes utilizados para describir redes complejas

Nombre de la centralidad	Símbolo	Fórmula ^a	Información
Distancia promedio	$C_{\text{dist}}(\text{Gene}_i, \text{Chr}_k)$	$= \text{avg}(\text{dist}(\text{gene}_i, \text{gene}_j))$	Distancia topológica promedio entre los genes del cromosoma y/o topológica entre los genes con orientación invertida.
Grado	$C_{\text{deg}}(\text{Gene}_i, \text{Chr}_k)$	$= \text{deg}(\text{gene}_i)$	Número de genes vecinos con orientación inversa en la vecindad del cromosoma (ventana de valor de corte).
Cercanía	$C_{\text{clo}}(\text{Gene}_i, \text{Chr}_k)$	$= \left(\sum_{\text{gene}_j \in V} \text{dist}(\text{gene}_i, \text{gene}_j) \right)^{-1}$	Proximidad de los genes vecinos con orientación inversa en la vecindad del (ventana de valor de corte).
Promedio de centralidad ^b	Símbolo	Fórmula ^a	Información
Distancia Promedio	$\langle C_{\text{dist}}(\text{Chr}_k) \rangle$	$= \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\text{avg}(\text{dist}(\text{gene}_i, \text{gene}_j)) \right)$	Valor esperado de la distancia (proximidad promedio) para un gen en el cromosoma k.
Closeness Promedio	$\langle C_{\text{clo}}(\text{Chr}_k) \rangle$	$= \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\sum_{\text{gene}_j \in V} (\text{dist}(\text{gene}_i, \text{gene}_j)) \right)^{-1}$ $= \frac{1}{n_k} \sum_{i=1}^{n_k} (C_{\text{clo}}(\text{Gene}_i, \text{Gene}_j))^{-1}$	Valor esperado de cercanía (proximidad promedio) para un gen en el cromosoma k.
Moving Promedio ^b	Símbolo	Fórmula ^a	Información

MA de Closseness	$\Delta C_{clo}(Gene_i, Chr_k)$	$= C_{clo}(Gene_i, Chr_k) - \langle C_{clo}(Chr_k) \rangle$	Deviación de la cercanía del gene i con respecto al valor esperado para todos los genes del mismo cromosoma k .
MA de Grado	$\Delta C_{deg}(Gene_i, Chr_k)$	$= C_{deg}(Gene_i, Chr_k) - \langle C_{deg}(Org) \rangle$	Deviación del grado del gen i con respect al valor esperado para todos los genes del organismo.

^a Todos los símbolos utilizados en estas fórmulas son muy comunes en la literatura de redes complejas y se deben explicar en detalle aquí. Sin embargo, $G = (V, E)$ es un gráfico conectado (fuerte) no dirigido con $n = |V|$ vértices o nodos; $dist(gene_i, gene_j)$ denota la longitud de una ruta más corta entre los nodos de $gene_i$ y $gene_j$. La matriz A es la matriz de adyacencia del grafo G ., para más detalles ver las referencias citadas. Software: CB = CentiBiN. ^b Ejemplos seleccionados de operadores utilizados de Moving Average (MA).

3.2.2.4. Datos de la función biológica de las proteínas relacionadas con RIFIN. En el conjunto de datos anterior, también se ha recogido la función biológica de las proteínas codificadas por los genes en cada cromosoma del organismo *P. falciparum*. Este organismo es el agente causal de la enfermedad mortal de la malaria (Goel *et al.*, 2015). Se ha centrado en los genes que codifican para la clase de proteínas llamadas RIFINs, presentes en este organismo y pertenecientes a la mayor familia conocida de proteínas expresadas en la superficie de eritrocitos y también es, naturalmente, inmunogénica. Las proteínas RIFIN se utilizaron para analizar las respuestas de anticuerpos de individuos que viven en un área de intensa transmisión de la malaria (Abdel *et al.*, 2002). Hay 149 genes en el genoma del parásito que codifican la proteína RIFIN. La información descargada de Mapviewer se procesó para asignar una clase de actividad biológica a cada proteína $R_{ik} = 1$ para proteínas relacionadas con RIFIN o $R_{ik} = 0$ en caso contrario. Se ha creado una lista de nombres de genes $Gene.name_{ik}$ para todos los genes en cada Chr_k . Luego, utilizando la siguiente función anidada en una hoja de cálculo, se han extraído los valores de R_{ik} :

$$R_{ik} = Si(ENCONTRAR("RIFIN"; Gene.name_{ik}); 1; 0) \quad (4)$$

Por ejemplo, el segundo gen se registra desde la posición de inicio = 39205 y la región de parada = 40430. Este gen tiene una orientación negativa $O_{ik} = -1$ y el símbolo es MAL1P4.02 y su descripción es RIFIN. Como el segundo gen pertenece a la clase de proteínas RIFIN entonces $R_{ik} = 1$. El último gen se registra desde la posición de inicio = 609110 hasta la región de parada = 616613, tiene una orientación negativa ($O_{ik} = -1$), el símbolo es PFA0765c y su descripción es proteína de membrana de eritrocitos 1 (PfEMP1). El último gen pertenece a la clase de no-RIFIN Proteínas ($R_{ik} = 0$). El archivo S5 de la Información de Apoyo muestra los valores para cada R_{ik} .

3.2.2.5. Modelos de ML. Se han utilizado los parámetros de las GOINs como entradas de modelos de ML para probar la capacidad de estas redes para codificar información microestructural del cromosoma relevante para la información de la actividad biológica.

Primero, se han calculado las centralidades de nodo C_t ($Gene_i, Chr_k$) para cada nodo ($Gene_i$) de cada grafo (Chr_k) inicialmente del modelo GOIN, más adelante en los modelos S-GOIN y R-GOIN. Después de eso, se han utilizado los valores de las centralidades de los nodos como insumos para llevar a cabo un GDA. El objetivo del GDA fue buscar un nuevo modelo capaz de discriminar genes que expresan proteínas similares a RIFIN de genes que codifican de otras proteínas en el proteoma de *P. falciparum*. La variable a predecir la presencia ($R_{ik}=1$) o de lo contrario ($R_{ik}=0$) de la proteína a la clase RIFIN. La salida de la ecuación GDA no es R_{ik} en sí, sino $S(R_{ik})$, que es una puntuación real de R_{ik} .

Se puede escribir la ecuación de GDA con los parámetros mencionado anteriormente en la siguiente forma:

$$\begin{aligned}
 S(R_{ik}) = a_0 + & \sum_{t=1}^5 b_t \cdot [C_t(Gene_i, Chr_k) - \langle C_t(Chr_k) \rangle] \\
 & + \sum_{t=1}^5 c_t \cdot [C_t(Gene_i, Chr_k) - \langle C_t(Orient_i, Chr_k) \rangle] \\
 & + \sum_{t=1}^5 d_t \cdot [C_t(Gene_i, Chr_k) - \langle C_t(Org) \rangle]
 \end{aligned} \tag{5}$$

$$S(R_{ik}) = a_0 + \sum_{t=1}^5 b_t \cdot \Delta C_t(Gene_i, Chr_k) \quad (6)$$

$$+ \sum_{t=1}^5 c_t \cdot \Delta C_t(Gene_i, Orient_i, Chr_k) + \sum_{t=1}^5 d_t \cdot \Delta C_t(Gene_i, Chr_k, Org)$$

Se ha utilizado el algoritmo GDA implementado en el software STATISTICA versión 12.0 para ajustar a un modelo lineal (Hill & Lewicki, 2006). Para este estudio, se ha usado como variable de salida R_{ik} y como entradas los diferentes operadores de “moving average” (ver los detalles en la Tabla 5). Los parámetros de la bondad de ajuste del modelo GDA utilizados aquí son: n = Número de casos, Chi cuadrado, p-value, especificidad (Sp) y sensibilidad (Sn), tanto para la muestra como para la validación de la matriz de clasificación. Por último, se ha realizado un análisis para calcular AUROC para los modelos (Hill & Lewicki, 2006). El archivo S6 de la Información de Apoyo muestra los resultados del Modelo lineal ML.

3.2.3. Resultados y discusión

3.2.3.1. GOINs de los cromosomas de *P. falciparum*. Se han construido dos tipos de grafos, las GOIN y las S-GOIN para todos los cromosomas de *P. falciparum* utilizando el software CentiBiN. En la Figura 21, se ilustran las GOIN y S-GOIN para el cromosoma I en la interfaz del software CentiBiN. A continuación, se realiza un triple análisis comparativo de los grafos GOIN y S-GOIN observados con los modelos Ęrdos-Renyi R-GOIN. El objetivo principal de este estudio es comparar las características topológicas generales de GOIN con modelos aleatorios con el fin de estudiar el grado de aleatoriedad de los patrones de orientación de los genes en el genoma de *P. falciparum*.

Al hacerlo, se construyen 14 R-GOIN lo más similares posible a los grafos S-GOIN respectivamente. También se calcularon varios parámetros topológicos de estas redes. En la Tabla 6, se resumen los resultados de este estudio para algunos cromosomas seleccionados.

Este estudio descriptivo se centró en los valores medios de distancia $\langle C_{dist} \rangle$ y la centralidad closeness $\langle^k C_{clo} \rangle$ entre los genes. Se ha centrado en este parámetro porque estas variables miden la distancia entre genes con inversión de orientación de genes, $O_{ik} \neq O_{jk}$. Después de una simple inspección visual, se puede concluir que los valores de los índices topológicos estudiados son muy diferentes en los grafos S-GOIN vs R-GOIN. Por ejemplo, la distancia topológica varía en el rango $\langle C_{dist} (Chr_1) \rangle = [15.71- 52.84]$ para los S-GOINs de los 14 cromosomas.

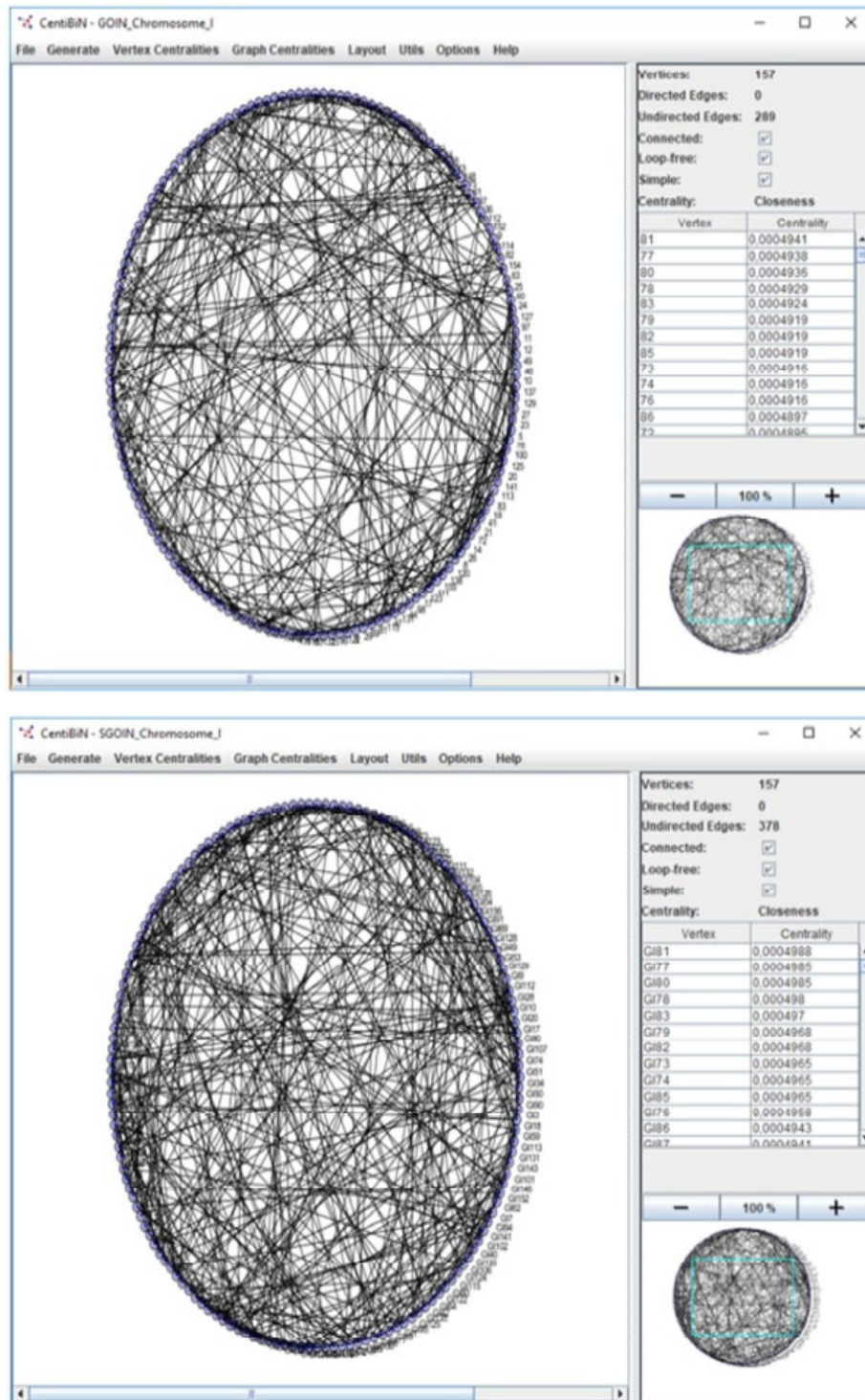


Figura 21. Ilustración del GOIN y S-GOIN para el cromosoma I en la interfaz del software CentiBiN.

Al contrario, el mismo parámetro varía en el rango $\langle C_{\text{diam}}(\text{Chr}_1) \rangle = [2.90-4.35]$ para las R-GOIN de los mismos cromosomas. Mientras tanto, los valores promedio de closeness

varían en el rango $\langle C_{clo} (Chr_1) \rangle = [0.00004-0.00038]$ (S-GOIN) vs. $\langle C_{clo} (Chr_1) \rangle = [0.00043-0.00191]$ (R-GOIN), (ver ejemplos seleccionados en Tabla 6). En el archivo S1 de la Información de Apoyo, se muestran resultados detallados para cada uno de los 14 cromosomas. De hecho, la Figura 22 ilustra las notables diferencias en los valores promedio de Closeness $\langle C_{clo} (Chr_k) \rangle$ para S-GOINs vs las redes aleatorias en los 14 cromosomas respectivamente.

Estos resultados parecen indicar que la distribución de la orientación de los genes en el S-GOIN no sigue un patrón aleatorio de acuerdo con el modelo de Erdős-Renyi (Junker, 2006).

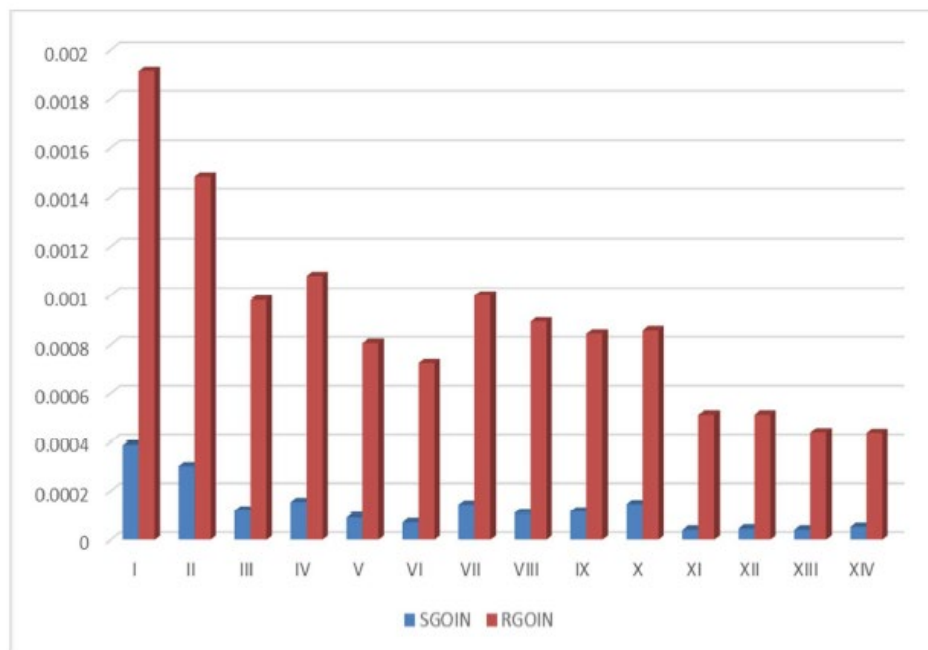
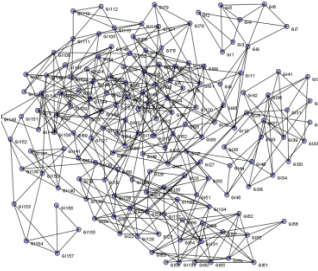
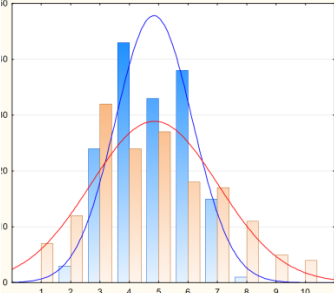
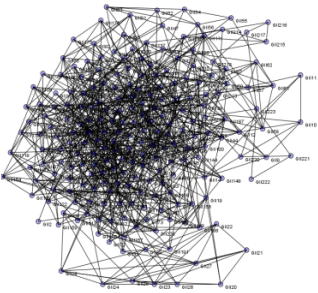
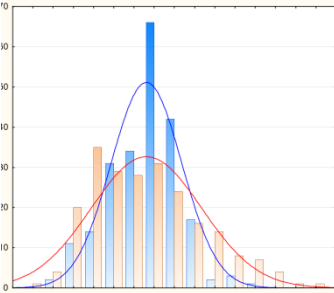
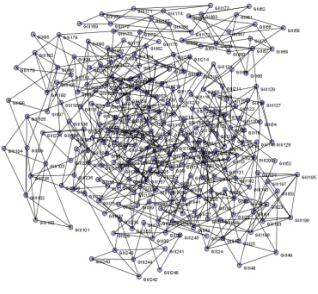
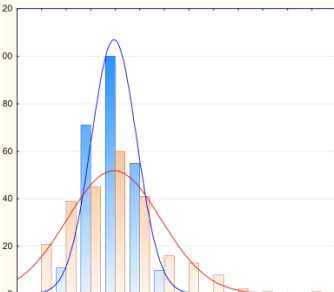


Figura 22. Promedio de proximidad de S-GOINs frente a R-GOINs de 14 cromosomas.

Los histogramas representados en la Tabla 6 muestran que las distribuciones de grados de nodos para diferentes cromosomas se asemejan a la distribución normal para las redes S-GOIN y R-GOIN (Ver los resultados completos para 14 cromosomas por separado en el archivo S1 de la Información de Apoyo). Sin embargo, se ha llevado a cabo el test de normalidad de Kolmogorov-Smirnov (KS) para la S-GOIN de todo el cariotipo de *P. falciparum* (todos los cromosomas). Como el valor de $D = 0.1629$ tiene asociado a $Lillieforsp < 0.01$ (inferior a 0.05) se debe rechazar la hipótesis de que el S-GOIN de *P. falciparum* sigue una distribución normal. Además, se ha realizado el test de normalidad de Shapiro-Wilks para cada una de las 14 redes GOIN + 14 S-GOIN = 28 redes estudiadas por separado. Se pueden ver los resultados completos de 28 Redes separadas en el archivo S1 de la Información de Apoyo.

Todos los valores del software Statistic están en el rango $W = 0.89$ a 0.98 con $p < 0.01$ (inferior a 0.05) para las 28 redes. Entonces, se deberían rechazar la hipótesis de que la inversión genética en GOIN y S-GOINs de *P. falciparum* siguen una distribución normal.

Tabla 6. Modelos GOIN / S-GOIN vs. R-GOIN (ejemplos seleccionados)

Grafo S-GOIN	Parám. ^a	Redes ^b			Distribuciones ^c				
		S-GOIN	GOIN	R-GOIN					
 <p>Cromosoma I</p>	n	157	157	157	 <table border="1" data-bbox="1173 907 1468 985"> <tr> <td>t-value</td> <td>p</td> </tr> <tr> <td>-6.30E-02</td> <td>0.94977</td> </tr> </table>	t-value	p	-6.30E-02	0.94977
	t-value	p							
	-6.30E-02	0.94977							
	L	378	289	379					
	$p(L_k)$	0.031	0.024	0.031					
	$\langle C_{dist} \rangle$	17.28	17.50	3.39					
	$\langle C_{clo} \rangle$	0.38	0.38	1.91					
$W(p)$	0.933 (<0.01)	0.936 (<0.01)	0.9587 (<0.01)						
 <p>Cromosoma II</p>	n	223	216	223	 <table border="1" data-bbox="1173 1344 1468 1422"> <tr> <td>t-value</td> <td>p</td> </tr> <tr> <td>8.20E-15</td> <td>1.00000</td> </tr> </table>	t-value	p	8.20E-15	1.00000
	t-value	p							
	8.20E-15	1.00000							
	L	734	615	734					
	$p(L_k)$	0.030	0.026	0.030					
	$\langle C_{dist} \rangle$	15.71	15.33	3.06					
	$\langle C_{clo} \rangle$	0.30	0.31	1.48					
$W(p)$	0.9569 (<0.01)	0.9584 (<0.01)	0.959 (<0.01)						
 <p>Cromosoma III</p>	n	247	230	247	 <table border="1" data-bbox="1173 1780 1468 1859"> <tr> <td>t-value</td> <td>p</td> </tr> <tr> <td>3.30E-15</td> <td>1.00000</td> </tr> </table>	t-value	p	3.30E-15	1.00000
	t-value	p							
	3.30E-15	1.00000							
	L	485	358	485					
	$p(L_k)$	0.016	0.014	0.016					
	$\langle C_{dist} \rangle$	35.76	33.68	4.19					
	$\langle C_{clo} \rangle$	0.11	0.13	0.98					
$W(p)$	0.8966 (<0.01)	0.9202 (<0.01)	0.9387 (<0.01)						

^a Param. = Parámetros de red son: n = número de nodos, L = número de enlaces, $p(L_k)$ = probabilidad de enlace, $\langle C_{dist} \rangle$ = Promedio distancia topológica, $\langle C_{clo} \rangle$ = Centralidad promedio de Closeness, parámetro estadístico de prueba Shapiro-Wilks $W(p)$ = . ^b GOINs

solamente tiene información sobre las inversiones, pero sin la distribución espacial del gen en el cromosoma, S-GOIN = GOINs con ambas inversiones y se incluye la distribución espacial del gen en el cromosoma y R-GOIN = *Érdos-Renyi* como modelo de red aleatorio de S-GOIN.
^b Valores en notación científica en escala $\times 10^{-3}$. ^c S-GOINs son de color azul y R-GOIN de color naranja.

De lo contrario, se comparan los grados de S-GOIN de todos los cromosomas (14 redes) frente a los respectivos valores de los R-GOINs (14 redes) obtenidos con el modelo ER. Para la comparación se utilizó una prueba de Student. ($t = -0.0292$ y $p = 0.98$). Esto confirma que se deben rechazar la hipótesis (H_0) de que ambas distribuciones tienen el mismo valor promedio de grado (Figura 23). En conclusión, los patrones de inversión de la distribución de la orientación de los genes (grados de nodo GOIN/S-GOINs) en *P. falciparum* no parecen tener una distribución aleatoria según modelo ER. Se pueden ver los resultados completos de estas 14 redes aleatorias separadas en el archivo S1 en la Información de Apoyo.

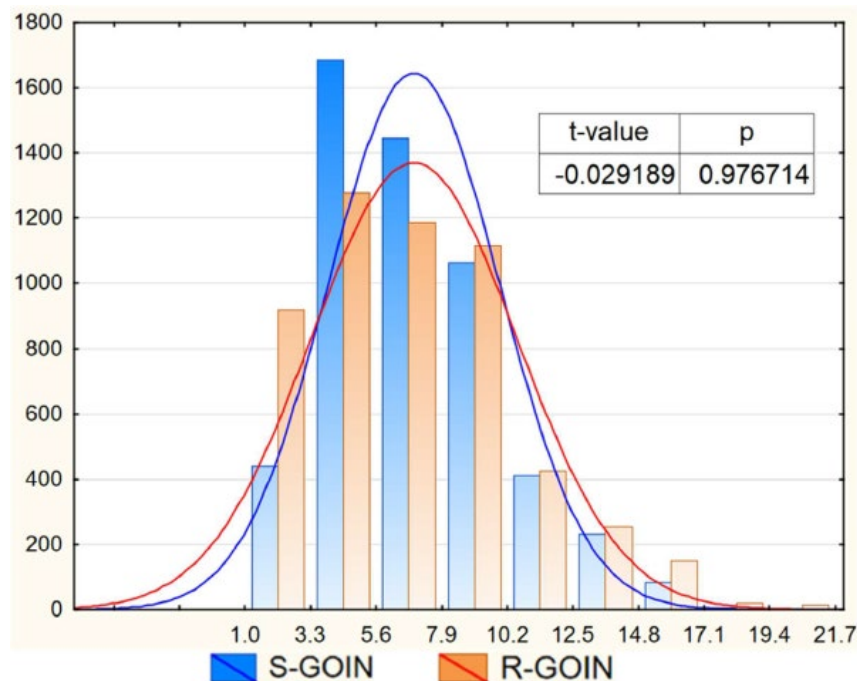


Figura 23. Grado de S-GOINs frente a R-GOINs de 14 cromosomas.

3.2.3.2. Comunidades en las GOINs de *P. falciparum*. Otro objetivo del estudio es hacer una evaluación preliminar de la capacidad de las metodologías de GOIN para encontrar los genes más relevantes y, o, agrupaciones de genes. Se refieren las agrupaciones de genes como pequeños grupos, comunidades o sub-grafos en la red compleja. Pueden ser co-expresadas

codificando proteínas con funciones relacionadas en el proteoma y, o, trabajo entre ellas para regular una función biológica específica (Haye *et al.*, 2014; Inoue *et al.*, 2007).

Cuando se construyen los 14 grafos de GOIN con la herramienta CentiBiN, se percibe que se crearon pequeñas comunidades aisladas (hasta 3) en pocos grafos. La Figura 24 muestra las comunidades capturadas del grafo número 2 en la herramienta CentiBiN. Se nota que estas pequeñas comunidades corresponden a genes consecutivos en el cromosoma. De hecho, al aplicar la fórmula (1) para obtener las matrices se observó que las comunidades se formaron cuando varios genes consecutivos mantienen la misma orientación (esto también depende del valor de corte). Por ejemplo, la GOIN del cromosoma II tiene 223 nodos (genes) y 2 comunidades. La primera comunidad $Comm_1$ (II) tiene 216 genes y la segunda comunidad $Comm_2$ (II) es un pequeño grupo de 7 genes. Así, el estudio se va a centrar solo en las comunidades más pequeñas o grupos de genes porque son más factibles para inspección visual.

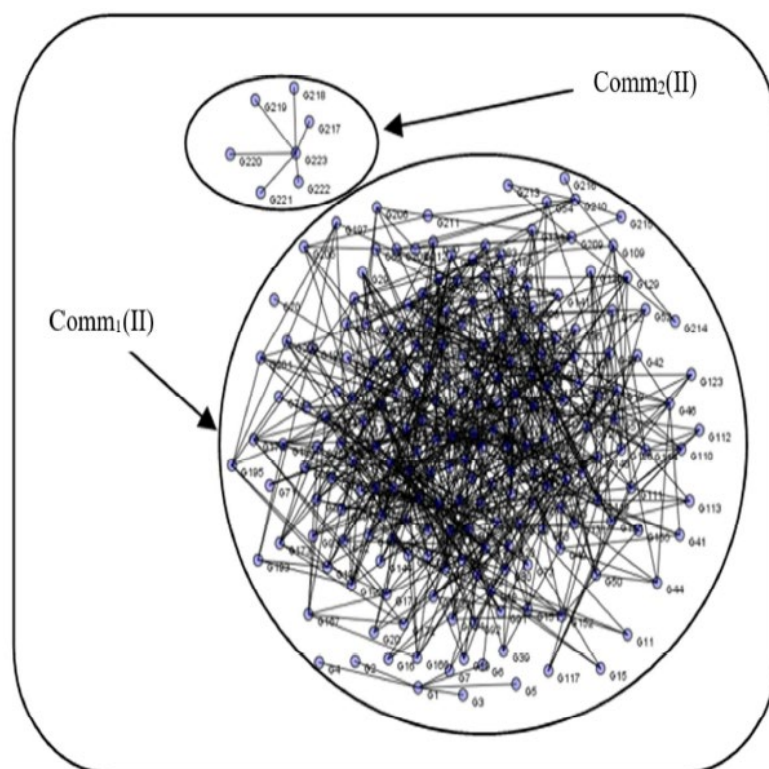


Figura 24. Ilustración de las comunidades del cromosoma II.

La comunidad más pequeña $Comm_2$ (II) envuelve los últimos genes en el cromosoma (desde gene217 hasta gene223). Estos genes presentan las siguientes actividades biológicas. TheGene₂₁₇: PfEMP1, Gene₂₁₈: proteína hipotética PFB1030w, Gene₂₁₉: RIFIN, Gene₂₂₀: RIFIN, Gene₂₂₁: PfEMP1, Gene₂₂₂: RIFIN y Gene₂₂₃: PfEMP1. Interesante, todos estos nodos

parecen formar un grupo de genes casi todos relacionados con la actividad biológica RIFIN (ver tabla 7). Del mismo modo, la comunidad Comm₂ (III) del cromosoma III tiene 5 codificaciones para la proteína RIFIN y 1 gen PfEMP1 de 11 genes con función conocida. También Comm₃ (VII) tiene 7 genes que codifica proteínas RIFIN, 1 gen PfEMP1 de 9 genes con función conocida. En consecuencia, las mismas comunidades GOINs parecen tener una tendencia a ser agrupaciones de secuencias similares a RIFIN que probablemente tengan algún papel biológico. De hecho, es bien sabido que todos ellos RIFIN, PfEMP1 y stevor están relacionados. Proteínas traficadas a la superficie de los eritrocitos (Lavazec *et al.*, 2006). De hecho, todas las secuencias similares a RIFIN contienen el PEXEL motif (una secuencia pentamérica RxLxE/Q/D, conocida como el elemento de exportación *Plasmodium* sp.) requerida para la correcta exportación y expresión de superficie o señal de destino de host (HT) que desempeña un papel central en la exportación de proteínas en la célula huésped (Mwakalinga *et al.*, 2012).

Tabla 7. Proteínas en las comunidades más pequeñas.

Crom (K)	n _k	Comunidad ^b		Rango de Genes en las Comunidades	Funciones biológicas ^a							
		Comm _q (K)	N _{Comm.}		A	B	C	D	E	F	G	
I	157	-	-	-	-	-	-	-	-	-	-	-
II	223	Comm ₁ (II)	216	1 - 216	-	-	-	-	-	-	-	-
		Comm ₂ (II)	7	217 - 223	3	3	-	-	-	-	-	1
		Comm ₂ (III)	13	1 - 13	5	1	-	1	-	4	2	
III	247	Comm ₁ (III)	230	14 - 243	-	-	-	-	-	-	-	-
		Comm ₃ (III)	4	244 - 247	1	1	-	1	-	-	-	1
IV	254	-	-	-	-	-	-	-	-	-	-	-
		Comm ₂ (V)	5	1 - 5	2	1	-	-	-	-	1	1
V	330	Comm ₁ (V)	325	6 - 330	-	-	-	-	-	-	-	-
VI	322	-	-	-	-	-	-	-	-	-	-	-
		Comm ₂ (VII)	90	1 - 90	3	4	7	-	3	17	56	
VII	294	Comm ₁ (VII)	192	91 - 282	-	-	-	-	-	-	-	-
		Comm ₃ (VII)	12	283 - 294	7	1	-	1	-	-	-	3
VIII	299	Comm ₁ (VIII)	286	1 - 286	-	-	-	-	-	-	-	-

		Comm ₂ (VIII)	13	287 - 299	2	3	-	-	-	3	5
IX	367	-	-	-	-	-	-	-	-	-	-
X	404	-	-	-	-	-	-	-	-	-	-
XI	490	-	-	-	-	-	-	-	-	-	-
		Comm ₁ (XII)	519	1 - 519	-	-	-	-	-	-	-
XII	530	Comm ₂ (XII)	11	520 - 530	7	1	-	2	-	-	1
XIII	677	-	-	-	-	-	-	-	-	-	-
XIV	771	-	-	-	-	-	-	-	-	-	-
Total de la función Biológica					30	15	7	5	3	25	70

^a Funciones biológicas: A = RIFIN, B = PfEMP1, C = proteína Cg, D = Stevor, E = proteína de choque térmico, F = Otras funciones, G = proteína indefinida o hipotética. ^b El "-" indica que no se formaron comunidades en estos cromosomas, por este motivo, solo se muestra información de las comunidades más pequeñas (Comm₂ y Comm₃)

3.2.3.3. Modelos predictivos GOIN-ML de función biológica de genes. En las secciones anteriores, se encuentran algunas pistas que indican que GOIN no parece ser aleatorio, lo que indica un posible papel biológico. También se encuentra que algunos pequeños sub-grafos GOIN parecen ser proteínas relacionadas con grupos de genes con función RIFIN en el proteoma del parásito. En consecuencia, se supone que se puede obtener un modelo computacional capaz de predecir las proteínas RIFIN en el proteoma utilizando solo parámetros numéricos de la estructura de GOIN y sin confiar sobre la secuencia de las proteínas. **El modelo reportado aquí es probablemente el primer modelo para la inferencia de la función de las proteínas basándose solo en la información de GOIN y no en la secuencia de los genes.** El mejor modelo lineal encontrado fue el siguiente:

$$S(R_{ik}) = -68035.949 \cdot [C_{clo}(Gene_i, Chr_k) - \langle C_{clo}(Chr_k) \rangle] - 4.896 \quad (7)$$

Donde $n = 4025$, $\chi^2 = 293.77$ y $p < 0.05$. La salida del modelo $S(R_{ik})$ es una función de puntuación de valor real útil para predecir cuando el gen (representado por el nodo n_{ik}) codifica una proteína con función RIFIN ($R_{ik} = 1$) o no ($R_{ik} = 0$). Cuando el valor de $S(R_{ik}) > 0$ se puede esperar que el gen representado por este nodo tenga una mayor propensión a codificar

una proteína con actividad biológica RIFIN ($R_{ik} = 1$). La única variable de entrada en el modelo es $\Delta C_{clo}(\text{Gene}_i, \text{Chr}_k) = C_{clo}(\text{Gene}_i, \text{Chr}_k) - \langle C_{clo}(\text{Chr}_k) \rangle$. Esta variable mide la desviación de Closeness $C_{clo}(\text{Gene}_i, \text{Chr}_k)$ de gene_i con respecto al valor esperado de proximidad $\langle C_{clo}(\text{Chr}_k) \rangle$ para todos los genes en el mismo Chromosome Chr_k (ver también Tabla 5). Se deben tener en cuenta los valores de $C_{clo}(\text{Gene}_i, \text{Chr}_k)$ cuantificar la proximidad del Gene_i a genes vecinos con orientación inversa en su vecindario en el cromosoma Chr_k . En consecuencia, según este modelo los genes con menor Closeness (cercanía) tienen menos genes vecinos con orientación invertida y mayor valor $S(R_{ik})$ de la función de puntuación que indica mayor propensión a codificar proteínas RIFIN en el proteoma de *P. falciparum*. Los parámetros estadísticos de este modelo son: n = número de nodos (genes) en entrenamiento, χ^2 son las estadísticas de Chi-cuadrado, y p es p-level de error. El modelo discriminó correctamente los RIFIN de otras proteínas en el proteoma de *P. falciparum* con alta especificidad y sensibilidad 70-85% (ver Tabla 8). En cualquier caso, los valores prefijados para hacer comparaciones en los modelos de clasificación no son valores de función de puntuación $S(R_{ik})$ pero sí los valores de probabilidad $p(R_{ik})$. Los valores $p(R_{ik})$ son las probabilidades posteriores con lo que el modelo predice que Gene_i codifica una proteína con función RIFIN. Compilamos en detalle información sobre los cromosomas, nodos, funciones biológicas, C_{clo} , R_{ik} observado, R_{ik} previsto y $p(L_k)$ para todos los 5365 casos estudiados en este trabajo en el archivo S6 en la Información de Apoyo. En la Tabla 9, se ofrecen ejemplos seleccionados de proteínas RIFIN predichas por el modelo.

Además, se intenta probar la robustez de las relaciones lineales entre $C_{clo}(\text{Gene}_i, \text{Chr}_k)$ y R_{ik} utilizando otro algoritmo. Para ello, se realiza un análisis de los resultados de la Red Neural Lineal ó “Linear Neuronal Network” (LNN). Los valores de S_p y S_n encontrados son similares o incluso ligeramente más altos que para GDA. Una ventaja del algoritmo LNN es la posibilidad de ejecutar un análisis de curva ROC. Se encuentra un valor de AUROC = 0.85 - 0.95 para las series de entrenamiento y validación externa (ver la Figura 25). Estos valores son notablemente más altos que los típicos AUROC = 0.5 de un clasificador aleatorio y cercanos a AUROC = 1 para un clasificador perfecto (Hill & Lewicki, 2006). Este resultado parece confirmar que la relación entre la orientación del gen y la función biológica de las proteínas no parece aleatoria al menos para RIFIN en el proteoma de *P. falciparum*.

Tabla 8. Resultados del análisis discriminante.

Algorit. ^a	Set	Clase	Parám. estadís. ^b	Valor (%)	$R_{ik} = 0$ $p = 0.45$	$R_{ik} = 1$ $p = 0.55$
GDA	Entrenar	0	S_p	84.2	3294	620

		1	Sn	72.1	31	80
	Validación	0	Sp	84.8	1103	198
		1	Sn	84.6	6	33
LNN	Entrenar	0	Sp	83.4	3263	651
		1	Sn	77.5	25	86
	Validación	0	Sp	84.2	1095	206
		1	Sn	89.7	4	35

^a GDA: Análisis Discriminante General, LNN: Red Neuronal Lineal.

^b Parámetros: Especificidad (Sp) and Sensibilidad (Sn)

Tabla 9. Principales genes codificando proteínas con altas probabilidades RIFIN según modelo

^a Num.	Gen	Orient	Crom.	Posición	Descripción	Observado	Probab.	Pred.
156	1	I	156	RIFIN	1	0.9998	1	
2	-1	I	2	RIFIN	1	0.9997	1	
4	1	I	4	RIFIN	1	0.9995	1	
152	1	I	152	RIFIN	1	0.9993	1	
376	1	II	219	RIFIN	1	0.9988	1	
377	1	II	220	RIFIN	1	0.9988	1	
379	1	II	222	RIFIN	1	0.9988	1	
382	-1	III	2	RIFIN	1	0.8639	1	
385	-1	III	5	RIFIN	1	0.8515	1	
386	1	III	6	RIFIN	1	0.8390	1	
880	1	IV	253	RIFIN	1	0.9307	1	
629	-1	IV	2	RIFIN	1	0.9303	1	
878	-1	IV	251	RIFIN	1	0.9189	1	
885	-1	V	4	RIFIN	1	0.7596	1	
1209	1	V	328	RIFIN	1	0.7521	1	
886	-1	V	5	RIFIN	1	0.7457	1	
1532	1	VI	321	RIFIN	1	0.6479	1	
1214	-1	VI	3	RIFIN	1	0.6338	1	
1525	1	VI	314	RIFIN	1	0.6259	1	
1535	1	VII	2	RIFIN	1	0.9255	1	
1536	-1	VII	3	RIFIN	1	0.9134	1	
1823	-1	VII	290	RIFIN	1	0.9067	1	
2123	1	VIII	296	RIFIN	1	0.7878	1	
2122	1	VIII	295	RIFIN	1	0.7702	1	
2045	-1	VIII	218	RIFIN	1	0.2039	0	

2128	-1	IX	2	RIFIN	1	0.8371	1
2129	-1	IX	3	RIFIN	1	0.8371	1
2131	-1	IX	5	RIFIN	1	0.8371	1
2893	1	X	400	RIFIN	1	0.9266	1
2894	1	X	401	RIFIN	1	0.9266	1
2895	1	X	402	RIFIN	1	0.9266	1
3383	1	XI	486	RIFIN	1	0.4675	0
3386	-1	XI	489	RIFIN	1	0.4675	0
3389	-1	XII	2	RIFIN	1	0.4931	0
3390	-1	XII	3	RIFIN	1	0.4931	0
3919	-1	XIII	2	RIFIN	1	0.4611	0
3921	1	XIII	4	RIFIN	1	0.4611	0
5358	1	XIV	764	RIFIN	1	0.5314	1
5360	1	XIV	766	RIFIN	1	0.5314	1

^aNodos con las mayores probabilidades del modelo

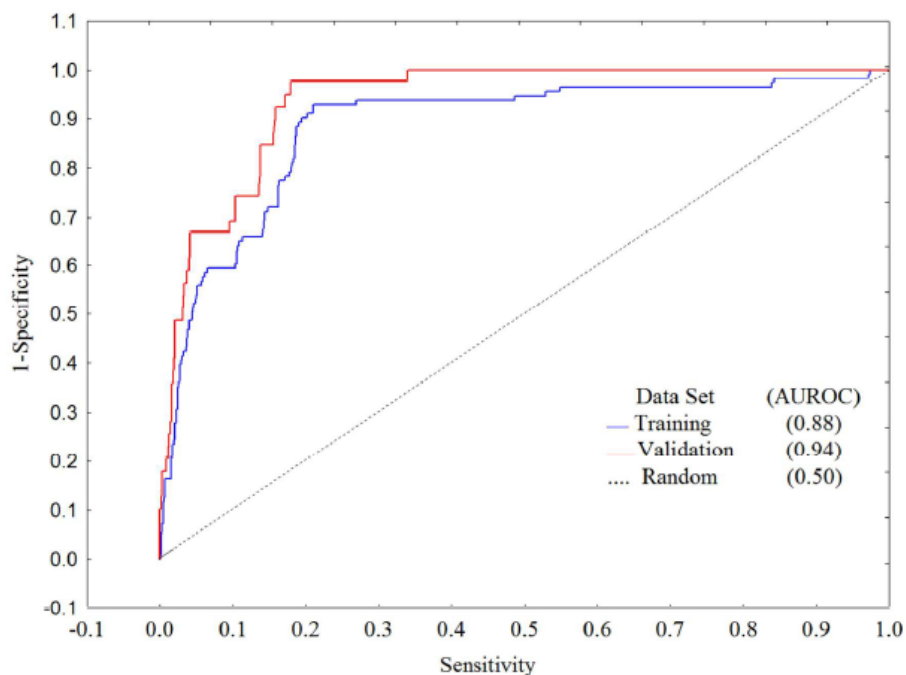


Figura 25. Resultados del análisis de la curva ROC para el modelo lineal.

3.2.4. Conclusiones

Se demuestra, por primera vez, cómo construir una representación de una red compleja de la distribución espacial y orientación de los genes en el cromosoma. Se acuñan este nuevo tipo de las redes como redes de inversión de orientación de los genes (GOINs). Se informa de algunos resultados que parecen confirmar que la orientación de los genes no sigue un patrón aleatorio en los cromosomas de *P. falciparum*. Los resultados también indican que

existe una relación entre la orientación del gen y la función biológica de las proteínas; al menos para RIFIN en el proteoma de *P. falciparum*.

3.2.5. Contenido asociado

Información de apoyo

La información de apoyo está disponible de forma gratuita en el sitio web de las Publicaciones de ACS en el DOI: 10.1021/acs.jproteome.7b00861.

Archivo S1. Tabla S5. Información del proteoma y del genoma de *P. falciparum* descargada de la herramienta Mapviewer. Tabla S6. Información completa de la Tabla 2 sobre los modelos GOIN/S-GOIN versus R-GOIN. Tabla S7. Histogramas y test de student de GOIN, R-GOIN y S-GOIN. (PDF)

Archivo S2. Archivos en formato.net para todos los GOIN de los 14 cromosomas de *P. falciparum*. (TXT)

Archivo S3. Archivos en extensión.net para todas las S-GOIN de los 14 cromosomas de *P. falciparum*. (TXT)

Archivo S4. Archivos en extension.net para todos los R-GOINs de los 14 cromosomas de *P. falciparum*. (TXT)

Archivo S5. Matrices GOIN, S-GOIN y R-GOIN, valores de las centralidades y los promedios calculados, y los valores de cada Rik para los genes de los diferentes cromosomas. (XLSX)

Archivo S6. Modelo de aprendizaje automático. (XLSX)

3.3. Capítulo 3. Modelos forestales aleatorios para sistemas de liberación de Nanopartículas decoradas con fármacos antimaláricos

Artículo 3

Ref. Diana V. Urista, Diego B. Carrué, Iago Otero, Sonia Arrasate, Viviana F. Quevedo-Tumaili, Marcos Gestal, Humbert González-Díaz and Cristian R. Munteanu. (2020). Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest. *Biology*. 9(8): 198. doi: 10.3390/biology9080198.

3.3.1. Introducción

Las NanoPartículas Decoradas con Fármacos ó “Drug-Decorated Nanoparticles” (DDNPs) se encuentran entre los nanomateriales más interesantes, con una amplia gama de aplicaciones médicas. Muchas de ellas se utilizan en sistemas de administración de fármacos para diferentes tipos de compuestos químicos. Estos sistemas presentan numerosas ventajas, ya que existen innumerables combinaciones de fármacos y nanopartículas que pueden ser eficaces en el tratamiento de diferentes enfermedades. Al mismo tiempo, tienen algunos puntos débiles. Por ejemplo, la síntesis de nanopartículas puede ser a veces costosa, o puede implicar mucho tiempo que puede aumentar con el número de muestras. Por este motivo, para mejorar la posibilidad de formar pares eficaces, se hace imprescindible la utilización de modelos de computación avanzada y AI.

Recientemente, algunas investigaciones se han centrado en encontrar DDNPs que muestren propiedades antimaláricas. Por ejemplo, las nanopartículas de plata y oro, que se sintetizaron a partir de extractos de hojas y cortezas de Myrtaceae, mostraron una eficaz actividad antiplasmódica (Dutta *et al.*, 2017), y las nanopartículas de ZnO (EPS-ZnO NPs) presentaron efectos funcionales contra los vectores de la malaria (Abinaya *et al.*, 2018). Por lo tanto, este estudio pretende diseñar un modelo computacional útil que permita una buena predicción de la actividad antimalárica de diversos pares de fármacos y nanopartículas.

Además, se ha desarrollado un nuevo método de fusión de datos en nanotecnología, ciencias biomoleculares, química y el análisis de big data que se ha propuesto en diferentes trabajos: integra la Teoría de Perturbación (PT) y el Aprendizaje Automático (ML) (Quevedo-Tumaili *et al.*, 2018; Ferreira da Costa *et al.*, 2018; Martínez-Arzate *et al.*, 2017; Liu *et al.*, 2017; González-Durruthy *et al.*, 2017; Ran *et al.*, 2016; Luan *et al.*, 2014; Kleandrova *et al.*, 2014), utilizando distintos operadores de PT para analizar los cambios en las variadas condiciones no estructurales y estructurales de una prueba a la vez (PTML). Algunos de estos

operadores PT representan la generalización de un enfoque clásico de la quimioinformática introducido por Corwin Hansch (Hansch, 2011). El autor se dio cuenta del importante potencial de utilizar metodologías predictivas para resolver cuestiones multivariantes en química médica. El enfoque clásico de Hansch permite buscar modelos con múltiples condiciones fisicoquímicas para predecir la actividad biológica de los compuestos, y estos modelos posiblemente incluyan términos cuadráticos y, o, lineales. En este proceso, que es un modelo de Relación de Energía Libre Lineal ó “Linear Free Energy Ratio” (LFER), la mayoría de los términos son parámetros fisicoquímicos relacionados con la energía libre de ionización del fármaco, de unión, de transporte, *etc.* Además, como se está fusionando la información (IF) de los fármacos y las nanopartículas, el modelo se convierte en un PTMLIF (PTML + IF).

Como ilustración, el término logarítmico (logP) del coeficiente de partición octanol/agua (P) se presenta como una estimación de la energía libre de transporte del fármaco y de la lipofilia molecular (Kubinyi, 1993). Se pueden aproximar los valores de logP mediante métodos de fragmentos químicos (tales como CLogP), o mediante métodos atómicos (como ALogP o XLogP) (Cho & Hermsmeier, 2002; Tetko *et al.*, 2001). Los términos logarítmicos de las constantes de acidez (pKa) están relacionados con la energía libre de ionización del fármaco. Además, para tener en cuenta más propiedades moleculares, se pueden utilizar diferentes parámetros como el Área de Superficie Polar (PSA). En general, para una molécula m_i , se pueden utilizar como entrada para el modelo varios tipos de propiedades moleculares, teniendo en cuenta medidas de polarizabilidad, lipofilia, electronegatividad, *etc.* (Tetko *et al.*, 2001; Zhang *et al.*, 2006). Se pueden definir estos modelos como:

$$f(\varepsilon_i) = \sum_{k=1}^{kmax} a_k \cdot D_k(m_i) + \sum_{k=1}^{kmax} b_k \cdot D_k(m_i)^2 + e_0 \quad (1)$$

Donde ε_i es la actividad biológica de la molécula m_i , $f(\varepsilon_i)$ es una función de la variable ε_i , $D_k(m_i)$ son los descriptores moleculares de m_i , y a_k y b_k son los coeficientes. Este modelo clásico sirve para tener en cuenta los cambios en la estructura química del fármaco/compuesto mediante los descriptores moleculares, pero no tiene en cuenta el resultado sobre la actividad del fármaco de las perturbaciones en múltiples condiciones experimentales (c_j). Estas incluyen las condiciones de ensayo o los cambios en la estructura química del fármaco, como c_0 = el parámetro biológico utilizado (CC_{50} : la relación de la concentración citotóxica del 50%, IC_{50} : concentración de inhibición, *etc.*), c_1 = organismo, c_2 = nombre de la célula, c_3 = organismo de ensayo, *etc.* Un ejemplo son los grandes conjuntos de datos que se encuentran en la base de datos pública ChEMBL (Davies *et al.*, 2015), (Papadatos & Overington, 2014; Bento *et al.*, 2014; Willighagen *et al.*, 2013; Hu & Bajorath, 2012; Wassermann *et al.*, 2011; Gaulton *et al.*,

2012). Se han utilizado los métodos PTML para analizar un gran conjunto de más de 50.000 ensayos preclínicos de fármacos. Estos ensayos incorporan fármacos dirigidos a *Plasmodium*. El modelo PTML procede del enfoque clásico LFER para la actividad de los fármacos. Se ha combinado el uso de ocho métodos de ML con la selección de características para obtener un clasificador más preciso para nuestra tarea.

3.3.2. Materiales y métodos

La Figura 26 presenta la metodología utilizada para construir el clasificador PTMLIF para la actividad antimalárica de las DDNPs. El flujo metodológico contiene los siguientes pasos:

- (1) Obtener 23 propiedades de las nanopartículas y fármacos/compuestos antipalúdicos de la literatura y las bases de datos públicas como descriptores moleculares iniciales;
- (2) Fusionar la información sobre las condiciones experimentales y las propiedades de los fármacos/compuestos y nanopartículas, utilizando una transformación centrada en el experimento de las características originales (operadores de Moving Average de Box-Jenkins);
- (3) Integrar los datos de fármacos/compuestos y nanopartículas en el conjunto de datos del estudio;
- (4) Construir los modelos PTMLIF de referencia utilizando los parámetros por defecto de los métodos ML
- (5) Mejorar el rendimiento del mejor clasificador utilizando sólo las características más importantes (selección de características).

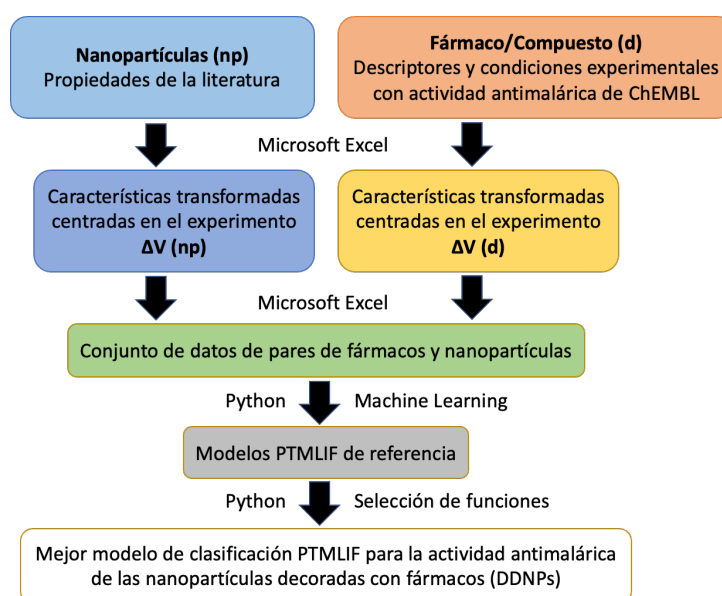


Figura 26. Flujo de trabajo para el desarrollo de Modelos PTML.

Los descriptores moleculares iniciales fueron Mw, PSA y ALOGP (3 descriptores para los compuestos/moléculas pequeñas de ChEMBL), y NMUnp, Lnp, Vnpu, Enpu, Pnpu, Uccoat, Uicoat, Hycoat AMRcoat, TPSA(NO)coat, TPSA(Tot)coat, ALOGPcoat, ALOGP2coat, SAtotcoat, SAaccoat, SAdoncoat, Vxcoat, VvdwMGcoat, VvdwZAZcoat y PDIcoat (20 descriptores para las nanopartículas). Se utilizaron las siguientes abreviaturas: Mw = peso molecular; PSA = Área de superficie polar; ALOGP = término logarítmico del coeficiente de partición octanol/agua; np = nanopartícula; npu = unidad elemental de la nanopartícula (Al, SiO₂, *etc.*); NMU = número de unidades monoméricas en la np; V = media del volumen atómico de Van der Waal para todos los átomos de la npu ($\langle V(\text{cm}^3/\text{mol}) \rangle$); E = electronegatividad; P(A3) = polarizabilidad atómica; L = np grande (datos experimentales); UC = nanopartículas sin recubrimiento; NMU = número de unidades monoméricas; HMT = Hexametilentetramina; TMAOH = Hidróxido de Tetrametilamonio; DMEM = medio de Eagle modificado por Dulbecco; coat = recubrimiento np; Uc = recuento de insaturación ; Ui = índice de insaturación; Hy = factor hidrofílico; AMR = refractividad molar de Ghose-Crippen; TPSA(NO) = área de superficie polar topológica utilizando contribuciones polares N,O; TPSA(Tot) = área de superficie polar topológica superficie polar topológica utilizando las contribuciones polares N,O,S,P; ALOGP2 = coeficiente de partición octanol/agua de Ghose-Crippen al cuadrado ($\log P^2$); SAtot = área superficial total de los descriptores tipo P_VSA; SAacc = área superficial de los átomos aceptores de los descriptores tipo P_VSA; SAdon = superficie de los átomos donantes de descriptores tipo P_VSA; Vx = volumen de McGowan; VvdwMG = volumen de van der Waals de McGowan volumen; VvdwZAZ = volumen de van der Waals de la ecuación de Zhao-Abraham-Zissimos; PDI = índice de densidad de empaquetamiento.

3.3.2.1. Preprocesamiento de datos en ChEMBL

Se han obtenido los resultados de varios ensayos preclínicos de ChEMBL. La medida experimental, $\epsilon_{ij}(d)$ utilizada para cuantificar la actividad biológica de la *i*-ésima molécula (m_i) sobre el *j*-ésimo objetivo, representa el resultado de cada ensayo. Los valores de $\epsilon_{ij}(d)$ dependen de la estructura del compuesto y también de ciertas condiciones límite que marcan las propiedades del ensayo $c_j(d) = (c_0(d), c_1(d), c_2(d), \dots, c_n(d))$. La primera $c_j(d)$ es $c_0(d)$ = la actividad biológica; se han utilizado fármacos con CC₅₀, EC₅₀ e IC₅₀. Las otras condiciones son $c_1(d)$ = organismo, $c_2(d)$ = nombre de la célula, c_3 = organismo del ensayo, $c_4(d)$ = cepa de ensayo, *etc.* (véase la Tabla 10). Se han utilizado técnicas de clasificación porque los valores $\epsilon_{ij}(d)$ no son números exactos en algunos casos. Además, se han discretizado los valores de esta manera: para el IC₅₀ y el EC₅₀ $f(v_{ij}(d))_{\text{obs}} = 1$ cuando $v_{ij} <$ valor de corte y la deseabilidad del

parámetro de actividad biológica $D(c_0(d)) = -1$ (véase la Tabla 11); para el CC_{50} , $f(v_{ij}(d))_{obs} = 1$ cuando $v_{ij} >$ valor de corte y la deseabilidad $D(c_0(d)) = 1$, si no $f(v_{ij}(d))_{obs} = 0$. La deseabilidad $D(c_0(d)) = 1$ o -1 denota que el parámetro medido disminuye o aumenta directamente con un efecto biológico, que puede ser deseado o no. Por último, se han calculado las desviaciones de cada condición para todos los fármacos/compuestos.

Tabla 10. Condiciones de ensayo ChEMBL (ejemplos seleccionados)

c₀ = Parámetro	n_j^a	c₀ = Parámetro	n_j
IC ₅₀ nM	30,981	IC ₅₀ ug·mL ⁻¹	4914
EC ₅₀ nM	10,337	CC ₅₀	11,629
c₁ = Organismo	n_j	c₁ = Organismo	n_j
<i>P. falciparum</i> D6	564	<i>P. falciparum</i> K1	6066
<i>P. falciparum</i>	33,463	<i>P. Yoelii</i> Yoelii	36
<i>P. berghei</i>	471	<i>P. cynomolgi</i>	15
c₂ = Nombre de la cédula	n_j	c₂ = Nombre de la cédula	n_j
Eritrocito	677	Huh-7	118
FM3A	14	L6	85
HeLa	19	MRC5	23
Hepatocito	28	Ovocito	5
HepG2	123	Vero	156
c₃ = Organismo de ensayo	n_j	c₃ = Organismo de ensayo	n_j
<i>P. falciparum</i>	31,587	<i>P. berghei</i> ANKA	6
<i>P. falciparum</i> D10	1147	<i>P. falciparum</i> 3D7	2606
<i>P. falciparum</i> FcB1/Columbia	330	<i>P. falciparum</i> NF54	938
<i>P. berghei</i>	461	<i>P. falciparum</i> FCR-3/Gambia	31
c₄ = Ensayo Strain	n_j	c₄ = Ensayo Strain	n_j
W2mef	39	W2	8591
NF54	929	TM91C235	474
W2/Indochina	31	VS1	25
W2-Mef	16	TM90C2B	68
c₅ = Curado por	n_j	c₅ = Curado por	n_j
Autocuración	38,150	Experto	2794
Intermediario	5288		

$c_6 = \text{Tipo de Ensayo}$	n_j	$c_4 = \text{Tipo de Ensayo}$	n_j
F	46,179	B	53

^a n_j indica el número de muestras para cada una de las condiciones

Tabla 11. Parámetros de actividad del compuesto (c_0).

$C_0 = \text{Actividad}$ (Unidades)	Parámetros Estadísticos ^a						Valor de corte
	<LogP>	<PSA>	n_0	n_1	p_1	d	
IC ₅₀ nM	4.128024	72.6311	30,981	8954	0.289	-1	100.0
EC ₅₀ nM	4.2390887	67.1602	10,337	1437	0.139	-1	100.0
IC ₅₀ ug·mL ⁻¹	4.0724379	75.0632	4914	4889	0.994	-1	325.0
CC ₅₀ nM	4.0650589	67.7534	11,629	11,608	0.998	1	100.0

^a Los parámetros <LogP> y <PSA> = Valor promedio de LogP y PSA para todos los fármacos m_i con valor reportado en el conjunto de datos ChEMBL. Esos parámetros son necesarios para el cálculo de Moving Average. Otros parámetros: n_0 = número de compuestos que muestran actividades diferentes, n_1 = número de compuestos considerado como positivo, $p_1 = n_1/n_0$ probabilidad de un compuesto considerado como positivo, $d = -1, 0, 1$ es la deseabilidad del parámetro, valor de corte = límite para el compuesto siendo tratado como activo o no.

3.3.2.2. Preprocesamiento de datos de nanopartículas

A partir de la literatura, se han recogido los resultados de muchas nanopartículas, y la medida ϵ_{ij} expresa el resultado de cada una de ellas. Los valores de $\epsilon_{ij}(\text{np})$ dependen de diferentes propiedades de la nanopartícula y también de algunas condiciones de contorno que delimitan las características del ensayo $c_j(\text{np}) = (c_0(\text{np}), c_1(\text{np}), c_2(\text{np}), \dots, c_n(\text{np}))$ (véase la Tabla 12). De nuevo, el primer $c_j(d) =$ la actividad biológica, y sólo se han utilizado nanopartículas con CC₅₀, EC₅₀ e IC₅₀, para que pudieran coincidir con las actividades biológicas de los fármacos/compuestos. Otras condiciones son $c_1(\text{np}) =$ nombre de la célula, $c_2(\text{np}) =$ forma de la nanopartícula $c_3(\text{np}) =$ medio de la nanopartícula y $c_5 =$ recubrimiento de la superficie.

Además, se han discretizado los valores de la misma manera que se hizo con los fármacos/compuestos (véase la Tabla 13. Al final, determinamos las desviaciones de cada c_j para todas las nanopartículas.

Tabla 12. Condiciones de ensayo de las nanopartículas decoradas (ejemplos seleccionados)

c₀ = Parámetro	n_j^a	c₀ = Parámetro	n_j
IC ₅₀ nM	29	CC ₅₀	113
EC ₅₀ nM	30		
c₁ = Línea celular	n_j	c₁ = Línea celular	n_j
A549(H)	23	BRL 3A (R)	4
Lycopersicon Esculentum	16	3T3 (M)	9
HepG2 (H)	15	CaCo-2 (H)	6
c₂ = Forma	n_j	c₂ = Forma	n_j
Esférica	61	Elíptica	21
Irregular	3	Pseudo-esférica	8
Porción	3	Poliédrica	3
Afilado	2	Piramidal	10
Vara	9		
c₃ = Ensayo Medio	n_j	c₃ = Ensayo Medio	n_j
Seco	118	RPMI	3
H ₂ O	44	1% Tritón X-100/H ₂ O	3
DMEM	3	H ₂ O/TMAOH	1
c₄ = Revestimiento	n_j	c₄ = Revestimiento	n_j
UC	125	11-ácido mercaptoundecanoico	3
PEG-Si (OMe) ₃	8	PVP	4
PVA	1	Fragmento de propilamonio	4
Citrato de sodio	17	Fragmento de undecilazida	2

^a n_j = el número de muestras para cada una de las condiciones; IC₅₀ = Mitad de la concentración inhibitoria máxima; EC₅₀ = Concentración de un fármaco que da la mitad de la respuesta máxima; CC₅₀ = Relación del 50% de la concentración citotóxica; A549 (H) = Célula de carcinoma de pulmón; HepG2 (H) = Células de cáncer de hígado humano; BRL 3A (R) = Células de hígado de rata de Búfalo; 3T3 (M) = Células de fibroblastos; CaCo-2 (H) = Células de carcinoma de colon humano; DMEM = Medio de águila modificado de Dulbecco; RPMI = Medio del Roswell Park Memorial Institute; TMAOH = Hidróxido de tetrametiamonio; UC = Sin recubrimiento; PEG-Si(OMe)₃ = Poli(etilenglicol) trimetoxisilil; PVP = Polivinilpirrolidona; PVA = alcohol polivinílico.

Tabla 13. Parámetros de actividad de la Nanopartícula (c_0)

C_0 =Actividad (Unidades)	Parámetros Estadísticos ^a						Valor de corte
	<LogP>	<PSA>	n_0	n_1	p_1	d	
EC ₅₀ uM	1.66	18.02	30	27	0.9	-1	25,422
IC ₅₀ uM	3.24	38.79	29	21	0.7241	-1	18,714
CC ₅₀ uM	1.63	24.97	113	21	0.1858	1	3099

^a Parámetros <LogP> y <PSA> = Valor promedio de LogP y PSA para todas las nanopartículas m_i . Estos parámetros son necesarios para el cálculo de moving average. Otros parámetros: n_0 = número de nanopartículas decoradas que mostraron cada actividad diferente, n_1 = número de nanopartículas consideradas como positivas, $p_1 = n_1/n_0$ probabilidad de que una nanopartícula se considere positiva, $d = -1, 0, 1$ es la deseabilidad del parámetro, valor de corte = límite para las DNP que se tratan.

3.3.2.3. Pre-procesamiento de los datos combinados

Una vez creadas ambas bases de datos, se combinan haciendo pares con las mismas condiciones experimentales, por ejemplo, un CC₅₀ con una nanopartícula CC₅₀.

Además, se utiliza el mismo método para discretizar cada par formado. Así, se utilizó un conjunto de datos de 107 características de entrada y 249.992 ejemplos para construir modelos de clasificación ML. Los casos de control positivo (1) y negativo (0) se asignaron como sigue: si la función de deseabilidad $d(c_0) = -1$, entonces $c_{ij} = 1$ cuando $\epsilon_{ij} < 100$ nM o $\epsilon_{ij} < \text{average } \langle \epsilon_{ij} \rangle$ para propiedades que no se miden en nM. Además, si $d(c_0) = 1, 0$, entonces $c_{ij} = 1$ cuando $\epsilon_{ij} > \text{valor average } \langle \epsilon_{ij} \rangle$. Una característica de entrada (prob = probabilidad) se creó como la probabilidad de c_0 para los pares compuesto-nanopartícula (recuento del número de pares compuesto-nanopartícula para cada tipo de actividad c_0 /número total de pares). El nombre de las características finales en el conjunto de datos tiene el formato [d_/np_] [descriptor original nombre] ([condición experimental]). Por ejemplo

- d_DP_{SA}(c_2) = diferencia (D) entre los valores originales del descriptor PSA y la media de los valores PSA en la condición experimental c_2 (para los fármacos/compuestos, d_);
- np_DL_{np}(c_4) = diferencia entre el valor de L_{np} y la media de los valores de L_{np} en la condición experimental c_4 (para nanopartículas, np_).

3.3.2.4. Métodos de ML

El estudio se realiza utilizando ocho clasificadores de ML “scikit-learn” para encontrar el mejor clasificador capaz de predecir la probabilidad de que un par nanopartícula-compuesto exprese altamente su actividad antimalárica:

1. K neighbors Classifier = KNN-k-vecinos más cercanos: uno de los clasificadores no paramétricos más conocidos en el campo del ML. Asigna una muestra no clasificada a la misma clase que la más cercana de las k del conjunto de entrenamiento (Cover & Hart, 1967);

2. SVC (lineal) = SVM-Clasificador de vectores de soporte lineal con núcleos lineales: los datos de entrada se mapean de forma no lineal en un espacio de mayor dimensionalidad, donde se puede establecer una superficie de decisión lineal (Hao & Ho, 2019).

3. SVC = SVC-Clasificador de vectores de soporte con núcleos RBF no lineales: los problemas reales tienden a soluciones no lineales, y SVM puede manejar esta limitación utilizando funciones de núcleo no lineales como la base radial gaussiana (RBF) (Patle & Chouhan, 2013);

4. Logistic Regression = LR-Regresión logística (Peduzzi *et al.*, 1996) es un modelo lineal que puede estimar la probabilidad de una respuesta binaria utilizando diferentes factores;

5. Linear Discriminant Analysis = LDA-Análisis discriminante lineal (Cristianini, 2004): es un método estadístico supervisado que se basa en la proyección de los datos a una dimensión inferior para maximizar la dispersión entre clases frente a la dispersión dentro de cada clase. Esta proyección facilita la separación de los datos;

6. Decision Tree Classifier = DT-Arbol de decisión utiliza una serie de reglas de decisión inferidas de las características como un árbol de reglas. Así, los caminos de la raíz a la hoja representan reglas de clasificación (Swain & Hauska, 1977);

7. Random Forest Classifier = RF-Random forest (Breiman, 2001) es un método de conjunto que agrega varios árboles de decisión (árboles paralelos). Cada árbol se genera utilizando una muestra “bootstrap” extraída aleatoriamente del conjunto de datos original utilizando un método de árbol de clasificación o regresión (CART) y la disminución de la Impureza de Gini (DGI) como criterio de división (Calle *et al.*, 2011). La RF se caracteriza por un bajo sesgo y una baja correlación entre los árboles individuales, y una alta varianza;

8. XGB Classifier = XGB-XGBoost un método de conjunto basado en árboles en el que se añaden clasificadores débiles para corregir errores (árboles secuenciales (Friedman, 2001)). Este clasificador ha demostrado un excelente rendimiento a través de los proyectos del concurso Kaggle (Chen *et al.*, 2016).

3.3.2.5. Flujo de trabajo de ML

Las características fueron estandarizadas mediante la eliminación de la media y el escalado a la unidad de varianza, utilizando el escalador estándar en scikit-learn. Se realizó una validación cruzada estratificada de 10 veces, conservando los porcentajes de muestras para cada clase. Como las muestras del conjunto de datos no estaban equilibradas, se calcularon los pesos de cada clase utilizando $N/(k \cdot n_i)$, donde N es el número total de muestras, k el número de clases y n_i el número de muestras que pertenecen a la clase i . Esto da como resultado pesos de 0,63778 para la clase 1 y 2,31448 para la clase 2. El rendimiento del modelo se midió mediante el área bajo el receptor de características operativas (AUC).

Dados los resultados obtenidos en la línea de base, el flujo de trabajo ha continuado sólo con el mejor modelo. A partir de este punto, se realizó una selección de características utilizando la disminución de la impureza media, que ya está implementado en “sklearn”. Esta métrica se calcula utilizando las disminuciones de impureza de gini ponderadas para todos los nodos, promediados en todos los árboles (Calle *et al.*, 2011). Por lo tanto, se realizó una selección de características utilizando ExtraTreesClassifier (Geurts *et al.*, 2006) con $n_estimadores = 100$, pesos de clase y CV de 10 veces (ver Feature-Selection.ipynb (D-Bcarrue, 2019)). Se elige este método basado en árboles para seleccionar las características más importantes porque los árboles adicionales (a veces denominados árboles aleatorios) ofrecen un mayor rendimiento en presencia de características ruidosas (Moore, 1987). El algoritmo de selección de características personalizado que se propone mantiene al menos una característica para cada condición experimental de los fármacos/compuestos y nanopartículas, y la característica de probabilidad (si el selector automático las elimina).

Los modelos lineales PTML más sencillos serán los primeros clasificadores que se probarán para conjuntos de datos complejos con múltiples características de BD (González-Díaz *et al.*, 2014; González-Díaz *et al.*, 2013). Se pueden aproximar los valores de la función $f(v_{ij}(d)$ y $v_{ij}(np))_{calc}$ para el i -ésimo par fármaco-nanopartícula en el j -ésimo ensayo preclínico con múltiples condiciones de ensayo c_j . Como entrada se utilizan operadores PT que también pueden ser operadores de Moving Average (MA) de Box-Jenkins (Casañola-Martin *et al.*, 2015; Tenorio-Borroto *et al.*, 2014). Los modelos lineales PTML tienen la siguiente forma genérica:

$$f(v_{ij}(d), v_{ij}(np))_{calc} = a_0 + a_1 \cdot f(v_{ij}(d), v_{ij}(np))_{expt} + \sum_{k=1, j=0}^{k_{max}, j_{max}} a_{kj} \cdot \Delta D_k(d_j, c_j) + \sum_{k=1, j=0}^{k_{max}, j_{max}} b_{kj} \cdot \Delta D_k(np_j, (c_j))$$

(2)

Se han proporcionado resultados adicionales para explicar las predicciones con el mejor modelo utilizando valores de Shapley (Roth, 1988) (SHAP_test.ipynb). Todos los scripts para

los AUC de referencia, la selección de características y el modelo están disponibles en un repositorio abierto en <https://github.com/d-bcarrue/NanoDrugsMalaria> (D-Bcarrue, 2019).

3.3.3. Resultados y discusión

En el presente trabajo, se ha creado un modelo PTML para predecir la actividad de los compuestos orgánicos ensamblados de algunas nanopartículas utilizadas contra la enfermedad de la malaria. Al hacerlo, se amplía la idea de análisis de Hansch y se buscan modelos con aplicaciones a la nanomedicina. Como prueba de concepto se investiga sobre un enorme conjunto de datos de fármacos descargados de ChEMBL, y otro conjunto de datos de nanopartículas. Estos conjuntos de datos contienen (véase el material y los métodos) los resultados de muchos ensayos farmacológicos experimentales.

El modelo supone que los cambios en la unión entre el fármaco y la nanopartícula se producen gracias a las perturbaciones en las condiciones de entrada de las nanopartículas y los fármacos. Se ha centrado la acción únicamente en un pseudo-constante de unión nanopartícula-fármaco-compuesto ($v_{ij}(d), v_{ij}(np)$), definida por nosotros, para cuantificar la probabilidad de que un par nanopartícula-fármaco-compuesto tenga una alta expresión de actividad contra la malaria.

Este modelo PTML comienza con un valor de referencia, $f((v_{ij}(d), v_{ij}(np))_{\text{expt}})$ y luego añade los efectos de las perturbaciones en la estructura del compuesto o en las condiciones del ensayo, y las propiedades de la nanopartícula y su recubrimiento. Otros términos de entrada utilizados aquí son los términos de perturbación ΔLogP y ΔPSA , que son similares a las funciones de Moving Average (MA) utilizadas en los modelos Box-Jenkins en series temporales (Box y Jenkins, 1970). Ejemplos de MA son las desviaciones de PSA y logP de compuestos/fármacos y nanopartículas con respecto a los valores esperados de estos parámetros para los ensayos en las mismas condiciones c_j . Por ejemplo, $\Delta\text{LogP} = \text{LogP}(m_i) - \text{LogP}(c_j)$, donde $\text{LogP}(c_j)$ es el average de $\text{LogP}(m_i)$ para todas las moléculas, m_i , en el mismo ensayo con un conjunto de condiciones c_j .

Utilizando ocho clasificadores ML, se han calculado los valores AUC (10 veces de CV). Los resultados son presentados en la Tabla 14. El mejor modelo se obtuvo con RF, y el AUC es de $0,9844 \pm 0,0007$. La Figura 27 representa el diagrama de caja para los valores AUC de referencia de los métodos ML (10 veces de CV). Los valores AUC para las 10 divisiones tienen rangos cortos, especialmente RF. Esto sugiere que los AUC de todos los métodos ML son estables dentro de cada pliegue. Además, la elevada diferencia entre el RF y los demás métodos (los gráficos de caja están lejos de superponerse) demostraron que es estadísticamente significativa.

Tabla 14. Área bajo las características operativas del receptor (AUC) para los modelos de clasificación de referencia.

Método ML	Clasificador	Media AUC + sd
KNN	Clasificador aproximados K	0.8994±0.0022
SVM Lineal	SVC (lineal)	0.8949±0.0019
SVM	SVC (rbf)	0.9223±0.007
LR	Regresión Lógica	0.8946±0.0013
LDA	Análisis de discriminación lineal	0.8939±0.0015
DT	Clasificador Árbol de Decisión	0.9277±0.0021
RF	Bosques Aleatorios	0.9844±0.007
XGB	Clasificador XGB	0.9242±0.0017

KNN = vecinos más cercanos; SVM lineal = clasificador de vectores de apoyo con núcleos lineales; SVM = clasificador de vectores de apoyo con núcleos no lineales; LR = regresión logística; LDA = análisis discriminante lineal; DT = árbol de decisión; RF = bosque aleatorio; XGB = XGBoost.

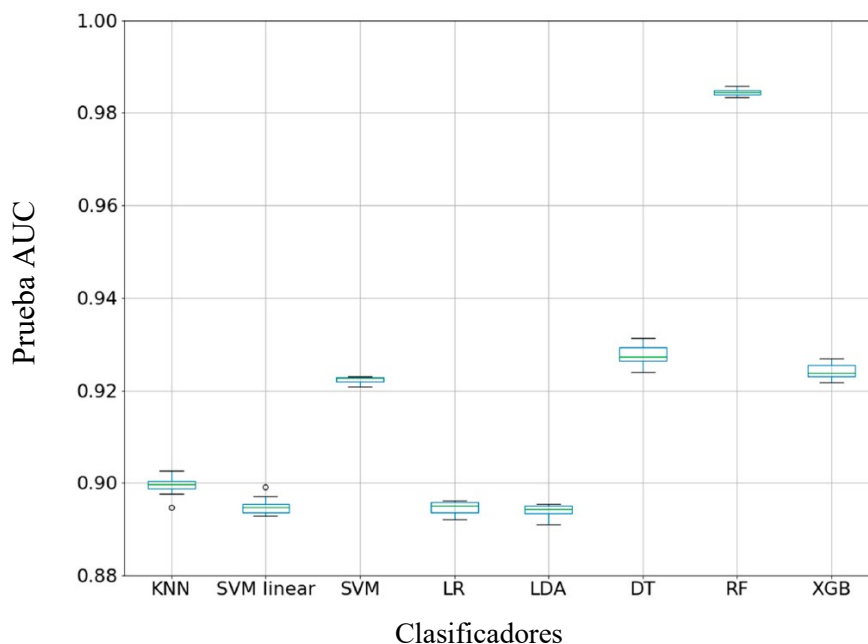


Figura 27. Diagrama de caja de los valores AUC de los clasificadores ML (CV de 10 veces).

En el siguiente paso, se ha reducido el número de características para mejorar el AUC del modelo RF.

Para ello, se realizó una selección de características utilizando ExtraTreesClassifier. Las 27 características se seleccionaron de una cantidad inicial de 107:

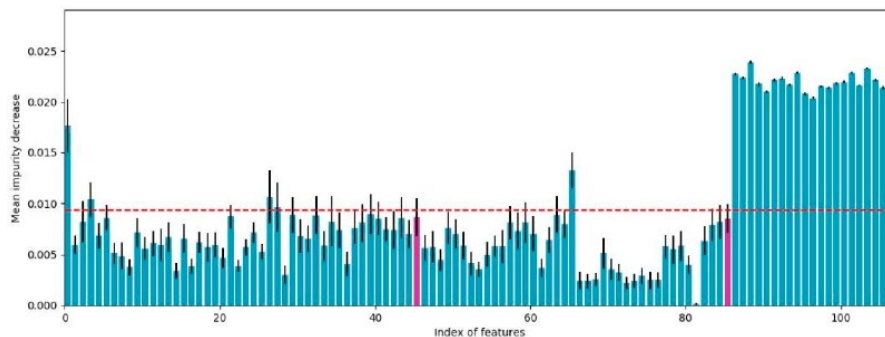
- Una característica del par np-compuesto: prob;
- 5 características np que utilizan 5 condiciones experimentales (c_0 - c_4): np_DVnpu(c_0), np_DUccoat(c_1), np_DVnpu(c_2), np_DPnpu(c_3), np_DPnpu(c_4);
- 21 características de fármacos/compuestos utilizando 7 condiciones experimentales (c_0 - c_6): d_DMw(c_0), d_DALOGP(c_0), d_DPSA(c_0), d_DMw(c_1), d_DALOGP(c_1), d_DPSA(c_1), d_DMw(c_2), d_DALOGP(c_2), d_DPSA(c_2), d_DMw(c_3), d_DALOGP(c_3), d_DPSA(c_3), d_DMw(c_4), d_DALOGP(c_4), d_DPSA(c_4), d_DMw(c_5), d_DALOGP(c_5), d_DPSA(c_5), d_DMw(c_6), d_DALOGP(c_6) y d_DPSA(c_6).

Notablemente, este es el primer modelo que combina tanto la Teoría de Perturbación como los MAs en un estudio QSBR de pares relevantes de nanopartículas-fármaco-compuesto utilizados como sistema de administración antimalárica. Se han determinado las perturbaciones más relevantes bajo diferentes condiciones experimentales, c_j , relacionadas con la propiedad antimalárica utilizando un RF. Casualmente, en este modelo la mayoría de los operadores utilizados son de tipo PSA y Δ LOGP. Por lo tanto, sólo miden las perturbaciones en el valor de Δ LOGP con respecto a otros subconjuntos, c_j , de fármacos y nanopartículas. El Δ LOGP es un parámetro relevante utilizado en la química medicinal porque está relacionado con la lipofilia y la capacidad de atravesar las membranas biológicas.

La Figura 28 muestra la reducción media de impurezas para cada una de las características tanto en el orden original como el clasificado. Esta reducción media de impurezas se obtuvo utilizando un clasificador Extra Trees con 100 árboles y clases ponderadas, y este modelo se aplicó en un CV estratificado de 10 veces. La línea horizontal discontinua indica el umbral utilizado como filtro de selección. Tras comprobar que la característica de probabilidad está presente, ya que esto es estrictamente necesario en la Teoría de la Perturbación, se comprueba que todas las condiciones experimentales se reflejan en el subconjunto seleccionado. Tras el filtrado, las condiciones experimentales, c_2 y c_4 , de las nanopartículas no estaban incluidas. Por lo tanto, se selecciona la característica con la mayor disminución media de impurezas para ambas condiciones experimentales, y se añade a la selección anterior, marcado en rosa. El gráfico sin clasificar presenta las características en el eje x en el orden en que se presentaron en el conjunto de datos. Para una mejor comparación de

las impurezas medias de las características seleccionadas (las que están por encima del valor de corte), también se presentó la versión ordenada de la trama.

(A) Sin clasificar



(B) Clasificado

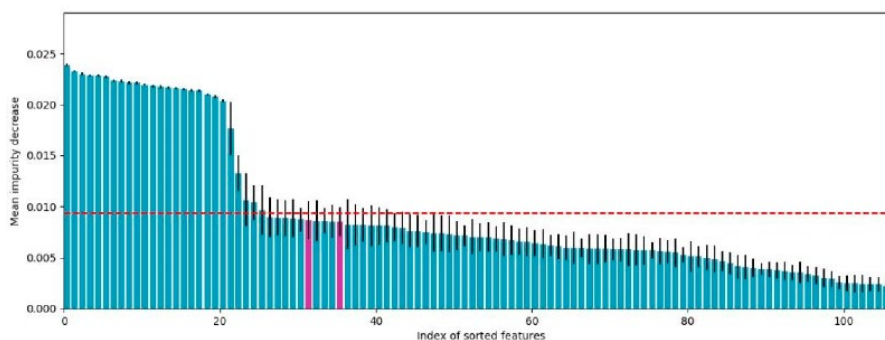


Figura 28. Selección de características mediante ExtraTrees: disminución de la impureza promedio por característica, en orden original y en orden descendente; las barras rosas representan las características reintroducidas tras el filtrado inicial.

Por lo tanto, con sólo 27 características seleccionadas (de las 107 iniciales), el AUC medio de la prueba para el clasificador RF aumentó a $0,9921 \pm 0,000244$ (de $0,9844 \pm 0,0007$). Este modelo muestra un muy buen rendimiento para un modelo PTML.

La selección de características mostró que el clasificador prefiere las perturbaciones (MAs) del término logarítmico del coeficiente de partición octanol/agua (ΔLOGP), el área superficial polar (PSA) y peso molecular (Mw) de los compuestos/fármacos en todas las condiciones experimentales, como el tipo de actividad (c_0), organismo (c_1), nombre de la célula (c_2), organismo de ensayo (c_3), cepa de ensayo (c_4), tipo de curación (c_5) y tipo de ensayo (c_6). En el caso de las nanopartículas, el modelo seleccionó las perturbaciones del promedio de volumen atómico de Van der Waal para todos los átomos de la np (V_{npu}) con el tipo de actividad (c_0) y con la forma (c_2), el recuento de insaturaciones (U_{coat}) en la línea celular (c_1),

la polarizabilidad atómica (P_{npu}) con el medio de ensayo (c_3) y el recubrimiento de la superficie (c_4). Así, se puede concluir que las perturbaciones de los siguientes descriptores moleculares bajo diferentes condiciones experimentales son importantes para las nanopartículas decoradas con fármacos antimaláricos: la polaridad tanto de los componentes/fármacos como de las nanopartículas, la masa de los compuestos/fármacos el volumen, la forma y la insaturación del recubrimiento de las nanopartículas.

Un modelo lineal es fácil de interpretar, pero no siempre es el más preciso. Por ello, los modelos complejos utilizan diferentes herramientas para evitar un modelo de "caja negra". Valores de Shapley y SHAP (explicaciones aditivas de Shapley) son la solución propuesta para el mejor modelo de RF. Los valores Shapley representan el promedio de la contribución marginal en todas las permutaciones, un método para cuantificar la contribución de las características al modelo final. Así, el método SHAP es capaz de explicar el resultado de un modelo de ML mediante:

- la interpretabilidad global: cuánto contribuye cada característica, positiva o negativamente, a la variable de salida;
- interpretabilidad local: cada caso/instancia recibe sus propios valores SHAP para explicar por qué un caso tiene una predicción específica, y la contribución de las características a esta instancia.

La interpretabilidad global se presenta mediante la correlación de las características con la variable de salida o el impacto positivo/negativo mediante los valores SHAP (Figura 29). El impacto promedio ordenado de las características en la salida del modelo para cada clase, y la interpretabilidad local para diferentes instancias/casos, se incluyen en el repositorio de GitHub con el nuevo script (SHAP_test.ipynb).

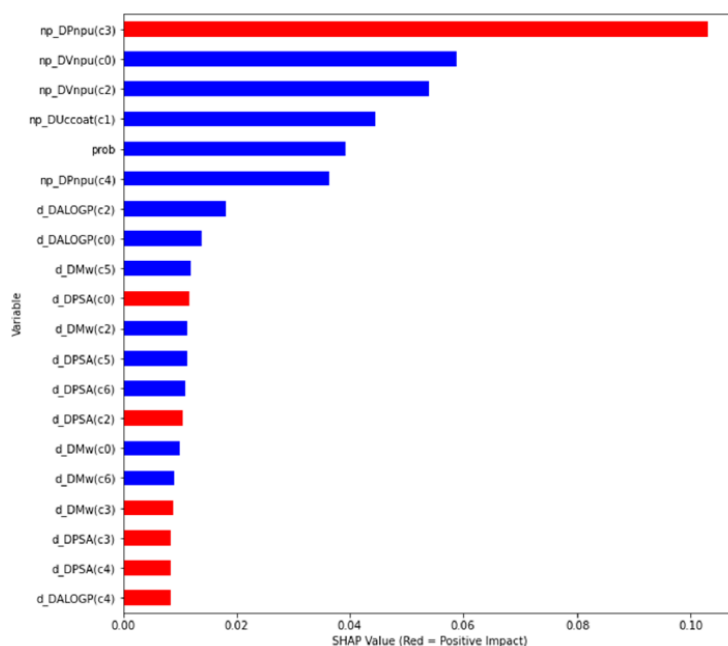


Figura 29. Impacto de la característica en la variable de salida para el mejor modelo basado en los valores medios de SHAP.

Esta figura presenta:

- La importancia de las características utilizando valores SHAP: las variables se clasifican en orden descendente.
- El impacto en el valor de la predicción utilizando los valores SHAP en el eje x;
- El color muestra si esa variable tiene un impacto positivo (en rojo) o negativo (en azul) en la variable de salida.

Así, se puede observar que la perturbación de los descriptores moleculares de las nanopartículas en condiciones experimentales tiene un gran impacto en la predicción del modelo para los fármacos/compuestos antipalúdicos portadores. Entre ellos se encuentran la perturbación de la polarizabilidad atómica (Pnp), la media del volumen atómico de Van der Waal para todos los átomos (Vnp) y la insaturación del recubrimiento (Uccoat). Para los compuestos/fármacos, el Δ LOGP tiene más impacto que el peso de la masa y el PSA. Así, se puede afirmar que las propiedades moleculares vinculadas a la polaridad son las que tienen mayor impacto en los portadores de fármacos/compuestos-nanopartículas contra la malaria.

La polarizabilidad atómica de las nanopartículas tiene un impacto más positivo en el resultado del modelo, y el volumen de las nanopartículas sólo tiene un impacto negativo: los portadores óptimos del fármaco-np contra la malaria deben ser nanopartículas con alta polarizabilidad atómica, pero de pequeño volumen. Además, los compuestos/fármacos deben tener áreas superficiales polares (PSA) más altas, con un impacto positivo, y una masa de peso más pequeña con un impacto negativo, en el resultado del modelo.

3.3.4. Conclusiones

Combinando las ideas de la Teoría de la Perturbación con el análisis QSAR de Hansch y la fusión de información, se ha desarrollado un modelo multiobjetivo PTMLIF que resulta útil para clasificar los fármacos en función de su constante a muchas nanopartículas diferentes y su capacidad para actuar contra el *Plasmodium*, que es la causa de la malaria en los seres humanos. Este modelo puede ayudar a ahorrar recursos experimentales y tiempo, ya que permite determinar qué nanopartículas decoradas con fármacos serían útiles y cuáles no. De este modo, se pueden probar sólo aquellas con mayor probabilidad de ser activas.

Las características transformadas de los fármacos y las nanopartículas se han utilizado como entrada para ocho métodos de ML. El mejor modelo de clasificación se ha obtenido utilizando Random Forest con sólo 27 características seleccionadas de fármacos y nanopartículas en todas las condiciones experimentales consideradas. La media del AUC de la prueba fue de $0,9921 \pm 0,000244$ (10 veces de CV). El rendimiento del modelo RF demostró el poder de la fusión de información de las características experimentales de los fármacos y las nanopartículas para la predicción de la probabilidad, relacionada con la actividad antimalárica de la nanopartícula-fármaco/compuesto. Todos los cálculos pueden reproducirse utilizando los scripts y el conjunto de datos incluidos en un repositorio abierto de GitHub en <https://github.com/d-bcarrue/NanoDrugsMalaria>.

3.4. Capítulo 4. Mapeo IFPTML de gráficos de fármacos con red estructural de proteínas y cromosomas frente a información de ensayos preclínicos para el descubrimiento de compuestos antipalúdicos

Artículo 4

Ref. Quevedo-Tumaili, V.; Ortega-Tenezaca, B.; González-Díaz, H. (2021). IFPTML Mapping of Drug Graphs with Protein and Chromosome Structural Networks vs. Pre-Clinical Assay Information for Discovery of Antimalarial Compounds. *International Journal of Molecular Sciences*. 22: 13066. doi.org/10.3390/ijms222313066

3.4.1. Introducción

La malaria es un importante problema de salud mundial con la mayoría de los casos reportados en diferentes regiones. En la actualidad, las áreas de riesgo para contraer esta enfermedad son África, América Central y del Sur, así como en algunas partes del Caribe, en Asia, Europa del Este y el Pacífico Sur. La Organización Mundial de la Salud (OMS) estimó 219 millones de casos de malaria notificados en todo el mundo en 2017. Es una infección de los glóbulos rojos por parásitos del género *Plasmodium* con las formas más graves y comunes causadas por *Plasmodium falciparum* (*P. falciparum* o *Pf*) y especies relacionadas como *Plasmodium vivax* (*P. vivax* o *Pv*), *Plasmodium malariae* (*P. malariae* o *Pm*) y *Plasmodium ovale* (*P. ovale* o *Po*). El más frecuente y mortal es el *Pf*. Según la OMS, la malaria durante el embarazo puede causar complicaciones importantes. La resistencia emergente del parásito a los medicamentos antipalúdicos disponibles plantea grandes desafíos para el tratamiento. Además, los costos han aumentado significativamente en los últimos años para determinar y desarrollar un nuevo medicamento, según el Centro de Tufts para el Estudio del Desarrollo de Medicamentos se estima el costo de su bolsillo en medicamentos aprobados por pares en \$ 1861 millones para medicamentos antipalúdicos (Alonso & Noor, 2017; Kalanon & McFadden, 2010; Gaillard *et al.*, 2018; DiMasi *et al.*, 2016).

En realidad, la base de datos ChEMBL enlista > 17750 ensayos preclínicos de compuestos antipalúdicos. El conjunto de datos de ChEMBL sobre compuestos antipalúdicos cubre múltiples parámetros de actividad biológica (inhibición, IC₅₀, actividad, *etc.*), diferentes ensayos únicos solo para la “proteína target” del organismo *Pf*, aplicado a diferentes genes sobre el proteoma. Además, la base de datos ChEMBL compila conjuntos de datos de ensayos preclínicos muy heterogéneos. Se pueden enriquecer los datos de ChEMBL con los datos de las bases de datos NCBI-GDV y UniProt para obtener información sobre las proteínas, los

cromosomas y los genes objeto de los medicamentos. Por ejemplo, UniProt incluye información relacionada con la secuencia de proteínas. Por último, NCBI-GDV incluye información relacionada con la secuencia de genes y la estructura del cromosoma (secuencia de ADN, adyacencia de genes, orientación, *etc.*). Esta información también puede ser relevante para la síntesis de proteínas con diferentes funciones en el *Pf* (Gaulton *et al.*, 2017; Gaulton *et al.*, 2012; Wolfsberg, 2010; Coordinators, 2018; UniProt Consortium, 2019; UniProt Consortium, 2018; Pundir *et al.*, 2017).

Además, los modelos IFPTML se han utilizado en la química médica, proteómica, nanotecnología, *etc.*, para modelar grandes conjuntos de datos con características de Big Data. Los modelos IFPTML (IF + PT + ML) combinan técnicas de IF con ideas de la PT y desarrollos ML. El modelado IFPTML también es útil para llevar a cabo la fusión de información de datos de diversas fuentes. Por ejemplo, Se pueden incluir datos sobre la secuencia de proteínas del GenBank, redes metabólicas, nanopartículas, o incluso información sobre datos de epidemiología en los condados de Estados Unidos, *etc.* (Gonzalez-Diaz *et al.*, 2013; Santana *et al.*, 2019; Nocedo-Mena *et al.*, 2019).

Para desarrollar modelos IFPTML, se necesita utilizar como variables de entrada parámetros capaces de cuantificar la información sobre las condiciones estructurales y experimentales de ensayo de todos los sistemas implicados (fármacos, proteínas, redes de genes, *etc.*). En este sentido, las medidas de información de la Entropía de Shannon introducidas por Claude E. Shannon podrían ser extremadamente útiles (Shannon, 1948). De hecho, Graham, Marrero-Ponce, Barigye y otros investigadores, han utilizado diferentes clases de valores de información de Shannon para medir cuantitativamente la información química y, o, biológicamente relevante (Graham, 2002; Graham *et al.*, 2004; Graham & Schacht, 2000; Graham & Schulmerich, 2004; Graham, 2005; Graham, 2007; Contreras-Torres *et al.*, 2019; Martinez-Lopez *et al.*, 2019; Valdes-Martini *et al.*, 2017; Ruiz-Blanco *et al.*, 2015; Ruiz-Blanco *et al.*, 2013; Barigye *et al.*, 2013). González-Díaz y Munteanu combinaron la idea de la entropía de Shannon con las cadenas de Markov para calcular los valores $Sh(\text{sys})_k$, las Entropías de Shannon estocásticas de orden k^{th} , y diferentes sistemas moleculares (Munteanu *et al.*, 2008).

En trabajos anteriores, se ha analizado el proteoma/genoma y los cromosomas de *Pf* utilizando datos de las bases de datos NCBI-GDV y UniProt (Quevedo-Tumaili *et al.*, 2018). Sin embargo, este trabajo previo no ha considerado la posibilidad de “mapear” estos datos frente a los ensayos preclínicos de los compuestos para avanzar en el diseño de nuevos antimaláricos. Además, no hay informes de modelos IFPTML para compuestos antimaláricos que consideren la información de las bases de datos NCBI-GDV, UniProt y ChEMBL al mismo tiempo. En

este trabajo, se ha desarrollado un modelo IFPTML de propósito general para la predicción de nuevos compuestos antimaláricos fusionando la información de las tres bases de datos diferentes. La Figura 30 ilustra los diferentes pasos que se incluyen en el flujo de trabajo general utilizado para obtener este modelo IFPTML.

En primer lugar, se descarga toda la información relevante de las bases de datos ChEMBL, NCBI-DVG y UniProt. Estos tres conjuntos de datos se fusionaron en uno solo utilizando técnicas de IF. Este nuevo conjunto de datos se limpió y pre-procesó aplicando diferentes criterios; por ejemplo, eliminando los ensayos preclínicos que no registran valores en las actividades biológicas.

A continuación, se calculó el $Sh(\text{syst})_k$ de los diferentes subsistemas implicados, como, por ejemplo, los fármacos, las secuencias de proteínas, los genes y los cromosomas, utilizando modelos de cadenas de Markov. Después, se utilizaron OTPs con forma de MAs para cuantificar las desviaciones en los parámetros estructurales $Sh(\text{syst})_k$ (parámetros numéricos) respecto a los cambios en las condiciones experimentales (variables categóricas). Esto permitió cuantificar en PTOs simples la información de la estructura y las condiciones experimentales de los ensayos de todos los subsistemas involucrados. Por último, se han entrenado, validado y comparado los modelos IFPTML.

También se discutió el papel de las diferentes fuentes de información. Este tipo de análisis abre una nueva vía para llevar a cabo la IF combinada con el modelado ML hacia el descubrimiento de nuevos compuestos antimaláricos utilizando ensayos preclínicos e información del proteoma.

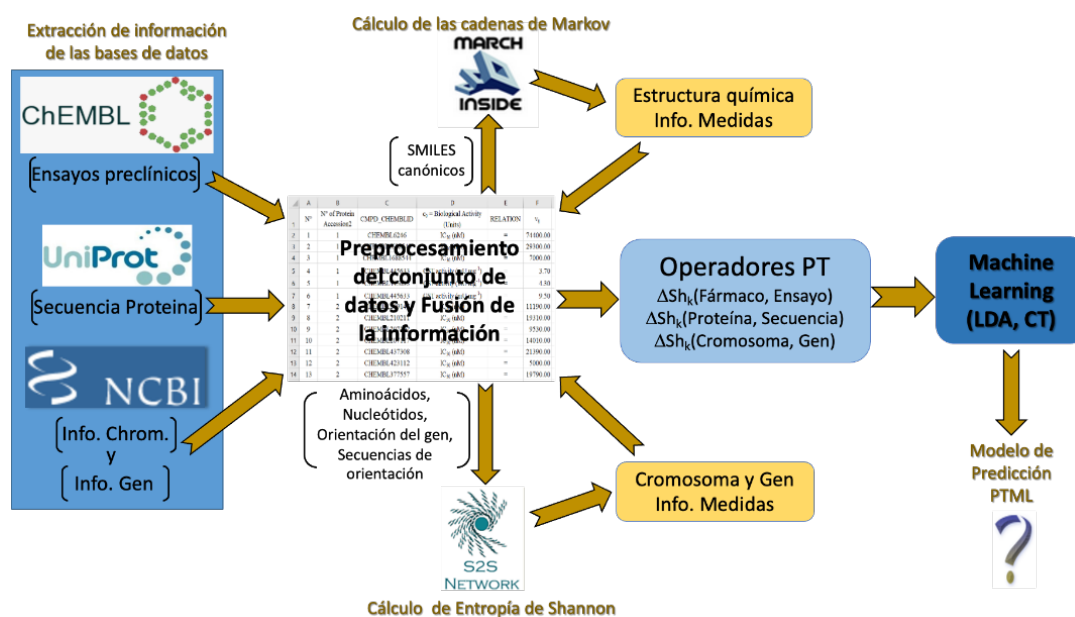


Figura 30. Flujo de trabajo general de los pasos dados en este trabajo.

3.4.2. Resultados

Se han desarrollado varios modelos IFPTML utilizando operadores PTOs y MMAs (Nocedo-Mena *et al.*, 2019). El modelo calculó la función de puntuación $f(v_{ij})_{\text{calc}}$ para el resultado del i -ésimo fármaco frente a la j -ésima proteína en las condiciones múltiples del ensayo preclínico definidas por las variables categóricas c_j . El primer modelo desarrollado fue el modelo lineal IFPTML-GDA. La ecuación (1) de este modelo es la siguiente:

$$\begin{aligned}
 f(v_{ij})_{\text{calc}} = & -20.12298 + 99.13885 \cdot f(v_{ij})_{\text{ref}} \\
 & +0.74880 \cdot \Delta\text{Sh}(\text{Drug}; \text{Csat})_{5c_{aj}} \\
 & -2.20919 \cdot \Delta\text{Sh}(\text{Drug}; \text{Hetero})_{5c_{aj}} \\
 & +3.36764 \cdot \Delta\text{Sh}(\text{Drug}; \text{Hx})_{1c_{aj}} \\
 & +2.39122 \cdot \Delta\text{Sh}(\text{Drug}; \text{Csat})_{1c_{pj}} \\
 & +2.25745 \cdot \Delta\text{Sh}(\text{Drug}; \text{Hetero})_{4c_{pj}} \\
 & -3.32408 \cdot \Delta\text{Sh}(\text{Drug}; \text{Hx})_{4c_{pj}} \\
 & -2.88041 \cdot \Delta\text{Sh}(\text{Drug}; \text{Csat})_{1c_{dj}} \\
 & +6.57931 \cdot \Delta\text{Sh}(\text{Drug}; \text{Halog})_{1c_{dj}} \\
 & -6.84622 \cdot \Delta\text{Sh}(\text{Drug}; \text{Halog})_{2c_{dj}} \\
 & -0.00877 \cdot \Delta\text{Sh}(\text{Chr}; \text{Gen})_{5c_{aj}} \\
 & +0.46021 \cdot \Delta\text{Sh}(\text{Prot}; \text{Seq})_{5c_{dj}}
 \end{aligned} \tag{1}$$

$$n = 17758\chi^2 = 6595.853 \quad p < 0.05$$

Las variables de este modelo IFPTML son el resultado de varios procedimientos de pre-procesamiento y post-procesamiento (después de obtener el modelo) de las variables de entrada/salida. Para la entrada, la salida del modelo es la función de puntuación $f(v_{ij})_{\text{calc}}$. Se trata de una función de valor real que sirve para cuantificar las posibilidades de que el i -ésimo fármaco dé un resultado positivo en el j -ésimo ensayo preclínico con variables categóricas $c_j = c_{aj}$, c_{pj} y c_{dj} (condiciones experimentales, *etc.*).

En la Figura 31, se detallan los procedimientos realizados para el pre-procesamiento y el post-procesamiento de las variables. Tras el procedimiento de pos-procesamiento, se han podido comparar las entradas con las salidas del modelo IFPTML para obtener la matriz de clasificación y medir su rendimiento.

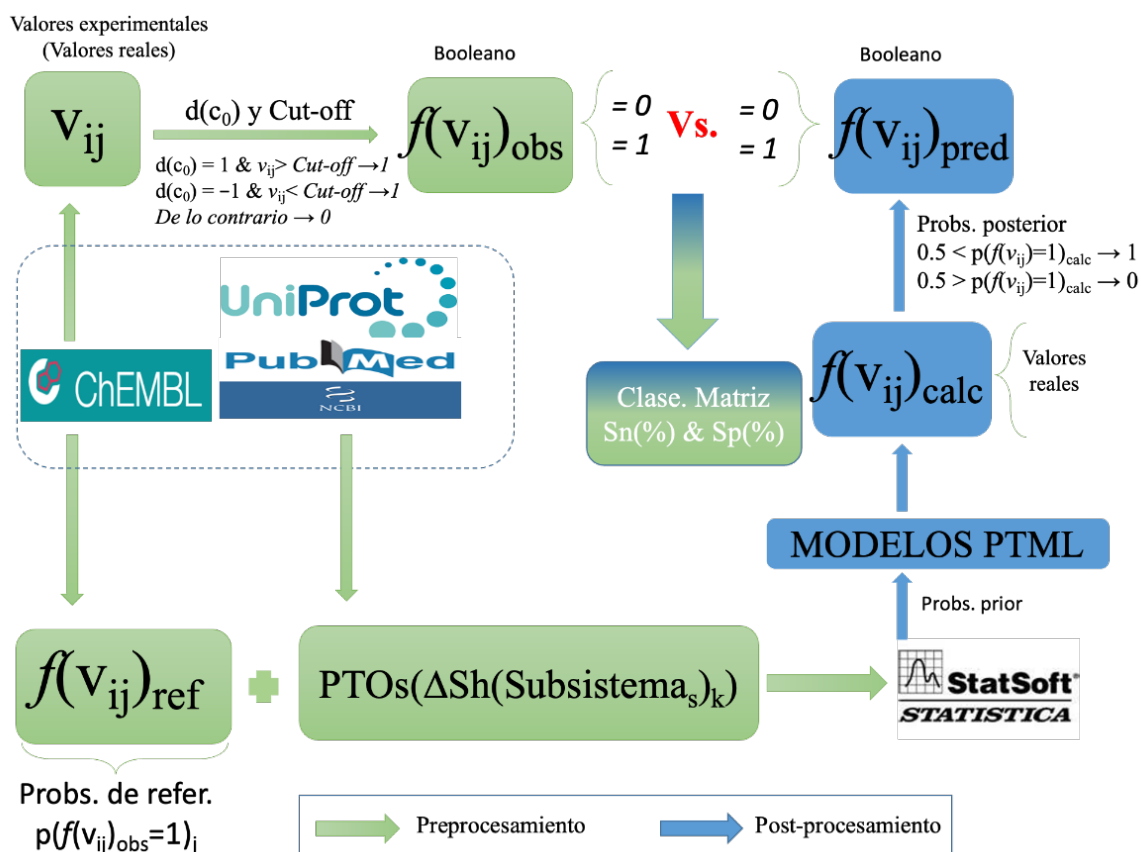


Figura 31. Variables pre-procesadas frente a post-procesadas.

Además, en la Tabla 15, se puede ver que el modelo está desequilibrado, con valores altos de Sp(%) y Accuracy Ac(%) > 98 en el entrenamiento y la validación, pero los valores de Sn(%) son bajos. Los demás parámetros estadísticos del modelo son los siguientes: n es el número de casos utilizados para entrenar el modelo, igual a 17.758; χ^2 es el estadístico Chi-cuadrado, igual a 6595,853; y p es el nivel p con un valor inferior a 0,05. En el modelo se introducen múltiples variables de entrada que codifican información relacionada con la estructura y las condiciones de ensayo del fármaco, utilizando una estrategia de selección de características por pasos hacia delante (Mendez *et al.*, 2019). El modelo también incluye variables que codifican información sobre la secuencia de la proteína, la secuencia del gen y la estructura del cromosoma, como $\Delta Sh(\text{Prot}; \text{Seq})_{sc_{dj}}$ y $\Delta Sh(\text{Chr}; \text{Gen})_{sc_{aj}}$. Sin embargo, parecen tener una contribución menor.

Tabla 15. Resultado del modelo IFPTML-GDA.

Conjuntos Observados ^a	Parámetro Estadístico ^b	Predicción Estadística	Conjuntos Predichos		
			n _j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Series de entrenamiento					
$f(v_{ij})_{obs} = 0$	Sp(%)	98.8	13,087	12,934	153
$f(v_{ij})_{obs} = 1$	Sn(%)	65.9	232	79	153
total	Ac(%)	98.3	13,319		
Series de validación externa					
$f(v_{ij})_{obs} = 0$	Sp(%)	98.7	4365	4310	55
$f(v_{ij})_{obs} = 1$	Sn(%)	66.2	74	25	49
total	Ac(%)	98.2	4439		

^a Las clases de clasificación observadas son dos: fármacos con un nivel de efecto biológico deseado $f(v_{ij})_{obs} = 1$ o $f(v_{ij})_{obs} = 0$ en caso contrario. ^b Sn (%) = Sensibilidad, Sp (%) = Especificidad y AC (%) = Precisión.

En la matriz de clasificación, se puede ver que el número de casos positivos $n(f(v_{ij}) = 1)$ obtenidos tras la aplicación de los valores de corte está muy desequilibrado con respecto al número de casos $n(f(v_{ij}) = 0)$ en la serie de control. De hecho, tenemos $n(f(v_{ij}) = 1) = 232$ en el entrenamiento y 74 en la validación frente a $n(f(v_{ij}) = 0) = 13.087$ en el entrenamiento y 4365 en la validación para el grupo de control. Se ha realizado un estudio de escaneo de los puntos de corte para comprobar si podía deberse a un valor muy restrictivo de los puntos de corte o no. Como se puede ver en la Tabla 16, el número de casos positivos $n(f(v_{ij}) = 1)$ no varía notablemente y es en todos los casos muy bajo para todos los rangos de valor de corte lo que es interesante para los usos de quimioterapia antimicrobiana. Por ejemplo, en el caso de la inhibición (%) el $n(f(v_{ij}) = 1) < 230$ para todos los valores de corte en el rango de inhibición (%) = 75-100. El número de casos positivos aumenta en el rango $n(f(v_{ij}) = 1) = 300-9700$ sólo para Inhibición (%) < 50%, que no es un rango clínicamente útil. En otras propiedades como IC50 (nM) y Ki (nM), el número de casos positivos $n(f(v_{ij}) = 1) < 140$, casos en todos los rangos de corte 1-100 nM y para todos los valores de corte en el rango Inhibición (%) = 75-100. Debido a todos estos problemas, se ha intentado probar también modelos IFPTML no lineales (véase la siguiente sección).

Tabla 16. Valores seleccionados de promedios de condiciones múltiples para diferentes combinaciones de condiciones de ensayo.

$c_0 = \text{Actividad}$ (Unidades)	Cut-off(c_0)								Total
	1	10	25	50	75	95	100	200	
Inhibición (%)	9785	1535	564	376	228	78	39	-	13,469
IC ₅₀ (nM)	2	29	49	81	101	108	110	133	3715
K _i (nM)	24	78	100	120	132	134	138	160	369
Otras Actividades	59	133	146	148	150	149	150	152	205
$n(f(v_{ij}) = 1)$	9870	1775	859	725	611	469	437	445	17,758
$n(f(v_{ij}) = 0)$	7888	15,983	16,899	17,033	17,147	17,289	17,321	17,313	

Uno de los modelos IFPTML no lineales encontrados fue el modelo de árbol de clasificación (CT)-IFPTML (IFPTML-CTUS), que es un modelo CT basado en una regla de división univariante (US) (Mendez *et al.*, 2019). En este modelo, las probabilidades previas con las que se predice que un compuesto es activo se fijaron en $\pi_1 = 0,5$. Estas probabilidades están perfectamente equilibradas en comparación con las probabilidades previas no equilibradas de $\pi_1 = 0,7$ utilizadas en el modelo GDA-IFPTML. En la Figura 32, se muestra el árbol de decisión para el modelo IFPTML-CTUS.

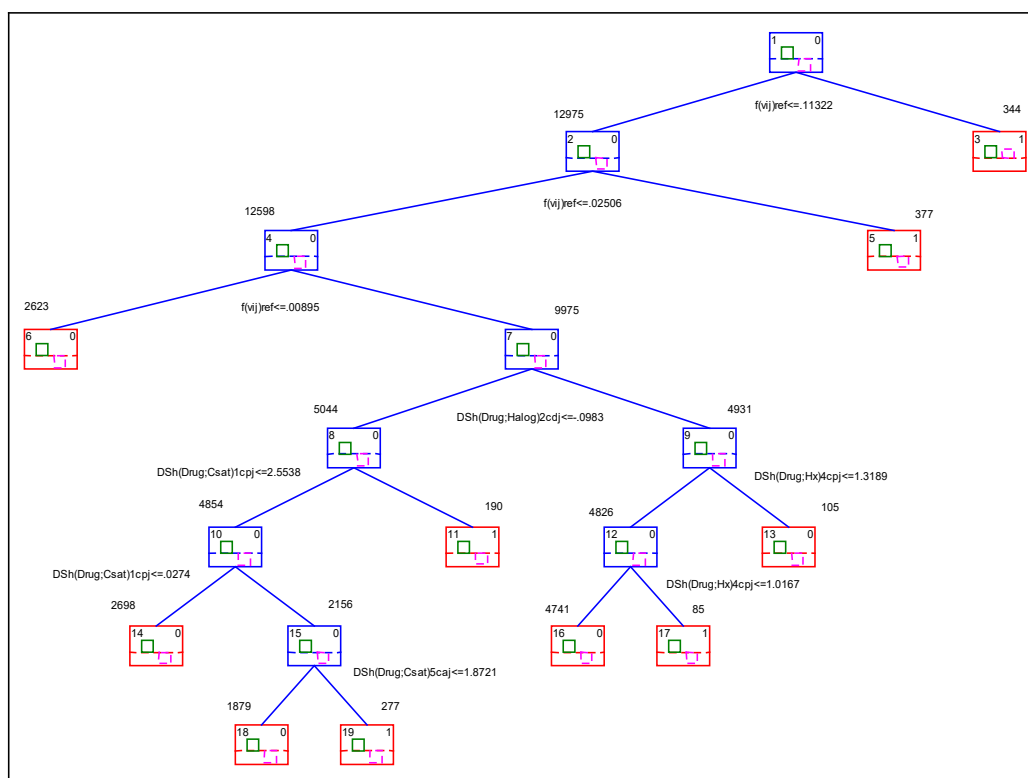


Figura 32. Árbol de decisión del modelo IFPTML-CTUS.

En la Tabla 17, se muestran los resultados y coeficientes de todas las variables en las diferentes reglas de división sobre el árbol de clasificación de este modelo. Las variables que se introdujeron en el modelo son $\Delta Sh_1 = \Delta Sh(\text{Drug}; \text{Halog})_2 c_{dj}$, $\Delta Sh_2 = \Delta Sh(\text{Drug}; \text{Csat})_1 c_{pj}$, $\Delta Sh_3 = \Delta Sh(\text{Fármaco}; \text{Hx})_4 c_{pj}$, $\Delta Sh_4 = \Delta Sh(\text{Fármaco}; \text{Csat})_1 c_{pj}$, $\Delta Sh_5 = \Delta Sh(\text{Fármaco}; \text{Hx})_4 c_{pj}$, $\Delta Sh_6 = \Delta Sh(\text{Fármaco}; \text{Csat})_5 c_{aj}$.

Tabla 17. Coeficientes del modelo IFPTML-CTUS.

Clase Nodo	Rama Izq.	Rama Der.	Control		Activo $n(f(v_{ij}) = 1)$	Clase Predic.	Separación Constante	Separación Variable
			$n(f(v_{ij}) = 0)$					
1	2	3	13,087		232	0	0.11321607	$f(v_{ij})_{refi}$
2	4	5	12,903		72	0	0.02505894	$f(v_{ij})_{refi}$
3			184		160	1		--
4	6	7	12,542		56	0	0.00895431	$f(v_{ij})_{refi}$
5			361		16	1		--
6			2623		0	0		--
7	8	9	9919		56	0	-0.0982586	$\Delta Sh(\text{Drug}; \text{Halog})_2 c_{dj}$
8	10	11	5006		38	0	2.55375728	$\Delta Sh(\text{Drug}; \text{Csat})_1 c_{pj}$
9	12	13	4913		18	0	1.318866	$\Delta Sh(\text{Drug}; \text{Hx})_4 c_{pj}$
10	14	15	4821		33	0	0.02739699	$\Delta Sh(\text{Drug}; \text{Csat})_1 c_{pj}$
11			185		5	1		--
12	16	17	4809		17	0	1.01671015	$\Delta Sh(\text{Drug}; \text{Hx})_4 c_{pj}$
13			104		1	0		--
14			2681		17	0		--
15	18	19	2140		16	0	1.87205633	$\Delta Sh(\text{Drug}; \text{Csat})_5 c_{aj}$
16			4726		15	0		--
17			83		2	1		--
18			1868		11	0		--
19			272		5	1		--

Otro modelo encontrado fue el IFPTML-CTLTC, que es un modelo IFPTML basado en CT pero que utiliza combinaciones lineales (LC) como reglas de división. En la Figura 33, se muestra el árbol de decisión del modelo IFPTML-CTLTC. En la Tabla 18, se muestran los coeficientes de todas las variables en las diferentes CL utilizadas como reglas de división.

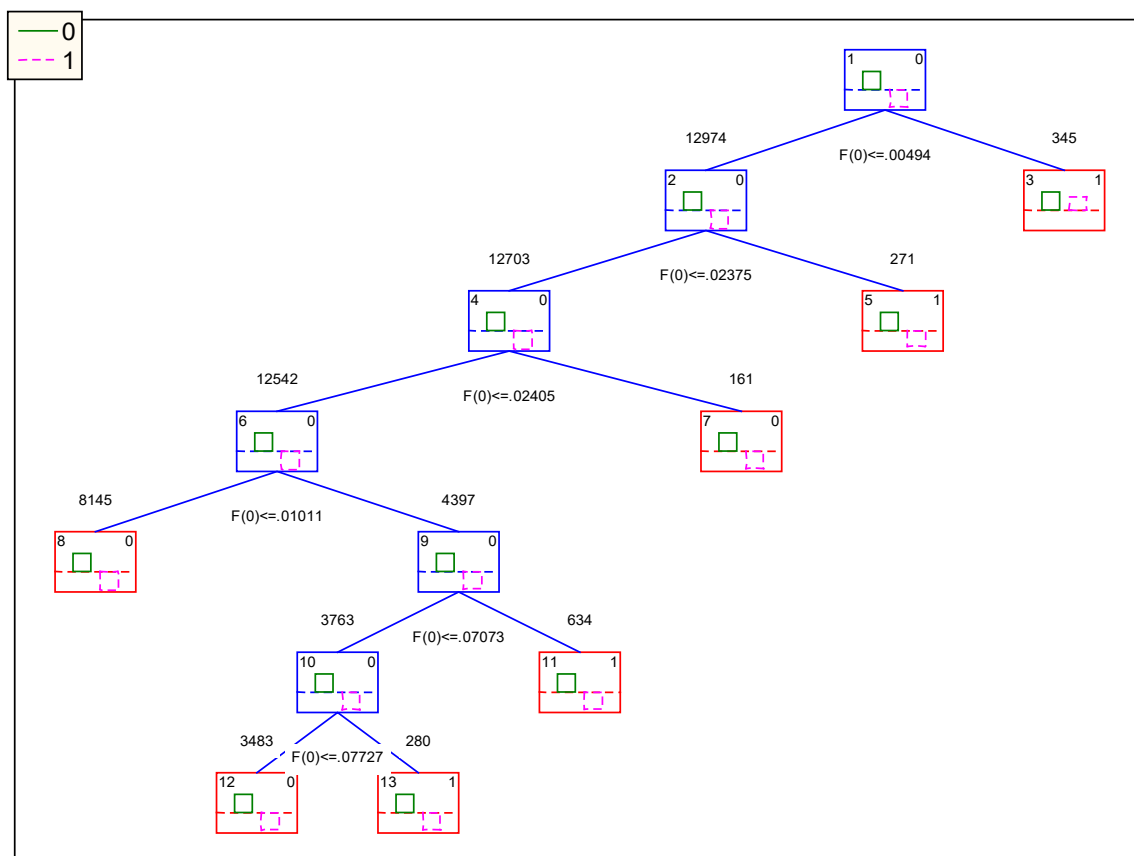


Figura 33. Árbol de decisión del modelo IFPTML-CTLC.

Tabla 18. Coeficientes del modelo IFPTML-CTLC.

Var	Coeff.	$f(v_{ij})_{01}$	$f(v_{ij})_{02}$	$f(v_{ij})_{03}$	$f(v_{ij})_{04}$	$f(v_{ij})_{05}$	$f(v_{ij})_{06}$	Prom.	S.D.
Dividir la const.	a_{00}	-0.005	-0.024	-0.024	-0.010	-0.071	-0.077	-0.04	0.03
$f(v_{ij})_{ref}$	a_{01}	0.044	0.762	0.751	0.818	2.678	2.881	1.32	1.17
$\Delta Sh(\text{Drug}; \text{Csat})_5 c_{aj}$	a_{02}	0.000	0.008	-0.001	-0.003	-0.008	-0.007	0.00	0.01
$\Delta Sh(\text{Drug}; \text{Hetero})_5 c_{aj}$	a_{03}	-0.001	-0.010	-0.042	-0.033	-0.103	-0.143	-0.06	0.06
$\Delta Sh(\text{Drug}; \text{Hx})_1 c_{aj}$	a_{04}	0.001	0.020	0.047	0.047	0.120	0.160	0.07	0.06
$\Delta Sh(\text{Drug}; \text{Csat})_1 c_{pj}$	a_{05}	0.001	0.014	0.020	0.023	0.083	0.093	0.04	0.04
$\Delta Sh(\text{Drug}; \text{Hetero})_4 c_{pj}$	a_{06}	0.001	0.009	0.036	0.028	0.078	0.109	0.04	0.04
$\Delta Sh(\text{Drug}; \text{Hx})_4 c_{pj}$	a_{07}	-0.001	-0.017	-0.038	-0.037	-0.092	-0.117	-0.05	0.04
$\Delta Sh(\text{Drug}; \text{Csat})_1 c_{dj}$	a_{08}	-0.001	-0.019	-0.016	-0.017	-0.065	-0.079	-0.03	0.03
$\Delta Sh(\text{Drug}; \text{Halog})_1 c_{dj}$	a_{09}	0.003	0.057	0.087	0.088	0.713	0.577	0.25	0.31
$\Delta Sh(\text{Drug}; \text{Halog})_2 c_{dj}$	a_{10}	-0.003	-0.059	-0.094	-0.095	-0.740	-0.609	-0.27	0.32
$\Delta Sh(\text{Chr}; \text{Gen})_5 c_{aj}$	a_{11}	0.000	0.000	0.002	0.002	0.039	0.075	0.02	0.03
$\Delta Sh(\text{Prot}; \text{Seq})_5 c_{dj}$	a_{12}	0.000	0.004	-0.002	-0.003	0.008	0.024	0.01	0.01

En primer lugar, se comparan los modelos en términos de rendimiento. En la Tabla 19, podemos ver una comparación de los tres modelos IFPTML desarrollados en esta investigación: GDA, CTUS y CTLC. El modelo IFPTML-GDA mostró el valor más bajo de S_n (%) = 65,9/66,2 y S_p (%) = 98,7/98,8 para el entrenamiento y la validación, respectivamente. Ambos modelos IFPTML-CT tienen probabilidades previas equilibradas $\pi_1 = 0,5$ con las que se predice que un compuesto es activo (comparado con $\pi_0 = 0,5$). Estos valores están perfectamente equilibrados, recordar que los modelos IFPTML-GDA presentan un importante desequilibrio en este sentido con $\pi_1 = 0,7$ (comparado $\pi_0 = 0,3$). Además, ambos modelos IFPTML-CT alcanzaron valores de S_n (%) y S_p (%) superiores al 80,0%. Los valores de IFPTML-CTUS son iguales a S_n (%) = 81,0/82,4 y S_p (%) = 91,7/91,6. El IFPTML-CTLC también presenta valores elevados de S_n (%) = 83,6/85,1 y S_p (%) = 89,7/89,8.

A continuación, sería conveniente comparar los modelos en términos de número de variables de entrada, LC y número de reglas de división. El IFPTML-GDA utiliza más de 10 variables de entrada, pero sólo una LC con una regla de división. Curiosamente, el modelo IFPTML-CTUS utiliza 5 variables de entrada y 9 constantes de división sin recurrir a las LC. Por el contrario, el IFPTML-CTLC es el modelo más complicado de los tres, con más de 10 variables de entrada y 6 LC, cada una con sus respectivas constantes de división. Por ejemplo, incluye información sobre la secuencia de la proteína en la variable $\Delta Sh(\text{Prot}; \text{Seq})_{5c_{aj}}$ e información sobre el gen y el cromosoma de esta proteína con la variable $\Delta Sh(\text{Chr}; \text{Gen})_{5c_{aj}}$. Según estos resultados, se puede decir que el último modelo es la mejor selección en términos de rendimiento e inclusión de información biológicamente relevante.

Por último, se deben comparar los modelos en cuanto a la relevancia de la información biológica incluida en las variables de entrada. El modelo IFPTML-GDA contiene información relevante sobre la estructura del fármaco, la secuencia de la proteína, *etc.* Por el contrario, el modelo IFPTML-CTUS no incluye información sobre la secuencia de la proteína, la secuencia del gen o la estructura del cromosoma. La falta de información sobre la secuencia de la proteína invalida el modelo IFPTML-CTUS para usos prácticos en la predicción de fármacos antimaláricos contra una diana proteica con cambios de secuencia específicos (mutaciones). De hecho, se ha descubierto que las mutaciones en el gen de la malaria son importantes en el desarrollo de mecanismos de resistencia a los fármacos (Nowotka *et al.*, 2017; Davies *et al.*, 2015). Por último, el modelo IFPTML-CTLC incluye variables biológicas relevantes relacionadas con la proteína diana, *etc.*, al igual que el modelo IFPTML-GDA. En general, el modelo IFPTML-CTLC es el más complejo, pero al mismo tiempo parece ser el más valioso

porque está equilibrado, tiene altos valores de Sn (%) y Sp (%), e incluye información biológica relevante.

Tabla 19. Comparación de modelos con diferentes algoritmos.

Algoritmo	Conjunto	Clase	Parám.Est adis.	Value (%)	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
IFPTML	Entrenar	$f(v_{ij})_{obs} = 0$	Sp	98.8	12,934	153
GDA		$f(v_{ij})_{obs} = 1$	Sn	65.9	79	153
$\pi_0 = 0.30$	Validación	$f(v_{ij})_{obs} = 0$	Sp	98.7	4310	55
$\pi_1 = 0.70$		$f(v_{ij})_{obs} = 1$	Sn	66.2	25	49
IFPTML	Entrenar	$f(v_{ij})_{obs} = 0$	Sp	91.7	12,002	1085
CTUS		$f(v_{ij})_{obs} = 1$	Sn	81.0	44	188
$\pi_0 = 0.50$	Validación	$f(v_{ij})_{obs} = 0$	Sp	91.6	3997	368
$\pi_1 = 0.50$		$f(v_{ij})_{obs} = 1$	Sn	82.4	13	61
	Entrenar	$f(v_{ij})_{obs} = 0$	Sp	89.8	11,751	1336
IFPTML		$f(v_{ij})_{obs} = 1$	Sn	83.6	38	194
CTLC						
$\pi_0 = 0.50$	Validación	$f(v_{ij})_{obs} = 0$	Sp	89.7	3917	448
$\pi_1 = 0.50$		$f(v_{ij})_{obs} = 1$	Sn	85.1	11	63

3.4.3. Discusión

3.4.3.1. Modelo lineal IFPTML con MMA

Para evaluar el rendimiento del modelo en términos de Especificidad Sp(%) y Sensibilidad Sn(%), el IFPTML-GDA transforma $f(v_{ij})_{calc}$ en la variable booleana $f(v_{ij})_{pred}$. La variable $f(v_{ij})_{pred} = 1$ cuando se predice que los compuestos son activos en este ensayo; $f(v_{ij})_{pred} = 0$ en caso contrario. Esta variable obtiene el valor $f(v_{ij})_{pred} = 1$ cuando la probabilidad posterior con el compuesto es activa $p(f(v_{ij}) = 1) \geq 0,5$. El algoritmo IFPTML-GDA puede estimar los valores de las probabilidades posteriores como una función sigmoideal $p(f(v_{ij}) = 1) = \pi_1 / (\pi_1 + \pi_0 \cdot \text{Exp}(-f(v_{ij})_{calc}))$ de las probabilidades previas π_1 y π_0 y los valores de la función de puntuación. En este modelo, las probabilidades “*a priori*” con las que se predice que un compuesto es activo se han fijado en $\pi_1 = 0,7$ (Mendez *et al.*, 2019). El deficiente número de compuestos activos en

el conjunto de datos ChEMBL justifica de alguna manera este valor relativamente alto de la probabilidad “*a priori*”, véase la siguiente discusión.

La principal ventaja de este algoritmo IFPTML es la obtención de un único modelo global. Esto significa que se ha construido un modelo unificado para la optimización de ensayos preclínicos de nuevos compuestos antimaláricos frente a las 28 secuencias de proteínas en muchas condiciones de ensayo diferentes c_j . De hecho, el modelo predice correctamente el resultado de 17.758 ensayos en total. Este modelo también podrá predecir nuevos compuestos antipalúdicos para nuevas secuencias de proteínas no incluidas en el conjunto de datos anterior. Por el contrario, si se construye un modelo para cada proteína objetivo, se tendrá que entrenar/validar un modelo para cada proteína. Esto significa que se tendrá que entrenar/validar un total de 28 modelos individuales, excluyendo todas las demás condiciones variables. En consecuencia, el algoritmo IFPTML puede ajustar un modelo, realizando el trabajo de 28 modelos clásicos. Además, cada modelo clásico debe ser entrenado con un número menor de ensayos. Por último, los modelos para una sola proteína no pueden predecir los resultados de un compuesto para otras proteínas y/o mutantes de proteínas, ya que no son sensibles a la secuencia.

3.4.3.2. Modelos IFPTML-CTUS e IFPTML-CTLC

Los modelos hicieron hincapié en las variables de entrada relacionadas con la información química sobre la estructura del fármaco y las condiciones de los ensayos.

3.4.3.3. Ejemplo de uso práctico del modelo IFPTML-CTLC

En esta sección, se ilustra el uso del modelo con un ejemplo práctico. Se selecciona la molécula con código ChEMBL264770. Consultar los detalles de este compuesto en el archivo de materiales complementarios en www.mdpi.com/xxx/s1. En la Figura 34, se representan gráficamente todos los pasos necesarios para procesar un compuesto conocido o nuevo con el presente modelo utilizando ChEMBL264770 como ejemplo. En esta figura, se ilustran las tres etapas principales del algoritmo y sus pasos más importantes. La etapa IF incluye los pasos (1) y (2), la etapa PT incluye sólo el paso (3), y la etapa ML incluye los pasos (4) y (5). En el paso (1), se descarga toda la información conocida sobre la molécula, la proteína diana, el gen, el cromosoma y, o, las condiciones del ensayo de tres bases de datos ChEMBL, UniProt y NCBI-GDV. En el caso de un nuevo compuesto, se desconoce el valor de la actividad biológica v_{ij} , pero se conoce el resto de la información sobre el ensayo. Esta información incluye variables numéricas y variables categóricas que codifican información sobre las condiciones

experimentales de los ensayos preclínicos o sobre la naturaleza y la calidad de los datos. Para la molécula CHEMBL264770, el parámetro de actividad es K_i (nM), el ID de acceso a Uniprot de la proteína diana es P39898, el organismo del ensayo es *P. falciparum*, la función ChEMBL es Enzima, el “mapeo” de la diana es una proteína, el nombre de la APD y la confianza están etiquetados como ND (No datos), el tipo de ensayo es B, el curado por Autocur, el número de la puntuación de confianza es 9, y SMILES Canonical. Otros datos descargados de la base de datos NCBI-GDV son la información biológica sobre las proteínas objetivo, los genes y los cromosomas. Así, para este ejemplo el nombre del gen en el cromosoma XIV es PF14_0075, la orientación del gen es 1 que significa positivo, la función de la proteína es plasmepsina, la recurrencia de nucleótidos del gen y las orientaciones de los genes en este cromosoma. Toda la información descargada de estas bases de datos se copió en un archivo .xlsx. En el paso (2), calculamos las entropías de Shannon de los fármacos, las secuencias de proteínas y el cromosoma para cuantificar la información estructural. Como entradas, se han utilizado los SMILES canónicos de los fármacos, la secuencia de las proteínas, la secuencia del gen y las GOIN de los cromosomas. Se utilizó el software MARCH-INSIDE para calcular la entropía de información de Shannon de los fármacos $Sh(\text{fármaco})$. Otras variables calculadas fueron las entropías de Shannon de la recurrencia de aminoácidos $Sh(\text{prot})$, de la recurrencia de nucleótidos $Sh(\text{gen})$, y de la orientación del gen en el cromosoma $Sh(\text{Chr})$. Estas variables se calcularon con la herramienta S2Snetwork. Después del paso (2), se ha terminado la fase IF y se entra en la fase PT. En el paso (3), se calculan los OTP con la forma de operadores de Media Móvil (MA). Hasta este momento, la limpieza y el preprocesamiento de los datos se habían realizado junto con los cálculos de los operadores aplicando la Teoría de la Perturbación. En el paso (4), se utiliza el software STATISTICA para ejecutar diferentes algoritmos de ML. Para la nueva molécula, se sustituyen los valores de los operadores $\Delta Sh(\text{Drug}_i)_k, \mathbf{c}_{aj}$, $\Delta Sh(\text{Prot}_i)_k, \mathbf{c}_{pj}$, etc., en estos modelos. Utilizando el modelo IFPTML-GDA, por ejemplo, se puede predecir un resultado de $p(f(v_{ij})=1) = 0,99$ para este ejemplo. Esto significa que el modelo predice que se espera que este compuesto tenga un valor $K_i < 10$ nM (punto de corte) con una probabilidad de 0,99. Finalmente, en el paso (5), se puede concluir que la $f(v_{ij})_{\text{pred}} = 1$ (el compuesto puede considerarse activo según este ensayo). Como este compuesto ya es conocido, se puede corroborar que esta predicción coincide con la clasificación observada $f(v_{ij})_{\text{obs}} = 1$ que proviene de un valor real de $K_i = 0,3$ nM. En el caso de un compuesto no ensayado previamente, habría que ensayar el compuesto para corroborar esta predicción.

ensayos en la Base de Datos ChEMBL). Por ejemplo, se registró una enzima ChEMBLID = "ChEMBL1697656" con su Nombre Preferido = "Glutación S-transferasa", Acceso UniProt = "Q8MU52", Tipo de Objetivo = "Proteína Única", Organismo = "*P. falciparum*", Compuestos = "4", y Puntos Finales = "6". Además, cada punto final proviene de un único ensayo con los siguientes campos principales CMPD ChEMBLID, Nombre de la molécula, SMILES, ID de actividad, Tipo de estándar, Relación, Valor estándar y Unidades estándar. Otros campos son Assay ID, Assay ChEMBLID, Assay Type, Description, Protein Accession (UniProt Accession), Journal, Year, Volume, y Issue, entre otros.

3.4.4.2. Conjunto de datos NCBI-GDV

El genoma de *Pf* utilizado fue reportado originalmente en la base de datos Mapviewer (Wolfsberg, 2010), (Coordinators, 2018). En la actualidad, este conjunto de datos está disponible en la nueva base de datos NCBI-GDV (<https://www.ncbi.nlm.nih.gov/genome/gdv/> (consultada en noviembre de 2017)) (Coordinators, 2018). Inicialmente, el genoma de *Pf* tenía 14 cromosomas diferentes. Cada cromosoma contiene una media de 383 genes. En este trabajo, se han utilizado solo 10 de estos 14 cromosomas porque las proteínas codificadas por los 4 cromosomas restantes no tienen ensayos biológicos reportados en ChEMBL.

Los genes tienen una posición de inicio y fin dentro del cromosoma. La base de datos informa de la posición (P_{ik}) de cada gen en el cromosoma y de una descripción de la función biológica. El conjunto de datos registró la secuencia biológica de nucleótidos de cada gen. Además, el conjunto de datos informa del símbolo, la orientación del gen, como positivo o negativo ($O_{ik} = 1$ u $O_{ik} = -1$). Se ha descubierto que esta información es de alguna manera relevante para la actividad biológica de algunas proteínas en el proteoma de *Pf*. En consecuencia, en este trabajo también se han utilizado las GOINs del proteoma de *Pf* ensambladas con información P_{ik} y O_{ik} en un trabajo anterior (Quevedo-Tumaili *et al.*, 2018).

3.4.4.3. Conjunto de datos UniProt

Se descargó la secuencia biológica de aminoácidos de las 28 proteínas registradas en ChEMBL en formato FASTA. El conjunto de datos se obtuvo de la base de datos UniProt (<https://www.uniprot.org/> (consultada en noviembre de 2018)) utilizando la herramienta browser protein (UniProt Consortium, 2019), (UniProt Consortium, 2018; Pundir *et al.*, 2017). A su vez, el formato FASTA tiene dos parámetros que fueron utilizados en este trabajo: cadena de características y secuencia de proteínas.

3.4.4.4. Fusión de información ChEMBL, NCBI-GDV y UniProt

Se ha construido un conjunto de datos basado en los tres conjuntos de datos anteriores. Para ello, se ha realizado un proceso de FI (Whittle *et al.*, 2006; Weininger, 1988; Toropov & Benfenati, 2007; Veselinovic *et al.*, 2013). Tras realizar el proceso IF, el conjunto de datos de trabajo creado contenía un total de 18.381 resultados (filas). Se han añadido los códigos SMILE canónicos y sus respectivos valores de Entropía de Shannon para cada compuesto químico.

Los códigos SMILES descargados de ChEMBL son un sistema de notación utilizado para codificar la información sobre la estructura química de los compuestos (Leone *et al.*, 2018). Las representaciones de tipo SMILES se han utilizado en gran medida en la quimioinformática (Pogany *et al.*, 2019; Toropova & Toropova, 2019; Zheng *et al.*, 2019; Prado-Prado *et al.*, 2011; Hill & Lewicki, 2006; Tilley & Rosenthal, 2019). También se ha agregado la secuencia de la proteína y las Entropías de Shannon en cada fila de acuerdo con el respectivo ID de acceso a la proteína.

Además, han añadido los parámetros de cada gen y los valores de la Entropía de Shannon para cada proteína.

3.4.4.5. Pre-procesamiento del conjunto de datos de trabajo

En primer lugar, se eliminan las filas en las que no se informaba de valores para las variables v_{ij} , PSA o ΔLOGP con el fin de limpiar el conjunto de datos. Por esta razón, las categorías de la variable cp4 se reducen a 19 Enzimas, 2 Transportadores, 1 Regulador Epigenético, 2 Otras Proteínas Citosólicas y 4 Proteínas No Clasificadas. El total de proteínas válidas de ChEMBL eran 28. Por lo tanto, los datos eliminados representan sólo un 3,4% de todo el conjunto de datos de trabajo. Además, todas las celdas vacías de tipo cadena fueron reemplazadas por la etiqueta ND (No Data). Al final, el conjunto de datos para obtener el modelo basado en el IFPTML tenía 17.758 filas. En la Figura 35, se ilustran los diferentes pasos dados para el pre-procesamiento de los datos y la realización del proceso IF.

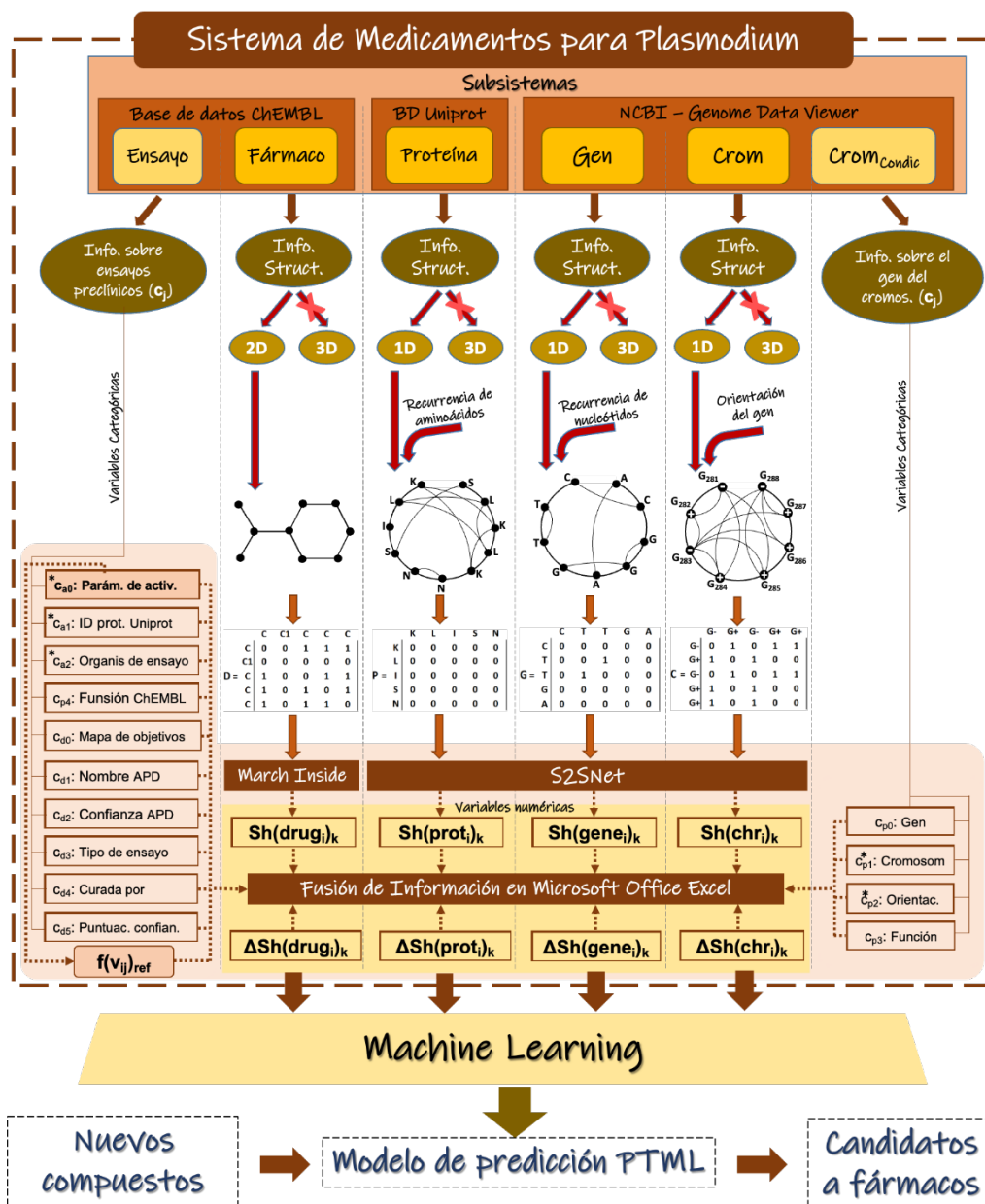


Figura 35. Desarrollo del modelo IFPTML y proceso de IF.

3.4.4.6. Modelos IFPTML de la Teoría de la Información de Shannon

En la Figura 35, se ilustran los detalles de los diferentes pasos dados para pre-procesar los datos y entrenar/validar el modelo IFPTML. En primer lugar, se realiza el proceso de IF, a continuación, se calculan los valores de $Sh(\text{Subsystems})_k$, los valores de la función $f(v_{ij})_{ref}$ y los valores de PTO (variables de entrada), y luego se procede a buscar los modelos IFPTML.

Véanse más detalles sobre el cálculo de las variables de entrada/salida en las siguientes secciones. El objetivo del modelo IFPTML es predecir una función $f(v_{ij})_{calc}$ de los valores observados $f(v_{ij})_{obs}$. Para desarrollar el modelo IFPTML, se han tenido en cuenta tanto la información estructural como la funcional para el cálculo de las variables de entrada. La

información estructural se refiere a la estructura química del fármaco, así como a las características estructurales de la proteína diana, el gen que codifica esta proteína diana y el cromosoma de este gen.

Se puede abordar el presente problema desde el punto de vista de la teoría de la información de Shannon y la teoría de los sistemas complejos. En este sentido, se puede cuantificar la información estructural/funcional relevante del sistema con los valores $Sh(\text{Syst})_k$ calculados mediante un enfoque de cadena de Markov (Munteanu *et al.*, 2008). Después, se calcula la propiedad externa del sistema $f(v_{ij})_{\text{calc}}$ como función de un valor de referencia $f(v_{ij})_{\text{ref}}$ y una función $f(Sh(\text{Syst})_k, c_j)$ de la información estructural y funcional. En la ecuación (2) se utiliza un enfoque aditivo IFPTML para incluir y separar las diferentes partes del sistema o subsistemas.

$$f(v_{ij})_{\text{calc}} = a_0 + a_1 \cdot f(v_{ij})_{\text{ref}} + \sum_{s=0, k=0, j=0}^{s_{\text{max}}, k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \text{PTO} \left[\text{Sh}(\text{Subsystem}_s)_{k,c_j} \right] \quad (2)$$

La función de referencia $f(v_{ij})_{\text{ref}}$ cuantifica el valor esperado de la probabilidad de actividad biológica para una medida de compuesto en determinadas condiciones experimentales especificadas por la partición c_j de variables categóricas. Los subsistemas considerados son el subsistema₀ = fármaco, el subsistema₁ = proteína, el subsistema₂ = gen y el subsistema₃ = cromosoma. La información sobre cada subsistema se cuantificará con los respectivos valores de la medida de información de la Entropía de Shannon de orden k^{th} para cada subsistema $Sh(\text{Subsistemas})_k$. Por ejemplo, $Sh(\text{Subsistema}_0)_k = Sh(\text{Droga})_k$ y $Sh(\text{Subsistema}_1)_k = Sh(\text{Prot})_k$, *etc.* El valor k^{th} puede registrar valores de 0 a 5. Además, el modelo IFPTML utiliza los PTOs para cuantificar la desviación (perturbaciones) en las variables continuas (parámetros estructurales, tiempo, concentración, *etc.*) con respecto a la información funcional codificada por las variables categóricas c_j (condiciones experimentales). Se pueden observar los detalles en las siguientes secciones (Nocedo-Mena *et al.*, 2019).

En este contexto, en la ecuación (3), se puede ilustrar la forma general de un modelo IFPTML para los casos lineales. En la Ecuación (4), se seleccionan los casos lineales por razones de simplicidad, pero en este trabajo, también se reportan modelos no lineales. Se puede ampliar la ecuación anterior del modelo para escribir una forma general del modelo IFPTML. Para ello, utilizamos MMA como operadores PTO de la siguiente manera.

$$f(v_{ij})_{\text{calc}} = a_0 + a_1 \cdot f(v_{ij})_{\text{ref}} + \sum_{s=0, k=0, j=0}^{s_{\text{max}}, k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \Delta\text{Sh}(\text{Subsystem}_s)_{k,c_j} \quad (3)$$

$$f(v_{ij})_{\text{calc}} = a_0 + a_1 \cdot f(v_{ij})_{\text{ref}} + \sum_{k=0, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \Delta\text{Sh}(\text{Drug})_{k,c_j} + \sum_{k=0, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \Delta\text{Sh}(\text{Prot})_{k,c_j} \quad (4)$$

$$+ \sum_{k=0, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \Delta\text{Sh}(\text{Gene})_{k,c_j} + \sum_{k=0, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{s,k,j} \cdot \Delta\text{Sh}(\text{Chr})_{k,c_j}$$

3.4.4.7. Variable de salida y función de referencia

En este trabajo, se desarrolla un modelo IFPTML para el estudio de los valores experimentales v_{ij} de la actividad biológica del i -ésimo fármaco en el j -ésimo ensayo preclínico de fármacos antipalúdicos reportados en la base de datos ChEMBL.

Debido al elevado número de parámetros biológicos con diferentes escalas y niveles de error, se discretizan para obtener la función booleana $f(v_{ij})_{\text{obs}}$ para desarrollar un modelo de clasificación. En primer lugar, se realiza el pre-procesamiento para limpiar el conjunto de datos, definir/calcular las variables de entrada y salida.

En concreto, los valores $f(v_{ij})_{\text{obs}}$ y $f(v_{ij})_{\text{ref}}$ se han calculado mediante funciones de Excel y se han añadido al conjunto de datos, véase la Tabla 20. Por ejemplo, para el cálculo del número de casos con un nivel específico de ca_0 (un parámetro específico de la actividad biológica) se ha utilizado la función COUNTIF.

El primer argumento de la sintaxis es $\text{Range}(ca_0)$ = celdas que contienen todos los valores de la variable categórica ca_0 (nombres de los parámetros de actividad biológica medidos en cada ensayo preclínico). El segundo argumento es $\text{Criteria}(ca_0)$ = celdas que contienen el valor de un único nivel de ca_0 (nombre de un parámetro específico de actividad biológica). La función recorre todo el $\text{Range}(ca_0)$ comparando $\text{Criteria}(ca_0)$ con la celda específica del $\text{Range}(ca_0)$.

Otros argumentos utilizados en las diferentes funciones son $\text{Range}(v_{ij})$ = celdas que contienen todos los valores de actividad biológica para todos los ensayos preclínicos (v_{ij}), $\text{Units}(ca_0)$ = las unidades de la actividad biológica medida (ca_0), $d(ca_0) = 1$ o -1 , y $\text{Range}(f(v_{ij})_{\text{obs}})$ = celdas que contienen el valor de $f(v_{ij})_{\text{obs}}$ (Nocedo-Mena *et al.*, 2019).

Tabla 20. Funciones más relevantes utilizadas en la etapa de pre-procesamiento de datos.

Variable	Sintaxis de las Funciones de Excel	Notas
$n_j(c_{a0})$	=COUNTIF (Range(c_{a0}), Criterias(c_{a0}))	Función que determina el número total de casos para cada actividad biológica en el conjunto de datos.
$\langle v_{ij}(c_{a0}) \rangle$	=AVERAGEIF (Range(c_{a0}), Criterias(c_{a0}), Range(v_{ij}))	Calcula el promedio de todos los valores estándar de la actividad biológica en el conjunto de datos. Se utiliza como argumento para la función valor de corte (c_{a0}).
valor de corte(c_{a0})	=IF(Units(c_{a0}) = %, 95, IF(Units(c_{a0}) = nM, 10, $\langle v_{ij}(c_{a0}) \rangle$)	El valor de valor de corte se utiliza para decidir si los compuestos son activos o no. Para los valores de Actividad (%) e Inhibición (%), el valor de corte (c_{a0}) = 95%. Del mismo modo, para el IC_{50} (nM), K_i (nM), y K_m (nM), el valor de corte (c_{a0}) = 10nM, <i>etc.</i>
$d(c_{a0})$	=OR($d(c_{a0}) = 1$, $d(c_{a0}) = -1$)	Indica que el parámetro medido aumenta o disminuye directamente con un efecto biológico deseado o no deseado.
$f(v_{ij})_{obs}$	=IF(AND($v_{ij} >$ valor_de_corte(c_{a0}), $d(c_{a0}) = 1$), 1, IF(AND($v_{ij} \leq$ valor_de_corte(c_{a0}), $d(c_{a0}) = -1$), 1, 0))	$f(v_{ij})_{obs} = 1$ para los compuestos activos de $(v_{ij})_{obs} = 0$ para el grupo de control según el conjunto de valores de valor de corte y deseabilidad utilizados para cada c_{a0} . Es la función utilizada como salida para entrenar el modelo IFPTML.
$n(f(v_{ij})=1)$	=COUNTIF(Range(c_{a0}), Criterias(c_{a0}), Range($f(v_{ij})_{obs}$, 1))	Función que determina el número total de cada actividad biológica en el conjunto de datos y $f(v_{ij})_{obs}$ igual a 1.
$f(v_{ij})_{ref}$	= $n(f(v_{ij})=1)/n_j(c_{a0})$	La función de referencia $f(v_{ij})_{ref} = p(f(v_{ij})=1/c_{a0})$ es la probabilidad con la que la función observada obtiene el valor $f(v_{ij})_{obs} = 1$, ensayo positivo. Se utiliza como primera variable de entrada del modelo IFPTML.

3.4.4.8. Medidas de entropía de Shannon

Las ecuaciones anteriores del IFPTML se introdujeron como variables $Sh(\text{Subsystems})_k$. Se han calculado los valores de las Entropías de Shannon $Sh(\text{Drug})_k$, $Sh(\text{Prot})_k$, $Sh(\text{Gene})_k$, y $Sh(\text{Chrom})_k$ para cuantificar la información de la estructura de los diferentes subsistemas. Se ha utilizado la herramienta MARKovCHains Invariants for Network Selection and DEsign (MARCH-INSIDE) para calcular los valores $Sh(\text{Drug})_k$ de los fármacos

(Zhao *et al.*, 2019). El software MARCH-INSIDE se utilizó para introducir los códigos SMILES de cada compuesto descargado de ChEMBL. Por otra parte, se utilizó la herramienta Sequences to Networks (S2SNet) (Munteanu *et al.*, 2008) para calcular los valores del índice de información $Sh(Prot)_k$, $Sh(Gene)_k$, y $Sh(Chrom)_k$ sobre la secuencia y la recurrencia de diferentes aminoácidos en las proteínas, nucleótidos en los genes, y genes en los cromosomas. Se utilizó el software S2SNet para introducir las secuencias de proteínas y genes descargadas de UniProt y NCBI-GDV, respectivamente. S2SNet también se utilizó para introducir un código de secuencia np (negativo/positivo) para expresar la orientación de la lectura y la posición de cada gen en el cromosoma.

Tanto MARCH-INSIDE (fármacos) como S2SNet (proteínas, genes y cromosomas) utilizan un grafo para representar las partes del subsistema (nodos) y las relaciones (enlace) entre ellas en la estructura del subsistema. Las partes de los subsistemas son átomos, aminoácidos, bases de nucleótidos o genes. Los enlaces entre ellos son enlaces químicos, enlaces peptídicos, secuencia de genes o posición de genes según el sistema. El software S2SNet también tiene en cuenta las relaciones de recurrencia con tipos específicos de aminoácidos, nucleótidos y la orientación de los genes. La Figura 36 ilustra algunos ejemplos de los gráficos utilizados para representar los diferentes subsistemas. Muestra el nombre, el gráfico de representación y una pequeña parte del gráfico con sus nodos y enlaces. Se pueden ver en esta figura, de abajo a arriba, el cromosoma XI representado por los genes y los enlaces a los pares de genes con orientación inversa. Los nodos del grafo del gen 285 con su grafo de representación en el cromosoma, y el grafo con sus nodos representados por los nucleótidos y los enlaces representados por la secuencia del gen por sus recurrencias. La proteína Q9NFSS tiene nodos con aminoácidos y enlaces con enlaces peptídicos y la recurrencia. Por último, el gráfico del fármaco CHEMBL510738 se representó con átomos (nodos) y enlaces químicos (enlaces).

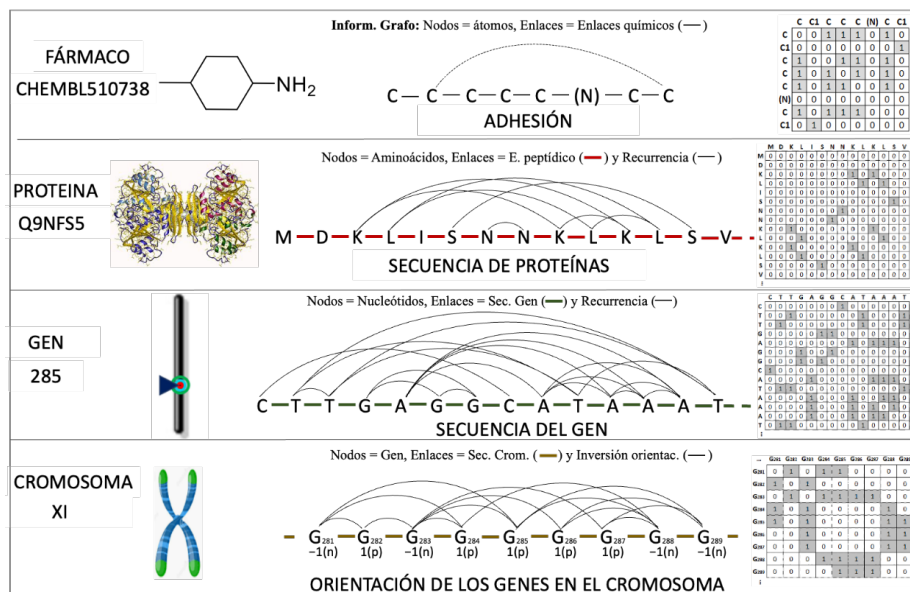


Figura 36. Ilustración de diferentes representaciones para representar sistemas moleculares múltiples.

Tanto MARCH-INSIDE como S2SNet asocian una matriz de adyacencia de nodos $A(\text{Subsistema}_s)$ a los respectivos grafos para llevar a cabo una representación numérica del sistema (véase la Figura 36). A continuación, ambos programas de computación avanzada transforman la matriz de adyacencia de cada subsistema $A(\text{Subsistema}_s)$ en una matriz de Markov $\Pi_1(\text{Subsistema}_s)$, no representada en la Figura 36. Después, ambas herramientas calculan las potencias naturales de orden k^{th} para cada matriz $\Pi_1(\text{Subsistema}_s)$. Por último, ambos programas utilizan las ecuaciones de Chapman-Kolmogórov para calcular las probabilidades absolutas ${}^a p(n/s)_k$ de cada nodo de un subsistema determinado (n/s) (Munteanu *et al.*, 2008; Zhao *et al.*, 2019). Con estas probabilidades y la ecuación (5), el software realiza el cálculo de los diferentes valores $\text{Sh}(\text{Drug})_k$, $\text{Sh}(\text{Prot})_k$, $\text{Sh}(\text{Gene})_k$ y $\text{Sh}(\text{Chrom})_k$.

$$\text{Sh}(\text{Subsystem}_s)_k = - \sum_{n=1}^{n_{\max}} {}^a p(n, s)_k \cdot \log({}^a p(n, s)_k) \quad (5)$$

3.4.4.9. Particiones de variables categóricas

Se crean dos particiones (subconjuntos) de variables categóricas del conjunto de datos ChEMBL para codificar toda la información funcional o no estructural. La primera partición de variables categóricas fue $\mathbf{c}_{\text{assay}_j}$ (abreviada como \mathbf{c}_{aj}). La segunda partición fue $\mathbf{c}_{\text{data}_j}$ (abreviada como \mathbf{c}_{dj}). Estas particiones contienen variables que codifican información sobre las condiciones experimentales de los ensayos preclínicos (\mathbf{c}_{aj}) o sobre la naturaleza y calidad de los datos (\mathbf{c}_{dj}). Estas variables categóricas incluyen información sobre 22 tipos de actividad

biológica (c_{a0}), 28 proteínas objetivo (c_{a1}) y 9 organismos del ensayo (c_{a2}), etc. También se crea otra partición ($c_{proj} = c_{pj}$) que incluye variables categóricas con información biológica sobre las proteínas objetivo, los genes y los cromosomas. Estas variables abarcan 32 genes (c_{p0}), 10 cromosomas (c_{p1}), la orientación de los genes (c_{p2}) y 31 funciones de las proteínas (c_{p3}). La tabla 21 muestra los detalles de estas particiones.

Tabla 21. Particiones y niveles (valores únicos) tomados por las variables de entrada categóricas (no ordenadas).

Partición (c_j)	Var.	Información	NL ^a	Niveles únicos
c_{assayj} (c_{aj})	c_{a0}	Actividad Biológica	22	Inhibition(%); IC ₅₀ (nM); K _i (nM); IC ₅₀ (ug.mL ⁻¹); BHIA ₅₀ (-); IC ₅₀ (mill equivalent); FC(-); K _{inact} (/min); Activity(%); VAR(-); Ratio(-); Ratio(/M/s); IC ₅₀ (molar ratio); Ratio IC ₅₀ (-); Mean(pM mg ⁻¹); GST activity (mU mg ⁻¹); K _m (nM); Ratio(/s/M); Activity(-); K _a (10 ³ /M/s); K _{cat} (/s); Inhibition(uM)
	c_{a1}	ID de acceso a la proteína de UniProt	28	Q8MU52; Q3HTL5; Q9NBA7; Q9NFS5; Q8T6J6; Q25856; P39898; Q9N6S8; Q0PJ46; Q6T755; Q8MMZ4; Q868D6; Q25917; Q9GSW0; Q9NAW4; O77078; Q9NAW2; Q9BJJ9; Q8T6B1; Q9N623; Q9XYC7; P05227; P11144; Q17SB2; O77239; Q9Y006; O96214; O97467
	c_{a2}	Organismo de ensayo	9	<i>Plasmodium falciparum</i> ; <i>Plasmodium falciparum</i> K1; <i>Plasmodium falciparum</i> NF54; <i>Plasmodium falciparum</i> Dd2; <i>Plasmodium</i> sp.; <i>Plasmodium yoelii</i> ; <i>Plasmodium berghei</i> ; <i>Leishmania mexicana</i> ; ND (No registered data)
c_{dataj} (c_{dj})	c_{d0}	Mapeo de objetivos	2	Proteína; Proteína homóloga
	c_{d1}	Nombre del APD	9	Peptidase C1; Pkinase; Peptidase S8; Asp; OMPdecase; Spermine synth; Sugar tr; Hist deacetyl
	c_{d2}	Confianza APD	2	ND (No hay datos); alto
	c_{d3}	Tipo de ensayo	2	Binding (B) = Datos que miden la vinculación del compuesto con una diana molecular. Functional (F) = Datos que miden el efecto biológico de un compuesto.

	<i>c_{d4}</i>	Nivel de curación de datos	3	Autocuración; Intermedio; Experto
	<i>c_{d5}</i>	Puntuación de confianza	2	8 = Objetivo de proteína única homóloga asignada. 9 = Objetivo de proteína individual directo asignado.
<i>c_{protj}</i> (<i>c_{pj}</i>)	<i>c_{p0}</i>	Gen	32	<i>PF140187; PF110161; PFB0325c; PF110301; PF100225; PF140341; PF140075; PF110165; PF130141; MAL13P1.214; PF140346; PFE0355c; PF140294; PF140125; PF110162; PFB0505c; PF140511; PF140076; PFE0370c; PF110147; PFB0330c; PFF0730c; PF140598; MAL7P1.27; PF11260c; PFB0100c; PF080054; PF140077; MAL13P1.185; PF140078; PFB0150c; PFE1455w</i>
	<i>c_{p1}</i>	Cromosoma	10	II; V; VI; VII; VIII; IX; X; XI; XIII; XIV
	<i>c_{p2}</i>	Orientación	2	Downstream = -1; Upstream = 1
	<i>c_{p3}</i>	Función de la proteína (UniProt)	31	Glutathione s-transferase, putative; Falcipain-2 precursor; Cysteine protease, putative; Spermidine synthase; Orotidine-monophosphate-decarboxylase, putative; Glucose-6-phosphate isomerase; Plasmepsin, putative; Falcipain 2 precursor; L-lactate dehydrogenase; phosphoethanolamine <i>N</i> -methyltransferase; cGMP-dependent protein kinase 1, beta isozyme, putative; Serine protease belonging to subtilisin family, putative; Mitogen-activated protein kinase 1; Deoxyhypusine synthase; Falcipain-3; Beta-ketoacyl-acyl carrier protein synthase III precursor, putative; Glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase; Plasmepsin 1 precursor; Subtilisin-like protease precursor, putative; Mitogen-activated protein kinase 2; Enoyl-acyl carrier reductase; Glyceraldehyde-3-phosphate dehydrogenase; Chloroquine resistance transporter, putative; Histone deacetylase; Knob associated histidine-rich protein; Heat shock 70 kDa protein; Plasmepsin 2 precursor; CDK-related protein kinase 6; HAP protein; Protein kinase, putative; Sugar transporter, putative

c_{p4}	Tipo de función de destino ChEMBL	5	Enzima; Transportador; Regulador epigenético; Otra proteína citosólica; Proteína no clasificada
----------	-----------------------------------	---	---

^a *NL = Número de niveles (valores únicos) que quedan después del pre-procesamiento.*

3.4.4.10. Operadores de la teoría de la perturbación (PTO)

Como se ha mencionado anteriormente, el modelo IFPTML utiliza PTOs para cuantificar la desviación (perturbaciones) en las variables continuas (parámetros estructurales, tiempo, concentración, *etc.*) con respecto a la información funcional codificada por las variables categóricas c_j (condiciones experimentales).

En este trabajo se seleccionan los operadores MMAs del tipo PTO($\text{Sh}(\text{Subsistema}_s)_k = \Delta\text{Sh}(\text{Subsistema}_s)_{k,c_j} = \text{Sh}(\text{Subsistema}_1)_k - \langle \text{Sh}(\text{Subsistema}_1)_{k,c_j} \rangle$ o $f(\text{Sh}(\text{Subsistema}_s)_k = \Delta\text{Sh}(\text{Subsistema}_s)_{k,c_j} = \text{Sh}(\text{Subsistema}_1)_k - \langle \text{Sh}(\text{Subsistema}_1)_{k,c_j} \rangle$). Estos operadores cuantifican la desviación (ganancia o pérdida de información) del valor específico $\text{Sh}(\text{Subsistema}_1)_k$ del subsistema respecto a la media $\langle \text{Sh}(\text{Subsistema}_1)_{k,c_j} \rangle$ (valor esperado) de información para todos los casos medidos en las mismas condiciones experimentales.

Se han utilizado tres particiones diferentes c_j de variables categóricas para codificar las condiciones experimentales y, o, la información no estructural (véase la sección siguiente).

Además, en esta etapa de pre-procesamiento de datos, se han calculado los operadores PT similares a los operadores Box-Jenkins MA que se utilizan como datos de entrada. En este contexto, c (con c en negrita) se refiere a un vector de combinaciones múltiples de variables categóricas al mismo tiempo. Las particiones de las variables categóricas utilizadas aquí son c_{assayj} , c_{protj} y c_{dataj} . Estas particiones son fusiones de variables categóricas relacionadas con el ensayo farmacológico (c_{assayj}), la naturaleza del objetivo del fármaco (c_{protj}), o sobre la naturaleza y, o, la precisión de los datos medidos (c_{dataj}).

Para simplificar, se abrevian estas particiones como $c_{\text{assayj}} = c_{aj}$, $c_{\text{protj}} = c_{pj}$, y $c_{\text{dataj}} = c_{dj}$. La partición $c_{aj} = (c_{a0}, c_{a1}, c_{a2})$ incluía las siguientes variables categóricas: actividad biológica (c_{a0}), el ID de acceso a la proteína de UniProt (c_{a1}) y el organismo de ensayo (c_{a2}).

En el archivo de Materiales Suplementarios que está disponible en línea en www.mdpi.com/xxx/s1 se detallan todos los conjuntos de datos fusionados de fármacos, secuencias únicas, proteínas, cromosomas, genes, valores de Entropías de Shannon y los valores de PTO, este proceso se denomina técnica IF.

La Tabla 22 muestra los detalles de los Operadores de la Teoría de la Perturbación.

Tabla 22. Variables de entrada de los modelos IFPTML desarrollados.

Tipo de Variable	Símbolo	Fórmula	Variables Categórica	Detalles
-	$f(v_{ij})_{ref}$	$n(f(v_{ij})_{expt} = 1) / n_j$	c_{a0}	Valor esperado de la probabilidad $p(f(v_{ij}) = 1)_{ref}$ para la actividad v_{ij} del tipo c_{a0} .
MMA_{caj}	$\Delta Sh(Drug_i)_{k,caj}$	$Sh(Drug_i)_k - \langle Sh(Drug)_{k,caj} \rangle$	c_{aj}	Variación (Δ) de la información de la estructura del fármaco en diferentes subconjuntos de variables categóricas múltiples relacionadas con el ensayo farmacológico c_{aj} .
MMA_{cdj}	$\Delta Sh(Drug_i)_{k,cdj}$	$Sh(Drug_i)_k - \langle Sh(Drug)_{k,cdj} \rangle$	c_{dj}	Variación (Δ) de la información de la estructura del fármaco en diferentes subconjuntos de variables categóricas múltiples relacionadas con la naturaleza y/o la presión de los datos medidos c_{dj} .
MMA_{cpj}	$\Delta Sh(Prot_i)_{k,cpj}$	$Sh(Prot_i)_k - \langle Sh(Prot)_{k,cpj} \rangle$	c_{pj}	Variación (Δ) de la información de la secuencia de la proteína, secuencia del ge e información sobre el cromosoma para diferentes subconjuntos de variables categóricas múltiples relacionadas con la naturaleza de la proteína objetivo c_{pj} .
	$\Delta Sh(Gene_i)_{k,cpj}$	$Sh(Gene_i)_k - \langle Sh(Gene)_{k,cpj} \rangle$		
	$\Delta Sh(Chrom_i)_{k,cpj}$	$Sh(Chrom_i)_k - \langle Sh(Chrom)_{k,cpj} \rangle$		

3.4.4.11. Entrenamiento y validación del modelo IFPTML

El primer paso para desarrollar los modelos IFPTML (Gonzalez-Diaz *et al.*, 2013; Santana *et al.*, 2019; Nocado *et al.*, 2019; Shannon, 1948; Graham, 2002; Graham *et al.*, 2004) fue descargar toda la información sobre los ensayos preclínicos, la estructura de los fármacos, las secuencias de proteínas, las secuencias de genes y la información de los cromosomas de las

bases de datos públicas (ChEMBL, UniProt, NCBI-GDV). El segundo paso fue realizar un pre-procesamiento de toda la información anterior para calcular la $f(v_{ij})_{obs}$ (variable dependiente) y la $f(v_{ij})_{ref}$. A continuación, se calcularon los valores $Sh(\text{Subsystems})_k$ (variables de entrada). Esto incluye un proceso de fusión de información que incluye datos de las diferentes bases de datos (ChEMBL, UniProt, NCBI-GDV). Una vez preparados los datos para el análisis, se ejecutan los desarrollos ML: GDA, Árbol de Clasificación (CT) con divisiones univariantes (CTUS) y CT con combinación lineal (CTLC) para buscar modelos IFPTML alternativos. Todos los modelos IFPTML se desarrollaron utilizando el software STATISTICA v. 12 (Mendez *et al.*, 2019).

3.4.5. Conclusiones

La predicción computacional de nuevos compuestos antimaláricos es un objetivo muy importante para la industria farmacéutica. Sin embargo, la enorme cantidad de información disponible de diferentes fuentes dificulta el análisis de datos para el descubrimiento de nuevos compuestos. El método IFPTML permitió llevar a cabo la fusión y el análisis de tres conjuntos de datos diferentes de las bases de datos ChEMBL, UniProt y NCBI-GDV para lograr este objetivo. El conjunto de datos ChEMBL contiene resultados de 17.758 ensayos únicos que incluyen descriptores numéricos (variables) para la estructura de los compuestos.

El desarrollo IFPTML tuvo éxito a la hora de tener en cuenta tanto la información numérica (parámetros estructurales) como la información categórica (múltiples condiciones experimentales) de los tres conjuntos de datos. Las medidas de entropía de Shannon Sh_k (variables numéricas) fueron útiles para cuantificar la información sobre la estructura de los fármacos, las secuencias de proteínas, las secuencias de genes y los cromosomas. Además, las MMA de diferentes particiones de variables categóricas del conjunto de datos ChEMBL fueron útiles para codificar múltiples condiciones experimentales de ensayos preclínicos e información sobre proteínas, genes y cromosomas diana. El modelo IFPTML-CTLC es el más complejo en términos de número de variables de entrada, número de LCs y número de reglas de división. Sin embargo, el modelo IFPTML-CTLC mostró un mejor rendimiento que el IFPTML-GDA e incluye más información biológicamente relevante que el modelo IFPTML-CTUS. Este modelo podría convertirse en una herramienta útil para la optimización de ensayos preclínicos de nuevos compuestos antimaláricos teniendo en cuenta la estructura del fármaco, la especie de *Plasmodium*, la secuencia de la proteína diana y otros múltiples parámetros.

3.4.6. Materiales complementarios: Los siguientes están disponibles en línea en www.mdpi.com/xxx/s1.

4. Conclusiones

1. Gracias a la adaptación “*ad hoc*” de técnicas y procedimientos de computación avanzada y AI para el propósito específico planteado como objetivo en la presente tesis doctoral, se ha comprobado que es posible definir redes complejas de tipo GOINs en forma de redes secuencia-recurrencia para codificar información sobre la disposición (locus) y orientación de genes en los cromosomas. Siendo el *P. falciparum* un organismo modelo adecuado para la construcción y estudio de las GOINs para todo el genoma de un organismo
2. También gracias a las citadas técnicas y procedimientos de computación avanzada y AI “*ad hoc*” se pudo probar que las redes tipo GOINs constuidas para el genoma del *P. falciparum* tienen una distribución no aleatoria. La relevancia biológica de este fenómeno no ha podido ser establecida completamente necesitándose desarrollar estudios adicionales.
3. Las comunidades dentro de las redes tipo GOINs contruidas mostraron cierta relevancia en la detección de clusters de genes que codifican para proteínas con actividad biológica similar del grupo de las RIFINs pero no se logró demostrar por el momento para otros tipos de proteínas.
4. Las centralidades de nodos de las redes GOINs son útiles como variables de entrada en estudios de ML para predecir función biológica de los genes que codifican para proteínas de tipo RIFINs. No se pudo demostrar su utilidad en el estudio de otros tipos de proteínas.
5. Es posible usar la matriz de Markov para calcular índices de información de Shannon de diferentes sistemas biomoleculares relacionados con el *P. falciparum* incluyendo: grafos moleculares de fármacos, redes de secuencia-recurrencia de genes y proteínas, y redes tipo GOINs obteniéndose por esta vía índices numéricos de estos sistemas en escalas comparables.
6. La estrategia NIFPTML permite desarrollar nuevos modelos predictivos para problemas biológicos complejos que involucren varios de los sistemas biomoleculares anteriores a la vez usando como variables de entrada los índices de Shannon de estos sistemas.
7. Se pudo corroborar experimentalmente algunos de los resultados predichos por los modelos NIFPTML desarrollados en colaboración con otros grupos de investigación mostrando la gran utilidad práctica de esta metodología.

5. Futuros desarrollos

1. Aprovechar las potencialidades de nuevos desarrollos de ML y Deep Learning para definir las redes complejas tipo GOINs en forma de redes de secuencia-recurrencia que incluyan no solo la información sobre la disposición (locus) y orientación de genes en el cromosoma sino información sobre distancia entre genes, interacción con proteínas histonas, etc..
2. Utilizar estas o similares técnicas de computación avanzada y AI para explorar el uso de cromosómica de otros tipos de redes y, o, matrices asociadas diferentes de la matriz de Markov (matrices de adyacencia, distancia topológica, incidencia, etc.)
3. Utilizar estas o similares técnicas de computación avanzada y AI para explorar construir GOINs para el genoma de otros organismos.
4. Calcular otros tipos de índices numéricos, TIs, parámetros grafo teóricos, invariantes, etc., de las GOINs y explorar su utilidad en estudios de ML.
5. Desarrollar nuevos modelos NIFPTML para otros problemas biológicos complejos (prognosis de cáncer, enfermedades hereditarias, etc.) no estudiados en esta tesis, usando parámetros de redes tipo GOINs como variables de entrada.
6. Desarrollar un software para la recopilación de información, construcción, representación, manipulación, cálculo de índices numéricos, y estudios de ML usando las redes tipo GOINs de los cromosomas.

6. Bibliografía

- Abdel-Latif, M. S., Khattab, A., Lindenthal, C., Kremsner, P. G. e Klinkert, M. Q. (2002). Recognition of variant Rifin antigens by human antibodies induced during natural *Plasmodium falciparum* infections. *Infect. Immun.*, 70 (12), 7013–21.
- Abinaya, M., Vaseeharan, B., Divya, M., Sharmili, A., Govindarajan, M., Alharbi, N.S., Kadaikunnan, S., Khaled, J.M. e Benelli, G. (2018). Bacterial exopolysaccharide (EPS)-coated ZnO nanoparticles showed high antibiofilm activity and larvicidal toxicity against malaria and zika virus vectors. *J. Trace Elem. Med. Biol.*, 45, 93–103.
- Abraira, V., Pérez, A. (1996) Métodos multivariantes en bioestadística, Ramón Areces, España. Pp. 341-348.
- Akuffo, H., Costa, C., van Griensven, J., Burza, S., Moreno, J. e Herrero, M. (2018). New insights into leishmaniasis in the immunosuppressed, *PLoS Negl Trop Dis*, 12, e0006375.
- Alonso, P. e Noor, A.M. (2017). The global fight against malaria is at crossroads. *Lancet*, 390, 2532–2534.
- Altman DG, Bland JM (June 1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ*. 308 (6943): 1552.
- Arie, S. (2017). Researchers and WHO clash over global threat of drug resistant malaria. *BMJ*, 359, j5127.
- Barbolla, I., Lete, E., e Sotomayor, N. (2019). Intramolecular mizoroki-heck reaction in the synthesis of heterocycles: strategies for the generation of tertiary and quaternary stereocenters, For examples of our work in the area, see, in: O. Attanasi, P. Merino, D. Spinelli (Eds.), Targets in Heterocyclic Systems, vol. 23 *Societá Chimica Italiana*, Roma, pp. 340-362 (Chapter 17).
- Barbolla, I., Sotomayor N. e Lete, E. (2019). Carbopalladation/suzuki coupling cascade for the generation of quaternary centers. Access to pyrrolo[1,2-b]isoquinoline, *J. Org. Chem.*, 84, 10183-10196.
- Barigye, S.J., Marrero-Ponce, Y., Martinez-Lopez, Y., Torrens, F., Artiles-Martinez, L.M., Pino-Urias, R.W. e Martinez-Santiago, O. (2013). Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. *J. Comput. Chem.* 34, 259–274.
- Bar-Yam, Y. (2002). General Features of Complex Systems. *Encyclopedia of Life Support Systems*.

- Bavelas, A. (1948). A mathematical model for group structures. *Human Organization*, 7, 16–30.
- Bender, E. e Williamson, S. (2010). Lists, Decisions and Graphs. With an Introduction to Probability.
- Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., *et al.* (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* 42, D1083–D1090.
- Bhat, S.Y., Jagruthi, P., Srinivas, A., Arifuddin, M. e Qureshi, I.A. (2020). Synthesis and characterization of quinoline-carbaldehyde derivatives as novel inhibitors for leishmanial methionine aminopeptidase, *Eur. J. Med. Chem.* 186 111860.
- Biemolt, J. e Ruijter, E. (2018). Advances in palladium- catalyzed cascade cyclizations, *Adv. Synth. Catal.* 360 3821-3871.
- Biggs, N., Lloyd, E. e Wilson, R. (1986). Graph Theory, 1736–1936. Oxford University Press.
- Bilbao-Ramos, P., Dea-Ayuela, M.A., Cardenas-Alegría, O., Salamanca, E., Santalla-Vargas, J.A., Benito, C. Flores, N. e Bolás-Fernández, F. (2017). Leishmaniasis in the major endemic region of Plurinational State of Bolivia: species identification, phylogeography and drug susceptibility implications, *Acta Trop.* 176 150-161.
- Blázquez-Barbadillo, C., Aranzamendi, E., Coya, E., Lete, E., Sotomayor, N. e González-Díaz, H. (2016). Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions, *RSC Adv.* 6, 38602-38610.
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.
- Bringmann, G., Bischof, S.K., Müller, S., Gulder, T., Winter, C., Stich, A., Moll, H., Kaiser, M., Brun, R., Dreher, J. e Bauman, K., (2010). QSAR guided synthesis of simplified antiplasmodial analogs of naphthylisoquinoline alkaloids, *Eur. J. Med. Chem.* 45 5370-5383.
- Brinton, W. (1939). Presentación gráfica. Pág. 43.
- Cabrera-Andrade, A., López-Cortés, A., Munteanu, C., Pazos, A., Pérez-Castillo Y., Tejera E., Arrasate S., González-Díaz, H. (2020). Perturbation-Theory Machine Learning (PTML) Multilabel Model of the ChEMBL Dataset of Preclinical Assays for Antisarcoma Compounds. *ACS Omega.* 5(42). 27211-27220.
- Calle, M. e Urrea, V. (2011). Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* 12, 86–89.
- Carral-Menoyo, A., Sotorrios, L., Ortiz-de-Elguea, V., Díaz-Andres, A., Sotomayor, N., Gomez-Bengo, E. e Lete, E. (2020). Intramolecular Palladium(II)- catalyzed 6-endo C-

- H Alkenylation directed by the remote N-protecting Group. Mechanistic insight and application to the synthesis of dihydroquinolines, *J. Org. Chem.* 85 2486-2503.
- Casañola-Martin, G.M., Le-Thi-Thu, H., Pérez-Giménez, F., Marrero-Ponce, Y., Merino-Sanjuán, M., Abad, C. e González-Díaz, H. (2015). Multi-Output model with box-jenkins operators of linear indices to predict multi-target inhibitors of ubiquitin-proteasome pathway. *Mol. Divers.* 19, 347–356.
- Chan, C., Yin, H., Garforth, J., McKie, J.H., Jaouhari, R., Speers, P., Douglas, K.T., Rock, P.J., Yardley, V., Croft, S.L. e Fairlamb, A.H. (1998). Phenothiazine inhibitors of Trypanothione reductase as potential antitrypanosomal and antileishmanial drugs, *J. Med. Chem.* 41 148-156.
- Chen, T. e Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17.
- Cho, S.J. e Hermsmeier, M.A. (2002). Genetic algorithm guided selection: Variable selection and subset selection. *J. Chem. Inf. Comput. Sci.* 42, 927–936.
- Claude, C. (1958). Théorie des graphes et ses applications. Paris: Dunod. English edition.
- Claussen, U. (2005). Chromosomics. *Cytogenet Genome Res.* 111:101-106.
- Contreras-Torres, E., Marrero-Ponce, Y., Teran, J.E., Garcia-Jacas, C.R., Brizuela, C.A. e Sanchez-Rodriguez, J.C. (2019). MuLiMs-MCoMPAs: A novel multiplatform framework to compute tensor algebra-based three-dimensional protein descriptors. *J. Chem. Inf. Model.* 60, 1042–1059.
- Coordinators, N.R. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46, D8–D13.
- Corral, A. G., e León, M. de. (2020). Las matemáticas de la biología: De las celdas de las abejas a las simetrías de los virus. Los Libros De La Catarata.
- Cortés, N., Posada-Duque, R.A., Álvarez, R., Alzate, F., Berkov, S., Cardona- Gómez, G.P. e Osorio, E. (2015). Neuroprotective activity and acetylcholinesterase inhibition of five Amaryllidaceae species: a comparative study, *Life Sci.* 122, 42-50.
- Cover, T. e Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 21–27.
- Coya, E., Sotomayor, N. e Lete, E. (2014). Intramolecular direct arylation and Heck reactions in the formation of medium sized rings. Selective synthesis of fused indolizine, pyrroloazepine and pyrroloazocine systems, *Adv. Synth. Catal.* 356, 1853-1865.

- Coya, E., Sotomayor, N. e Lete, E. (2015). Enantioselective palladium catalyzed Heck- Heck cascade reactions. Ready access to the tetracyclic core of lycorane alkaloids, *Adv. Synth. Catal.* 357, 3206-3214.
- Cristianini, N. (2004). Fisher Discriminant Analysis (Linear Discriminant Analysis). In Dictionary of Bioinformatics and Computational Biology; Wiley: Hoboken, NJ, USA.
- Cristianini, N., Shawe-Taylor, J. (2000). An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, chap-6. <http://www.supportvector.net>.
- Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L. e Overington, J.P. (2015). ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, 43, W612–W620.
- D-Bcarrue. D-Bcarrue/Nano Drugs Malaria. Available online: <https://github.com/d-bcarrue/NanoDrugsMalaria> (accessed on 11 April 2019).
- Díaz-Alarcia, R., Yanez-Pérez, V., MunetaArrate, I., Arrasate, S., Lete, E., Meana, J.J. e González-Díaz, H. (2019). Big data challenges targeting proteins in GPCR signaling pathways; combining PTML-ChEMBL models and [35S]GTPgammaS binding assays, *ACS Chem. Neurosci.* 10 4476e4491.
- DiMasi, J.A., Grabowski, H.G. e Hansen, R.W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* 47, 20–33.
- Dorlo, T.P., Balasegaram, M., Beijnen, J.H. e de Vries, P.J. (2012). Miltefosine: a review of its pharmacology and therapeutic efficacy in the treatment of leishmaniasis, *J. Antimicrob. Chemother.* 67 2576-2597.
- Duardo-Sanchez, A. (2010). Study of Criminal Law Networks with Markov-Probability Centralities. In Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks; González-Díaz, H., Munteanu, C. R., Eds.; *Transworld Research Network*: Kerala, India, 205–212.
- Duardo-Sanchez, A. (2011). Criminal Law Networks, Markov Chains, Shannon Entropy and Artificial Neural Networks. In Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neuro sciences; González-Díaz, H., Prado-Prado, F.J., García-Mera, X., Eds.; *Transworld Research Network*: Kerala, India, 107– 114.
- Duardo-Sanchez, A., Munteanu, C. R., Riera-Fernandez, P., Lopez-Diaz, A., Pazos, A. e Gonzalez-Diaz, H. (2014). Modeling complex metabolic reactions, ecological systems,

- and financial and legal networks with MIANN models based on Markov-Wiener node descriptors. *J. Chem. Inf. Model.* 54 (1), 16–29.
- Duenas, M., Mastrandrea, R., Barigozzi, M. e Fagiolo, G. (2017). Spatio- Temporal Patterns of the International Merger and Acquisition Network. *Sci. Rep.*, 7 (1), 10789.
- Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control.* 14: 200–230.
- Dutta, P.P., Bordoloi, M., Gogoi, K., Roy, S., Narzary, B., Bhattacharyya, D.R., Mohapatra, P.K. e Mazumder, B. (2017). Antimalarial silver and gold nanoparticles: Green synthesis, characterization and “*in vitro*” study. *Biomed. Pharmacother.* 91, 567–580
- Estrada, E. (2006). Protein bipartivity and essentiality in the yeast protein- protein interaction network. *J. Proteome Res.* 5 (9), 2177–84.
- Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 6 (1), 35–40.
- Euler, L. (1736). *Solutio problematis ad geometriam situs pertinentis.* *Academiae Sci. I. Petropolitanae* 8, 128-140.
- Evidente, A. e Kornienko, A. (2009). Anticancer evaluation of structurally diverse Amaryllidaceae alkaloids and their synthetic derivatives, *Phytochemistry Rev.* 8, 449-459.
- Faudree, R., Flandrin, E., Ryjáček, Z. (1997). Claw-free graphs — A survey. *Discrete Mathematics.* 164 (1–3): 87–147.
- Fawcett T., 2004, ROC Graphs: Notes and Practical Considerations for Researchers. Technical report. Palo Alto (USA): HP Laboratories.
- Ferreira da Costa, J., Silva, D., Caamano, O., Brea, J.M., Loza, M.I., Munteanu, C.R., Pazos, A., García-Mera, X. e González-Díaz, H. (2018). Perturbation theory/machine learning model of ChEMBL data for dopamine targets: docking, synthesis, and assay of new 1-Prolyl-1-leucyl-glycinamide peptidomimetics, *ACS Chem. Neurosci.* 9, 2572-2587.
- Fertin, G., Raspaud, A., Reed, B. (2004). Star coloring of graphs. *Journal of Graph Theory,* 47 (3): 163–182.
- Fisher, R.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(7), 179–188 (1936)
- Fowler, J. H. e Jeon, S. (2008). The authority of Supreme Court precedent. *Social Networks* 30, 16–30.

- Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S. e Wahlbeck, P. J. (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15 (3), 324–346.
- Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 1189–1232.
- Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 16. 906-914.
- Gaillard, T., Boxberger, M., Madamet, M. e Pradines, B. (2018). Has doxycycline, in combination with anti-malarial drugs, a role to play in intermittent preventive treatment of *Plasmodium falciparum* malaria infection in pregnant women in Africa? *Malar. J.* 17, 469.
- Garcia-Bernardo, J., Fichtner, J., Takes, F. W. e Heemskerk, E. M. (2017). Uncovering Offshore Financial Centers: Conduits and Sinks in the Global Corporate Ownership Network. *Sci. Rep.*, 7 (1), 6246.
- Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., *et al.* (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrian-Uhalte, E., *et al.* (2017). The ChEMBL database in 2017. *Nucleic. Acids Res.* 45, D945–D954.
- Geurts, P., Ernst, D. e Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 3–42.
- Ghanat Bari, M., Ung, C. Y., Zhang, C., Zhu, S. e Li, (2017). H. Machine Learning-Assisted Network Inference Approach to Identify a New Class of Genes that Coordinate the Functionality of Cancer Networks. *Sci. Rep.* 7 (1), 6993.
- Goel, S., Palmkvist, M., Moll, K., Joannin, N., Lara, P., Akhouri, R. R., Moradi, N., Ojemalm, K., Westman, M., Angeletti, D., Kjellin, H., Lehtio, J., Blixt, O., Idestrom, L., Gahmberg, C. G., Storry, J. R., Hult, A. K., Olsson, M. L., von Heijne, G., Nilsson, I. e Wahlgren, M. (2015). RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. *Nat. Med.* 21 (4), 314–7.
- Gómez D., Martínez B. (2001). Relación entre análisis discriminante lineal y regresión lineal múltiple, Inst. de Investigación FCM. Disponible en la web: http://sisbib.unmsm.edu.pe/bibvirtualdata/libros/Matematicas/4to_taller_2001/relacion_analisis.pdf

- Gonzalez-Diaz, H. e Riera-Fernandez, P. (2012). New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *J. Chem. Inf. Model.* 52 (12), 3331–40.
- Gonzalez-Díaz, H., Arrasate, S., Gómez-SanJuan, A., Sotomayor, S., Lete, E., Besada-Porto, L., Ruso, J.M. (2013). General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry, *Curr. Top. Med. Chem.* 13, 1713-1741.
- González-Díaz, H., Arrasate, S., Sotomayor, N., Lete, E., Munteanu, C.R., Pazos, A., Besada-Porto, L., e Ruso, J.M. (2013). MIANN models in medicinal, physical and organic chemistry, *Curr. Top. Med. Chem.* 13(5), 619-641.
- Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F. M. e Uriarte, E. (2008). Proteomics, networks and connectivity indices. *Proteomics* 8 (4), 750–78.
- Gonzalez-Diaz, H., Herrera-Ibata, D. M., Duardo-Sanchez, A., Munteanu, C. R., Orbegozo-Medina, R. A. e Pazos, A. (2014). ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J. Chem. Inf. Model.* 54 (3), 744–55.
- González-Díaz, H., Pérez-Montoto, L.G. e Ubeira, F.M. (2014). Model for vaccine design by prediction of b-epitopes of iedb given perturbations in peptide sequence, *in vivo* process, experimental techniques, and source or host organisms. *J. Immunol. Res.* 1–15.
- Gonzalez-Diaz, H., Riera-Fernandez, P., Pazos, A. e Munteanu, C. R. (2013). The Rucker-Markov invariants of complex Bio-Systems: applications in Parasitology and Neuroinformatics. *BioSystems*, 111 (3), 199– 207.
- González-Díaz, H., Viña, D., Santana, L., de Clercq, E., e Uriarte, E. (2006). Stochastic entropy QSAR for the *in silico* discovery of anticancer compounds: Prediction, synthesis, and “*in vitro*” assay of new purine carbanucleosides. *Bioorganic & Medicinal Chemistry*, 14(4), 1095-1107.
- González-Durruthy, M., Alberici, L.C., Curti, C., Naal, Z., Atique-Sawazaki, D.T., Vázquez-Naya, J.M., González-Díaz, H. e Munteanu, C.R. (2017). Experimental-Computational study of carbon nanotube effects on mitochondrial respiration: In silico nano-QSPR machine learning models based on new raman spectra transform with Markov-Shannon entropy invariants. *J. Chem. Inf. Model.*, 57, 1029–1044.
- González-Durruthy, M., Monserrat, J.M., Rasulev, B., Casañola-Martín, G.M., Barreiro Sorrivias, J.M., Paraíso-Medina, S., Maojo, V., González-Díaz, H., Pazos, A. e Munteanu, C.R. (2017). Carbon nanotubes’ effect on mitochondrial oxygen flux

- dynamics: Polarography experimental study and machine learning models using star graph trace invariants of raman spectra. *Nanomaterials*, 7, 386.
- González-Durruthy, M., Werhli, A.V., Seus, V., Machado, K.S., Pazos, A., Munteanu, C.R., González-Díaz, H., e Monserrat, J.M. (2017). Decrypting strong and weak single-walled carbon nanotubes interactions with mitochondrial voltage-dependent anion channels using molecular docking and perturbation theory. *Sci. Rep*, 7, 13271.
- Graham, D.J. (2002). Information content in organic molecules: Structure considerations based on integer statistics. *J. Chem. Inf. Comput. Sci.*, 42, 215.
- Graham, D.J. (2005). Information content and organic molecules: Aggregation states and solvent effects. *J. Chem. Inf. Model.*, 45, 1223–1236.
- Graham, D.J. (2007). Information content in organic molecules: Brownian processing at low levels. *J. Chem. Inf. Model.*, 47, 376–389.
- Graham, D.J. e Schacht, D. (2000). Base Information content in organic molecular formulae. *J. Chem. Inf. Comput. Sci.*, 40, 942.
- Graham, D.J. e Schulmerich, M.V. (2004). Information content in organic molecules: Reaction pathway analysis via brownian processing. *J. Chem. Inf. Comput. Sci.*, 44, 1612–1622.
- Graham, D.J., Malarkey, C. e Schulmerich, M.V. (2004). Information content in organic molecules: Quantification and statistical structure via brownian processing. *J. Chem. Inf. Comput. Sci.*, 44, 1601–1611.
- Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. (2009). A Novel Connectionist System for Improved Unconstrained Handwriting Recognition (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31 (5): 855–868. CiteSeerX 10.1.1.139.4502.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- Hansch, C. (2011). The Advent and Evolution of QSAR at Pomona College. *J. Comput. Aided Mol. Des.*, 495–507.
- Hanson, S., Adelman, J. e Ullman, B. (1992). Amplification and molecular cloning of the ornithine decarboxylase gene of *Leishmania donovani*, *J. Biol. Chem.* 267 2350-2359.
- Hao, J. e Ho, T.K. (2019). Machine learning made easy: A review of scikit-learn package in Python programming language. *J. Educ. Behav. Stat.*, 107699861983224.
- Harary F. (1969) *Graph Theory*. Westview Press: MA.

- Haye, A., Albert, J. e Rooman, M. (2014). Modeling the *Drosophila* gene cluster regulation network for muscle development. *PLoS One*, 9 (3), e90285.
- He, M., Qu, C., Gao, O., Hu, X. e Hong, X. (2015). Biological and pharmacological activities of Amaryllidaceae alkaloids, *RSC Adv.* 5 16562-16574.
- Hendrickx, S., Caljon, G. e Maes, L. (2019). Need for sustainable approaches in anti-leishmanial drug discovery, *Parasitol. Res.* 118 2743-2752.
- Herrera-Ibata, D. M., Pazos, A., Orbegozo-Medina, R. A., Romero-Duran, F. J. e Gonzalez-Diaz, H. (2015). Mapping chemical structure- activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of U.S. counties. *BioSystems*, 132–133, 20–34.
- Hill, T. e Lewicki, P. (2006). STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining, StatSoft: *Tulsa, OK*, Vol. 1, p 813.
- Hu, Y. e Bajorath, J. (2012). Growth of ligand-target interaction data in ChEMBL is associated with increasing and activity measurement-dependent compound promiscuity. *J. Chem. Inf. Model.*, 52, 2550–2558.
- Hurst, L. D., Williams, E. J. e Pal, C. (2002). Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* 18 (12), 604–6.
- Ingrosso, D., Fowler, A.V., Bleibaum, J. e Clarke, S. (1989). Sequence of the D-aspartyl/L-isoaspartyl protein methyltransferase from human erythrocytes. Common sequence motifs for protein, DNA, RNA, and small molecule S-Adenosylmethionine-Dependent methyltransferases, *J. Biol. Chem.* 264, 20131-20139.
- Inoue, L. Y., Neira, M., Nelson, C., Gleave, M., Etzioni, R. (2007). Cluster-based network model for time-course gene expression data. *Biostatistics*, 8 (3), 507–25.
- Jones, N.G., Catta-Preta, C.M.C., Lima, A.P.C.A. e Mottram, J.C. (2018). Genetically validated drug targets in *Leishmania*: current knowledge and future prospects, *ACS Infect. Dis.* 4 467-477.
- Junker, B. H., Koschuetzki, D. e Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. *BMC Bioinf.* 7 (1), 219.
- Kalanon, M. e McFadden, G. (2010). Malaria, *Plasmodium falciparum* and its apicoplast. *Biochem. Soc. Trans.* 38, 775–782.
- Khaparde, S., Kale, P., y Agarwal, S. (1991). Application of artificial neural network in protective relaying of transmission lines. Proceedings of the First International Forum on Applications of Neural Networks to Power Systems. 122-125.

- Kemp, L.E., Yamamoto, M. e Soldati-Favre, D. (2013). Subversion of host cellular functions by the apicomplexan parasites, *FEMS Microbiol. Rev.* 37 607-631.
- Khare, S., Nagle, A.S., Biggart A., *et al.*, (2016). Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness, *Nature* 537, 229-233.
- King, R. Chemical Applications of Topology and Graph Theory, *Elsevier*, 1983
- Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J.M., Melo, A., Speck-Planche, A. e Cordeiro, M.N.D.S. (2014). Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.*, 73, 288–294.
- Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J.M., Speck-Planche, A. e Cordeiro, M.N.D.S. (2014). Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ. Sci. Technol.*, 48, 14686–14694.
- Kubinyi, H. (1993). QSAR: Hansch Analysis and Related Approaches. In *Methods and Principles in Medicinal Chemistry*; Wiley: Hoboken, NJ, USA.
- Kustatscher, G., Grabowski, P. e Rappsilber, J. (2017). Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* 13 (8), 937.
- Lage, S., Martínez Estíbalz, U., Sotomayor, N. e Lete, E. (2009). Intramolecular palladium-catalyzed direct arylation vs. Heck reactions: synthesis of pyrro-loisoquinolines and isoindoles, *Adv. Synth. Catal.* 351, 2460-2468.
- Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G. e Mann, M. (2002). Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419, (6906), 537–42.
- Lavazec, C., Sanyal, S. e Templeton, T. J. (2006). Hypervariability within the Rifin, Stevor and Pfmc-2TM superfamilies in *Plasmodium falciparum*. *Nucleic Acids Res.*, 34 (22), 6696–707.
- Leone, C., Bertuzzi, E.E., Toropova, A.P., Toropov, A.A. e Benfenati, E. (2018). CORAL: Predictive models for cytotoxicity of functionalized nanozeolites based on quasi-SMILES. *Chemosphere*, 210, 52–56.
- Lin, H. H., Zhang, L. L., Yan, R., Lu, J. J. e Hu, Y. (2017). Network Analysis of Drug-target Interactions: A Study on FDA-approved New Molecular Entities Between 2000 to 2015. *Sci. Rep.*, 7 (1), 12230.

- Liu, Y., Tang, S., Fernandez-Lozano, C., Munteanu, C.R., Pazos, A., Yu, Y.-Z., Tan, Z. e González-Díaz, H. (2017). Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. *Expert Syst. Appl.*, 306–316.
- Loiseau, P.M., Cojean, S. e Schrével, J. (2011). Sitamaquine as a putative antileishmanial drug candidate: from the mechanism of action to the risk of drug resistance, *Parasite* 18 115-119.
- Luan, F., Kleandrova, V.V., González-Díaz, H., Ruso, J.M., Melo, A., Speck-Planche, A. e Cordeiro, M.N.D.S. (2014). Computer-Aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale*, 6, 10623–10630.
- Martínez Estíbalez, U., Sotomayor, N., Lete, E., (2009). Intramolecular carbolithiation reactions for the synthesis of 2,4-disubstituted tetrahydroquinolines. Evaluation of TMEDA and (-)-sparteine as ligands in the stereoselectivity, *Org. Lett.* 11 1237-1240.
- Martínez-Arzate, S.G., Tenorio-Borroto, E., Barbabosa Pliego, A., Díaz-Albiter, H.M., Vázquez-Chagoyán, J.C. e González-Díaz, H. (2017). PTML model for proteome mining of B-cell epitopes and theoretical-experimental study of Bm86 protein sequences from Colima, Mexico. *J. Proteome Res.*, 16, 4093–4103.
- Martinez-Lopez, Y., Marrero-Ponce, Y., Barigye, S.J., Teran, E., Martinez-Santiago, O., Zambrano, C.H. e Torres, F.J. (2019). When global and local molecular descriptors are more than the sum of its parts: Simple, but not simpler? *Mol. Divers.*, 24, 913–932.
- Mashaghi, A., *et al.* (2004). Investigation of a protein complex network. *European Physical Journal.* 41 (1): 113–121.
- Mauri, A., Consonni, V., Todeschini, R., (2017). Molecular Descriptors. Handbook of Computational Chemistry. *Springer International Publishing.* pp. 2065–2093. ISBN 978-3-319-27282-5.
- Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Hochreiter, S. *et al.* (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci*, 9(24), 5441-5451.
- McCulloch, W.S., Pitts, W. A. (1990). A Logical Calculus of Ideas Inmanent in Nervous Activity. *Reprint Bltn Mathcal Biology*, 52, 99–115
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Felix, E., Leach, A. R. *et al.* (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*, 47(D1), D930-D940.

- Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magarinos, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., *et al.* (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.*, 47, D930–D940.
- Minkin, V. (1999). Glossary of terms used in theoretical organic chemistry. *Pure and Applied Chemistry*, 71(10), 1919-1981.
- Minodier, P. e Parola, P. (2007). Cutaneous leishmaniasis treatment, *Trav. Med. Infect. Dis.* 5 150-158.
- Moore, D.H. (1987). Classification and regression trees, by Leo Breiman, Jerome, H., Friedman, Richard, A. Olshen, and Charles, J. Stone. Brooks/Cole Publishing, Monterey, 1984, 358 Pages, 27.95. *Cytometry*, 534–535.
- Mukherjee, D., Yousuf, Md, Dey, S., Chakraborty, S., Chaudhuri, A., Kumar, V., Sarkar, B., Nath, S., Hussain, A., Dutta, A., Mishra, T., Roy, B.G., Singh, S., Chakraborty, S., Adhikari, S. e Pal, C. (2020). Targeting the Trypanothione reductase of tissue-residing Leishmania in hosts' reticuloendothelial system: a flexible water soluble ferrocenylquinoline-based preclinical drug candidate, *J. Med. Chem.* 63, 15621-15638.
- Munteanu, C.R., Gonzalez-Diaz, H., Borges, F. e de Magalhaes, A.L. (2008). Natural/random protein classification models based on star network topological indices. *J. Theor. Biol.*, 254, 775–783.
- Mwakalinga, S. B., Wang, C. W., Bengtsson, D. C., Turner, L., Dinko, B., Lusingu, J. P., Arnot, D. E., Sutherland, C. J., Theander, T. G. e Lavstsen, T. (2012). Expression of a type B RIFIN in *Plasmodium falciparum* merozoites and gametes. *Malar. J.* 11, 429.
- Nagle, A., Biggart, A., Be, C., Srinivas, H. *et al.* (2020). Discovery and characterization of clinical candidate LXE408 as a Kinetoplastid-selective proteasome inhibitor for the treatment of leishmaniasis, *J. Med. Chem.* 63, 10773-10781.
- Nair, J.J. e van Staden, J. (2019). Antiplasmodial lycorane alkaloid principles of the plant family Amaryllidaceae, *Planta Med.* 85 637e647.
- Nair, J.J., van Staden, e J. Bastida, J. (2016). Apoptosis-inducing effects of Amaryllidaceae alkaloids, *Curr. Med. Chem.* 23 161e185.
- Newman, M. (2003). The Structure and Function of Complex Networks. *SIAM Rev.* 45, 167–256.
- Nirmalan, N., Flett, F., Skinner, T., Hyde, J. E. e Sims, P. F. (2007). Microscale solution isoelectric focusing as an effective strategy enabling containment of hemeoglobin-derived products for high-resolution gel-based analysis of the *Plasmodium falciparum* proteome. *J. Proteome Res.* 6 (9), 3780–7.

- Nocedo-Mena, D., Cornelio, C., Camacho-Corona, M.R., Garza-González, E., Waksman, N.H., Arrasate, S., Sotomayor, N., Lete, E. e González-Díaz, H. (2019). Modeling antibacterial activity with machine learning and fusion of chemical structure information with microorganism metabolic networks, *J. Chem. Inf. Model.* 59, 1109-1120.
- Nowotka, M.M., Gaulton, A., Mendez, D., Bento, A.P., Hersey, A. e Leach, A. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opin. Drug. Discov.* 12, 757–767.
- Ortalli, M., Varani, S., Cimato, G., Veronesi, R., Quintavalla, A., Lombardo, M., Monari, M. e Trombini, C. (2020). Evaluation of the pharmacophoric role of the O-O bond in synthetic antileishmanial compounds: comparison between 1,2-dioxanes and tetrahydropyrans, *J. Med. Chem.* 63 13140e13158.
- Ortega-Tenezaca, B., Quevedo-Tumaili, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C.R., Cordeiro, M.N.D.S. e González-Díaz, H. (2020). PTML multi-label algorithms: models, software, and applications, *Curr. Top. Med. Chem.* 20 2326-2337.
- Ortiz de Elguea, V., Sotomayor, N. e Lete, E. (2015). Two consecutive Pd(II)-promoted C-H alkenylation reactions for the synthesis of substituted 3-alkenylquinolones, *Adv. Synth. Catal.* 357, 463-473.
- Osorio, E.J., Robledo, S.M. e Bastida, J. (2008). Alkaloids with antiprotozoal activity, in: G.A. Cordell (Ed.), *The Alkaloids*, vol. 66, Elsevier, San Diego, 113-190.
- Papadatos, G. e Overington, J.P. (2014). The ChEMBL database: A taster for medicinal chemists. *Future Med. Chem.*, 6, 361–364.
- Parikh, R., Mathai, A., Parikh, S., Chandra S., G, Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology.* 56(1): 45–50.
- Patle, A. e Chouhan, D.S. (2013). SVM Kernel functions for classification. In 2013 International Conference on Advances in Technology and Engineering (ICATE), IEEE: New York, NY, USA, pp. 1–9.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. e Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49, 1373–1379.

- Pereira, M.F., Rochais, C. e Dallemagne, P. (2015). Recent advances in phenanthroindolizidine and phenanthroquinolizidine derivatives with anti-cancer activities, *Anti Canc. Agents Med. Chem.* 15, 1080-1091.
- Ping, Y., Li, Y., Zhu, J. e Kong, W. (2019). Construction of quaternary stereocenters by palladium-catalyzed carbopalladation-initiated cascade reactions, *Angew. Chem. Int. Ed.* 58, 1562e1573.
- Pogany, P., Arad, N., Genway, S. e Pickett, S.D. (2019). De novo molecule design by translating from reduced graphs to SMILES. *J. Chem. Inf. Model.*, 59, 1136–1146.
- Ponte-Sucre, A., Gulder, T., Wegehaupt, A., Albert, C., Rikanović, C., Schaefflein, L., Frank, A., Schultheis, M., Unger, M., Holzgrabe, U., Bringmann, G. e Moll, H. (2009). Structure-activity relationship and studies on the molecular mechanism of leishmanicidal N,C-coupled arylisoquinolinium salts. *J. Med. Chem.* 52, 626-636.
- Prabhu, G., Agarwal, S., Sharma, V., Madurkar, S.M., Munshi, P., Singh, S. e Sen, S. (2015). A natural product based DOS library of hybrid systems, *Eur. J. Med. Chem.* 95, 41-48.
- Prado-Prado, F., García-Mera, X., Abeijón, P., Alonso, N., Caamaño, O., Yáñez, M., Gárate, T., Mezo, M., González-Warleta, M., Muiño, L., *et al.* (2011). Using entropy of drug and protein graphs to predict FDA drug-target network: Theoretic-experimental study of MAO inhibitors and hemoglobin peptides from *Fasciola hepatica*. *Eur. J. Med. Chem.* 46, 1074–1094.
- Proulx, S. R., Promislow, D. E. L. e Phillips, P. C. (2005). Network thinking in ecology and evolution (PDF). *Trends in Ecology and Evolution.* 20 (6): 345–353.
- Pundir, S., Martin, M.J. e O'Donovan, C. (2017). UniProt Protein Knowledgebase. *Methods Mol. Biol.*, 1558, 41–55.
- Quevedo-Tumaili, V.F., Ortega-Tenezaca, B., e Gonzalez-Diaz, H. (2018). Chromosome gene orientation inversion networks (GOINs) of *Plasmodium* proteome. *J. Proteome Res.*, 17, 1258–1268.
- Quevedo-Tumaili, V., Ortega-Tenezaca B., González-Díaz, H. (2021). IFPTML Mapping of Drug Graphs with Protein and Chromosome Structural Networks vs. Pre-Clinical Assay Information for Discovery of Antimalarial Compounds. *Int. J. Mol. Sci.* 2021, 22(23), 13066.
- Ran, T., Liu, Y., Li, H., Tang, S., He, Z., Munteanu, C.R., González-Díaz, H., Tan, Z. e Zhou, C. (2016). Gastrointestinal spatiotemporal mRNA expression of ghrelin vs growth hormone receptor and new growth yield machine learning model based on perturbation theory. *Sci. Rep.* 6, 30174.

- Randic M., Zupan J., Vikić-Topić D. (2007). On representation of proteins by star-like graphs. *J Mol Graph Model.* 290-305.
- Rangwala, S. H., Kuznetsov, A., Ananiev, V., Asztalos, A., Borodin, E., Evgeniev, V., Schneider, V. A., *et al.* (2021). Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). *Genome Research*, 31(1), 159-169.
- Ranjan, R., Chugh, M., Kumar, S., Singh, S., Kanodia, S., Hossain, M. J., Korde, R., Grover, A., Dhawan, S., Chauhan, V. S., Reddy, V. S., Mohammed, A. e Malhotra, P. (2011). Proteome analysis reveals a large merozoite surface protein-1 associated complex on the *Plasmodium falciparum* merozoite surface. *J. Proteome Res.*, 10 (2), 680–91.
- Riera-Fernandez, P., Munteanu, C. R., Escobar, M., Prado-Prado, F., Martín-Romalde, R., Pereira, D., Villalba, K., Duardo-Sanchez, A. e Gonzalez-Diaz, H. (2012). New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. *J. Theor. Biol.*, 293, 174–88.
- Roth, A.E. (1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, Cambridge University Press: Cambridge, UK.
- Ruiz-Blanco, Y.B., Marrero-Ponce, Y., Paz, W., Garcia, Y. e Salgado, J. (2013). Global stability of protein folding from an empirical free energy function. *J. Theor. Biol.*, 321, 44–53.
- Ruiz-Blanco, Y.B., Paz, W., Green, J. e Marrero-Ponce, Y. (2015). ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinform.* 16, 162.
- Sak, H., Senior, A., Beaufays, F. (2014). Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling (PDF).
- Sam-Yellowe, T. Y., Florens, L., Wang, T., Raine, J. D., Carucci, D. J., Sinden, R. e Yates, J. R. (2004). 3rd Proteome analysis of rhoptry-enriched fractions isolated from *Plasmodium* merozoites. *J. Proteome Res.* 3 (5), 995–1001.
- Sander, E. L., Wootton, J. T. e Allesina, S. (2017). Ecological Network Inference From Long-Term Presence-Absence Data. *Sci. Rep.* 7 (1), 7154.
- Santana, R., Zuluaga, R., Gañán, P., Arrasate, S., Onieva, E. e González-Díaz, H. (2020). PTML model of ChEMBL compounds assays for vitamin derivatives, *ACS Comb. Sci.* 22 129-141.

- Santana, R., Zuluaga, R., Gañán, P., Arrasate, S., Onieva, E. e González-Díaz, H. (2020). Predicting coated-nanoparticle drug release systems with perturbation- theory machine learning (PTML) models, *Nanoscale*, 12 13471-13483.
- Santana, R., Zuluaga, R., Gañán, P., Arrasate, S., Onieva, E. e González-Díaz, H. (2019). Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale*, 11, 21811–21823.
- Sastry, G.M., Inakollu, V.S. e Sherman, W. (2013). Boosting virtual screening enrichments with data fusion: Coalescing hits from two-dimensional fingerprints, shape, and docking. *J. Chem. Inf. Model.*, 53, 1531–1542.
- Schwing, A., Pomares, C., Majoor, A., Boyer, L., Marty, P. e Michel, G. (2019). Leishmania infection: misdiagnosis as cancer and tumor-promoting potential, *Acta Trop.* 197, 104855.
- Shannon, C.E. A (1948). Mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423.
- Sharma, K., Gopalakrishnan, B., Chakrabarti, A. S. e Chakraborti, A. (2017). Financial fluctuations anchored to economic fundamentals: A mesoscopic network approach. *Sci. Rep.*, 7 (1), 8055.
- Staderini, M., Piquero, M., Abnegózar, M.A., Nachér-Vázquez, M., Romanelli, G., López-Alvarado, P., Rivas, L., Bolognesi, M.L. e Menéndez, J.C. (2019). Structure-activity relationships and mechanistic studies of novel mitochondria-targeted, leishmanicidal derivatives of the 4- aminostyrylquinoline scaffold, *Eur. J. Med. Chem.* 171 38-53.
- Sundar, S. e Chakravarty, J. (2013). Leishmaniasis: an update of current pharmaco-therapy, *Expet Opin. Pharmacother.* 14, 53-63.
- Sundar, S. e Singh, A. (2018). Chemotherapeutics of visceral leishmaniasis: present and future developments, *Parasitology* 145, 481-489.
- Swain, P.H. e Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Trans. Geosci. Electron.* 15, 142–147.
- Swets, J., 1996. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Lawrence Erlbaum Associates, Inc.
- Taylor, M.C., Kelly, J.M., Chapman, C.J., Fairlamb, A.H. e Miles, M.A. (1994). The structure, organization, and expression of the *Leishmania donovani* gene encoding Trypanothione reductase, *Mol. Biochem. Parasitol.* 64 293-301.

- Tenorio-Borroto, E., Ramirez, F.R., Speck-Planche, A., Cordeiro, M.N.D.S., Luan, F. e Gonzalez-Diaz, H. (2014). QSPR and Flow Cytometry Analysis (QSPR-FCA): Review and new findings on parallel study of multiple interactions of chemical compounds with immune cellular and molecular targets. *Curr. Drug Metab.* 15, 414–428.
- Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. e Villa, A.E. (2001). Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds. *J. Chem. Inf. Comput. Sci.*, 41, 246–252.
- Thomas, M.G., De Rycker, M., Ajakane, M., Albrecht S., *et al.* (2019). Identification of GSK3186899/DDD853651 as a preclinical development candidate for the treatment of visceral leishmaniasis, *J. Med. Chem.* 62, 1180e1202.
- Thomas, M.G., De Rycker, M., Wall, R.J., Spinks, D., Epemolu, O., Manthri, S., Norval, S., Osuna-Cabello, M., Patterson, S., Riley, J., Simeons, F.R.C., Stojanovski, L., Thomas, J., Thompson, S., Naylor, C., Fiandor, J.M., Wyatt, P.G., Marco, M., Wyllie, S., Read, K.D., Miles, T.J. e Gilbert, I.H. (2020). Identification and optimization of a series of 8-hydroxy naphthyridines with potent “*in vitro*” antileishmanial activity: initial SAR and assessment of *in vivo* activity, *J. Med. Chem.* 63 9523-9539.
- Tilley, L. e Rosenthal, P.J. (2019) Malaria parasites fine-tune mutations to resist drugs. *Nature*, 576, 217–219.
- Todeschini, R., Consonni, V., (2000). Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry. Wiley. ISBN 978-3-527-29913-3.
- Toropov, A.A. e Benfenati, E. (2007). SMILES as an alternative to the graph in QSAR modelling of bee toxicity. *Comput. Biol. Chem.*, 31, 57–60.
- Toropova, A.P. e Toropov, A.A. (2019). Quasi-SMILES: Quantitative structure-activity relationships to predict anticancer activity. *Mol. Divers.*, 23, 403–412.
- Tuomisto, H. (2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*. 33: 2–22.
- UniProt Consortium. (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, 46, 2699.
- UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47, D506–D515.
- UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, 49(D1), D480-D489.

- Upadhyay, A., Chandrakar, P., Gupta, S., Parmar, N., Singh, S.K., Rashid, M., Kushwaha, P., Wahajuddin, M., Sashidhara, K.V. e Kar, S. (2019). Synthesis, biological evaluation, structure-activity relationship, and mechanism of action studies of quinoline-metronidazole derivatives against experimental visceral leishmaniasis. *J. Med. Chem.* 62, 5655-5671.
- Valdes-Martini, J.R., Marrero-Ponce, Y., Garcia-Jacas, C.R., Martinez-Mayorga, K., Barigye, S.J., Vazd'Almeida, Y.S., Pham-The, H., Perez-Gimenez, F. e Morell, C.A. (2017). QuBiLS-MAS, open source multiplatform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *J. Cheminform.* 9, 35.
- Vazquez-Prieto, S., Gonzalez-Diaz, H., Paniagua, E., Vilas, R. e Ubeira, F. M. A (2014). QSPR-like model for multilocus genotype networks of *Fasciola hepatica* in Northwest Spain. *J. Theor. Biol.*, 343, 16–24.
- Veselinovic, A.M., Milosavljevic, J.B., Toropov, A.A. e Nikolic, G.M. (2013). SMILES-based QSAR model for arylpiperazines as high-affinity 5-HT(1A) receptor ligands using CORAL. *Eur. J. Pharm. Sci.*, 48, 532–541.
- Wassermann, A.M. e Bajorath, J. (2011). BindingDB and ChEMBL: Online compound databases for drug discovery. *Expert Opin. Drug Discov.* 6, 683–687.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36.
- Whittle, M., Gillet, V.J., Willett, P. e Loesel, J. (2006). Analysis of data fusion methods in virtual screening: Similarity and group fusion. *J. Chem. Inf. Model.* 46, 2206–2219.
- Whittle, M., Gillet, V.J., Willett, P. e Loesel, J. (2006). Analysis of data fusion methods in virtual screening: Theoretical model. *J. Chem. Inf. Model.* 46, 2193–2205.
- Willett, P. (2013). Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.* 53, 1–10.
- Willighagen, E.L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A.J., Tkachenko, V., Hastings, J., Chen, B. e Wild, D.J. (2013). The ChEMBL database as linked open data. *J. Cheminform.* 5, 23.
- Witten, H., Frank, E., Hall, A., Pal, J. (2011). Minería de datos: herramientas y técnicas prácticas de ML, 3ª edición. Morgan Kaufmann, San Francisco (CA).
- Wolfsberg, T. G. (2011). Using the NCBI Map Viewer to browse genomic sequence data. *Curr Protoc Hum Genet*, Chapter 18, Unit18 15.

- World Health Organization (WHO)/Department of Control of Neglected Tropical Diseases. (2014). Leishmaniasis in high-burden countries: an epidemiological update based on data reported in. https://www.who.int/leishmaniasis/resources/who_wer9122/en/. (Accessed 10 February 2021).
- Wyllie, S., Thomas, M., Patterson S., *et al.* (2018). *Nature* 560 192e197.
- Yang, Y., Luo, T., Li, Z., Zhang, X. e Yu, P. S. (2017). A Robust Method for Inferring Network Structures. *Sci. Rep.* 7 (1), 5221.
- Zhan, G., Zhou, J., Liu, J., Huang, J., Zhang, H., Liu, R. e Yao, G., (2017). Acetylcholinesterase inhibitory alkaloids from the whole plants of *zephyranthescarinata*, *J. Nat. Prod.* 80, 2462e2471.
- Zhang, S., Golbraikh, A. e Tropsha, A. (2006). Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J. Med. Chem.* 49, 2713–2724.
- Zhao, L., Pi, L., Qin, Y., Lu, Y., Zeng, W., Xiang, Z., Qin, P., Chen, X., Li, C., Zhang, Y., *et al.* (2019). Widespread resistance mutations to sulfadoxine-pyrimethamine in malaria parasites imported to China from Central and Western Africa. *Int. J. Parasitol. Drugs Drug Resist.* 12, 1–6.
- Zheng, S., Yan, X., Yang, Y. e Xu, J. (2019). Identifying structure-property relationships through SMILES syntax analysis with self-attention mechanism. *J. Chem. Inf. Model.* 59, 914–923.