Implementación y Optimización de Algoritmos para Aprendizaje Automático con Teoría de Perturbaciones.

Autor: Delfín Bernabé Ortega Tenezaca

Tesis doctoral UDC / 2022

Director: Prof. Dr. Cristian Robert Munteanu - UDC

Co-Director: Dra. Aliuska Duardo Sánchez - Ikerdata

Programa de doctorado en Tecnologías de la Información y las Comunicaciones

UNIVERSIDADE DA CORUÑA

UNIVERSIDADE DA CORUÑA

Prof. Dr. D. Cristian Robert Munteanu, Profesor Titular del Departamento 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, España.

Dna. Aliuska Duardo Sánchez, Legal Affairs. Director, IKERDATA S.L., ZITEK, Universidad del País Vasco, UPV/EHU, Rectorate Building, N0 6, Leioa, España.

HACEN CONSTAR QUE:

La memoria de investigación **"IMPLEMENTACIÓN Y OPTIMIZACIÓN DE ALGORITMOS PTML CON PROGRAMACIÓN PARALELA"** ha sido realizada por **D. DELFIN BERNABE ORTEGA TENEZACA**, bajo nuestra dirección en el Programa de doctorado en TECNOLOGÍAS DE LA INFORMACIÓN Y LASCOMUNICACIONES, y constituye la Tesis que presenta para optar al Grado de Doctora de la Universidade da Coruña.

A Coruña, 19 de junio de 2022

Prof. Dr. Cristian Robert Munteanu	**Dra. Aliuska Duardo Sánchez**
Director de Tesis	**Co-Director de Tesis**

A mi esposa y a mis hijos

a mi madre

a mis hermanas y querida familia

AGRADECIMIENTOS

En primer lugar, deseo expresar mi agradecimiento a los directores de esta tesis los Profesores Dr. Cristian Robert Munteanu y Dra. Aliuska Duardo Sánchez, por toda la ayuda y dedicación que han brindado a este trabajo. Asimismo, quiero agradecer a sus colaboradores de la Universidade da Coruña, especialmente al grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos – Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR). También agradecer a sus colaboradores de la Universidad del País Vasco, especialmente al Departamento de Química Orgánica e Inorgánica.

Deseo expresar también todo mi agradecimiento a mi madre, a mis hermanas, a mi esposa y a mis hijos por su apoyo incondicional a lo largo de este trayecto y por su confianza.

A mis 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

Na actualidade acumúlanse unha inxente cantidade de datos relacionados con sistemas complexos de moi variada natureza: biomoleculares, económicos, sociais, etc. Estes sistemas son de gran relevancia en diferentes áreas como as ciencias biomoleculares, a enxeñaría biomédica e as ciencias sociais e xurídicas. As técnicas de Intelixencia Artificial (IA) e/ou Machine Learning (ML) poden ser útiles para predicir propiedades de interese nestes sistemas. Para iso, son necesarios polo menos dous pasos principais. O primeiro refírese a recoller información similar de moitos casos de sistemas coñecidos para poder adestrar modelos de IA/ML. O segundo paso indispensable está relacionado coa cuantificación numérica da información estrutural, as condicións externas ao sistema e as propiedades do mesmo a predicir. Neste segundo paso, defínense as variables numéricas de entrada e saída para adestrar os algoritmos AI/ML. Desafortunadamente, os sistemas complexos están formados xeralmente por varios subsistemas, e a información sobre o sistema no seu conxunto ou as súas partes non se pode atopar na mesma fonte. Non obstante, é habitual atopar información sobre cada un dos subsistemas e as súas propiedades en diversas fontes dispersas. Para resolver este problema, desenvolveuse o algoritmo NIFPTML = NI + IF + PT + ML. Estes algoritmos implican as seguintes etapas. Na etapa NI (Network Invariant) utilízanse redes complexas para representar diferentes sistemas e/ou os seus subsistemas e calcúlanse os invariantes destas redes para cuantificar a súa estrutura. Na seguinte etapa, é necesario utilizar técnicas de fusión de información (IF) de diversas fontes para obter un conxunto de datos enriquecido. Posteriormente, os operadores da Teoría da Perturbación (PT) procesan a información cuantificando as perturbacións/desviacións nas variables estruturais con respecto aos valores esperados para diferentes subconxuntos de variables categóricas.

Finalmente, en Machine Learning (ML), adestran diferentes algoritmos de IA/ML, que permiten atopar modelos preditivos. Alí aplicáronse os algoritmos NIFPTML e os resultados publicáronse na literatura. Desafortunadamente, non hai unha aplicación de software amigable para os usuarios habituais destes algoritmos. Polo tanto, os desenvolvedores de algoritmos NIFPTML necesitan utilizar varias ferramentas diferentes para cada unha das etapas. Por outra banda, hai un descoñecemento das implicacións legais do desenvolvemento de algoritmos computacionais como o NIFPTML na investigación científica nestas áreas. Nesta tese propoñemos desenvolver (programar) unha versión beta dun software, SOFT.PTML, no que se implementan por primeira vez os algoritmos NIFPTML nunha mesma aplicación. Ademais, demostrarase a utilidade deste programa aplicado a diferentes problemas prácticos dos ámbitos mencionados, como son: o deseño de fármacos, o descubrimento de nanomateriais, o estudo dos ordenamentos xurídicos. Por último, realizarase unha análise das implicacións legais do desenvolvemento e aplicación deste tipo de algoritmos na investigación.

RESUMEN

En la actualidad se ha acumulado una ingente cantidad de datos relacionados con sistemas complejos de muy variada índole: biomoleculares, económicos, sociales, etc. Estos sistemas son de gran relevancia en diferentes áreas como las ciencias biomoleculares, ingeniería biomédica, y las ciencias sociales y jurídicas. Las técnicas de Inteligencia Artificial (IA) y/o Machine Learning (ML) pueden ser útiles para predecir propiedades de interés de estos sistemas. Para ello se necesitan al menos dos pasos principales. El primero se refiere a recopilar información similar de muchos casos de sistemas conocidos para poder entrenar los modelos AI/ML. El segundo paso indispensable está relacionado con la cuantificación numérica de información estructural, de las condiciones externas al sistema, y de las propiedades del mismo a ser predichas. En este segundo paso se definen las variables numéricas de entrada y salida para entrenar los algoritmos AI/ML. Desafortunadamente los sistemas complejos están compuestos por lo general por varios subsistemas no encontrándose información del sistema como un todo o de sus partes en la misma fuente. No obstante, si es habitual encontrar en varias fuentes dispersas información sobre cada uno de los subsistemas y sus propiedades. Para resolver esta problemática se ha desarrollado el algoritmo NIFPTML = NI + IF + PT + ML. Estos algoritmos involucran las siguientes etapas. En la etapa NI (Network Invariant) se usan redes complejas para representar distintos sistemas y/o sus subsistemas y se calculan las invariantes de estas redes para cuantificar su estructura. En la siguiente etapa es necesario utilizar técnicas de Fusión de Información (IF) de diversas fuentes para obtener un conjunto de datos enriquecido. Posteriormente 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. Por último, en la etapa de Aprendizaje Automático (ML) se entrenan distintos algoritmos AI/ML permitiendo encontrar modelos predictivos.

Los algoritmos NIFPTML han sido ampliamente utilizados y los resultados publicados en la literatura científica. Desafortunadamente, no existe una aplicación de software de fácil manejo (user-friendly) para los usuarios habituales de estos algoritmos. Por lo tanto, los desarrolladores de algoritmos NIFPTML necesitan utilizar varias herramientas diferentes para cada una de las etapas. Por otra parte, existe desconocimiento de las implicaciones jurídicas del desarrollo de algoritmos computacionales como los NIFPTML en investigación científica en estas áreas.

En esta tesis nos proponemos desarrollar (programar) una versión beta de un software, al que hemos llamado SOFT.PTML, en el que se implementan por primera vez algoritmos NIFPTML en una misma aplicación. Además, se demostrará la utilidad de este programa aplicándolo a distintos problemas prácticos en las áreas mencionadas: diseño de fármacos, descubrimiento de nanomateriales, estudio de sistemas jurídicos. Por último, se aportará un análisis de las implicaciones jurídicas del desarrollo y aplicación de este tipo de algoritmos en investigación.

ABSTRACT

Currently, a huge amount of data related to complex systems of a very varied nature has been accumulated: biomolecular, economic, social, etc. These systems are of great relevance in different areas such as biomolecular sciences, biomedical engineering, and social and legal sciences. The techniques of Artificial Intelligence (AI) and/or Machine Learning (ML) can be useful to predict properties of interest in these systems. For this, at least two main steps are needed. The first refers to collecting similar information from many cases of known systems to be able to train AI/ML models. The second indispensable step is related to the numerical quantification of structural information, the conditions external to the system, and the properties of the same to be predicted. In this second step, the numeric input and output variables are defined to train the AI/ML algorithms. Unfortunately, complex systems are generally made up of various sub-systems, and information about the system as a whole or its parts cannot be found in the same source. However, it is common to find information on each of the sub-systems and their properties in various scattered sources. To solve this problem, the algorithm NIFPTML = NI + IF + PT + ML has been developed. These algorithms involve the following stages. In the NI stage (Network Invariant) complex networks are used to represent different systems and/or their sub-systems and the invariants of these networks are calculated to quantify their structure. In the following stage, it is necessary to use Information Fusion (IF) techniques from various sources to obtain an enriched set of data. Later, the operators of the Perturbation Theory (PT) process the information by quantifying the perturbations/deviations in the structural variables with respect to the expected values for different subsets of categorical variables. Finally, in Machine Learning (ML), different AI/ML algorithms are trained, allowing predictive models to be found.

The NIFPTML algorithms have been applied there and the results published in the literature. Unfortunately, there is no user-friendly software application for regular users of these algorithms. Therefore, the developers of NIFPTML algorithms need to use several different tools for each of the stages. On the other hand, there is a lack of knowledge of the legal implications of the development of computational algorithms such as the NIFPTML in scientific research in these areas.

In this thesis we propose to develop (program) a beta version of a software, SOFT.PTML, in which NIFPTML algorithms are implemented for the first time in the same application. In addition, the usefulness of this program applied to different practical problems in the aforementioned areas will be demonstrated, such as: the design of drugs, the discovery of nanomaterials, the study of legal systems. Lastly, an analysis of the legal implications of the development and application of this type of algorithm in research will be provided.

INDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE ABREVIATURAS

AAE: Average Atomic Electronegativity

AAP: Average Atomic Polarizability

Ac: Accuracy

AMV: Average Molar Volume

APS: Average Particle Size

AUROC: Receiver Operating Characteristic Curve

FRAMA: Framework Moving Average

GRNs: Gene Regulatory Networks

HTS: High-Throughput Screening

IDE: Integrated Development Environment

IF: Information Fusion

IFPTML: Information Fusion Perturbation Theory Machine Learning

KDA: Kernel Discriminant Analysis

LAGA:

LDA: Linear Discriminant Analysis

LOGR: Logistic Regression

MA: Moving Average

MMA: Media Moving Average

ML: Machine Learning

NCBI: National Center for Biotechnology Information

MNs: Metabolic Reaction Networks

NPs: Nanopartícles

PINs: Protein Interaction Networks

PT: Perturbation Theory

PTO: Perturbation Theory Operators

PTML: Perturbation Theory Machine Learning

QSAR: Quantitative Structure-Activity Relationship

QSAR-Co: Cheminformatics with conditions

RF: Random Forest

Sn: Sensitivity

Sp: Specificity

SVM: Super Vector Machine

UnitProt: Universal Protein Resource.

INTRODUCCIÓN

INTRODUCCIÓN

1.1. Artículo 1 (Revisión - Introducción)

Ref. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C. R., Cordeiro, M., & González-Díaz, H. (2020). PTML Multi-Label Algorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337.

La modelización de QSAR es un enfoque computacional ampliamente utilizado que tiene como objetivo predecir la/s respuesta/s final/es (por ejemplo, la actividad, la propiedad o la toxicidad) de las sustancias químicas sobre la base de sus características de codificación (descriptores), y está desempeñando un papel cada vez más importante en el diseño de fármacos o materiales.

Cualquier valor de respuesta de un compuesto químico puede variar considerablemente cuando se determina utilizando diferentes protocolos experimentales o cuando se aplica el mismo protocolo experimental, pero en diferentes condiciones, como de laboratorio, ambientales, temporales, e incluso si se emplean diferentes medidas biológicas como IC_{50}, EC_{50}, K_i, *etc.* (Kalliokoski *et al.*, 2013; Eriksson *et al.*, 2003). Determinar el valor de respuesta de nuevos compuestos químicos es una tarea especialmente importante en Química Médica, pero, al mismo tiempo, muy exigente tanto en términos de tiempo como de recursos. En la actualidad, se realizan estudios sobre modelos quimio-informáticos para predecir las propiedades fisicoquímicas de pequeñas moléculas orgánicas, proteínas, proteomas y sistemas complejos. Es útil para reducir el tiempo y los recursos de investigación en los laboratorios. Diferentes autores han aplicado la combinación de PT (Teoría de Perturbación ó Perturbation Theory en idioma

inglés), y ML (Aprendizaje Automático ó Machine Learning en idioma inglés) para obtener modelos PTML (Teoría de Perturbación con Aprendizaje Automático ó Perturbation Theory Machine Learning en idioma inglés) sobre sistemas biológicos(Arrasate & Duardo-Sanchez, 2018).

En este caso, cabe destacar la base de datos ChEMBL, que hoy en día es un recurso muy reconocido en el campo del descubrimiento de fármacos y la investigación en química medicinal. De hecho, esta base de datos conserva y almacena datos estandarizados de bioactividad, moléculas, objetivos y fármacos recuperados de múltiples fuentes, así como de la literatura primaria de química medicinal (Davies *et al.*, 2015). Incluye, además, múltiples condiciones de ensayos, como diferentes parámetros experimentales, ensayos biológicos, proteínas diana, líneas celulares, organismos de ensayo*, etc*. Otras bases de datos que existen y comprenden dicha información son NCBI y UnitProt, ambas permiten fusionar su información con la procedente de ChEMBL en un conjunto de datos para un objeto de estudio. UniProt, por ejemplo, es un recurso completo de datos de secuencias y anotaciones de proteínas que actúan sobre los fármacos (Pundir *et al.*, 2016). Por otra parte, el NCBI proporciona un amplio conjunto de recursos en línea para la información y los datos biológicos, incluida la base de datos de secuencias de ácidos nucleicos GenBank y la base de datos de citas PubMed, y los resúmenes de las revistas de ciencias de la vida publicadas (NCBI Resource Coordinators, 2016).

Se analizan varios estudios recientes que han aplicado herramientas como el modelado de PT, técnicas de ML y la técnica de IF (Fusión de Información ó Information Fusion en idioma inglés). Hay que tener en cuenta que estas herramientas se pueden utilizar de forma independiente o combinada para resolver un problema particular de tipo combinatorio. Normalmente, se recurre a las siguientes combinaciones: PTML (PT + ML) o IFPTML (IF + PT

+ ML). Esto permite realizar un estudio racional de datos complejos para extraer relaciones útiles y predecir nuevas sustancias químicas. El modelado PT, por ejemplo, permite predecir la/s respuesta/s de un punto final de una consulta relativa a un compuesto químico o sistema de materiales bajo múltiples condiciones experimentales y/o teóricas, basándose en la/s respuesta/s de un sistema de referencia conocido (Ferreira da Costa *et al.*, 2018). Para ello, la PT se combina con el enfoque de MA (Media Móvil o Moving Average en idioma inglés) de Box-Jenkins, fusionando las características únicas de los sistemas, y simplificando las dificultades para manejar toda la información. En cuanto a las herramientas de ML, éstas se han utilizado en la investigación de fármacos o materiales desde los años 90, proporcionando soluciones rápidas y precisas a una gran cantidad de problemas. En cuanto a la combinación de estas últimas, es decir, el modelado predictivo PTML se han aplicado ampliamente en química médica, proteómica, nanotecnología*, etc*., para hacer frente a grandes conjuntos de datos heterogéneos con numerosas características (Blazquez-Barbadillo *et al.*, 2016; Casañola-Martin *et al.*, 2016; Romero-Durán *et al.*, 2016; Kleandrova *et al.*, 2014; Luan F. *et al.*, 2014; Alonso *et al.*, 2013).

Recientemente, se han lanzado tres soluciones de software para automatizar el proceso de obtención de modelos mediante PT, ML e IF, a saber, QSAR-Co (Ambure *et al.*, 2019), LAGA y FRAMA (Bernabe Ortega-Tenezaca & González-Díaz, 2017). QSAR-Co es un software útil para abordar algunas de las cuestiones críticas que suelen descuidarse durante el desarrollo de modelos robustos basados en la clasificación de múltiples objetivos. LAGA es un software desarrollado para el diseño de fármacos recurriendo tanto a la teoría de la perturbación como a las técnicas de aprendizaje automático. FRAMA ha sido desarrollado para permitir el cálculo de descriptores y la configuración de condiciones multietiqueta para resolver varios problemas de diseño (Ambure *et al.*, 2019; Bernabe Ortega-Tenezaca & González-Díaz, 2017). Estos últimos

incluyen, por ejemplo, la simplificación del análisis de cualquier conjunto de datos con un gran número de características, el cribado de la actividad de nuevos compuestos químicos, *etc*.

En general, se plantea como objetivo, la exploración de los avances realizados en los últimos años, sobre las técnicas utilizadas para establecer modelos predictivos PT/ML/IF para la química medicinal u otras aplicaciones, prestando especial atención a la disponibilidad de software de fácil uso que permita agilizar su utilización.

Objetivos

1.2. Objetivos

1.2.1. En esta tesis nos proponemos desarrollar (programar) una versión beta de un software -SOFT.PTML-, en el que se implementan por primera vez algoritmos NIFPTML en una misma aplicación.

1.2.2. Demostrar la utilidad del programa desarrollado aplicándolo a distintos problemas prácticos en las áreas diversas como: el diseño de fármacos, el descubrimiento de nanomateriales, y el estudio de sistemas jurídicos complejos.

1.2.3. Analizar las implicaciones jurídicas del desarrollo y aplicación de este tipo de algoritmos en investigación científica.

Hipótesis de trabajo

1.3. Hipótesis de trabajo

El desarrollo (programación) de una versión beta de un software -SOFT.PTML-, en el que se implementen por primera vez algoritmos NIFPTML en una misma aplicación, facilitará el trabajo a los usuarios/desarrolladores de estos algoritmos a campos como el diseño de fármacos, descubrimiento de nanomateriales, estudio de sistemas jurídicos. Además, el análisis de las implicaciones jurídicas del desarrollo y aplicación de este tipo de algoritmos en investigación permitirá su desarrollo y uso acorde a las exigencias de la legislación vigente.

Fundamentos teóricos

FUNDAMENTOS TEÓRICOS

1.4. Artículo 1 (Revisión - Fundamentos)

Algoritmos multietiqueta PTML: Modelos, software y aplicaciones Resultados y Discusión

Ref. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C., Cordeiro, M., & González-Díaz, H. (2020). PTML Multi-Label Algorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337. doi: https://doi.org/10.2174/1568026620666200916122616

1.5. Metodología general del PTML

La metodología general se desarrolla principalmente en dos etapas. La primera etapa comprende el preprocesamiento de los datos. Después de recuperar la información de interés de una base de datos, se preprocesa según criterios valiosos. La segunda etapa se refiere a la aplicación de técnicas de modelización. Esto es útil para buscar modelos predictivos para conjuntos de datos complejos con múltiples características de Big Data. Finalmente, la metodología permite desarrollar modelos lineales PTML para predecir la actividad biológica o clasificar compuestos como activos o no en actividad biológica*, etc*. (Bediaga *et al.*, 2018).

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

La salida del modelo $f(v_{ij})_{calc}$ es una función de puntuación del valor v_{ij} de la actividad biológica del i-ésimo fármaco en las diferentes combinaciones y condiciones de ensayo c_j. La

tabla 1 muestra los tipos de operadores MA, detallando las condiciones que se utilizan en cada uno, así como los símbolos, la fórmula del operador y la información del mismo.

Tabla 1. Operadores PT utilizados a menudo como entradas en los modelos PTML

Tipos de MA[a]	Condiciones[a] (c_j)	Símbolo	Operador de Fórmula	Información de operador
Dif.	-	ΔD_{kij}	$\Delta D_k = D_{ki} - D_{kj}$	Obtiene la diferencia entre dos descriptores.
MA	Monoescala	$\Delta D_1(c_0)$	$D_{ki} - \langle D_{ki}(c_0)\rangle$ . . . $D_{ki} - \langle D_{ki}(c_j)\rangle$. $D_{ki} - \langle D_{ki}(c_j)\rangle$ $[\Delta D_k(c_j)]^n$	ΔALOGP(c_j) tiene en cuenta la variabilidad en la estructura química, y las condiciones de ensayo c_j en términos de desviación (Δ) de la Hidrofobicidad del compuesto ($ALOGP_i$) respecto al valor esperado ($\langle ALOGP(c_j)\rangle$) para una condición de ensayo. Desviación (Δ) del D_1 = $ALOGP_i$ o D_2 = PSA_idel i-ésimo compuesto respecto al respectivo valor esperado ($\langle ALOGP(c_j)\rangle$) o ($\langle PSA(c_j)\rangle$) para un subconjunto dado de condiciones de ensayo múltiples c_j
MMA			$\Delta D_k(c_g) - \Delta D_{k'}(c_j') = [D_{ki} - \langle D_k(c_j)\rangle] - [D_{k'i} - \langle D_{k'}(c_j')\rangle]$	Covarianza entre las desviaciones de los descriptores k, y k' con las condiciones c_j, y c_j'.
COV			$\Delta D_k(c_g) - \Delta D_{k'}(c_j') = [D_{ki} - \langle D_k(c_j)\rangle] \cdot [D_{k'i} - \langle D_{k'}(c_j')\rangle]$	
Momentos			$[D_{ki} - \langle D_{ki}(c_0)\rangle]^q$	Los momentos miden la desviación de la media. Según la naturaleza de n (par o impar), consideran la asimetría de la desviación. El efecto de la perturbación sobre la salida depende del valor de n (cuanto mayor es n, más fuerte es la perturbación).

[a]Operadores MA han sido usados en la ecuación (2) y los operadores MMA en la ecuación (3).

1.6. Modelos PTML en química médica

Se revisan varios modelos con diferentes aplicaciones en Química Médica. La Tabla 2, resume los principales resultados de estos modelos. Todos los resultados de los parámetros estadísticos como Sensibilidad, Especificidad y Precisión de los modelos PTML revisados son superiores al 70%. El valor medio más alto es el 89,0% para el entrenamiento y la validación en

el modelo PTML propuesto por *Nocedo et al*. con un n_j máximo igual a 56377. El valor medio más bajo es del 75,1% en el modelo PTML de *Ferreira et al*. con un n_j mínimo igual a 18258; casualmente, los valores son presentados por los mismos autores en ambos casos.

1.6.1. Modelización de la actividad anticancerígena de PTML

Bediaga et al. informaron de la modelización mediante PTML de la actividad anticancerígena contra múltiples tipos de cáncer de un conjunto diverso de compuestos químicos, recuperados de la base de datos ChEMBL. El conjunto de datos utilizado incluía 35.565 compuestos químicos diferentes probados en ensayos preclínicos bajo varias condiciones experimentales (c_j): > 70 tipos de medidas de efectos biológicos (c_0), > 300 dianas farmacológicas diferentes (c_1), > 230 líneas celulares (c_2), y 5 organismos de prueba (c_4) o mapeo de dianas (c_5). En conjunto, comprendía datos procedentes de un total de 45.833 ensayos para la leucemia, 6.227 ensayos para el cáncer de mama, 2.499 ensayos para el cáncer de ovario, 3.499 para el cáncer de colon, 3.159 para el cáncer de pulmón, 2.750 para el cáncer de próstata, *etc*. Los autores optaron por un enfoque basado en la clasificación y derivaron dos tipos de modelos, a saber, un modelo lineal simple recurriendo al análisis lineal discriminante (LDA), y un modelo no lineal construido mediante redes neuronales artificiales (RNA). El modelo utilizó operadores PT basados en la media móvil multicondición para capturar toda la complejidad del conjunto de datos. La hipótesis de una relación lineal entre los operadores PT y la clasificación como compuestos anticancerígenos en diferentes combinaciones de condiciones de prueba se confirma (Bediaga *et al.*, 2018). El resultado del modelo PTML es el siguiente:

$$f(v_{ij})_{calc} = -5.9193 + 14.311 \cdot f(v_{ij})_{ref} - 0.07986 \cdot \Delta D_1(c_0) - 0.08724 \cdot \Delta D_1(c_1) - 0.00628 \cdot \Delta D_1(c_2) + 0.03577 \cdot \Delta D_1(c_4) + 0.00651 \cdot \Delta D_1(c_5) \quad (2)$$

$$N = 87701 \chi^2 = 42891.07 p < 0.05$$

Donde, $f(v_{ij})_{calc}$ es el valor de la función que puede predecir el *i*-ésimo compuesto en el *j*-ésimo ensayo preclínico con múltiples condiciones de ensayo $c_j = (c_0, c_1, c_2, \ldots, c_n)$. El modelo tiene dos tipos de variables de entrada: la función de valor esperado $f(v_{ij})_{ref}$, y los operadores $PT\Delta D_k$ similares a los operadores Box-Jenkins MA utilizados como entrada (Bediaga, Arrasate, & González-Díaz, 2018). El modelo PTML-LDA presentó valores de área bajo la curva operativa del receptor (AUROC) de 0,872, valores de especificidad (Sp) de 90,2% / 90,1%, sensibilidad = Sn de 70,6% / 71,4, y precisión general = Ac (%) = 87,7 /87,8 en series de entrenamiento / validación, respectivamente.

Tabla 2. Modelado PTML-LDA para diferentes actividades anticancerígenas.

Tipos de Cáncer	Cancers	PT[a]	ML[b]	NV[b]	Casos[c]	Sn (%)[d]	Sp (%)[d]
mama	1	MA	LDA	>10	24285 (total)	>90	>90
vejiga	1	MA	LDA	n.a.	n.a.	>90	>90
cerebro	1	MA	LDA	n.a.	n.a.	>90	>90
Colon rectal	1	MA	LDA	>10	1237 (entren.)	>90	>90
mama	1	MA	LDA	>10	2272 (total)	>85	>95
Próstata	1	MA	LDA	>10	1250 (entren.)	>85	>95
Múltiple	>10	MA	LDA	>10	87701 (entren.)	>70	90
Cáncer		MMA	LDA	3		>70	>90
			ANN	4		>80	>80

[a] Operadores PT utilizados, MA = Media Móvil, MMA = Media Móvil Multicondición. [b] Método ML utilizado, y NV = Número de variables de entrada, n.a. = no disponible. [c] Número total de casos en el entrenamiento, y/o series de validación. [d] Valores aproximados para las series de entrenamiento y validación.

Tabla 3. Resultados del modelo y variables de entrada analizadas.

Actividad Biológica	Set[a]	Set Obs.[a]	ParamStat..[a]	PredStat..[a]		Pred. Sets[a]	
					n_j	$f(v_{ij})_{pred}=0$	$f(v_{ij})_{pred}=1$
Antibacteriano	t	$f(v_{ij})_{obs}=0$	Sp	90.3	30181	27248	2933
		$f(v_{ij})_{obs}=1$	Sn	88.1	96667	11464	85203
Actividad		Total	Ac	88.7	126848		
	v	$f(v_{ij})_{obs}=0$	Sp	90.3	10030	9062	968
		$f(v_{ij})_{obs}=1$	Sn	88.1	32252	3842	28410
		Total	Ac	88.6	42282		
Antiretroviral	t	$f(v_{ij})_{pred}=1$	Sp	73.05	83748	18825	22573
Actividad		$f(v_{ij})_{pred}=0$	Sn	86.61	21735	2910	61175
		Total	Ac	75.84	105483		
	v	$f(v_{ij})_{pred}=1$	Sp	73.1	27959	20439	7520
		$f(v_{ij})_{pred}=0$	Sn	87.17	6370	92	6278
		Total	Ac	75.98	34329		
Dopamine Pathway Targets	t	$f(v_{ij})_{obs}=0$	Sp	70.1	37080	26005	11075
		$f(v_{ij})_{obs}=1$	Sn	83.9	4001	644	3357
		total	Ac	71.5	41081		
	v	$f(v_{ij})_{obs}=0$	Sp	70.2	12364	8675	3689
		$f(v_{ij})_{obs}=1$	Sn	83.3	1329	222	1107
		Total	Ac	71.4	13693		

[a]Set = Entrenamiento (t) o Validación (v). Set Obs= conjuntos observados, ParamStat = parámetro estadístico, PredStat = estadística predicha, PredSets = conjuntos predichos. [b]Sn = Sensibilidad (%), Sp = Especificidad (%), y Ac = Precisión (%).

1.6.2. Modelización PTML de la actividad antibacteriana

Nocedo-Mena *et al.*, reportaron, por primera vez, el aislamiento y caracterización de terpenos de la planta *Cissusincisa*. Se obtuvieron los resultados de la base de datos ChEMBL que

comprende 160.000 resultados de ensayos preclínicos de actividad antimicrobiana para 55931 compuestos con más de 365 parámetros de actividad biológica, 90 cepas de bacterias, 25 especies bacterianas y el conjunto de datos de Leong y Barabási que incluye 40 MRNs de microorganismos. Los investigadores combinaron el PTML con la técnica IF para desarrollar el primer modelo PTMLIF. El mejor modelo lineal encontrado, presentó valores de especificidad (%) = 90,31 / 90,40, y sensibilidad (%) = 88,14 / 88,07 en las series de entrenamiento/validación, como se muestra en la Tabla 3. La Tabla 4 muestra una comparación entre el PT-LDA obtenido, y parte de la literatura, como el modelo ANN y BLR. Finalmente, se determinó experimentalmente la actividad antibacteriana de los terpenos. Los compuestos más activos fueron el phytol, y la α-amirina con MIC = 100 μg/ml, *Enterococcusfaecium* resistente a la vancomicina, y *Acinetobacterbaumannii* resistente a los carbapenems. El modelo fue útil para predecir la actividad de estos compuestos contra otros microorganismos con diferentes NRMs para encontrar otros objetivos potenciales (Nocedo-Mena D. , *et al.*, 2019). El resultado del modelo PTMLIF es el siguiente:

$$f(v_{ij})_{calc} = -5.683 + 14.434 \cdot f(v_{ij})_{ref} - 16.426 \cdot Sh_1(Drug) + 24.818 \cdot DSh_1(Assay)_{c1} + 0.211 \cdot DSh_2(Assay)_{c1} + 1.882 \cdot DSh_1(Assay)_{c0} - 107.050 \cdot Sh_1(MRN)_{c2} + 155.395 \cdot Sh_2(MRN)_{c2} \quad (3)$$

$$n = 126848 x^2 = 122496.8 p < 0.05$$

Donde, $f(v_{ij})_{calc}$ es el valor de la función que puede predecir la actividad biológica del *i*-ésimo compuesto ensayado en el *j*-ésimo ensayo preclínico con condiciones $c_j = (c_0, c_1)$ contra la

s-ésima especie bacteriana con MRN_s. El modelo tiene cuatro tipos de variables de entrada. El primer tipo es la función de valor esperado $f(v_{ij})_{expt}$. El segundo tipo incluye los valores $Sh_k(Drug_i)$ que se utilizan para cuantificar la estructura de los compuestos químicos. Y los dos últimos tipos de operadores PT son el término $\Delta Sh_k(Assay_j)c_j$ y el término $\Delta Sh_k(MRN_s)c_j$ (Nocedo-Mena D. , *et al.*, 2019). En la Tabla 3, se ha presentado un estudio comparativo de este modelo con muchos otros modelos de la literatura, incluyendo modelos PTML y no PTML. Es muy importante destacar que este modelo es el único que es capaz de tener en cuenta la estructura del MRN del patógeno.

Tabla 4. Modelos PTML para la actividad antibacteriana vs otros modelos no PTML.

Tipo Cmpd.[a]	***N***[b]	***Var.***[b]	***Técn.***[c]	***Precis. (%)***	***Val.***[d]	***Multi especies***[e]	***Familia de fármacos***[f]	***MT***[g]	***Net.***[h]
HSC	83,605	6	LDA	88.6	i	MBS	>10	Si	Si
Péptido	3,592	4	LDA	96	i	MBS	>10	Si	No
Péptido	2,488	6	LDA	90	i	*Gram + bacteria*	>10	Si	No
HSC	30,181	6	LDA	90	i	*F. necrophorumP.intermedia*	>10	Si	No
HSC	54,000	6	ANN	90	i	*Pseudomonas spp*	>10	Si	No
Nano	300	7	LDA	77.7	i	MBS	>10	Si	No
HSC	37,800	5	LDA	95	i	No	>10	Si	No
HSC	11,576	4	ANN	97	i	*Streptococcusspp*	>10	Si	No
ATD	12,000	4	LDA	90	i	*Mycobacteriumspp*	>10	Si	No
HSC	667	7	LDA	92.9	i	No	>10	No	No
HSC	661	6	LDA	92.6	ii	No	8	No	No
HSC	661	6	BLR	94.7	ii	No	8	No	No
HSC	661	62	ANN	-	iii	No	8	No	No
HSC	352	7	LDA	91	i	No	9	No	No
HSC	111	7	LDA	94	i	No	3	No	No
HSC	111	7	ANN	89	i	No	3	No	No
HSC	-	8	LDA	>90	i	No	-	No	No
HSC	972	8	LDA	86.8	i	No	>5	No	No
HSC	458	2	LDA	~85	i	No	-	No	No
HSC	433	6	LDA	~86	i	No	>8	No	No

[a] Tipo de compuesto: HSC = Serie heterogénea de compuestos, fármaco antituberculoso = fármacos antituberculosos. [b] Número total de casos en las series de entrenamiento y/o validación, y Var. = Número de variables incluidas en el modelo. [c] Técnicas empleadas: LDA = Análisis Discriminante Lineal, ANN = Redes Neuronales Artificiales, BLR = Regresión Logística Binaria. [d] Métodos de validación: i) series de predicción externas, ii) dejar un 30% de validación cruzada, y iii) técnica de resustitución100 veces promediada. Además, nótese que los métodos ii) y iii) son métodos de validación cruzada. [e]Multi Especies: Cepa bacteriana múltiple (MBS), Fusobacteriumnecrophorum, Prevotellaintermedia. [f] Familia de fármacos: Aquí sólo se consideran las familias ampliamente representadas. [g] MT = Multi-target: Modelos que pueden predecir más de un tipo de actividad biológica

(MIC, IC_{50}, MBC*, etc*.). [h] Net. = MRNs: Modelos capaces de tener en cuenta los cambios en los MRNs de diferentes microorganismos.

1.6.3. Modelización PTML de la actividad antirretroviral

Vásquez-Domínguez *et al.*, propusieron el desarrollo de un nuevo modelo predictivo que define las proteínas diana de los nuevos compuestos antirretrovirales. ChEMBL registra más de 140.000 ensayos preclínicos experimentales de antirretrovirales (VIH, HTLV, SIV, HBV, MLV, RSV, FeLV) para 56.105 compuestos, cubriendo combinaciones con 359 parámetros de actividad biológica, 55 accesiones de proteínas, 83 líneas celulares, 64 organismos de ensayo y 773 subtipos o cepas. Incluye 150.148 ensayos experimentales preclínicos para el virus del VIH, 1.188 para el virus HTLV, 84 para el virus de la inmunodeficiencia simia, 370 para el virus de la leucemia murina, 119 para el virus del sarcoma de Rous, 1.581 para el MMTV*, etc*. Además, ha incluido 5.277 ensayos para el virus de la hepatitis B. El modelo PTML desarrollado ha alcanzado valores considerables en sensibilidad (%) 73,05/73,10, especificidad (%) 86,61/87,17, y precisión (%) 75,84/75,98 en las series de entrenamiento/validación, como se muestra en la Tabla 3,se comparan modelos alternativos de PTML con diferentes operadores de PT como covarianza, momentos exponenciales y términos(Vásquez-Domínguez *et al.*, 2019; Speck-Planche & Cordeiro, 2017; Speck-Planche & Cordeiro, 2017). El modelo desarrollado aplicado a los ARVs calcula la probabilidad de interacción de una molécula i con diferentes retrovirus bajo un conjunto de condiciones múltiples de ensayo c_j. La presencia de coinfecciones con el VIH y el VHB en los pacientes es común. El VIH prolonga la viremia del VHB, aumenta las tasas de cronicidad y también el riesgo de cirrosis y de morbilidad relacionada con el hígado. Por este motivo, el tratamiento de ambas infecciones debe coordinarse (Levy, 2006). Algunos estudios han encontrado fármacos ARV eficaces y con una actividad significativa en el tratamiento de

ciertos tipos de HBV resistentes en pacientes coinfectados por HIV/VHB (Benhamou, 2004). El grupo de Yang sugiere que, en caso de coinfección, la terapia antirretroviral debería incluir agentes con actividad contra el VIH y el VHB (Yang *et al.*, 2014). Los operadores de PT multicondición se calcularon utilizando medias móviles combinatorias o múltiples (MMA). La ecuación del modelo PTML con LDA se describe a continuación:

$$f(v_{ij})_{calc} = -16.6473 + 10.6828 \cdot f(v_{ij})_{ref} + 5.4195 \cdot D_1 - 5.0349 \cdot \Delta D_1(c_j) + 0.0512 \cdot \Delta D_2(c_j) \quad (4)$$

$$N = 140{,}644 \chi^2 = 37{,}710.77 p < 0.05$$

Donde $f(v_{ij})_{calc}$ es el valor de la función que calcula la probabilidad de interacción de una molécula i con diferentes retrovirus bajo un conjunto de condiciones múltiples de ensayo c_j aplicadas a los tratamientos ARV. El modelo tiene tres tipos de variables de entrada. El primer tipo es la función de valor de referencia $f(v_{ij})_{ref}$ que representa el valor de la actividad biológica de la molécula *m* bajo el subconjunto de condiciones múltiples c_j. El segundo y el tercer tipo, es decir, ΔD_k, y $\Delta D_k(c_j)$ efectos de perturbación en la estructura de la molécula se añaden a la ecuación (Vásquez-Domínguez *et al.*, 2019).

1.7. Modelos PTML de fármacos dirigidos a las vías dopaminérgicas

Ferreira da Costa *et al.*, diseñaron un modelo destinado a predecir las interacciones fármaco-proteína (DPI) para las proteínas diana implicadas en las vías dopaminérgicas. El conjunto de datos consta de un total de 50.000 casos. El presente trabajo informa de la síntesis orgánica, la caracterización química y el ensayo farmacológico de una nueva serie de compuestos peptidomiméticos de *L-prolyl-L-leucyl-glycinamide* (PLG) (Ferreira da Costa *et al.*,

2018). Se muestran los resultados generales de los subconjuntos de entrenamiento y validación. En la serie de entrenamiento, el modelo presenta valores elevados de Especificidad=Sp(%)=72,8, Sensibilidad=Sn(%)=72,4, y Exactitud general=Ac(%)=72,7 como se muestra en la Tabla 3. El modelo fue estable en las series de validación externa con valores de Sp(%)=72,7, Sn(%)=71,4, y Ac(%)=72,6.

El mejor modelo lineal encontrado fue el siguiente

$$f(v_{ij})_{calc} = -10.780386430140 - 0.000000000020 \cdot f(v_{ij})_{ref} + 0.440071875560 \cdot \Delta D_1(c_0) + 0.465335484664 \cdot \Delta D_1(c_1) - 0.541834505781 \cdot \Delta D_1(c_2) - 0.127705300409 \cdot \Delta D_1(c_3) - 0.114637007349 \cdot \Delta D_1(c_4) - 0.095637330548 \cdot \Delta D_2(c_5) - 0.054733584740 \cdot \Delta D_2(c_6) - 0.056732915285 \cdot \Delta D_2(c_7) \quad (5)$$

$$n = 41082 \chi^2 = 5564.0 p - level < 0.05$$

Donde $f(v_{ij})_{calc}$ es el valor de la función que predice el DPI para las proteínas objetivo implicadas en las vías de la dopamina. El modelo tiene nueve tipos de variables de entrada. El primer tipo es la función de valor referido $f(v_{ij})_{ref}$. Los otros tipos son $\Delta D_k(c_j)$, que representan los efectos sobre la actividad biológica de las perturbaciones de la hidrofobicidad del fármaco para ocho condiciones diferentes (Luan F. *et al.*, 2014). En la Tabla 5, se han presentado algunos de los compuestos sintetizados, ensayados farmacológicamente y estudiados con los modelos PTML y las técnicas de docking.

1.8. Modelos PTML para otras aplicaciones

1.8.1. Análisis de las redes de inversión de la orientación de los genes cromosómicos

Quevedo-Tumailli *et al*. definieron un nuevo tipo de red compleja denominada GOIN que codifica patrones de inversión de corto y largo alcance de la orientación de pares de genes en el cromosoma de *Plasmodium falciparum (Pf)*. Estas redes tienen una media de 383 nodos (genes) y 1314 enlaces (pares de genes con orientación inversa). Se encontraron algunas comunidades de genes que codifican proteínas relacionadas con RIFIN. El modelo PTML discrimina el tipo RIFIN de otras proteínas. Se utilizaron como valores de entrada los parámetros de las GOIN y las centralidades. El modelo presenta valores de sensibilidad y especificidad del 70-80% en las series de entrenamiento y validación externa, respectivamente. En conclusión, la relevancia biológica de la inversión de la orientación de los genes no depende directamente de la información de la secuencia genética (Quevedo-Tumailli, Ortega-Tenezaca, & González-Díaz, 2018). El mejor modelo lineal encontrado fue el siguiente:

$$f(v_{ij})_{calc} = -68035.949 \cdot f(v_{ij})_{ref} - 4.896$$
$$n = 4025 \chi^2 = 293.77 p < 0.05 \quad (6)$$

Donde $f(v_{ij})_{calc}$ es el valor de la función que predice RIFIN en las 5365 proteínas del proteoma de *Pf*. La entrada es una variable de centralidad llamada closeness C_{clo}. Su centralidad mide la desviación de los genes $_i$ en el cromosoma $_k$ con respecto al valor medio esperado de cercanía para todos los genes en el mismo cromosoma $_k$ (Quevedo-Tumailli, Ortega-Tenezaca, & González-Díaz, 2018). Esta $f(v_{ij})_{ref}$ es igual a $C_{clo}(Gene_i,Chr_k) - \langle C_{clo}(Chr_k) \rangle$.

1.8.2. Minería del proteoma de los epítopos de las células B

Martínez-Arzate *et al*. desarrollaron un modelo PTML para descubrir nuevos epítopos de células B útiles para el diseño de vacunas, y para predecir puntuaciones de epítopos inmunogénicos en diferentes condiciones experimentales. El modelo utiliza como entrada la secuencia del péptido *q*, y la actividad del epítopo. La información recuperada contiene cambios estructurales en 83.683 secuencias peptídicas (Seq) determinadas en ensayos experimentales reportados en la base de datos IEDB, y que involucran a 1.448 organismos epitópicos (Org), 323 organismos huéspedes (Host), 15 tipos de procesos *in vivo* (Proc) (Blay, Yokoi, & González-Díaz, 2018) técnicas experimentales (Tech), además de 505 aditivos adyuvantes (Adj). El modelo tiene una precisión, sensibilidad y especificidad entre el 71 y el 80% para el entrenamiento, y series de validación externa (Martínez-Arzate, *et al.*, 2017). La ecuación de este modelo se presenta como sigue (Ferreira da Costa *et al.*, 2018).

$$\begin{aligned} S_{qr}(c_i, c_j) = {} & -1.96618 \cdot \varepsilon_r - 0.00368 \cdot {}^{q}\theta_2(Seq) - 0.01357 \cdot {}^{q}\theta_0(Org) \\ & - 0.08383 \cdot {}^{q}\theta_0(Tech) + 0.00463 \cdot \Delta\theta_5(Seq) + 0.00404 \cdot \Delta\theta_0(Adj) \\ & + 0.0025 \cdot \Delta\theta_0(Org) + 0.00089 \cdot \Delta\theta_0(Host) - 0.00095 \\ & \cdot \Delta\theta_5(Proc) + 0.00438 \cdot \Delta\theta_0(Tech) + 7.95575 \end{aligned} \tag{7}$$

Donde $Sqr(c_i,c_j)$ es el valor de la función para la actividad del epítopo del péptido q predicho en condiciones experimentales. El modelo tiene nueve tipos de variables de entrada. Las primeras entradas($^{q}\theta_k$) corresponden a la medida de información de entropía de Shannon de la secuencia, el organismo del epítopo y las técnicas utilizadas para determinar la inmunogenicidad. Las otras entradas correspondientes ($\Delta\theta_k$) evalúan la información relativa a la

variación o perturbación de la secuencia, los aditivos adyuvantes, el organismo epítopo, el organismo expuesto al antígeno, el tipo de proceso *in vivo* y las técnicas (Martínez-Arzate, *et al.*, 2017).

1.8.3. Modelización de subclases de enzimas para la extracción del proteoma de microorganismos productores de biocombustibles

Concu *et al.*., desarrollaron un modelo para predecir un conjunto de enzimas que pertenecen al estipe de la levadura Pichia. Se ha aplicado a un conjunto de datos de 19.187 enzimas que representan las 59 subclases presentes en el Banco de Datos de Proteínas (PDB). Además, los autores desarrollaron modelos PTML basados en ANN para predecir pares enzima-enzima de secuencias plantilla-consulta con una precisión, especificidad y sensibilidad superiores al 90% tanto para la serie de entrenamiento como para la de validación (Concu *et al.*, 2019). La forma general de este modelo PTML se presenta en siguiente ecuación.

$$s_{qr}(c_i, c_j) = e_0 + e_1 \cdot \varepsilon_r(c_j) + \sum_{k=0}^{k=5} {}^{k}_{2}e \cdot \Delta TI_k(q,r) + \sum_{k=0}^{k=5} {}^{k}_{3}e \cdot \Delta TI_k(q,j) \sum_{k=0}^{K=5} {}^{k}_{4}e \cdot \Delta\Delta TI_k(i,j,r) + \sum_{k=0}^{k=5} {}^{k}_{5}e \cdot \nabla\nabla_k(i,j,q,r)$$

Donde la función representa el valor de la puntuación de las proteínas de consulta para la actividad enzimática de la clase c_i en comparación con la actividad enzimática $\varepsilon_r(c_j)$ de referencia. La primera variable de entrada del modelo $\varepsilon_r(c_j)$ cuantifica la presencia o ausencia de

la actividad enzimática de la subclase c_j. Los otros tipos de variables son valores PT como $\Delta TI_k(q,r)$, $\Delta TI_k(q,j)$, $\Delta\Delta TI_k(i,j,q,r)$, y $\nabla\nabla TI_k(i,j,q,r)$ (Concu *et al.*, 2019).

1.8.4. Diseño de materiales de zeolita para la desilicación

Se aplicó un modelo PTML para el estudio en zeolitas, y representa los efectos de la desilicación, logrando una precisión de R2 = 0,98 en la validación externa siendo útil para el diseño racional de nuevos materiales (Blay, Yokoi, & González-Díaz, 2018). La ecuación de este modelo se presenta en la ecuación 8.

$$\varepsilon_k(m_i, c_j)_{new} = -0.22881 + 1.02864 \cdot \varepsilon_k('m_i, 'c_j)_{ref} + 0.02792 \cdot V_1 + 67.29053 \cdot V_{10} + 0.01264 \cdot \Delta\Delta V_1(c_1, c_2, c_3, c_5, c_6, c_7, c_8) + 0.05753 \cdot \Delta\Delta V_7(c_1, c_2, c_3, c_5, c_6, c_7, c_8) \quad (8)$$

$$n_{tot} = 4975, R^2train = 0.980, R^2val = 0.985, F(1.3730) = 228700, p < 0.05$$

Donde, la salida del modelo PTML representa los efectos de la decepción, logrando una precisión. V_k son los valores de las variables de entrada que se utilizan para calcular los valores de los operadores PT, como las medias móviles, y los operadores PT multicondición (Concu, D S Cordeiro, Munteanu, & González-Díaz, 2019). En un operador PT multicondición, se utiliza la misma idea de la media móvil: $\Delta V_k(c_j) = V_k - \langle V_k(c_j) \rangle$ (Blay *et al.*, 2018).

1.9. Software disponible para la modelización de PTML

1.9.1. Estrategia multi-programa vs. Estrategia de software específico

Para el desarrollo de modelos PTML y manejo de la gran cantidad de información de los descriptores y operaciones PT, es necesario el empleo de software informático que permita realizar los cálculos respectivos. Se pueden encontrar software de uso general y especifico.

1.1.1.1. Software para PTML no específico

Microsoft Excel

Microsoft Excel es una herramienta de pago que forma parte de la suite ofimática de Microsoft Office. Se ha creado para el sistema operativo Windows inicialmente y en los últimos años se ha extendido a la plataforma ARM de MAC. Su uso principal es para el manejo de hojas de cálculo. Mediante fórmulas, programación de macros y entorno visual basic, se pueden realizar operaciones complejas tales como el cálculo de operaciones PT que contribuyen generar posteriormente un modelo.

STATISTICA

Statistica es un paquete de software estadístico de pago, desarrollado por la empresa StatSoft Europe. Inicialmente fue desarrollado en común acuerdo con la participación de profesores universitarios. Funciona sobre el sistema operativo Windows y permite realizar cálculos de estadística más comunes, técnicas multivariantes y modelos avanzados de regresión lineal y no lineal. En la actualidad incursiona en Análisis de Recursos Humanos, Análisis de clientes, Minería de datos, Análisis predictivo, Big Data, Reportes

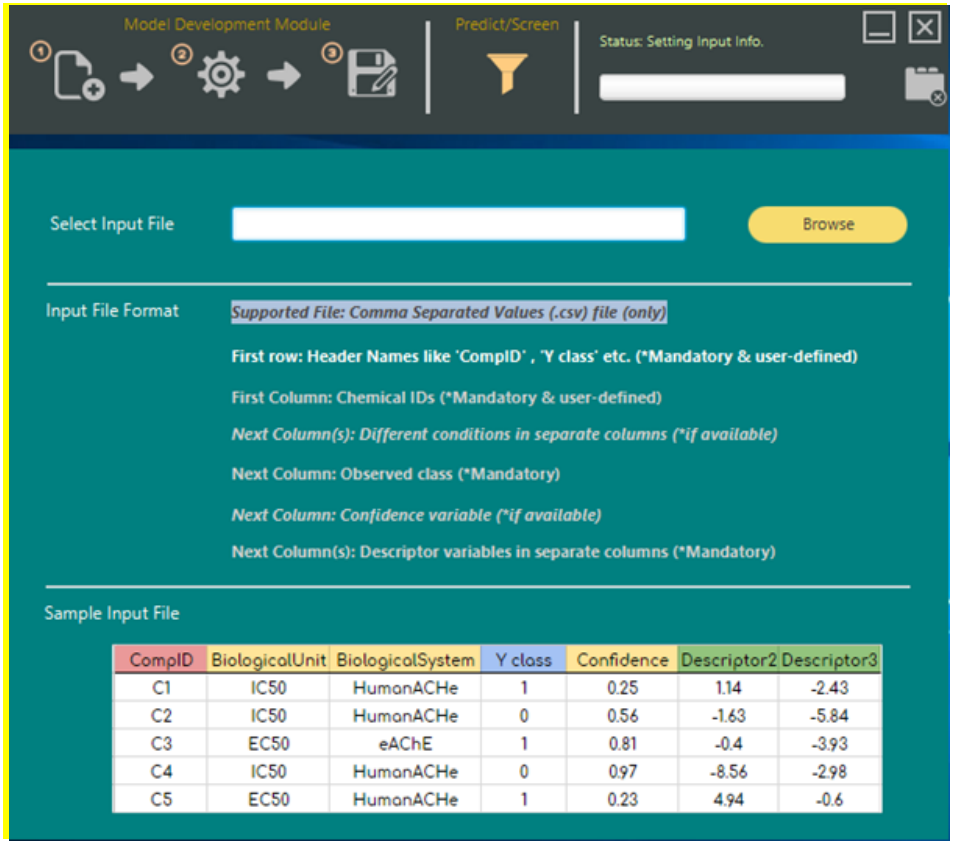
WEKA

Weka es un paquete multiplataforma de software libre, creado en la Universidad de Waicato para el desarrollo de ML, Minería de datos, entre otros. Contiene opciones de visualización y algoritmos para el análisis de datos.

1.1.1.2. Software Específico de PTML

Software QSAR-Co

El software QSAR-Co versión 1.0.0 es una nueva aplicación útil para abordar algunas cuestiones críticas que normalmente se descuidan durante el desarrollo de modelos quimio-informáticos convencionales basados en la clasificación. Se trata de un software autónomo QSAR-Co, disponible de forma gratuita, que permite realizar estudios basados en la clasificación, teniendo en cuenta diferentes condiciones experimentales según el caso. Cabe destacar que QSAR-Co es una forma abreviada de "*Cheminformatics with conditions*", siendo esta última una de las características clave de este software, aunque también se pueden desarrollar modelos quimio-informáticos simples basados en la clasificación sin condiciones. Otra razón que motivó el desarrollo de este software fue proporcionar una plataforma distinta para derivar modelos quimio-informáticos basados en la clasificación siguiendo todas las directrices recomendadas por la OCDE (Organisation for Economic Co-operation and Development (OECD), 2007), es decir, modelos quimio-informáticos robustos. El software consta de dos módulos: 1) el módulo de desarrollo de modelos y 2) el módulo de screening/predicción.



CompID	BiologicalUnit	BiologicalSystem	Y class	Confidence	Descriptor2	Descriptor3
C1	IC50	HumanACHe	1	0.25	1.14	-2.43
C2	IC50	HumanACHe	0	0.56	-1.63	-5.84
C3	EC50	eAChE	1	0.81	-0.4	-3.93
C4	IC50	HumanACHe	0	0.97	-8.56	-2.98
C5	EC50	HumanACHe	1	0.23	4.94	-0.6

Figura 1: Software QSAR-Co

El software 'QSAR-Co' versión 1.0.0 es una herramienta independiente que se puede descargar gratuitamente en la página web de QSAR-Co. Tiene dos módulos ('desarrollo de modelos' y 'cribado/predicción') que están disponibles en el software; ahora se discutirán todos los pasos y funcionalidades asociadas en cada módulo. En el módulo de 'desarrollo de modelos', el software proporciona todos los pasos básicos que intervienen en el desarrollo de un modelo quimio-informático basado en la clasificación, que también incluye el examen y el tratamiento de los datos de entrada para varias condiciones experimentales, si procede.

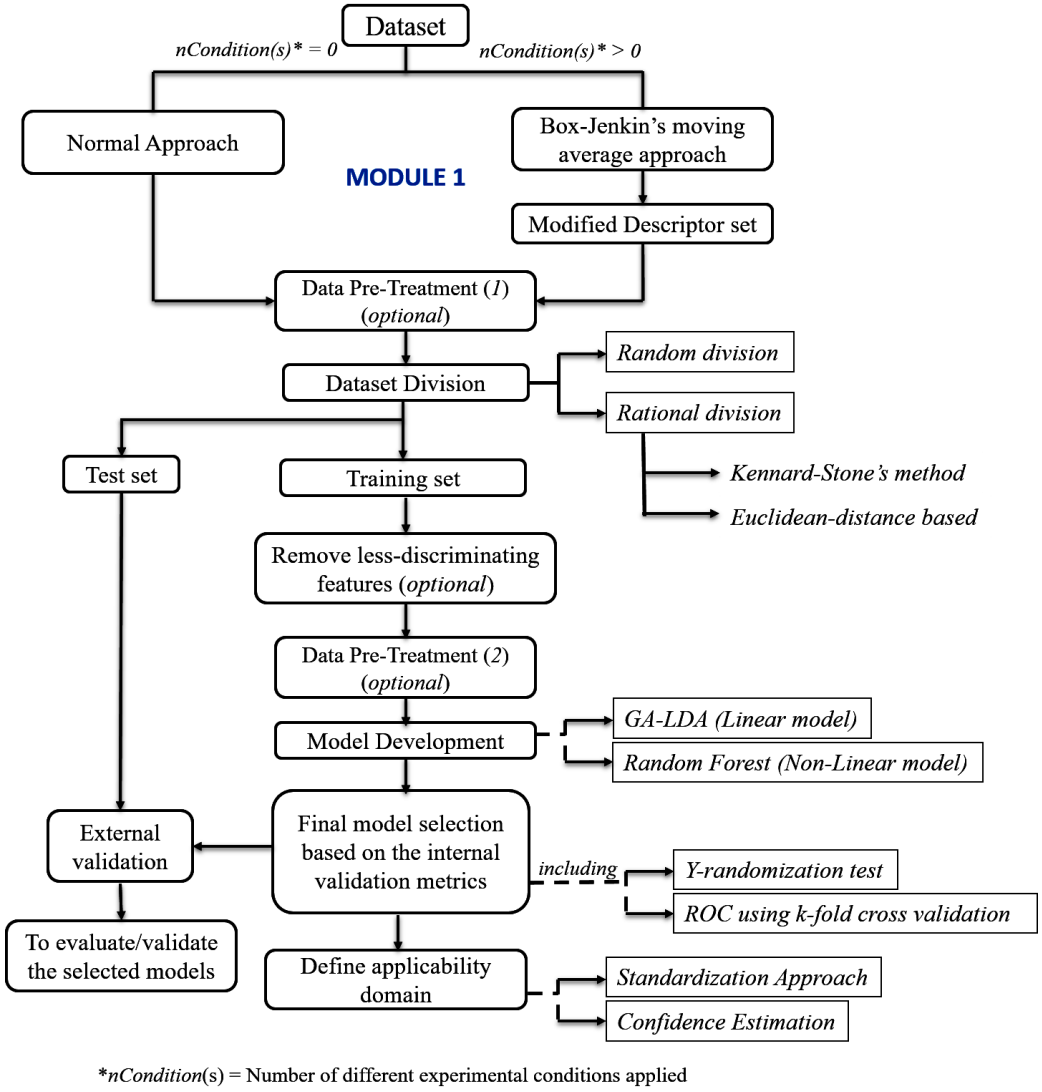


Figura 2: QSAR-Co - Flujo de trabajo para el desarrollo del modelo

El software permite calcular operadores de medias móviles Box-Jenkins para descriptores moleculares. El enfoque se ha discutido previamente en detalle (Speck-Planche & Cordeiro, 2013; Speck-Planche & Cordeiro, 2014; Speck-Planche & Cordeiro, 2017; Speck-Planche & Cordeiro, 2014). Calcula los descriptores de media móvil ($D(D_i)c_j$) para un descriptor molecular D_i de compuestos individuales '*i*'. El término derivado $D(D_i)c_j$ se denomina operador de Box-Jenkin (Speck-Planche & Cordeiro, 2017; Speck-Planche A. *et al.*, 2016; Kennard & Stone, 1969), y estos descriptores modificados capturan la información relativa tanto a estructuras

químicas y elementos específicos de la condición experimental (c_j) bajo la cual se ensayaron las muestras. Estos descriptores modificados son calculados por el software QSAR-Co y utilizados en los pasos posteriores de desarrollo del modelo quimio-informático. Opcionalmente, se puede realizar un pre-tratamiento de los datos para eliminar los descriptores no informativos que pueden no tener una contribución significativa en la construcción del modelo. También se puede hacer una división del conjunto de datos en conjuntos de entrenamiento y de prueba, de modo que, en los pasos posteriores, el conjunto de entrenamiento se emplee para el desarrollo del modelo y la selección del mismo, mientras que el conjunto de prueba se emplee para la validación del modelo. Existe la opción de repetir la misma división aleatoria para reproducir el desarrollo del modelo utilizando el mismo valor de semilla en los ajustes. En los enfoques racionales, el software proporciona dos técnicas, a saber, el algoritmo de Kennard-Stone y el método de división basado en la distancia euclidiana (Venkatasubramanian & Sundaram, 2002). El software también elimina los descriptores menos discriminantes. QSAR-Co también ofrece el "Algoritmo Genético"(Rogers & Hopfinger, 1994) como técnica de selección de variables para desarrollar modelos de Análisis Discriminante Lineal (LDA, en idioma inglés Linear Discriminant Analysis). El algoritmo genético (GA, en idioma inglés Genetic Algorithm) es una técnica bien conocida que se utiliza a menudo en el desarrollo de modelos quimio-informáticos basados en la regresión (Hemmateenejad, Akhond, Miri, & Shamsipur, 2003; Hasegawa, Miyashita, & Funatsu, 1997; Ambure & Roy, 2016; Gramatica, Chirico, Papa, Cassani, & Kovarich, 2013; Gao, 2001), así como para desarrollar modelos quimio-informáticos basados en la clasificación(Sutherland, O'Brien, & Weaver, 2003; Snedecor & Cochran, 1967). En la actualidad, el software proporciona dos técnicas de ML para desarrollar modelos robustos de Química basados en la clasificación, el LDA (Breiman, 2001), y el Bosque Aleatorio (Hall, *et*

al., 2009) son algoritmos de aprendizaje automático supervisado que consiste en una colección o un conjunto de predictores de árboles de decisión simples. En este software, se ha utilizado la biblioteca java Weka versión 3-9-3 (Wilks, 1932) para realizar el Random Forest. Las métricas de validación como la λ de Wilk (Fawcett, 2006) proporcionan una medida de la importancia de la discriminación alcanzada. Se puede diseñar una matriz de confusión utilizando la información de la clase de respuesta real y la predicha obtenida del modelo en evaluación; el software también proporciona parámetros como la sensibilidad, la especificidad, la razón de Fisher*, etc*. (Fisher, 1937), y realiza el análisis de la curva de características operativas del receptor (Fisher, 1937)(ROC), y la prueba de aleatorización Y(Roy *et al.*, 2015). Además, el software también realiza un análisis del dominio de aplicabilidad (AD) (Hill & Lewicki, 2006). Por último, en el software, en el módulo 2, se puede realizar un análisis predictivo de nuevos compuestos químicos.

LAGA

LAGA es un nuevo software de código abierto para el diseño de fármacos que utiliza la teoría de la perturbación y técnicas de aprendizaje automático. Predice el éxito de los ensayos preclínicos mediante una función discriminante. Para hacer este modelo accesible a los usuarios, se ha desarrollado una interfaz gráfica en la que el usuario puede introducir todas las variables de entrada y en la que, al final, se presentan los resultados en forma de probabilidad de éxito.

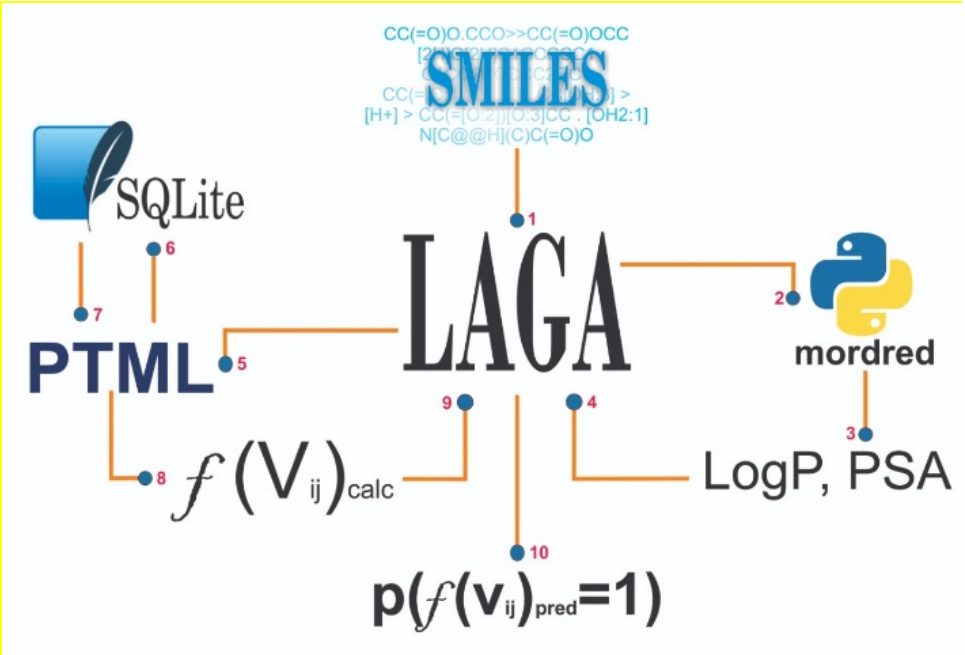


Figura 3**:** LAGA - Diagrama de flujo de ejecución.

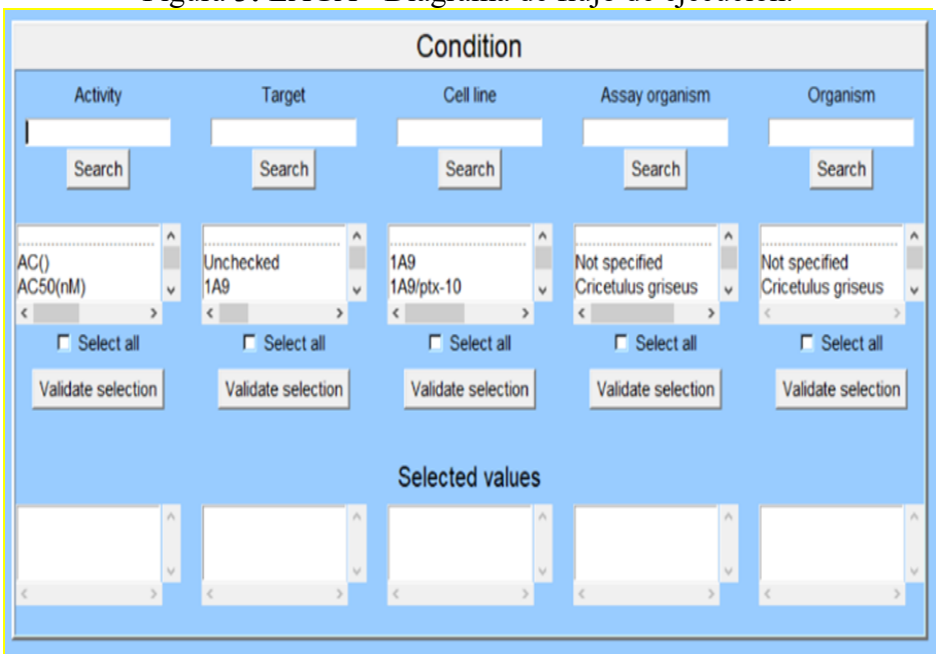


Figura 4: Software LAGA

Hay que tener en cuenta que LAGA favorece la predicción de resultados positivos, ya que se ha elegido una probabilidad alta *a priori*.

Trabajo experimental

Capítulo 1

TRABAJO EXPERIMENTAL

CAPÍTULO 1, ALGORITMOS MULTIETIQUETA PTML: MODELOS, SOFTWARE Y APLICACIONES

1.10. Artículo 1

Ref. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C., Cordeiro, M., & González-Díaz, H. (2020). PTML Multi-Label Algorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337. doi: https://doi.org/10.2174/1568026620666200916122616

1.11. Software SOFT.PTML

SOFT.PTML Studio v.1.0. SOFT.PTML es un software de propósito general para el cálculo de variables de PTOs y el desarrollo de modelos PTML. Fue escrito para sistemas operativos Microsoft Windows. Esta herramienta fue desarrollada utilizando el Entorno de Desarrollo Integrado (IDE) del lenguaje de programación C# de Microsoft Visual Studio .Net 2017, sobre .Net Framework 4.6.1. Utiliza bibliotecas de software libre como Accord.NET, Accord.MachineLearning 3.8.0, Accord.Math 3.8.0, Accord.MathCore 3.8.0, Accord.Statistic 3.8.0, MathNet.Numerics 3.20.0 y ExcelDataReader 2.1.2.3. Todas estas bibliotecas están disponibles a través del gestor de paquetes NuGet. SOFT.PTML fue probado en un ordenador con Intel® Core ™ i7-7500U CPU @ 2,70Ghz 2,90Ghz, con 16,0 GB de RAM, disco duro de estado sólido y el sistema operativo Windows 10 Pro. SOFT. PTML dispone de tres módulos para el tratamiento de ficheros, operaciones de variables y ML de la PT.

Modelización de PTML. Esto es útil para buscar modelos predictivos para conjuntos de datos complejos con múltiples características de Big Data. Finalmente, la metodología permite

desarrollar modelos lineales de PTML para predecir la actividad biológica o clasificar compuestos como activos o no activos en términos de actividad biológica, *etc.* (Bediaga, Arrasate, & González-Díaz, 2018). El caso del modelo lineal PTMLIF más simple se puede expresar con la siguiente ecuación general:

$$f\left(\mathrm{v_{ij}}/\mathbf{c}_{\mathrm{qij}}\right)_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f\left(\mathrm{v_{ij}}/\mathbf{c}_{\mathrm{qij}}\right)_{ref} + \sum_{q=1,k=1,j=0}^{q_{max},k_{max},j_{max}} \mathrm{a}_{qkj} \cdot \mathrm{PTO}(\mathbf{D}_{\mathrm{qki}}, \mathbf{V}_{\mathrm{qki,}}\mathbf{c}_{\mathrm{qij}}) \quad (9)$$

$$f\left(\mathrm{v_{ij}}/\mathbf{c}_{\mathrm{qij}}\right)_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f\left(\mathrm{v_{ij}}/\mathbf{c}_{\mathrm{qij}}\right)_{ref} + \sum_{k=1,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \mathrm{PTO}(\Delta\mathbf{D}_{\mathrm{qki}}(\mathbf{c}_{\mathrm{ij}}), \Delta\mathbf{V}_{\mathrm{qki}}(\mathbf{c}_{\mathrm{ij}})) \quad (10)$$

La salida del modelo $f(v_{ij})_{calc}$ es una función de puntuación del valor v_{ij} de la actividad biológica del i-ésimo fármaco en las diferentes combinaciones de condiciones del ensayo c_j. Sin embargo, se pueden construir modelos PTML más complicados y generales utilizando el concepto de operadores PT (PTO). Los PTOs son funciones que actúan sobre una o varias series de parámetros estructurales D_k continuos y variables continuas V_k y variables literales o condiciones c_j para medir la desviación del nuevo caso con respecto a un caso o población de referencia.

Módulo 1. Carga de datos y guardado de datos. El módulo 1 incluye la carga de datos desde archivos Excel, y guarda el archivo resultante con las operaciones realizadas disponibles. El módulo 2 contiene la fusión de datos, el preprocesamiento de los mismos y el cálculo de las

variables de entrada de los Operadores de Perturbación-Teoría (PTML). El módulo 3 está dedicado al desarrollo del modelo PTML propiamente dicho. Por el momento, se dispone de la implementación de cinco algoritmos de aprendizaje automático Regresión Logística (LOGR), Análisis Discriminante Lineal (LDA), Análisis Discriminante Kernel (KDA), Random Forest (RF), y Super Vector Machine (SVM). Adicionalmente permite predecir nuevos casos.

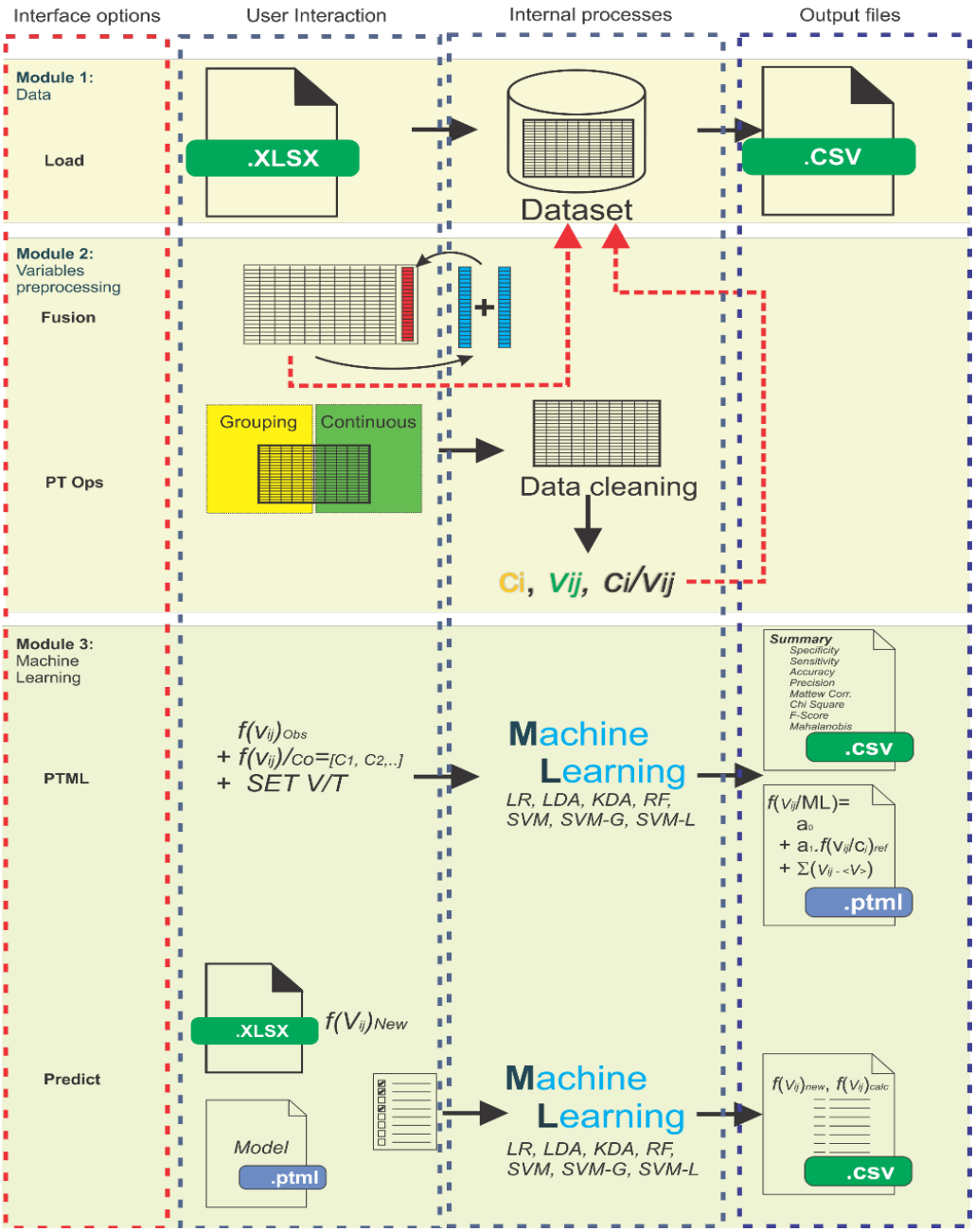


Figura 5: SOFT PTML - Flujo de trabajo

Paso 1. Carga del conjunto de datos. SOFT.PTML v1.0.0. soporta formatos de archivos xlsx. Inicialmente se requiere abrir una hoja de cálculo que contenga los datos a calcular. La información se carga en la memoria del ordenador, y de ella depende el tamaño soportado por la base de datos. El archivo a cargar debe contener los nombres de las variables en la primera fila. Una vez cargado el archivo, los datos de la hoja electrónica se convierten internamente en un conjunto de datos generales y, como resultado, se muestran las primeras 25 filas de datos.

Paso 2. Guardar los resultados. Una vez realizados los cálculos de la teoría de la perturbación, los resultados se pueden guardar en un archivo csv.

Módulo 2. Preprocesamiento de variables.

Paso 1. Fusión de Vars. Uno de los principales puntos fuertes de SOFT.PTML v1.0.0. consiste en generar nuevas variables categóricas a partir de las variables disponibles en el conjunto de datos general. La generación de variables de fusión se forma a partir de la unión de dos o más columnas de datos. Las variables seleccionadas para la fusión se almacenan en una lista de variables que deben ser procesadas. Se realiza un procesamiento por lotes de la lista de nuevas variables de fusión. Las nuevas variables de fusión generadas estarán disponibles para futuros cálculos.

Paso 2. Preprocesamiento de datos PT. La segunda opción de preprocesamiento de datos permite seleccionar las variables de agrupación y las variables continuas con las que es posible procesar los datos perdidos. Si hay datos perdidos dentro de las variables continuas, se puede optar por sustituirlos por una etiqueta "MD" o eliminar los casos. Dentro de las variables continuas es posible sustituir los datos perdidos por la media de la columna o eliminar los casos. Tanto en las variables de agrupación como en las continuas, la opción por defecto es no procesar.

Una vez realizado el tratamiento de los datos perdidos, existen opciones para realizar cálculos dentro de las variables de agrupación y continuas previamente seleccionadas. Los cálculos seleccionados se almacenan en una lista que luego se procesará por lotes. Los nuevos cálculos se convertirán en nuevas columnas disponibles dentro del conjunto de datos general. Principalmente, la metodología general se desarrolla en dos etapas. La primera etapa es el preprocesamiento de datos. Se trata de obtener los resultados de muchos ensayos preclínicos a partir de una base de datos. Posteriormente, la información es preprocesada según el criterio del experto en el área. El software utiliza como entrada dos tipos de variables continuas: D_{qki}(descriptores moleculares) y V_{qki} (variables externas como temperatura, tiempo, *etc.*) para describir el sistema. El software también utiliza como entrada variables literales o condiciones de ensayo c_{qij} (nombre de la propiedad medida, como el nombre de la proteína, organismo, línea celular) para describir las condiciones discretas en las que el sistema está siendo estudiado. El núcleo de la aplicación SOFT.PTML Studio es el software FRAMA v1.0.0. Se trata de un nuevo software independiente para el modelado PTML de múltiples etiquetas. La segunda etapa consiste en aplicar las técnicas de modelización. SOFT.PTML utiliza una versión generalizada de los modelos PTML clásicos. El modelo lineal generalizado de PTML basado en los PTOs puede expresarse como sigue.

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{ref} + \sum_{k=0,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \mathrm{PTO}\left[\Delta \mathrm{D_k}(\mathbf{c_j})\right] \qquad (11)$$

Podemos hablar de dos tipos de TDF: (1) los PTOs de referencia (REF) y (2) los PTOs de desviación (DEV). Los PTOs de tipo REF se utilizan como función de referencia $f(v_{ij})_{ref}$ y suelen ser la primera variable en un modelo PTML. Se utilizan para caracterizar el estado de referencia

utilizado como punto de partida para realizar la inferencia PT. Por otro lado, los PTOs de tipo DEV son los operadores propiamente dichos, se utilizan para medir la desviación de las variables de entrada (estructurales, funcionales, condiciones*, etc.*) del caso real con respecto al caso/grupo de referencia. En la Tabla 1 se muestran diferentes tipos de PTOs utilizados en los modelos PTML publicados anteriormente en la literatura. Todos estos PTOs han sido implementados en la herramienta SOFT.PTML versión 1.0. Algunos ejemplos de otros tipos de modelos PTML que utilizan diferentes PTOs aparecen en la siguiente, Ec. (12) Modelo PTML Doble MA (DMA), Ec. (14) Modelo de momentos PTML (MOM), Ec. (13) Modelo de covarianza PTML (COV):

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{ref} + \sum_{k=0,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \Delta\Delta\mathrm{D_k}(\mathrm{c_j}, \mathrm{c_{j'}}) \quad (12)$$

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{ref} + \sum_{k=0,j=0,n=1}^{k_{max},j_{max},nmax} \mathrm{a}_{kj} \cdot \left[\Delta\mathrm{D_k}(\mathbf{c_j})\right]^n \quad (13)$$

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{ref} + \sum_{k=0,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \Delta\mathrm{D_k}(\mathrm{c_j}) \cdot \Delta\mathrm{D_{k'}}(\mathrm{c_{j'}}) \quad (14)$$

Tabla 5: PTOs usado como entrada en los modelos PTML

Tipo PTO[a]	Sub-tipo PTO[a]	Partición de condiciones[a] (c_j)	Fórmula PTO	SOFT.PTML Cálculo de la secuencia	Información PTO
REF $f(v_{ij})_{ref}$	REF	-	= vi'ref	-	Valor de la actividad biológica vi'ref de un compuesto de referencia, generalmente medido en las mismas condiciones c_j que v_{ij}.
	AVG	c_j	$\langle v_{ij} \rangle = \sum_{i=1}^{imax} \frac{v_{ij}}{n_j}$	D_{ki} =><$D_{ki}(c_j)$>	Media (AVG) de los valores de actividad biológica vi'ref para un subconjunto de compuestos de referencia, normalmente medidos en las mismas condiciones c_j que v_{ij}.

Tipo PTO[a]	Sub-tipo PTO[a]	Partición de condiciones[a] (c_j)	Fórmula PTO	SOFT.PTML Cálculo de la secuencia	Información PTO
DEV ΔD_k	PROB	c_j	$= n(f(v_{ij})_{obs}=1)/n_j$	c_{ij} => Count(n_j) => v_{ij} , d_0 , cutoff =>Count(n($f(v_{ij})_{obs}=1$)) => $f(v_{ij})_{ref} = (f(v_{ij})_{obs}=1)$ = Count(n($f(v_{ij})_{obs}=1$))/ Count(n_j)	Valor de probabilidad (PROB) $p(f(v_{ij})=1)_{expt}$ esperado para el subconjunto de compuestos de referencia de la actividad v_{ij}, normalmente medido en las mismas condiciones c_j que v_{ij}.
	DIF	-	$\Delta D_k = D_{ki} - D_{ki`}$	Dki , Dki` => Dki - Dki`	Diferencia (DIF) entre el descriptor del compuesto de consulta Dki y el descriptor Dki` de un compuesto químico utilizado como referencia.
	MA	c_j	= Dki - <Dki(cj)>	D_{ki} , c_j =><$D_{ki}(c_j)$> => D_{ki} - <$D_{ki}(c_j)$>	Tiene en cuenta la variabilidad de la estructura química, expresada como D_k, con respecto a su valor esperado <$D_{ki}(c_j)$> para una determinada condición de ensayo c_j. Secuencia de cálculo.
	MMA	$c_j = [c_0, \ldots c_{kmax}]$	= D_{ki} - <$D_{ki}(c_j)$>	$c_0, c_1, \ldots c_n$ => Fusion($c_0, c_1, \ldots c_n$) => c_j => D_{ki} , c_j =><$D_{ki}(c_j)$> => D_{ki} - <$D_{ki}(c_j)$>	Tiene en cuenta la variabilidad de la estructura química, expresada como D_k, con respecto a su valor esperado <$D_{ki}(c_j)$> para un subconjunto dado de condiciones de ensayo múltiples c_j.
	MOM	$c_j = [c_0, \ldots c_{kmax}]$	=[$\Delta D_k(c_j)$]n = [D_{ki} - <$D_{ki}(c_j)$>]n	$c_0, c_1, \ldots c_n$ => Merge($c_0, c_1, \ldots c_n$) => c_j => D_{ki} , c_j =><$D_{ki}(c_j)$> => D_{ki} - <$D_{ki}(c_j)$> => NumPower[D_{ki} - <$D_{ki}(c_j)$>]n	Los momentos miden la desviación de la media. Según la naturaleza de n (par o impar), tienen en cuenta la asimetría de la desviación. El efecto de la perturbación sobre la salida depende del valor de n (cuanto mayor sea n, más fuerte será la perturbación).
	COV	c_j	= $\Delta D_k(c_j) \cdot \Delta D_{k'}(c_{j'})$	D_{ki} , c_j =><$D_{ki}(c_j)$> => D_{ki} - <$D_{ki}(c_j)$> => $\Delta D_k(c_j)$ => $D_{ki'}$, $c_{j'}$ =><$D_{ki'}(c_{j'})$> => $D_{ki'}$ - <$D_{ki'}(c_{j'})$> => $\Delta D_{k'}(c_{j'})$ => $\Delta D_k(c_j) \cdot \Delta D_{k'}(c_{j'})$	Covarianza entre las desviaciones de los descriptores D_k, y $D_{k'}$ con respecto a dos subconjuntos de condiciones c_j, y $c_{j'}$.
	XPRO	c_j and $c_{j'}$	= ΔDk(cj)·ΔDk'(cj') - ΔDk(cj´)·ΔDk'(cj)		El producto cruzado (XPRO) mide el efecto cruzado en las desviaciones de dos descriptores D_k, y $D_{k'}$ con respecto a dos subconjuntos diferentes de condiciones c_j, y $c_{j'}$.
	DMA	$c_j = [c_k, c_{k'}]$	= $\Delta D_k(c_j) - \Delta D_k(c_{j'})$ = [D_{ki} - <$D_k(c_j)$>] – [$D_{ki'}$ - <$D_k(c_{j'})$>]		Doble MA (DMA) del mismo descriptor Dk para dos compuestos medidos en dos subconjuntos de condiciones c_j y $c_{j'}$. $\Delta\Delta D_k(c_j, c_{j'}) = \Delta D_k$ cuando $c_j = c_{j'}$ (compuestos diferentes pero mismo ensayo) y $\Delta\Delta D_k(c_j, c_{j'})$ = <$D_{ki}(c_j)$> - <$D_{ki}(c_j)$> cuando $D_k = D_{k'}$ (mismos compuestos diferente ensayo).

[a]Las condiciones del ensayo (variables literales que delimitan las condiciones de contorno del ensayo biológico) pueden ser c_j simples o c_j múltiples (matriz de condiciones).

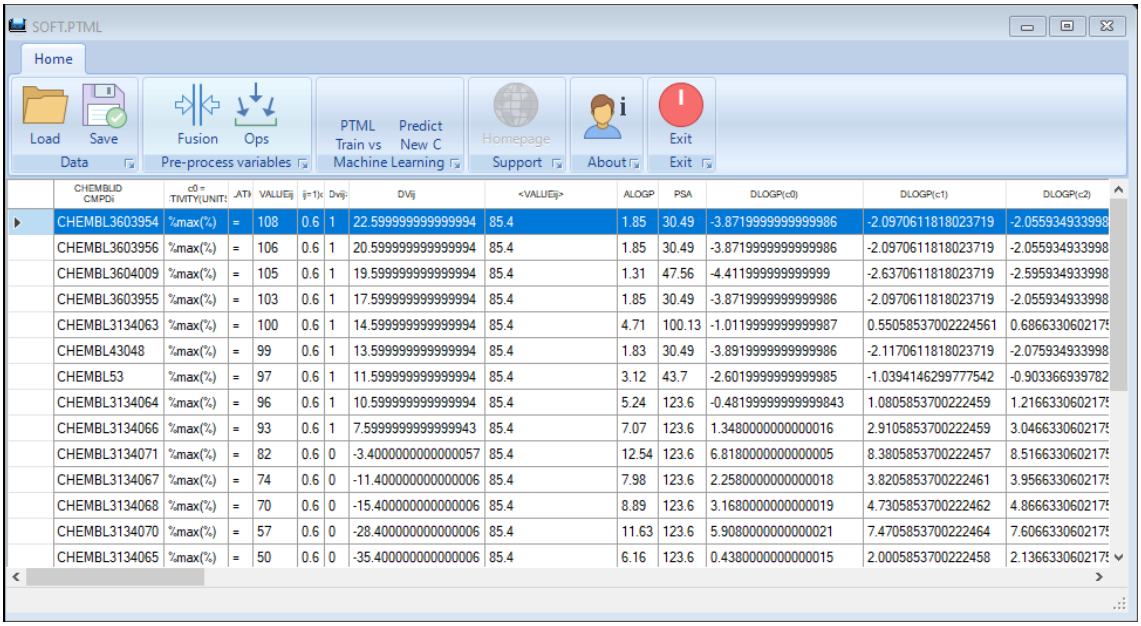
Módulo 3. Desarrollo del modelo.

Paso 1. Aprendizaje automático. Se muestra el conjunto de variables generadas a partir del conjunto de datos cargados previamente y las operaciones realizadas en el módulo PTOs. Para la generación del modelo, se requiere configurar el procedimiento. Selección de un valor discreto alojado en una variable dependiente (salida), la variable de referencia con sus correspondientes descriptores, la variable de validación. Si el conjunto de datos no tiene una variable de validación establecida, se puede generar una nueva. Antes de utilizar los algoritmos de aprendizaje automático, el usuario establecerá un patrón de secuencia para los valores generales de entrenamiento y validación. La secuencia de caracteres debe contener las letras T = entrenamiento, V = validación. La proporción de caracteres en la secuencia será determinada por el usuario. El valor por defecto es TTTV, lo que significa que el 75% de los datos se utilizarán como datos de entrenamiento, y el 25% restante como validación. A continuación, se seleccionan uno o varios algoritmos que se aplicarán al conjunto de datos seleccionado. SOFT.PTML 1.0.0 contiene cinco técnicas de Aprendizaje Automático para el desarrollo de modelos basados en PTML como son: Regresión Logística, Análisis Discriminante Lineal (LDA), Análisis Discriminante Kernel, Bosque Aleatorio y Máquina de Vectores de Soporte. El algoritmo de la máquina de vectores de soporte contiene dos variantes, que incluyen el uso de un núcleo de Gauss y de Laplaciano.

Predicción de nuevos casos. Antes de utilizar la opción de predicción de nuevos casos, el usuario debe poseer el archivo con el modelo PTML y el archivo xlsx con los nuevos casos a predecir. PTML.SOFT muestra una ventana independiente del conjunto de datos cargado inicialmente. Permite cargar un fichero de extensión ptml que contiene el modelo entrenado. Muestra el resumen del archivo seleccionado que contiene entre los datos más relevantes, las

principales medidas estadísticas con las que se generó. A continuación, se carga un fichero Excel que contiene los nuevos casos a predecir. El botón de procesamiento le pedirá que seleccione una carpeta para almacenar los resultados. Finalmente, como resultado se genera un archivo CSV con los resultados de la predicción del conjunto de datos introducidos.

Uso del software SOFT.PTML Studio. En la Figura 9 se muestra el ejemplo de carga del archivo "Chembldopamine receiver example.xls" de 4MB, con 54.776 filas, y 48 columnas. La primera fila contiene los nombres de las variables(Ferreira da Costa, *et al.*, 2018).



CHEMBLID CMPDi	c0 = TIVITY(UNITS	.ATI	VALUEij	ij=1)x	Dvij	DVij	<VALUEij>	ALOGP	PSA	DLOGP(c0)	DLOGP(c1)	DLOGP(c2)
CHEMBL3603954	%max(%)	=	108	0.6	1	22.599999999999994	85.4	1.85	30.49	-3.8719999999999986	-2.0970611818023719	-2.055934933998
CHEMBL3603956	%max(%)	=	106	0.6	1	20.599999999999994	85.4	1.85	30.49	-3.8719999999999986	-2.0970611818023719	-2.055934933998
CHEMBL3604009	%max(%)	=	105	0.6	1	19.599999999999994	85.4	1.31	47.56	-4.411999999999999	-2.6370611818023719	-2.595934933998
CHEMBL3603955	%max(%)	=	103	0.6	1	17.599999999999994	85.4	1.85	30.49	-3.8719999999999986	-2.0970611818023719	-2.055934933998
CHEMBL3134063	%max(%)	=	100	0.6	1	14.599999999999994	85.4	4.71	100.13	-1.0119999999999987	0.55058537002224561	0.686633060217
CHEMBL43048	%max(%)	=	99	0.6	1	13.599999999999994	85.4	1.83	30.49	-3.8919999999999986	-2.1170611818023719	-2.075934933998
CHEMBL53	%max(%)	=	97	0.6	1	11.599999999999994	85.4	3.12	43.7	-2.6019999999999985	-1.0394146299777542	-0.903366939782
CHEMBL3134064	%max(%)	=	96	0.6	1	10.599999999999994	85.4	5.24	123.6	-0.48199999999999843	1.0805853700222459	1.216633060217
CHEMBL3134066	%max(%)	=	93	0.6	1	7.5999999999999943	85.4	7.07	123.6	1.3480000000000016	2.9105853700222459	3.046633060217
CHEMBL3134071	%max(%)	=	82	0.6	0	-3.4000000000000057	85.4	12.54	123.6	6.8180000000000005	8.3805853700222457	8.516633060217
CHEMBL3134067	%max(%)	=	74	0.6	0	-11.400000000000006	85.4	7.98	123.6	2.2580000000000018	3.8205853700222461	3.956633060217
CHEMBL3134068	%max(%)	=	70	0.6	0	-15.400000000000006	85.4	8.89	123.6	3.1680000000000019	4.7305853700222462	4.866633060217
CHEMBL3134070	%max(%)	=	57	0.6	0	-28.400000000000006	85.4	11.63	123.6	5.9080000000000021	7.4705853700222464	7.606633060217
CHEMBL3134065	%max(%)	=	50	0.6	0	-35.400000000000006	85.4	6.16	123.6	0.4380000000000015	2.0005853700222458	2.136633060217

Figura 6: Software FRAMA (versión 1.0)

El tiempo de carga de la información hasta FRAMA es directamente proporcional al tamaño del fichero, y a la cantidad de información. La operación de unión de variables, permite crear nuevas variables de agrupación mediante la unión de dos o más variables existentes preferidas por el usuario. Se realiza mediante un procesamiento por lotes en la cola de operaciones seleccionadas. Dicho procesamiento añade las nuevas variables al conjunto de

variables disponibles en el contexto de trabajo. La figura 7 muestra un ejemplo de las variables disponibles previamente cargadas desde la hoja de cálculo.

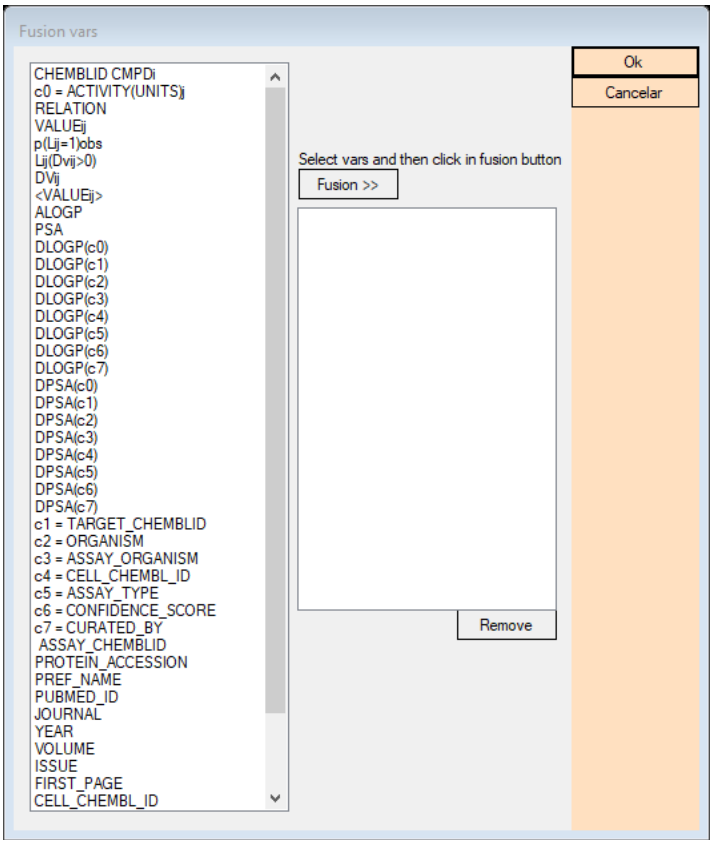


Figura 7: FRAMA - Unión de variables.

Una vez cargado el fichero y las variables de unión, se realiza la selección y clasificación de las variables de agrupación y de las variables continuas. Se permite comprobar si existen datos anómalos como nulos o vacíos y tomar una decisión en cuanto a su tratamiento. En las variables de agrupación, los valores anómalos pueden ser sustituidos por el valor "MD". En las variables continuas se puede sustituir por la media de los valores de la columna. También es posible en ambos tipos de variables eliminar casos.

Tras el pretratamiento de los datos, las variables seleccionadas pueden someterse a operaciones por lotes (véase la figura 8). En función de la naturaleza de las variables, se pueden

realizar operaciones de agrupación de variables como identidad, cuenta, probabilidad, entropía de Shannon. Operaciones de transformación de variables continuas como: identidad, exponencial, valor absoluto, potencia numérica, logaritmo, probabilidad máxima mínima, puntuación z. Operaciones básicas de variables continuas como: suma, producto, diferencia, división. Operaciones paramétricas como: máximo, mínimo, media, suma, desviación típica, multiplicaciones por una constante. Finalmente, operadores entre variables de agrupación, y variables continuas como media móvil, suma por agrupación, desviación estándar, probabilidad mínima máxima, Z-Score (Hill & Lewicki, STATISTICS Methods and applications. A comprehensive reference for science, industry and data mining, 2006).

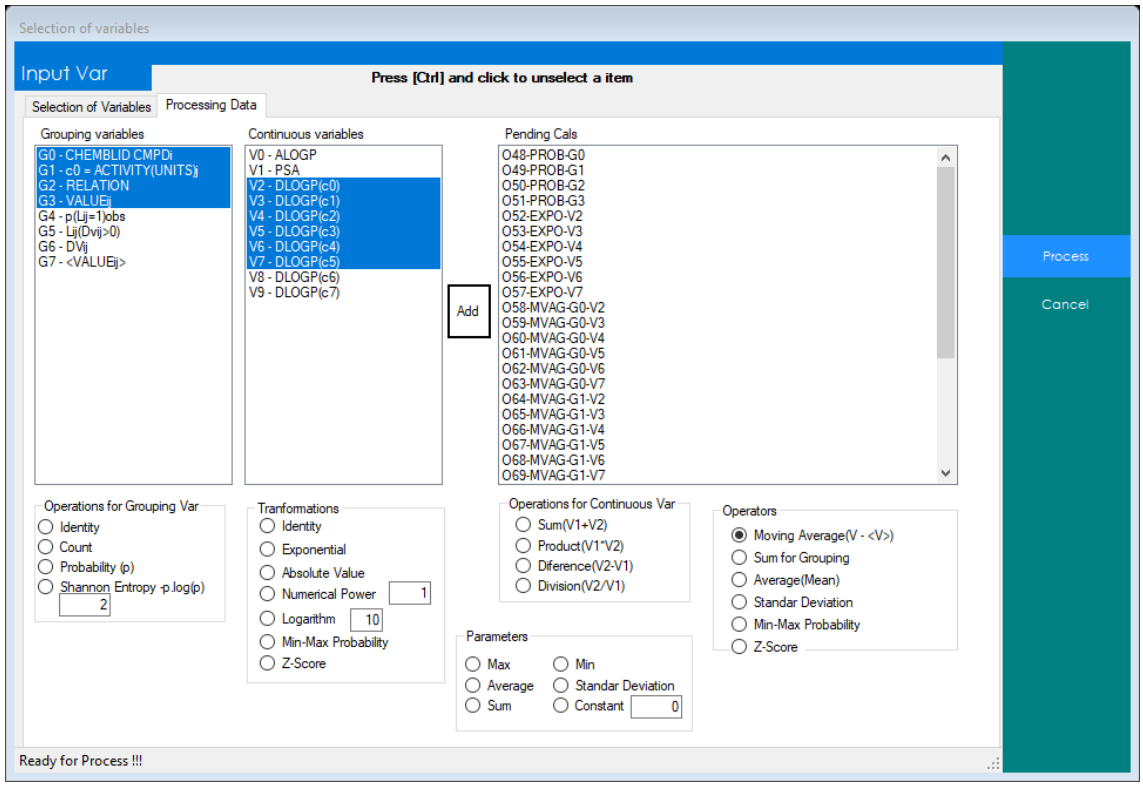


Figura 8: FRAMA - operaciones.

Es posible establecer los valores de entrenamiento y validación utilizando un patrón de caracteres. La letra T se utiliza para representar el entrenamiento, y la V para la validación. El

patrón por defecto es TTTV que expresa el 75% de los valores para la formación, y el 25% para los valores de validación. SOFT.PTML 1.0.0 permite realizar la operación de análisis de regresión lineal, basándose en la selección de variables de entrada independientes, y la variable dependiente de salida. Los resultados del procesamiento en un archivo CSV. El archivo resultante contiene los nombres de las variables, los valores de las constantes, el error estándar, el estadístico T, el valor p (T), el estadístico F, el valor p (F), límite superior de confianza, límite inferior de confianza, índice, intercepción, TTest, FTest. La información sobre la regresión se muestra con R2 (r-cuadrado), r2 ajustado, F-Test, Z-test, Chi-Squared Test. Es posible generar variables de media móvil multietiquetas y operadores que se utilizarán en la PTML.

3.1.2. Conclusión

En conclusión, podemos mostrar la similitud de los resultados obtenidos mediante PTML en los modelos seleccionados. Se presentan rangos similares en cuanto a las variables de estudio. Hemos revisado que existe un limitado desarrollo de software destinado a la automatización de PTML que contiene propiedades de cálculo como las medidas de dispersión aplicadas a los descriptores. La metodología en PTML permite establecer medidas de dispersión sobre descriptores de las propiedades fisicoquímicas de diversos organismos. Esta revisión muestra como los trabajos desarrollados en PTML permiten obtener nuevos componentes químicos, predecir actividades biológicas y el desarrollo de estudios experimentales paralelos en múltiples que luego pueden ser sintetizados y desarrollar pruebas farmacológicas, entre otros. Los valores de sensibilidad, especificidad y precisión son superiores al 70%. En general, los estudios comienzan con la selección de la información de las fuentes de datos diversos. El modelo PTML con el valor medio más alto del 89% es el estudio de actividad antibacteriana de Nocedo-Mena *et al*.

Capítulo 2

CAPÍTULO 2, APLICACIONES: MAPEO IFPTML DE LA ACTIVIDAD ANTIBACTERIANA DE LAS NANOPARTÍCULAS FRENTE A LAS REDES METABÓLICAS DE LOS PATÓGENOS.

2. Artículo 2

Ref. Ortega-Tenezaca, B., & González-Díaz, H. (2020). IFPTML mapping of nanoparticle antibacterial activity *vs.* pathogen metabolic networks. *Nanoscale*, *13*, 1318–1330. doi: https://doi.org/10.1039/D0NR07588D

2.1. INTRODUCCIÓN

Las nanopartículas (NPs) tienen múltiples aplicaciones en las ciencias biomédicas, incluyendo su uso en sistemas de liberación de fármacos anticancerígenos o contra la malaria,(Zelepukin *et al.*, 2019; Najer *et al.*, 2016; Li *et al.*, 2015; Brunetti *et al.*, 2015; Cai & Yao, 2013; De Angelis *et al.*) entrega y/o silenciamiento de genes, (Wang *et al.*, 2010; Stanwix, 2012; Dizaj *et al.*, 2014; Schulze *et al.*, 2018; Wang *et al.*, 2019) en imágenes médicas para la investigación del cáncer(Ahn *et al.*, 2010; Betzer *et al.*, 2014; Hsu *et al.*, 2018)*, etc*. Recientemente, los investigadores se han centrado no sólo en las NPs como sistema de liberación de fármacos, sino también en la actividad biológica de las NPs en sí mismas (Nabil *et al.*, 2020; Caron *et al.*, 2013). En este trabajo nos centraremos específicamente en las NPs con actividad antibacterianas (Ruparelia *et al.*, 2008; Pramanik *et al.*, 2012; Azam A. *et al.*, 2012; Azam A. *et al.,* 2012; Hossain & Mukherjee, 2013; Botequim *et al.*, 2012; Taglietti *et al.*, 2012; Hossain & Mukherjee, 2012; Premanathan *et al.*, 2011; Inbaraj *et al.*, 2011; Hsu D. D., 2013; Zhao *et al.*, 2013; Zhen *et al.*, 2020; Arasoglu, 2016; Elizabeth *et al.*, 2014; Wong *et al.*, 2015). El uso del cribado de alto rendimiento (HTS) y técnicas similares para ensayar múltiples muestras de NPs

automáticamente tiene el potencial de acelerar el proceso de descubrimiento de NPs (Holden *et al.*, 2013; Tanaka *et al.*, 2017; Mu *et al.*, 2016; Hubble *et al.*, 2015; Shlar *et al.*, 2015). En cualquier caso, el descubrimiento de nuevas NPs sigue siendo un proceso experimental costoso y que requiere mucho tiempo. Las principales razones son el elevado número de combinaciones de tamaños, formas y composiciones de NPs frente al elevado número de combinaciones de especies de bacterias y pruebas biológicas. Además, la aparición de bacterias Multirresistentes (MDR) ponen más presión sobre los investigadores que buscan NPs con actividad antibacteriana (Fischbach & Walsh, 2009). Las bacterias MDR presentan mutaciones en proteínas específicas (probablemente como respuesta a condiciones de estrés) que pueden provocar cambios en cascada en su genoma, proteoma, metabolismo y susceptibilidad/resistencia de las bacterias a fármacos antibacterianos o NPs en última instancia. Estos cambios pueden estudiarse cuantificando la estructura topológica de las Redes Reguladoras de Genes (GRNs), las Redes de Interacción de Proteínas (PINs) y/o las Redes de Reacción Metabólica (MNs), de las bacterias. Estas complejas redes pueden verse como grandes colecciones de nodos (genes, proteínas, enzimas, metabolitos*, etc*.) interconectados por enlaces (reacciones metabólicas, procesos de transporte, señalización*, etc*.) (Jeong *et al.*, 2000). De hecho, Nagar *et al*., estudiaron el efecto de las condiciones de estrés de pH, temperatura y antibióticos sobre los PINs de E. coli (Nagar *et al.*, 2016). En otro esfuerzo importante, Larocque *et al.*., publicaron iMLTC806cdf, un conjunto de datos ampliamente reconstruido de GRN, PIN y/o MNs de la cepa patógena de C. difficile 630 (Larocque *et al.*, 2014). Todo esto nos habla de la importancia de tener en cuenta la estructura de los MNs de las especies/cepas de bacterias patógenas en el descubrimiento de nuevas NPs con actividad antibacteriana.

En este contexto, las técnicas computacionales pueden desempeñar un papel importante en la reducción del tiempo y los costes en el descubrimiento de nuevas NPs con actividad biológica (Duncan & Bevan, 2015). En concreto, las técnicas de tipo quimio informático basadas en el análisis de datos, la estadística, el ML y/o la AI han cobrado impulso en la nanotecnología en los últimos años (Manganelli *et al.*, 2016; Toropova *et al.*, 2015; Toropova *et al.*, 2016; Rybinska-Fryca *et al.*, 2020; Le *et al.*, 2016; Ahmadi *et al.*, 2020; Ojha *et al.*, 2019; Sizochenko *et al.*, 2018; Tasi *et al.*, 2018;Villaverde *et al.*, 2018; Sizochenko *et al.*, 2017; Manganelli & Benfenati, 2017; Puzyn *et al.*, 2011). No obstante, muchos de estos modelos se centran en la toxicidad de las NPs y sólo unos pocos trabajos importantes se han dedicado a la modelización de la actividad antibacteriana de las NPs (Puzyn *et al.*, 2011; Toropov *et al.*, 2012). Sin embargo, es importante mencionar que no sólo los modelos de NPs, sino también otros modelos de actividad de NPs no tienen en cuenta las múltiples condiciones de ensayo que hay que tener en cuenta en los ensayos biológicos. Sizochenko *et al*., informaron de uno de los trabajos pioneros sobre la predicción de la toxicidad de NPs contra varias especies simultáneamente (Sizochenko *et al.*, 2018). De todos modos, muchos modelos NPs no tienen en cuenta la actividad de la NP contra diferentes especies y cepas de bacterias. En nuestra opinión, esta situación se debe a dos factores principales. Por un lado, no hay suficientes informes experimentales sobre la actividad biológica de la NP para ser estudiada con técnicas de ML/AI. Gejewicz *et al.*., discutieron el efecto de este factor en el desarrollo actual de modelos ML en nanotecnología (Gajewicz, 2017). Por otro lado, la naturaleza del método computacional puede ser inadecuada (Jagiello *et al.*, 2016). Hasta donde sabemos no existen estudios computacionales que incluyan la actividad antibacteriana de las NPs y la estructura de los MNs de las bacterias objetivo hasta la fecha.

De hecho, muchos modelos de quimio informática basados en ML utilizan como entrada sólo los descriptores estructurales/moleculares, *D*k, del sistema en estudio y omiten todas las demás variables y factores no estructurales. En este contexto, González-Díaz *et al*., introdujeron el enfoque de aprendizaje automático (ML) basado en la teoría de la perturbación (PT) y la fusión de la información (IF), conocido como el enfoque de la fusión de la información IFPTML (Gonzalez-Diaz *et al.*, 2013). La idea general es predecir los valores de salida de la función $f(v_{ij})_{calc}$ para la *j*-ésima propiedad múltiple del sistema de consulta (S_i) en estudio con un único modelo. Para ello, IFPTML comienza con el valor de una función de referencia medida experimentalmente $f(v_{ij})_{ref}$ para sistemas ya conocidos. A continuación, en la fase de preprocesamiento de PT se recogen/calculan las variables estructurales del sistema de consulta S_i y de los sistemas de referencia. Posteriormente, IFPTML calcula los valores de los operadores PTOs utilizados para medir las desviaciones o perturbaciones en las variables del sistema de consulta S_i con respecto a todas las variables estructurales y no estructurales de los sistemas de referencia. Por último, en la fase de ML el algoritmo IFPTML realiza el entrenamiento/validación del modelo con las técnicas de ML actuales. Los modelos IFPTML se han utilizado en diferentes disciplinas para predecir la actividad biológica de fármacos, proteínas, materiales, NPs y sistemas más complejos como redes moleculares y sociales complejas. Por ejemplo, el IFPTML se ha utilizado para mapear información química y biológica sobre ensayos preclínicos de cócteles de fármacos contra el VIH y sus proteínas objetivo frente a información socioeconómica y epidemiológica de todos los condados de EE.UU. En Nanotecnología, el IFPTML se ha utilizado para estudiar la toxicidad de las NPs frente a diferentes especies en múltiples condiciones experimentales (González-Díaz *et al*., 2014; Herrera-Ibatá *et al*., 2015). IFPTML también se ha utilizado para estudiar sistemas complejos de co-distribución de NPs,

incluyendo sobre NPs agentes de recubrimiento, derivados vitamínicos de co-terapia y fármacos anticancerígenos (Santana *et al*., 2019; Santana *et al*., 2020a; Santana *et al*., 2020b).

Recientemente, Speck-Planche *et al*., informaron de muchos modelos PTML de la toxicidad y la actividad antibacteriana de las NPs frente a múltiples especies en diferentes condiciones (Concu *et al*., 2017; Luan *et al*., 2014; Speck-Planche *et al.*, 2015f). En concreto, su modelo PTML para la actividad de las NPs fue un importante paso adelante. Este modelo fue probablemente el primero que tuvo en cuenta múltiples clases de NPs, agentes de recubrimiento, propiedades biológicas y especies y cepas de bacterias (Speck-Planche *et al*., 2015). Sin embargo, el modelo de Speck-Planche para la actividad de las NPs sólo es capaz de hacer predicciones para aquellas especies/cepas bacterianas ya presentes en el conjunto de datos. El análisis no incluye una fase de IF que tenga en cuenta la información sobre las MNs de la especie bacteriana objetivo. En consecuencia, este modelo no puede predecir la actividad de las NPs contra otras especies/cepas de bacterias no presentes en el conjunto de datos y/o con diferente estructura de MNs. Curiosamente, Nocedo *et al*., publicaron un modelo IFPTML que tiene en cuenta la estructura de las MNs de las bacterias patógenas junto con múltiples especies, condiciones de ensayo*, etc.,* (Nocedo-Mena *et al*., 2019). Desafortunadamente, el modelo de Nocedo *et al*., sólo se aplica a compuestos orgánicos de bajo peso molecular y es inútil en la predicción de la actividad de las NPs. En consecuencia, en este trabajo desarrollamos el primer modelo IFPTML para la actividad de las NPs que incluye no sólo múltiples tipos de NPs, agentes de recubrimiento, condiciones de ensayo y propiedades biológicas, sino también la estructura de las MNs de las especies/cepas bacterianas implicadas. Al hacerlo, también probamos por primera vez nuestro software SOFT.PTML para el desarrollo de modelos IFPTML en Nanotecnología. SOFT.PTML tiene una interfaz fácil de usar, diseñada específicamente para facilitar el

preprocesamiento de datos y el entrenamiento y validación de modelos IFPTML. El nuevo modelo encontrado será capaz de predecir la probable actividad de las NPs frente a nuevas especies o cepas de bacterias con cambios en la topología de las MNs.

2.2. MATERIALES Y MÉTODOS

2.2.1. Pasos del análisis de datos de IFPTML NP y MN

El procedimiento de análisis de datos IFPTML utilizado aquí implica tres pasos (PT + ML + IF). Estos pasos generales son: (1) adquisición de datos, (2) preprocesamiento de los datos PT y cálculo de los PTOs, (3) fusión IF de los conjuntos de datos NPs y MNs y, (4) entrenamiento y validación del modelo ML. En el paso (1) obtuvimos los resultados de muchos ensayos preclínicos de NPs, así como información complementaria sobre NPs y MNs de dos conjuntos de datos ya publicados (Nocedo-Mena *et al*., 2019; A. Speck-Planche *et al*., 2015f) Como resultado obtenemos los conjuntos de datos NPs-set y MN-set. En el paso (2) escalamos todos los parámetros de NPs y MNs a medidas de información de entropía de Shannon. Este tipo de parámetros fueron introducidos en la Teoría de la Información por Claude E. Shannon (Shannon, 1948). Se han utilizado para cuantificar la cantidad de información de una señal o sistema, incluyendo estructuras moleculares, secuencias de proteínas, redes complejas*, etc*., (Graham, 2002, 2005; Graham *et al.*, 2004; Riera-Fernández *et al*., 2012; Strait & Dewey, 1996). También calculamos los PTOs de todas las variables de entrada en cada conjunto de datos de forma independiente. En el paso (3) realizamos el proceso de IF con el conjunto de datos NP y el conjunto de datos MN para obtener un conjunto de datos de trabajo más amplio. En el paso (4) realizamos el entrenamiento y la validación de varios algoritmos de ML para obtener nuestro modelo IFPTML. En las siguientes secciones informamos de los detalles de cada uno de los

pasos anteriores. En la Figura 9 ilustramos el flujo de trabajo del IFPTML integrando todos los pasos anteriores.

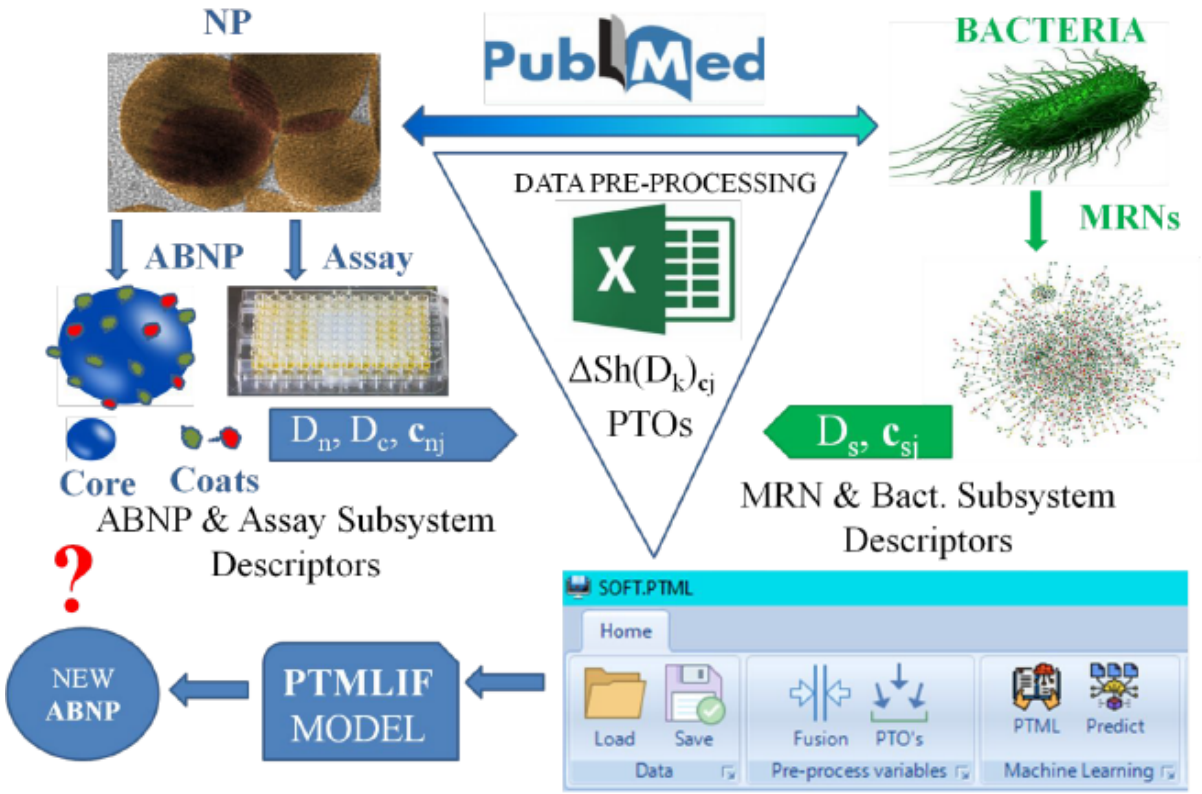


Figura 9. IFPTML - Flujo de trabajo usado en esta sección

2.2.2. Conjunto de datos de nanopartículas (NP-set).

Utilizamos un conjunto de datos previamente reportado con los resultados de N_n = 300 ensayos preclínicos de NPs metálicas, de sales metálicas y de óxidos metálicos contra diferentes especies de bacterias (s) (Speck-Planche *et al.*, 2015f). Las NPs metálicas tienen un núcleo de: oro (Au), plata (Ag) o cobre (Cu). Los núcleos de las NPs de sal metálica están hechos de sulfuro de cadmio (II) (CdS) o yoduro de cobre (I). Las NPs de óxido metálico incluyen: óxido de cadmio(II) (CdO), óxido de zinc (ZnO), óxido de cobre(II) (CuO), óxido de lantano(III) (La_2O_3), óxido de aluminio (Al_2O_3), óxido de hierro(III) (Fe_2O_3), óxido de estaño(IV) (SnO_2), óxido de

titanio(IV) (TiO_2), óxido de hierro(II, III) (Fe_3O_4) y dióxido de silicio (SiO_2). Los ensayos de estos 15 nanomateriales implicaron múltiples condiciones experimentales c_{nj}. Enumeramos todas las condiciones específicas de un ensayo como un vector $c_{nj}= [c_{nj},c_{nj},c_{nj},\ldots,c_{nmax}]$. Estas condiciones de ensayo incluyen la medición de 1 de los 4 posibles parámetros de actividad antibacteriana, frente a 1 de las 34 posibles especies de bacterias (incluyendo diferentes cepas). Otras etiquetas o condiciones experimentales consideradas son la selección de al menos 1 de 3 formas de NPs y la realización del experimento en 1 de 4 posibles intervalos de tiempo. Los datos originales se descargaron de la base de datos OCHEM (https://ochem.eu/home/show.do) (Sushko *et al*., 2011) y de otras fuentes (Azam *et al*., 2012a; Azam *et al.*, 2012b; Botequim *et al*., 2012; Hossain & Mukherjee, 2012, 2013; Hu, Cook *et al.*, 2009; Inbaraj *et al*., 2011; Pramanik *et al.*, 2012; Premanathan *et al.*, 2011; Ruparelia *et al.*, 2008; Taglietti *et al*., 2012; Zhao *et al*., 2013). El conjunto de datos también incluía información sobre los parámetros fisicoquímicos de las NPs y los agentes de recubrimiento utilizados. El tamaño de las NPs de este conjunto de datos se sitúa en el rango de 3 a 98 nm, que está dentro del rango de 1 a 100 nm definido para las nanopartículas (Vert *et al*., 2012). Casi todas las NPs estudiadas se sitúan en la mitad de tamaño más pequeño de este rango, incluyendo 194 NPs con 0-10 nm y 104 NPs con 10-50 nm. Sólo hay 2 NPs con un tamaño >50 nm, véase la figura 10 (A. Speck-Planche *et al*., 2015).

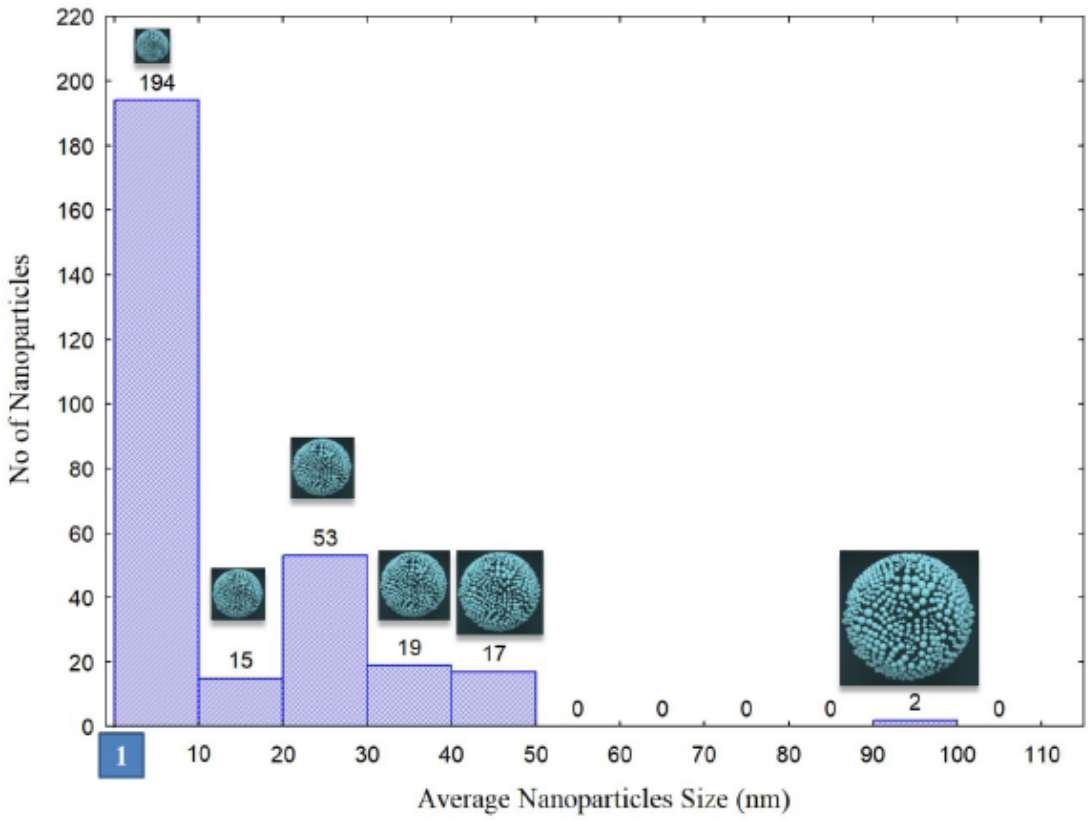


Figura 10. Histograma de distribución de las nanopartículas 1-100 (nm)

2.2.3. Conjunto de datos de MNs bacterianas (MN-set)

Los datos fueron publicados por el grupo de Barabasi como archivos ASCII comprimidos en gzip (Jeong *et al*., 2000). Para cada reacción, los eductos y los productos fueron considerados como nodos conectados a los complejos educto-educto temporales y a las enzimas asociadas. Los nombres, abreviaturas y enlaces de todas las redes estudiadas. Información detallada, los índices topológicos, los nombres completos y los códigos de >20 especies de bacterias estudiadas aquí aparecen en la información de apoyo.

2.2.4. Escala de entropía de Shannon de la información estructural de las NPs.

El conjunto original de NPs contiene diferentes parámetros físico-químicos experimentales/teóricos para caracterizar los detalles de la estructura/composición de las NPs. Estos parámetros son el volumen molar medio (AMV), la Electronegatividad Atómica Promedio (AAE), y la Polarizabilidad Atómica Promedio (AAP). Estas propiedades fisicoquímicas se obtuvieron del sitio web Chemicool Periodic Table (http://www.chemicool.com/elements) (Hsu, 2013). El cuarto parámetro fue el tamaño medio de las partículas (APS) expresado en nanómetros (nm). Sin embargo, para llevar a cabo el proceso de FI haciendo una fusión de las NPs y MNs en el mismo conjunto de datos de trabajo, decidimos expresar toda la información en la misma escala. En consecuencia, la información original se transformó en una escala de entropía de Shannon previamente a la fusión, véase la Tabla 6. La información sobre el núcleo de la NP y los agentes de recubrimiento se ha escalado utilizando las siguientes fórmulas para calcular los valores de entropía de Shannon.

$$p(D_k) = \frac{1}{(1 + exp(-D_k/1000)} \qquad (15)$$

$$Sh(D_k) = -p(D_k) \times log(p(D_k)$$

El valor de 1000 en la relación $D_k/1000$ se utilizó como valor de escalado. Se utilizó el mismo tipo de operadores $Sh_k(D_k)$ para escalar todos los descriptores que cuantifican la información sobre la estructura de los diferentes subsistemas. Esto significa que aplicamos el mismo operador Sh_k (D_k) a los descriptores estructurales del núcleo de la NP(D_{kn}). Después, obtuvimos los valores de entropía Sh_k (D_k). Con los valores Sh_k (D_{kn}) podemos calcular los PTOs de los ensayos de NPs utilizados como entrada para el modelo IFPTML. Los PTOs calculados

aquí tienen la forma de Masmulticondición por analogía con los informes anteriores. La fórmula de estos PTOs es la siguiente $\Delta Sh(D_{kn}) = Sh_{kn} - \langle Sh_{kn} \rangle_{cn}$. En la Tabla 6 también se muestran ejemplos seleccionados de los valores medios$\langle Sh_{kn} \rangle_{cn}$ para diferentes subconjuntos de condiciones de ensayo de NPs, c_n.

Tabla 6. Medidas de información y promedios de entropía de Shannon NP (ejemplos seleccionados)

Tipo NP	NP	Forma	$Sh_1(MW_n)$	$Sh_2(AMV_n)$	$Sh_3(AAE_n)$	$Sh_4(AAP_n)$	$Sh_5(APS_n)$
	ZnO	Acicular	0.148	0.150	0.150	0.150	0.150
	ZnO	N/A	0.148	0.150	0.150	0.150	0.149
	CuO	N/A	0.148	0.150	0.150	0.150	0.149
	La_2O_3	N/A	0.137	0.150	0.150	0.150	0.149
	Al_2O_3	N/A	0.147	0.150	0.150	0.150	0.149
Óxido	Fe_2O_3	N/A	0.145	0.150	0.150	0.150	0.149
	SnO2	N/A	0.145	0.150	0.150	0.150	0.149
	TiO2	N/A	0.148	0.150	0.150	0.150	0.149
	SiO2	N/A	0.148	0.150	0.150	0.150	0.149
	CdO	Esférica	0.146	0.150	0.150	0.150	0.150
	Fe3O4	Esférica	0.141	0.150	0.150	0.150	0.150
	Cul	N/A	0.143	0.150	0.150	0.150	0.150
	CdS	Esférica	0.145	0.150	0.150	0.150	0.150
Metal	Au	Esférica	0.143	0.150	0.150	0.150	0.151
	Ag	Esférica	0.147	0.150	0.150	0.150	0.150
	Cu	Esférica	0.148	0.150	0.150	0.150	0.150
NP	Org.	Cepa	Forma	Valores medios			
Tipo	cn_1	cn_2	cn_3	$\langle Sh_1(AMV_n) \rangle$	$\langle Sh_1(AEE_n) \rangle$	$\langle Sh_1(AAP_n) \rangle$	$\langle Sh_1(APS_n) \rangle$
	EC	K-12	Esférica	0.146	0.150	0.150	0.150
	EC	MDR		0.143	0.150	0.150	0.150
	EC	ATCC 10536		0.147	0.150	0.150	0.150
	EF	VCM-R		0.147	0.150	0.150	0.150
Todo	SA	ATCC 9144	Acicular	0.148	0.150	0.150	0.150
	EC	ATCC 10536		0.148	0.150	0.150	0.150
	PA	ATCC 9027		0.148	0.150	0.150	0.150

2.2.5. Escala de entropía de Shannon del tiempo de ensayo de las NPs y de la información estructural del recubrimiento.

Las NPs estudiadas aquí vienen con información adicional aparte de los descriptores de la forma del núcleo de la NP, el volumen*, etc*. En concreto, pueden presentar diferente número de agentes de recubrimiento monomérico (mono) o polimérico (poli) con N_{coat}= 0 (desnudo), 1 (recubrimiento simple) ó 2 (recubrimiento doble). Además, los diferentes ensayos de NPs pueden tener diferentes tiempos de duración t(h) = 2, 5, 18 o 24 horas. Se considera que esta información es complementaria y relevante, pero su grado de variabilidad es bajo. En primera instancia, intentamos cuantificar esta información utilizando como PTOs los respectivos MA de cada uno de los valores de entropía $\Delta Sh(D_{1c})$, $\Delta Sh(D_{2c})$ y $\Delta Sh(t)$. La fórmula del MA se deduce por analogía con los casos anteriores $\Delta Sh(D_{1c}) = Sh(D_{1c}) - \langle Sh(D_{1c})_{cc} \rangle$, $\Delta Sh(D_{2c}) = Sh(D_{2c}) - \langle Sh(D_{1c})_{cc} \rangle$ y $\Delta Sh(t) = Sh(t) - \langle Sh(t)_{cc} \rangle$. En la Tabla 7 mostramos los valores individuales de $Sh(D_{kn})$ y los valores medios $\langle Sh(D_{kn})_{cc} \rangle$ para cada descriptor D_{kn} de los agentes de recubrimiento. Estos MAs cuantifican la variabilidad en el primer agente de recubrimiento, el segundo agente de recubrimiento (si lo hay), y el tiempo de ensayo, respectivamente. Sin embargo, los valores de varianza de estos MAs eran demasiado bajos para ser incluidos en el análisis ML. En consecuencia, decidimos codificar toda esta información en un tipo modificado de PTOs basado en múltiples medidas de información de entropía de Shannon $\Delta Sh(D_{1c},D_{2c},t)$. El uso de muchos tipos diferentes de PTOs en el análisis IFPTML aplicado a la Nanotecnología fue discutido en la literatura anteriormente (Santana *et al*., 2019; Santana *et al*., 2020a; Santana, *et al*., 2020b). El PTO $\Delta Sh(D_{1c},D_{2c},t)$ en su conjunto tiene un mayor grado de varianza que cada MA individual; lo cual es un factor importante a incluir en el algoritmo ML. La fórmula de este operador es la siguiente:

$$\Delta Sh(D_{1c}, D_{1c}, t) = \Delta Sh(t) - [\Delta Sh(D_{1c}) + \Delta Sh(D_{2c})] \quad (16)$$

Tabla 7. Medidas de información de la entropía de Shannon para agentes de recubrimiento NPs.

Sistemas de recubrimiento			Información numérica de los sistemas de recubrimiento			
N_{agent} cc1	$Poly_{cc2}$	Sistema de recubrimiento	Agente de recubrimiento 01 $Sh(LOGP_{1c})$	$Sh(PSA_{1c})$	Agente de recubrimiento02 $Sh(LOGP_{2c})$	$Sh(PSA_{2c})$
Doble	Mono	PDT/Mel	0.148	0.151	0.146	0.151
		PDT/Ach	0.148	0.151	0.150	0.151
		PDT/CQ	0.148	0.151	0.150	0.150
		PDT/DMB	0.148	0.151	0.147	0.150
		PDT/CPB	0.148	0.151	0.147	0.151
		PDF/G	0.148	0.151	0.148	0.151
Simple		PDT	0.148	0.151	0.151	0.151
		Maltosa	0.143	0.151	0.151	0.151
		Lactosa	0.143	0.151	0.151	0.151
		Glutatión	0.145	0.151	0.151	0.151
		Glucosa	0.147	0.151	0.151	0.151
		DMA	0.151	0.150	0.151	0.151
		Galactosa	0.147	0.151	0.151	0.151
	Poly	PVP	0.150	0.151	0.151	0.151
		PGA	0.147	0.151	0.151	0.151
Ninguna	Ninguna	Ninguna	0.151	0.151	0.151	0.151
N_{agent} cc_1	Poly. cc_2	Sistema de recubrimiento	Agente de recubrimiento01 $Sh(LOGP_{1c})$	$Sh(PSA_{1c})$	Agente de recubrimiento02 $Sh(LOGP_{2c})$	$Sh(PSA_{2c})$
Doble	Mono	I	0.148	0.150	0.148	0.150
Simple	Mono	II	0.146	0.151	0.151	0.151
Simple	Poly	III	0.148	0.151	0.151	0.151
Ninguna	Ninguna	IV	0.151	0.151	0.151	0.151

2.2.6. Escala de entropía de Shannon de la información estructural local de MNs.

Como hemos mencionado antes, el mismo tipo de operadores $Sh_k(D_k)$ puede utilizarse para diferentes subsistemas. En primer lugar, calculamos los parámetros N_{ms} número de metabolitos (m), o $D_{ks} = <L_{ins}>$ grado medio de entrada, $D_{ks} = <L_{outs}>$ grado medio de salida para todos los metabolitos en el MN del s-ésimo organismo. El cálculo de estos parámetros se llevó a cabo con el software MI-NODES (Duardo-Sánchez *et al*., 2014) desarrollado por nuestro grupo y verificado con el software CentBin (Junker *et al.*, 2006). A continuación, utilizando la Ecuación 16 también aplicamos el mismo operador de probabilidad $p(D_k)$ a los descriptores

estructurales de los y MNs (D_{ks}). Después obtuvimos los valores de entropía respectivos $Sh(D_{ks})$ descriptores $Sh(N_{ms})$, $Sh(L_{ins})$ y $Sh(L_{outs})$ de MN del s-ésimo organismo utilizando la ecuación 17. Estos valores se han calculado en este trabajo por primera vez para este conjunto de MNs.

2.2.7. Escala de entropía Markov-Shannon de la información estructural de alto orden de los MNs.

En cualquier caso,N_{ms}, $<L_{ins}>$, y $<L_{outs}>$ son descriptores topológicos locales que sólo dan cuenta de la información del nodo (metabolito en cuestión) y de los nodos directamente vinculados a él precursores directos (eductos) para el caso de$<L_{ins}>$ y productos directos (aductos) para el caso de$<L_{outs}>$ (Jeong *et al.*, 2000). En consecuencia, también calculamos las medidas de información de Shannon para cuantificar la información estructural de orden superior de los MNs (metabolitos distantes en el MN). Sin embargo, en este caso particular, el operador no se aplica al descriptor local *per se*. En este caso, aplicamos el operador a las probabilidades obtenidas de un cálculo de la cadena de Markov. De este modo, calculamos los valores de entropía Sh_k de orden k-ésima para la especie s-ésima. Los valores Sh_k miden la información de conectividad en el MN de la s-ésima especie para todos los metabolitos y sus vecinos (sustratos o productos) situados a una distancia (número de reacciones) $\leq$ k. En concreto, los valores $Sh_k(\pi_1)$ miden la información para cada m-ésimo nodo (metabolito) en los MNs que son vecinos directamente conectados (sustratos o productos) al metabolito de consulta (k = 1). Además, $Sh_k(\pi_2)$ mide la información para los metabolitos que son vecinos del vecino directo (productos o sustratos conectados indirectamente a distancia dos k = 2).

Estos valores se han calculado en este trabajo por primera vez para este conjunto de MNs. Se pueden ver también los nombres de los organismos, los códigos de dos letras y los valores de

$Sh_k(\pi_1)$ y $Sh_k(\pi_2)$ para todos los MNs estudiados. En este archivo también explicamos en detalle los aspectos técnicos del cálculo de estos parámetros. La fórmula específica utilizada para calcular estos valores de $Sh(\pi_1)$ y $Sh_k(\pi_2)$ de las MNs se han reportado en la literatura y se representan y explican también en información de apoyo (Ahmadi *et al.*, 2020).

2.2.8. IF de los conjuntos de datos NPs vs. MNs (conjunto W).

En el proceso de FI, comenzamos con el conjunto NPs y añadimos los valores del conjunto MNs para crear el nuevo conjunto de datos de trabajo (conjunto W). de trabajo (conjunto W). Por lo tanto, cada fila del nuevo conjunto W está compuesta por una fila (caso de ensayo NP) del conjunto NPs y una fila (caso MN) del conjunto MNs. El nuevo conjunto W contiene un total de 5.327 casos (casos NPs *vs.* MNs) incluyendo todas las etiquetas de los casos y las condiciones experimentales c_{nj} y c_{sj} de los conjuntos originales. El nuevo conjunto W incluye todos los valores de $\Delta Sh(D_{kn})_{cnj}$ y $\Delta Sh(D_{ks})_{csj}$ utilizados para cuantificar la información de las variables de entrada. El conjunto W también incluye las funciones objetivo f(n,j,s) que debe ajustar el modelo y las funciones auxiliares utilizadas para definirlo.

2.2.9. Preprocesamiento de los valores observados de los parámetros biológicos de NPs.

Los parámetros $v_{nj}(c_{n0})$ se utilizan para cuantificar la actividad biológica del n-ésimo sistema NP en el j-ésimo ensayo con condiciones codificadas por el vector $c_{nj}=(c_{n0,}c_{n1},\ldots,c_{nmax})$. Esta variable se refiere a los valores numéricos v_{nj} de diferentes parámetros experimentales con nombre $c_{n0}(IC_{50}(\mu M)$, MIC(μM), etc.), véase la Tabla 5. Sin embargo, como tenemos múltiples parámetros c_{no} con diferentes unidades y errores decidimos transformar todos los valores $v_{nj}(c_{n0})$ en la función booleana $f(n,s,j)_{obs}$ =1ó 0. Esta variable de salida $f(n,s,j)_{obs}$ es la función objetivo que debe ajustar el modelo (mencionada en la sección anterior). Para definir esta función

utilizamos los valores originales de la actividad biológica $v_{nj}(c_{n0})$ y los parámetros cutoff(c_{n0}) y deseabilidad d(c_{n0}). El parámetro cutoff(c_{n0}) es un valor umbral utilizado para delimitar los NPs con efectos fuertes frente a los débiles. El parámetro deseabilidad (c_{n0}) obtiene el valor d(c_{n0}) = 1 cuando los parámetros c_{n0} deben minimizarse para obtener un NP óptimo, es decir, IC_{50}(μM) o d(c_{n0}) = 0 en caso contrario. Con estos parámetros obtuvimos los valores de la función f(n,s,j) como sigue: $f(n,s,j)_{obs}$ = 1 cuando $v_{nj}(c_{n0})$ >cutoff(c_{n0}) y deseabilidad d(c_{n0}) = 1. El valor es también $f(n,s,j)_{obs}$ = 1 cuando $v_{nj}(c_{n0})$ <cutoff(c_{n0}) y deseabilidad d(c_{n0}) = –1, $f(n,s,j)_{obs}$ = 0 en caso contrario. El valor $f(n,s,j)_{obs}$ = 1 señala un fuerte efecto deseable del NP sobre la especie o cepa bacteriana, mientras que $f(n,s,j)_{obs}$=0 indica un efecto débil con respecto al corte utilizado (Schulze *et al*., 2018). Una vez obtenidos los valores de $f(n,s,j)_{obs}$ para todo el conjunto de NP, contamos el número total de casos n(c_{n0}) y el número de casos positivos $n_1(c_{n0})$, casos con $f(n,s,j)_{obs}$=1, para cada propiedad (c_{n0}). Con estos parámetros calculamos los valores de la función de referencia $f(c_{n0})_{ref}$ =$n_1/n(c_{n0})$. Esta definición nos permite interpretar la función de referencia como la probabilidad previa $f(c_{n0})_{ref}$ =p(f(v_{ij}))= $1/c_{n0})_{ref}$ para que un NP tenga buenos valores de los diferentes parámetros c_{n0}. En la Tabla 8 representamos los valores de la función de referencia, el corte y otros parámetros utilizados para las diferentes propiedades biológicas.

Tabla 8. Función de referencia, límite y otros valores de medidas de los efectos biológicos NPs.

c_{n0}(Unidades)	Concepto	Cutoff(c_{n0})[a]	d(c_{n0})	n(c_{n0})$_j$	n_1(c_{n0})	f(c_{n0})$_{ref}$
IC_{50}(μM)	Concentración necesaria para la inhibición del 50% del crecimiento de la bacteria.	≤ 89.98	-1	164	96	0.59
MIC(μM)	Concentración mínima inhibitoria, es decir, la concentración mínima necesaria para impedir el crecimiento visible de la bacteria.	≤ 125.15	-1	123	55	0.45
MBC(μM)	Concentración bactericida mínima, es decir, la concentración mínima necesaria para matar completamente la bacteria.	≤ 173.03	-1	9	1	0.11
MBE(n.u.)	Fracción de efecto microbicida, ver detalles en la referencia anterior(Speck-Planche *et al.*, 2015f)	≥ 0.96	1	4	2	0.50

[a]Condición bajo la cual un NP fue considerado como activo, ver detalles en la referencia anterior (Speck-Planche *et al*., 2015f). El término (n.u.) = sin unidades = adimensional.

2.2.10. Modelo lineal IFPTML.

Las ideas del método PTML quimio-informática se han ampliado aquí para encontrar el nuevo modelo IFPTML. Las salidas del modelo son los valores de la función de puntuación $f(n,c,j,s)_{calc}$. Se calculan para el par formado por la $q^{th}NP_q$ ensayada en el *j*-ésimo ensayo preclínico con condiciones $c_{nj}=(c_{n0},c_{n1},c_{n2},\ldots,c_{njmax})$ contra la s-ésima especie bacteriana con MNs. Los modelos lineales IFPTML probados para este sistema tienen la siguiente ecuación general:

$$f(n,c,j,s)_{calc} = a_0 + a_1 \cdot f(c_{n0})_{ref} + \sum_{k=1,j=1}^{k=k\max,\, j=j\max} a_{kj} \cdot \Delta Sh_k(NP_q)_{cn_j} + \sum_{k=1,j=1}^{k=k_{\max},\, j=j_{\max}} a_{kj} \cdot \Delta Sh_k(MN_s)_{cd_j} \qquad (17)$$

El modelo IFPTML parte del valor esperado de la actividad biológica $f(v_{ij})_{expt}$ y suma el efecto de la información química relacionada con la estructura del fármaco y los efectos acumulados debidos a cambios o perturbaciones (operadores PT) en las condiciones de ensayo o la cepa bacteriana utilizada. Los operadores PT utilizados aquí son similares a los operadores de media móvil (MA) de Box-Jenkins utilizados en trabajos anteriores (Ojha*et al.*, 2019). Los otros operadores PT incluidos en este modelo son MA calculados para una condición a la vez. En el

modelo IFPTML tenemos dos tipos de operadores PT debido al proceso IF. Un tipo de operadores PT son los términos $\Delta Sh_k(NP_q)_{cj}$ y el otro tipo son los términos $\Delta Sh_k(MN_s)_{cj}$. Calculamos estas variables como sigue: $\Delta Sh_k(NP_q)_{cj}=Sh_k(NP_q)-<Sh_k(NP_q)_{cj}>$ ó $\Delta Sh_k(MN_s)_{cj}=Sh_k(MN_s)-<Sh_k(MN_s)_{cj}>$, respectivamente. En consecuencia, los términos $\Delta Sh_k(NP_n)_{cj}$ dan cuenta de la desviación de la información fisicoquímica $Sh_k(NP_n)$ del NP respecto al valor esperado de $<Sh_k(NP_q)_{cj}>$ (valor medio) para todos los NPs ensayados en las mismas condiciones en NPset (Ahmadi *et al.*, 2020). Por analogía, los términos $\Delta Sh_k(MN_s)_{cj}$ cuantifican la desviación de la información metabólica de la bacteria $Sh_k(MN_s)$ del valor esperado de $<Sh_k(MN_s)_{cj}>$ para todas las bacterias en el conjunto MNR con las mismas condiciones c_j.

2.2.11. Entrenamiento y validación de los modelos IFPTML.

Para entrenar y validar el modelo, todos los casos del conjunto W se asignaron al azar a series de entrenamiento (conjunto = t) ó de validación (conjunto = v) utilizando la función f(conjunto) = 1 (conjunto = t) o = 0 (conjunto = v). El conjunto t se utilizó para entrenar el modelo IFPTML y el conjunto v para validarlo. En primer lugar, se utilizó el algoritmo de análisis discriminante lineal (LDA) para encontrar el modelo preliminar. Se utilizó el procedimiento Forward Step-Wise (FSW) como estrategia de selección de variables para una selección automática de las características de entrada. El programa utilizado fue STATISTICA 6.0.81 Después, se utilizó una heurística de selección guiada por expertos (EGS) para volver a entrenar el modelo LDA con las características más importantes seleccionadas por FSW y otras características que faltaban. A continuación, utilizamos el software SOFT.PTML para desarrollar modelos alternativos de IFPTML con fines de comparación (Ortega-Tenezaca *et al*., 2020). SOFT.PTML tiene una interfaz fácil de usar especialmente diseñada para el desarrollo de

modelos IFPTML. Con este segundo software desarrollamos modelos IFPTML alternativos basados en otras técnicas de ML como: Regresión Logística (LOGR), Bosque Aleatorio (RF) y Máquina de Vectores de Apoyo (SVM). La calidad de todos los modelos IFPTML encontrados se evaluó calculando la Sensibilidad (Sn), la Especificidad (Sp), la Precisión (Ac), la Chi-cuadrado ($\chi 2$) y el nivel p (Hanczar *et al*., 2010; Huberty&Olejnik, 2006).

2.3. RESULTADOS Y DISCUSIÓN

2.3.1. Modelo PTML NP vs. MN

Como mencionamos en la introducción las técnicas de ML se están aplicando para resolver múltiples problemas prácticos en Nanotecnología (Alafeef *et al.*, 2020; Barnard & Opl et al, 2019; Bian *et al*., 2019; He *et al.*, 2019; Sun *et al.*, 2017a; Yan *et al.*, 2018). En este trabajo nos centramos en el uso del algoritmo IFPTML para mapear los ensayos preclínicos de las NPs frente a la estructura de las MNs. Durante la fase de preprocesamiento fuimos capaces de construir un conjunto de casos de NPs *vs.*MNs que implican múltiples condiciones de ensayo. Tras calcular los PTOs (variables de entrada) y reescalar la función objetivo, decidimos ajustar diferentes modelos IFPTML. En primera instancia, utilizamos el software STATISTICA 6.0 para ejecutar el algoritmo LDA con el fin de buscar el modelo preliminar IFPTML (Hill & Lewicki, 2005). Probamos diferentes estrategias de selección de variables para encontrar modelos alternativos. Resumimos los resultados de los diferentes modelos IFPTML encontrados. De acuerdo con la regla heurística de la navaja de Occam (principio de parsimonia) debemos utilizar las características mínimas, pero aún relevantes para resolver el problema (ni más ni menos) (Van Den Berg, 2018).

En consecuencia, encontramos el mejor resultado con una heurística de selección guiada por expertos (EGS). La EGS incluye tanto las características seleccionadas por Forward Stepwise (FSW) como las características relevantes que faltan. Se muestra un resumen de los parámetros estadísticos del nuevo modelo IFPTML-LDA obtenido mediante la heurística EGS. Este modelo incluye todas las variables importantes y también tiene valores similares o ligeramente mejores de los parámetros de control Sn, Sp, Ac, $\chi 2$, y el nivel p (Hanczar *et al*., 2010; Huberty & Olejnik, 2006). La ecuación de este último modelo lineal IFPTML-LDA encontrado usando la heurística EGS en el software STATISTICA se representa en la información de apoyo. También resumimos los parámetros estadísticos de todos los modelos alternativos desarrollados utilizando diferentes estrategias.

A continuación, utilizamos el software SOFT.PTML para desarrollar modelos alternativos de IFPTML con fines de comparación (Ortega-Tenezaca *et al*., 2020). Con este segundo software, diseñado específicamente para que podamos realizar experimentos de IFPTML, utilizamos las técnicas de ML LOGR, RF y SVM. Estas técnicas ML han sido ampliamente utilizadas en Quimio-informática y Nanotecnología con fines de clasificación (Bian *et al*., 2019; Findlay *et al.*, 2018; Mallawaarachchi *et al.*, 2019; Sun *et al*, 2017) En la Tabla 8 resumimos los resultados de los dos mejores modelos IFPTML encontrados. El modelo IFPTML-LOGR encontrado es similar al modelo IFPTML-LDA reportado anteriormente. Ambos modelos son lineales con las mismas variables. La ecuación del modelo IFPTML-LOGR encontrado utilizando el software SOFT.PTML fue la siguiente:

$$LOGR: f(n,c,j,s)_{calc}$$
$$= 70.5646 + 1.7600 \cdot f(c_{n0})_{ref} + 455.52464 \cdot \Delta Sh(AMVn)_{cn_j}$$
$$+ 22.3380 \cdot \Delta Sh(APSn)_{cn_j} + 91.1031 \cdot \Delta\Delta Sh(t,c1,c2)_{cn_j}$$
$$+ 2203.9229 \cdot \Delta Sh(\pi_1)_{cs_j} + 2434.2109 \cdot \Delta Sh(\pi_2)_{cs_j} + 137.4794$$
$$\cdot \Delta Sh(L_{in})_{cs_j} - 1644.1100 \cdot \Delta Sh(L_{out})_{cs_j} \quad (18)$$

$$N_{train} = 3213 X^2 = 2265.75 p - level < 0.05$$

El modelo PTML-LOGR tiene Ac = 97-96% en las series de entrenamiento/validación frente a Ac = 81-82% del modelo PTML-LDA. Podemos concluir que el modelo PTML-LOGR es notablemente más preciso (15% más) que el modelo PTML-LDA. En este sentido, decidimos seleccionar el modelo PTML-LOGR como nuestro mejor modelo lineal PTML. Además, podemos ver en la Tabla 8 que el modelo no lineal PTML-RF superó a todos los modelos lineales con valores de Sn, Sp y Ac en el rango 96-99,5%. Esto hace que este modelo sea la mejor opción para los estudios de predicción, pero su carácter no lineal lo hace un poco más complejo. Por último, encontramos que el SVM mostró valores muy altos de Sn, Sp, y Ac = 100% en el entrenamiento, pero resultó ser totalmente desequilibrado (Sp = 100% y Sn = 0%) en las series de validación. Como resultado, descartamos el modelo IFPTML-SVM para su uso práctico.

En definitiva, los presentes resultados demuestran que es posible buscar modelos predictivos IFPTML para NP frente a bacterias con diferentes MNs.

Estos resultados también validan el uso de SOFT.PTML para construir este tipo de modelos.

Tabla 9: Resumen de los resultados del modelo IFPTML-EGS

Soft.	Algor.	Clase Observada	Clasificación predicha			
STAT	LDA	$f(n,c,j,s)_{obs}=0$	Sp	81.1	2896	674
		$f(n,c,j,s)_{obs}=1$	Sn	90.1	42	384
		Total	Ac	82.1		
		Validación	Stat.	(%)	$f(n,c,j,s)_{pred}=0$	$f(n,c,j,s)_{pred}=1$
		$f(n,c,j,s)_{obs}=0$	Sp	81.7	976	218
		$f(n,c,j,s)_{obs}=1$	Sn	92	11	126
		Total	Ac	82.8		
PTML SOFT	LOGR	Entrenamiento	Stat.	(%)	$f(n,c,j,s)_{pred}=0$	$f(n,c,j,s)_{pred}=1$
		$f(n,c,j,s)_{obs}=0$	Sp	99.3	2905	21
		$f(n,c,j,s)_{obs}=1$	Sn	79.4	59	228
		Total	Ac	97.5		
		Validación	Stat.	(%)	$f(n,c,j,s)_{pred}=0$	$f(n,c,j,s)_{pred}=1$
		$f(n,c,j,s)_{obs}=0$	Sp	99.3	1826	12
		$f(n,c,j,s)_{obs}=1$	Sn	80.8	53	223
		Total	Ac	96.9		
	RF	Entrenamiento	Stat.	(%)	$f(n,c,j,s)_{pred}=0$	$f(n,c,j,s)_{pred}=1$
		$f(n,c,j,s)_{obs}=0$	Sp	99.5	2912	14
		$f(n,c,j,s)_{obs}=1$	Sn	96.9	9	278
		Total	Ac	99.3		
		Validación	Stat.	(%)	$f(n,c,j,s)_{pred}=0$	$f(n,c,j,s)_{pred}=1$
		$f(n,c,j,s)_{obs}=0$	Sp	99.5	1826	9
		$f(n,c,j,s)_{obs}=1$	Sn	98.6	4	272
		Total	Ac	99.4		

2.3.2. Estudio PTML de la resistencia de las NP-Bacterias frente a la topología metabólica de las MNs.

El estudio de las MNs de aquellas bacterias con alta resistencia a la acción de las NPs puede dar pistas para el futuro diseño de nuevas NP con actividad antibacteriana específica. Como hemos mencionado antes, los valores de $p(f(n,c,j,s)=1)_{pred}$ son las probabilidades con las que se predice que una determinada NP es activa contra las bacterias con una determinada MN. Podemos interpretar estas probabilidades como una medida de la susceptibilidad bacteriana a las NP. En consecuencia, desde el punto de vista de los MNs, se predice que aquellas bacterias con valores bajos de $p(f(n,c,j,s)=1)_{pred}$ son muy resistentes a la acción de la n-ésima NP en el j-ésimo ensayo. Por consiguiente, un valor medio bajo $p(f(n,c,j,s)=1)_{avg}=<p(f(n,c,j,s)=1)_{pred}>$ (media de todos los valores de $p(f(n,c,j,s)=1)_{pred}$) para la *s*-ésima bacteria frente a la misma NP en diferentes ensayos indica que esta especie debería ser muy resistente a esta NP en particular, independientemente del ensayo seleccionado. Para comparar la estructura de las MNs de diferentes bacterias frente a los valores predichos de $p(f(n,c,j,s)=1)_{avg}$ podríamos utilizar un único parámetro numérico de la estructura metabólica de las MNs (topología de la red). En este trabajo utilizamos 3 parámetros numéricos relacionados con la estructura metabólica de los MNs, N_{ms}, $<L_{ins}>$, y $<L_{outs}>$. Utilizamos estos parámetros para calcular un único parámetro que fusiona toda esta información. Vamos a llamar a este parámetro como el índice de desequilibrio anabolismo-catabolismo (ACUs) de las MNs de la especie de bacteria s-ésima especie de bacteria (ver Ecuación 19).

$$ACU_s = \alpha \cdot \frac{(< L_{ins} > - < L_{outs} >)}{N_{ms}} \qquad (19)$$

Los parámetros utilizados para construir ACUs tienen el siguiente significado estructural en términos de teoría de grafos. El parámetro N_{ms} = número de nodos, $<L_{ins}>$= grado medio de entrada y $<L_{outs}>$ = grado medio de salida de todos los nodos del gráfico. El número de nodos coincide con el número de metabolitos (N_{ms}) en el MN del organismo *s*-ésimo. El índice L_{ins} se utiliza para contar todas las flechas que llegan (entran) a un nodo en un grafo complejo. En el contexto de los MN, él L_{ins} es el número de metabolitos que son precursores (eductos) del metabolito de consulta (m). Por analogía, L_{outs} es el número de metabolitos que son productos (aductos) de una reacción metabólica con el metabolito de consulta como precursor (Jeong *et al.*, 2000; Ravasz *et al.*, 2002; Vidal *et al.*, 2011).Es así que usamos $<L_{ins}>$ como una medida del Anabolismo y $<L_{outs}>$ como una medida del Catabolismo de los MNs de esta bacteria. En consecuencia, podemos utilizar la diferencia ($<L_{outs}>-<L_{ins}>$) como una medida del desequilibrio del metabolismo anabólico frente al catabólico en la red. No hemos encontrado una referencia directa a este parámetro específico en la literatura.

Sin embargo, parámetros similares basados en las diferencias entre L_{out} y L_{in} se han utilizado antes para medir el desequilibrio del flujo en las redes (Bean *et al.*, 2017). El número α = 10 se utiliza aquí como factor de escala para transformar las ACUs en la misma escala que $<p(f(n,c,j,s)=1)_{pred}>$ para una mayor comparación. Representamos los valores de $p(f(n,c,j,s)=1)_{avg}$,N_{ms}, $<L_{ins}>$, $<L_{outs}>$ y AUCs de todos los MNs estudiados junto con otra información biológicamente relevante, ver Información de apoyo. Estos valores deben tomarse con precaución, recordando que estamos comparando valores medios y que una especie y/o cepa puede ser susceptible a un determinado NP en un ensayo específico. En consecuencia, recomendamos utilizarlos sólo como una guía general para descubrir tendencias sobre el comportamiento de las NP frente a diferentes especies de bacterias. En este sentido, una

inspección más detallada de las predicciones para todos los pares de NP frente a MN debería ofrecer una imagen más precisa.

Para dar esta imagen más cercana, decidimos comparar los valores observados y calculados de la probabilidad $p(f(n,c,j,s)=1)_{ns}$. La $p(f(n,c,j,s)=1)_{ns}$ son los valores de probabilidad de éxito del nº NP en todos los ensayos con el mismo sº MN de una especie bacteriana determinada. Este estudio nos dará también una visión más cercana del poder predictivo de los modelos IFPTML-LOGR e IFPTML-RF. Los valores de $p(f(n,c,j,s)=1)_{ns}$ son esencialmente diferentes de $p(f(n,c,j,s)=1)_{avg}$. Los valores $p(f(n,c,j,s)=1)_{avg}$ son el valor medio de las probabilidades predichas para grupos amplios de especies bacterianas. Los valores de $p(f(n,c,j,s)=1)_{ns}$ son tanto los valores observados como los predichos para pares específicos de NPs frente a MNs. Podemos obtener este parámetro como $p(f(n,c,j,s)=1)_{ns}= n(f(n,c,j,s)=1/n,s)/n(n,s)$. En esta fórmula $n(f(n,c,j,s)=1/n,s)$ es el número de casos de éxito. El parámetro *n(n,s)* es el número total de casos dado que el par NPs *vs.* MNs ha sido utilizado en los ensayos preclínicos.

Obtuvimos tanto la versión observada como la calculada de $p(f(n,c,j,s)=1)_{ns}$ utilizando los dos modelos IFPTML. En la Tabla 9 representamos casos seleccionados de las probabilidades observadas frente a las calculadas y los correspondientes valores de ACUs. En la Tabla 9 mostramos todos los NPn con resultados positivos para al menos una especie y casos seleccionados de pares de NPn *vs.* MNs con resultados negativos con $n(n,s) > 50$ casos en el conjunto de datos. En particular, tanto el modelo PTML-LOGR como el PTML-RF son muy precisos para descartar cerca del 100% de los casos de resultados negativos para los pares NPn *vs.* MNs (casos en negrita).

Sin embargo, PTML-RF es más de un 15% mejor que PTML-LOGR en la identificación de los casos positivos confirmados experimentalmente para los pares NPn *vs.* MNs.

Tabla 10. Probabilidades NPn *vs.* MNs observadas *vs.* calculadas por IFPTML y valores ACUs.

Valores experimentales observados			
NP *vs.* MN	n(*n,s*)	n(f(*n,c,s,j*)=1/*n,s*)	p(f(*n,c,s,j*)=1/*n,s*)
NP *vs.*CuI	4	0	0.00
NP *vs.* Ag	4	0	0.00
NP *vs.* Cu	4	0	0.00
NP *vs.*CuO	18	0	0.00
NP *vs.* Fe_2O_3	2	0	0.00
NP *vs.*ZnO	2	0	0.00
EC *vs.* Au	220	219	1.00
PA *vs.* Au	238	224	0.94
PA *vs.*CuI	4	0	0.00
Bs *vs.*CuI	11	0	0.00
EC *vs.* Fe_3O_4	8	8	1.00
EC *vs.*CdS	16	16	1.00
PA *vs.* Ag	71	56	0.79
PA *vs.*ZnO	16	11	0.69
EC *vs.* Ag	72	24	0.33
EF *vs.* Au	199	4	0.02
EF *vs.* Ag	100	1	0.01
HP *vs.*CuO	36	0	0.00
PG *vs.* Au	43	0	0.00
SC *vs.* Au	28	0	0.00
SC *vs.*CuO	45	0	0.00
PN *vs.* Au	91	0	0.00
PN *vs.*CuI	23	0	0.00

Curiosamente, observamos que casi todas las bacterias estudiadas presentan ACUs> 0,5, sólo MP tiene ACUs< 0,5, y ningún MNs tiene ACUs< 0 (incluso cuando esto es

matemáticamente posible). Sinceramente, no sabemos si esto es una pista biológica relevante o el resultado de un conjunto limitado de MNs en el análisis. Con el fin de ganar un poco dentro de la posible sistematización de los presentes resultados decidimos volver a analizarlos teniendo en cuenta el significado biológico de los ACUs y los valores $p(f(n,c,j,s)=1)_{ns}$. Para este estudio se consideró que un determinado MNs presenta un metabolismo desequilibrado del lado anabólico si ACUs> 0,5. Los MNs tienen un metabolismo equilibrado Anabólico-Catabólico si ACUs = 0. Los MNs tienen un metabolismo casi equilibrado si ACUs está en el rango de -0,5 a 0,5 (+/- 5% de desviación alrededor de 0). Por último, se considera que un MNs presenta un metabolismo desequilibrado del lado catabólico si ACUs< -0,5. La elección de estos valores aquí es empírica y debe ser vista sólo como una herramienta práctica para la sistematización/visualización preliminar de la información estructural-funcional sobre la interacción NP *vs.* MNs. A efectos de sistematización podemos dividir este gráfico en cuatro cuadrantes (Q) denominados QI, QII, QIII y QIV. En una primera aproximación al problema están delimitados por los cuartiles de ACUs y $p(f(n,c,j,s)=1)_{ns}$ dentro del rango 0-1. Por lo tanto, el primer cuadrante (QI) incluye todas las bacterias resistentes a NP ($p(f(n,c,j,s)=1)_{ns}<0.5$) con MNs desequilibradas por el lado anabólico (ACUs> 0,5). El segundo cuadrante (QII) incluye todas las bacterias susceptibles a NP ($p(f(n,c,j,s)=1)_{ns}>0.5$) con MNs desequilibradas por el lado anabólico (ACUs> 0,5). El tercer cuadrante (QIII) debe incluir todas las bacterias susceptibles a NP ($p(f(n,c,j,s)=1)_{ns}>0.5$) y con MNs casi equilibrados (ACUs = rango 0 - 0,5). El último cuadrante (QIV) debe incluir todas las bacterias susceptibles a NP ($p(f(n,c,j,s)=1)_{ns}<0.5$) con MNs casi equilibrados (rango ACUs = 0 - 0,5). En la Figura 11 representamos un gráfico (>5000 pares) de los valores de $p(f(n,c,j,s)=1)_{ns}$ de la susceptibilidad bacteriana del PN frente al índice de ACUs de los MNs de todas las bacterias estudiadas aquí.

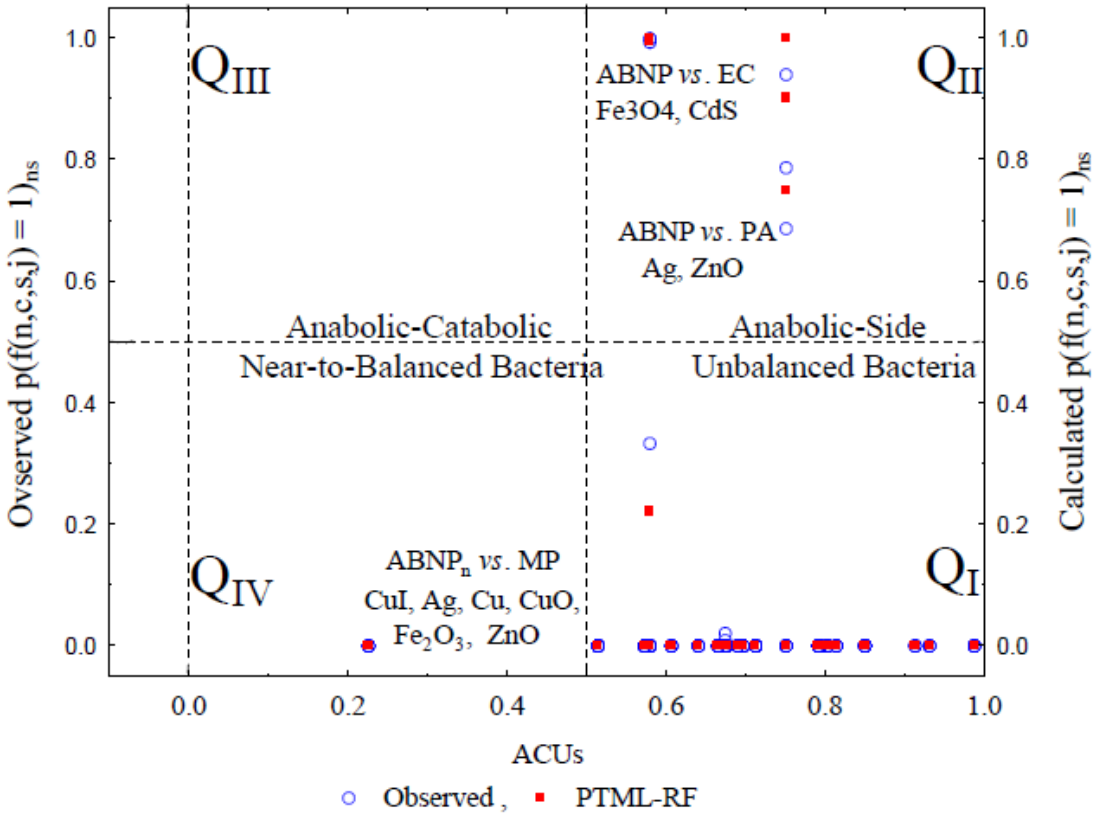


Figura 11. Susceptibilidad/resistencia bacteriana NP vs metabolismo de los MNRs

En la Figura 11, podemos confirmar visualmente que no hay MNs fuera de carta con ACUs< 0,0 (lado catabólico metabolismo desequilibrado), entre >20 especies de bacterias analizadas. Además, podemos ver que el QIII es un cuadrante vacío. Sólo la bacteria Mycoplasmapneumonia (MP) tiene ACUs< 0,5. Se predice que esta bacteria es resistente ($p(f(n,c,j,s)=1)_{ns}<0.5$) a las NPs con núcleos de CuI, Ag, Cu, CuO, Fe_2O_3 y ZnO, según el modelo. En consecuencia, para aumentar la actividad biológica de las NPs frente a las MPs podría ser necesario modificar el núcleo de las NPs. Por ejemplo, recientemente el núcleo de NPAg fusionado con *Zingiberzerumbet* obtenido por síntesis verde ha demostrado una actividad de las NP muy interesante frente a MP (Yang *et al*., 2019). Por otro lado, la mayor probabilidad de actividad de las NPs predicha fue frente a *Escherichiacoli* (EC) y *Pseudomona aeruginosa*

(PA) (QII). Estos resultados también están de acuerdo con los resultados experimentales de las NPs frente a EC y PA comunicados anteriormente (Azam *et al*., 2012; Hossain &Mukherjee, 2013; Inbaraj *et al*., 2011; Zhao *et al*., 2013). Podemos concluir también que es necesario seguir investigando para aumentar la eficacia de las NPs frente a las bacterias de QI.

2.4. Conclusiones

Las NPs son una opción muy interesante para frenar las infecciones causadas por especies y/o cepas de bacterias MDR. Los modelos IFPTML pueden ayudar a reducir los costes y el tiempo en el descubrimiento de nuevas NPs. La inclusión de información sobre la estructura de la MN para las bacterias objetivo aumenta el rango de aplicación de los modelos IFPTML para diseñar NP frente a otras especies de bacterias. Los modelos IFPTML-RF e IFPTML-LOGR parecen ser los más generales y precisos para la predicción de NP hasta la fecha. El software SOFT.PTML es una interfaz fácil de usar para el desarrollo de modelos IFPTML en nanotecnología.

Capítulo 3

CAPÍTULO 3. APLICACIONES. PREDICCIÓN IFPTML DE COMPUESTOS ANTI-LEISHMANIA

3. Articulo

Ref. Prediction of Antileishmanial Compounds: General Model, Preparation, and Evaluation of 2-Acylpyrrole Derivatives. Santiago C, Ortega-Tenezaca B, Barbolla I, Fundora-Ortiz B, Arrasate S, Dea-Ayuela MA, González-Díaz H, Sotomayor N, Lete E. J Chem Inf Model. 2022 Aug 22;62(16):3928-3940. doi: 10.1021/acs.jcim.2c00731.

3.1. Introducción

La leishmaniasis es una enfermedad parasitaria, causada por patógenos protozoarios del género *Leishmania*, que puede presentar diferentes manifestaciones clínicas, como la leishmaniasis cutánea (CL), la visceral o kala- azar (VL), la leishmaniasis dérmica post-kala-azar (PKDL) y la mucocutánea (MCL). Como todas las enfermedades desatendidas, la leishmaniasis sigue siendo un importante problema sanitario mundial, ya que es endémica en unos 100 países con más de 350 millones de personas en riesgo (OMS, 2020). El tratamiento de la leishmaniasis se basa principalmente en unos pocos fármacos: antimoniales pentavalentes (ampB), paromomicina, pentamidina, anfotericina rtBliposomal, fluconazol y miltefosina, dependiendo de la especie etiológica, del tipo de infección y también de la región geográfica, debido al creciente número de cepas resistentes. Además, el uso de estos fármacos está asociado a una serie de efectos secundarios graves relacionados con su toxicidad (Gupta *et al*., 2021; Brindha *et al*.; 2021, Jones *et al*., 2018; Nagle *et al*, 2014). Por lo tanto, es evidente la necesidad de identificar nuevos compuestos antileishmanios eficaces con quimiotipos distintos a los prototipos en uso

clínico. En este contexto, los heterociclos de nitrógeno se consideran andamios privilegiados, ya que aproximadamente el 60% de los fármacos de moléculas pequeñas aprobados por la FDA estadounidense contienen un heterociclo de nitrógeno (Vitaku *et al*., 2014). En particular, el núcleo de pirrol ha atraído nuestra atención porque este motivo está incrustado en una variedad de productos naturales (por ejemplo, prodigeninas, (Papirredy *et al*, 2011) bromopirrol (Parra *et al*, 2018) y alcaloides de espiroindimicina (Zhang *et al*, 2021)) con actividad antiparasitaria (Albino *et al*., 2020). En cuanto a los derivados sintéticos, los piridinil pirroles 1 y 2 han demostrado ser inhibidores de la caseína quinasa 1 que bloquean el crecimiento de los promastigotes de *Leishmania*major*in vitro* (Alloco *et al*, 2006). Los 1,2-diarilpirroles 3 han sido identificados como una nueva clase de compuestos activos contra la estancia de amastigotes de *Leishmania*donovani mediante la inhibición de la tripanotión reductasa (Baiocco *et al*, 2013). Por otra parte, los derivados de 2-acilpirrol 4 también mostraron perfiles prometedores contra *Leishmania* (Figura 12) (Rizvi *et al*, 2010). Además, se ha informado de que la pirrol-indolinona SU11652, un análogo de Sunitinib, se dirige al nucleósido difosfato quinasa de los parásitos de *Leishmania* (Vieira *et al*, 2017). Recientemente hemos sintetizado 2-acilpirroles mediante la acilación C-H radical catalizada por Pd(II) de derivados de pirrol. Este protocolo eficiente y flexible nos permitió recopilar una pequeña biblioteca de 2-acilpirroles 5, sustituidos de forma variable en el anillo de arilo, y con un anillo de pirimidina (serie 5a) o piridina (serie 5b) unido al átomo de nitrógeno del núcleo de pirrol (Figura 12). Estas características estructurales hacen que nuestros derivados de pirrol sean candidatos interesantes para ser probados como potenciales compuestos antileishmanios.

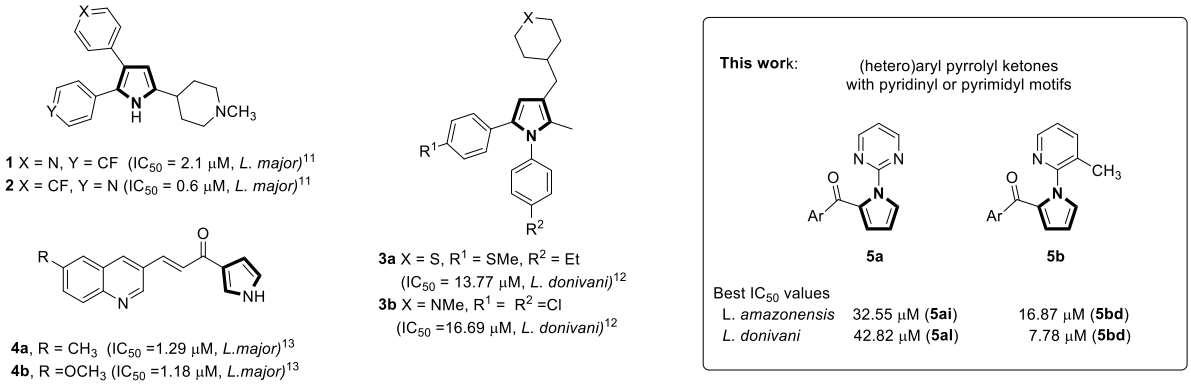


Figura 12. Actividad antileishmanial de algunos compuestos sintéticos con el motivo pirrol.

En este contexto, la modelización quimio-informática puede convertirse en una opción prometedora para reducir el coste de desarrollo y aumentar la probabilidad de encontrar nuevos éxitos antileishmanios. Curiosamente, los modelos quimio-informáticos clásicos hacen hincapié en acelerar el descubrimiento de fármacos antiparasitarios reduciendo el número de compuestos que hay que ensayar mediante pruebas de ensayo y error. Sin embargo, además del gran número de compuestos que hay que ensayar, hay otros factores que pueden influir en la ralentización de este proceso. Por ejemplo, el gran número de combinaciones de parámetros biológicos (MIC, IC_{50}, pK_i, *etc*.), especies de parásitos, estadios de parásitos, proteínas diana, *etc*. aumenta notablemente el tiempo y el coste por compuesto a ensayar. Desgraciadamente, los modelos clásicos de la quimio-informática no consiguen realizar una optimización multiobjetivo de los compuestos antiparasitarios, ya que ignoran todos estos factores. La razón principal podría ser la dificultad para codificar múltiples condiciones de contorno (parámetro, proteína, línea celular, especie, estadio del parásito, *etc*.) de los ensayos y la necesidad de obtener esta información de muchas fuentes de datos diferentes. Nuestro grupo ha reportado recientemente el primer modelo PTML [(Teoría de la Perturbación (PT) + Aprendizaje Automático (ML)] capaz de explicar un conjunto de datos muy grande de ensayos preclínicos de compuestos anti-leishmania, así como la

predicción, síntesis y ensayo de nuevas pirroloisoquinolinas frente a diferentes especies de *Leishmania* (Barbolla *et al*, 2021). Sin embargo, el proceso de desarrollo de este primer modelo PTML y su posterior uso para la predicción de nuevos éxitos antileishmanianos fue laborioso.

Por otro lado, acuñamos el término IFPTML = Fusión de Información (IF) + Teoría de la Perturbación (PT) + Aprendizaje Automático (ML) para diseñar un nuevo algoritmo de optimización multiobjetivo de compuestos. En algunos trabajos falta la etapa de IF y se utiliza el término PTML solo como caso particular (Simón-Vidal *et al*, 2018; Díez-Alarcia *et al*, 2019). Estos modelos de IFPTML se han utilizado en química medicinal, proteómica, metabolómica y nanotecnología (Nocedo-Mena *et al*., 2019; Santana *et al*, 2020).

IFPTML comienza con una primera fase (IF + PT) dedicada a realizar la fusión de la información procedente de diferentes fuentes y/o la transformación de las variables originales en operadores PT (PTO). Estos PTOs son nuevas variables de entrada útiles para codificar información sobre múltiples condiciones de ensayo procedentes de diferentes fuentes. Por ejemplo, los PTOs pueden utilizarse para codificar información sobre objetivos proteicos, líneas celulares, redes metabólicas microbianas de los organismos objetivo, portadores de nanopartículas del fármaco*, etc*. (Nocedo-Mena *et al*., 2019; Santana *et al*, 2020). A continuación, el flujo de trabajo del IFPTML entra en la fase de ML que utiliza algoritmos clásicos de ML. Hasta hace poco, los investigadores que pretendían entrenar los modelos PTML necesitaban ejecutar un software diferente para cada una de las fases del algoritmo (IF, PT y ML). (Nocedo-Mena *et al*., 2019; Santana *et al*, 2020). Este fue también el caso de nuestro anterior modelo IFPTML para compuestos antileishmaniales (Barbolla *et al.*, 2021). Necesitábamos utilizar una hoja de cálculo para ejecutar la primera fase, un software ML para buscar el modelo y una hoja de cálculo de nuevo para ejecutar las predicciones. Este problema

llamó la atención de los desarrolladores de software de quimioinformática sobre la necesidad de nuevas plataformas para unificar los diferentes pasos del análisis IFPMTL. De hecho, hemos introducido la herramienta QSAR-Co para resolver algunos problemas en este sentido. QSAR-Co ejecuta conjuntamente las etapas PT y ML del algoritmo (Ambure*et al*, 2019). Sin embargo, QSAR-Co no puede ejecutar procedimientos de FI para calcular PTOs de varias etiquetas o funciones de referencia que codifican múltiples condiciones de ensayo al mismo tiempo. Además, QSAR-Co solo calcula una clase de operadores PT, llamados promedios móviles de una sola condición. En consecuencia, necesita tantos PTOs como condiciones límite estén presentes en el problema, lo que implica un número notablemente mayor de variables a explorar con respecto a los PTOs multietiqueta utilizados en los algoritmos IFPTML. (Nocedo-Mena *et al*., 2019; Santana *et al*, 2020). Estos PTOs han demostrado ser muy útiles para reducir la dimensionalidad del problema, por ejemplo, en el caso del conjunto de datos de ensayos preclínicos antileishmania de ChEMBL (Barbolla *et al.*, 2021). Por lo tanto, introducimos la herramienta de estudio SOFT.PTML, que tiene la posibilidad de calcular PTOs multi-etiqueta, incluyendo funciones de referencia multi-etiqueta/multi-condición, medias móviles, co-varianzas*, etc.* (Ortega-Tenezaca *et al*, 2020; Ortega-Tenezaca *et al*, 2021). De hecho, SOFT.PTML ya ha sido utilizado con éxito en la modelización de IFPTML en nanotecnología y química medicinal con este fin (Ortega-Tenezaca, González-Díaz, 2021). En el presente trabajo, reportamos por primera vez el uso de SOFT.PTML para buscar modelos de IFPTML para compuestos antileishmaniales, realizando un estudio comparativo de diferentes algoritmos de ML. También llevamos a cabo un estudio predictivo de una serie de 2-acilpirroles, previamente reportados por nuestro grupo (Santiago *et al*, 2020), junto con la preparación experimental de nuevas muestras para el ensayo y su prueba leishmanicida *in vitro*. Algunos de los 2-acilpirroles

ensayados se comparan favorablemente con respecto a la miltefosina (compuesto de referencia) en términos de actividad y toxicidad. Este trabajo abre una nueva línea de investigación experimental centrada en la síntesis y optimización de compuestos antileishmania como derivados de 2-acilpirrol. Además, sentará las bases para el desarrollo de modelos IFPTML más rápidos y fáciles de usar para otras enfermedades tropicales desatendidas.

3.2. MATERIALES Y MÉTODOS

3.2.1. Métodos computacionales

Base del modelo IFPTML. En este trabajo, proponemos un modelo IFPTML para calcular los valores de la función de puntuación de la actividad biológica antileishmania $f(v_{ij})_{calc}$ del i-ésimo compuesto de consulta en el j-ésimo ensayo con múltiples condiciones límite $c_j = [c_0, c_1, c_2, c_{jmax}]$. En el caso de los modelos de clasificación (utilizados en este trabajo) $f(v_{ij})_{calc}$ obtiene valores adimensionales utilizados para la puntuación de la propensión del i-ésimo compuesto a alcanzar un determinado nivel de los valores de actividad biológica v_{ij} (véase la siguiente sección) (Martínez-Arzate *et al*, 2017). En consecuencia, los valores de $f(v_{ij})_{calc}$ pueden utilizarse directamente para comparar la propensión relativa de dos compuestos diferentes a alcanzar un determinado nivel de actividad biológica en el jº ensayo en comparación con un valor umbral de corte j. También pueden utilizarse para comparar el comportamiento de un mismo compuesto en dos ensayos diferentes. Para calcular los valores de salida $f(v_{ij})_{calc}$, el modelo IFPTML utiliza dos tipos de variables de entrada (a) funciones de referencia $f(v_{ij})_{ref}$ y (b) operadores de teoría de la perturbación $PTO_k(c_j)$. Así, el modelo IFPTML comienza con los valores de una función de referencia $f(v_{ij})_{ref}$, que se utilizan para caracterizar/identificar el tipo de actividad biológica que queremos modelar. A continuación, el modelo añade los valores de las

funciones $PTO_k(c_j)$ utilizadas para medir el efecto de las perturbaciones sobre el resultado de la actividad biológica. Las funciones $PTO_k(c_j)$ cuantifican las perturbaciones/desviaciones en la estructura del compuesto i-ésimo y/o en las condiciones de ensayo c_j en comparación con un conjunto de compuestos de referencia (Martínez-Arzate, *et al*, 2017). En la siguiente sección, se explica el preprocesamiento de los datos brutos para construir las funciones $f(v_{ij})_{ref}$ y $PTO_k(c_j)$. Los modelos lineales IFPTML tienen la siguiente forma:

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{k=1,j=0}^{k_{max},j_{max}} a_{kj} \cdot PTO_k(\mathbf{c}_j) \quad (1) \qquad (20)$$

Preprocesamiento de datos. Se utilizó la herramienta SOFT.PTML para preprocesar un conjunto de datos ChEMBL de ensayos preclínicos de compuestos candidatos a la lucha contra la leishmaniosis15. Esta herramienta también se utilizó para realizar el entrenamiento/validación de modelos alternativos de IFPTML. Estos datos incluyen 145.851 valores de actividad biológica (v_{ij}) para i-ésimos compuestos probados en j-ésimos ensayos preclínicos con condiciones experimentales límite (etiquetas) $c_j = [c_0, c_1, c_2, \ldots c_n]$. Los valores v_{ij} se expresan como diferentes parámetros de actividad antileishmania con la etiqueta c_0, como IC_{50}, Ki, EC_{50}, *etc*. Se utilizó la función de discretización SOFT.PTML para convertir todos los valores de actividad biológica v_{ij} en una función objetivo booleana $f(v_{ij})_{obs}$= 1 ó 0. La función implementada en el software es $f(v_{ij})_{obs}$= 1 IF (v_{ij}>$cutoff_j$ AND $d(c_0) = 1$) OR (v_{ij}>$cutoff_j$ AND $d(c_0) = 1$) ELSE $f(v_{ij})_{obs} = 0$. Los diferentes valores de $cutoff_j$ son los valores de cutoff utilizados para los diferentes parámetros de actividad biológica con etiqueta c_0 (IC_{50}, K_i, EC_{50}, *etc*.) en diferentes ensayos j. El parámetro de deseabilidad $d(c_0) = 1$ o -1 tiene que ser maximizado o minimizado para obtener un efecto biológico óptimo.

Fusión de información (IF). Como se ha mencionado anteriormente, todas las condiciones experimentales límite del ensayo de actividad biológica se expresan mediante el vector c_j = $[c_0,c_1,c_2,...c_{maxj}]$. En este trabajo, los miembros de este vector se reagruparon, creando dos particiones cI y cII. Estas particiones codifican independientemente la información sobre las condiciones experimentales de los ensayos preclínicos c_I= $[c_0,c_1,c_2,c_3,c_4,c_5]$ y la información sobre la naturaleza y calidad de los datos c_{II}=$[c_6,c_7,c_8,c_9,c_{10}]$. Esta información se expresa también como un vector D_{ki}=$[D_{1i},D_{2i},D_{3i},...D_{maxi}]$. Los elementos de este vector son D_{1i}= $ALOGP_i$, el coeficiente de partición *n*-octanol/agua, D*2i*=PSA*i*, el Área de Superficie Polar topológica, y D_{3i}=NVLR, el Número de Violaciones a la regla de Lipinski de la estructura del i-ésimo compuesto. Los valores de los descriptores moleculares D_k se descargaron de ChEMBL y/o se calcularon con el software DRAGON (Mauri *et al*, 2006) para los nuevos compuestos. A continuación, se llevó a cabo el proceso de IF calculando los operadores PT multicondición $PTO_k(D_{ki},c_j)$. Cada variable $PTO_k(D_{ki},c_j)$ es la expresión de la fusión de la información estructural de uno o varios elementos de D_i y uno o varios elementos de c_j. Esto significa que un $PTO_k(D_{ki},c_j)$ es una función u operador (PTO_k) calculado para llevar a cabo la fusión de la información estructural D_i del i-ésimo compuesto y las condiciones c_i del j-ésimo ensayo preclínico. El PTO_k puede tener muchas formas diferentes y/o puede calcularse para diferentes subconjuntos (dominios) de los vectores Dk_i y c_j. En primer lugar, calculamos operadores del tipo $PTO(D_{ki},c_I)$=$PTO(D_{1i},c_I)$, $PTO(D_{2i},c_I)$*, etc*., ó $PTO_k(D_{ki},c_{II})$=$PTO(D_{1i},c_{II})$, $PTO(D_{2i},c_{II})$*, etc*. Estos PTO calculados pueden adoptar la forma de operadores de media móvil (MA) multicondición: $\Delta D_k(c_I)=D_{ki}-\langle D_k(c_I)\rangle$ o $\Delta D_k(c_I)=D_{ki}-\langle D_k(c_{II})\rangle$. Miden la desviación ($\Delta$) de la estructura del i-ésimo compuesto expresada por D_{ki} con respecto a los valores medios/esperados $\langle D_k(c_j)\rangle$ para todos los compuestos ensayados en las mismas condiciones c_I o c_{II}.

SOFT.PTML en el cribado *in silico* de nuevos compuestos. El nuevo modelo se utilizó para estudiar una serie de 28 derivados de di(hetero)aryl cetona sintetizados en nuestro grupo. En primer lugar, se generaron los códigos SMILE de los 28 compuestos utilizando ChemDraw 18.2 (Evans, 2014). A continuación, se calcularon los valores de las variables de entrada D_1 = LOGP, D_2 = PSA y D_3 = NVLR. A continuación, estos valores se sustituyeron en el modelo SOFT.PTML para obtener las probabilidades de actividad de cada compuesto en diferentes ensayos biológicos. Se realizó una simulación de la respuesta biológica de 28 compuestos + 1 control (miltefosina) en muchos ensayos preclínicos diferentes. Estos ensayos incluyeron un total de >50 parámetros de actividad biológica diferentes [K_i(nM), IC_{50}(nM), Inhibiton (%)*, etc*.)], 35 proteínas diana (P00374 Dihidrofolato reductasa, Q0GKD7 Farnesil pirofosfato sintasa*, etc*.), 28 líneas celulares (J774, HL-60, Jurkat*, etc*.), 40 organismos de ensayo (L. donovani, L. major, L. amazonensis*, etc*.) y 2 etapas de desarrollo del organismo (amastigotes y promastigotes). En total, predecimos el resultado de los 29 compuestos en 249 ensayos preclínicos diferentes cada uno.

3.2.2. Métodos experimentales.

Síntesis de los 2-(hetero) aroilpirroles 5. Los acilpirroles 5aa-an y 5ba-bs se sintetizaron siguiendo un procedimiento previamente desarrollado por nosotros (Figura 14) (Santiago, 2020). La acilación de la 2-(1*H*-pirrol-1-il) pirimidina **6a** o de la 3-metil-2-(1H-pirrol-1-il) piridina **6b** con los correspondientes aldehídos (hetero)aromáticos utilizando $Pd(OAc)_2$ como catalizador, TBHP como oxidante y ácido piválico como aditivo, en tolueno seco como disolvente. Las reacciones se llevaron a cabo en tubos sellados a 60ºC (para **6a**) o 120ºC (para **6b**) durante 1,5-7 h.

3.2.3. Detalles de los ensayos preclínicos

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

Ensayo de susceptibilidad *in vitro* de los promastigotes. El ensayo biológico se llevó a cabo siguiendo protocolos previamente publicados (Bilbao-Ramos, 2012; Dea-Ayuela *et al*, 2016). De forma concisa, se han cultivado promastigotes (2,5 x 105 parásitos/pocillo) de fase logarítmica en placas de plástico de 96 pocillos. Se han preparado nuevas muestras de los compuestos según el protocolo descrito anteriormente y posteriormente se han disuelto soluciones del compuesto químico a ensayar en DMSO a 50 mg/mL. Se realizaron diluciones seriadas 1:2 de los compuestos en medio de cultivo fresco (100, 50, 25, 12,5, 6,25, 3,12, 1,56 y 0,78 µg/mL) hasta un volumen final de 200 µL. También se incluyó un control de crecimiento y un control de señal-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 ºC, se añadieron 20 µL de una solución de resazurina de 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 fluorímetro (Infinite 200Tecan i-Control). La inhibición del crecimiento (%) se calculó por 100 - [(RFU pozos tratados - RFU señal-ruido)/(RFU no tratados - RFU señal-ruido) x 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 (IS) 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 (Bilbao-Ramos *et al*, 2012). Brevemente, se sembraron 5 × 10^4 macrófagos J774 y promastigotes estacionarios en una proporción de 1:5 en cada pocillo de una placa de microtitulación, se suspendieron en 200µL de medio de cultivo y se incubaron durante 24 h a 33 °C en cámara de CO_2 al 5%. Tras 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, el sobrenadante se sustituyó por 200 µL/pocillo de medio de cultivo que contenía diluciones seriadas dobles de los compuestos de ensayo, como en el ensayo de promastigotes. También se incluyó el control de crecimiento y la señal de ruido. Tras la incubación durante 48 h a 37 °C, 5% de CO_2, el medio de cultivo se sustituyó por 200 µL/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 min. A continuación, las placas se centrifugaron a 3.500g durante 5 minutos y la solución de lisis se sustituyó por 200 µL/pocillo de medio para insectos de Schneider. A continuación, 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 µL/pocillo de resazurina 2,5mM y se incubaron

durante otras 3 h. 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 (Galiana-Roselló *et al*, 2013). Se sembraron líneas celulares de macrófagos J774 (5 × 104 células/pozo) en microplacas de fondo plano de 96 pozos con 100µL 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 200µL de medio y se expusieron durante otras 24 h. También se incluyeron controles de crecimiento y de señalización. Después, se añadió un volumen de 20µL de la solución de resazurina de 2,5mM, y las placas se devolvieron a la incubadora durante otras 3 h 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.

3.3. RESULTADOS Y DISCUSIÓN

Modelo SOFT.PTML. Como hemos mencionado en la introducción, el algoritmo IFPMTL es útil para buscar modelos predictivos para la optimización multiobjetivo de compuestos. De hecho, ya hemos utilizado modelos IFPTML para el estudio de nuevas pirroloisoquinolinas frente a diferentes especies de *Leishmania* (Barbolla *et al.*, 2021). Sin

embargo, todos los pasos del análisis IFPTML debían realizarse en diferentes programas informáticos y/o utilizando diferentes operaciones manuales. En este trabajo, describimos por primera vez el uso de nuestro software SOFT.PTML para el desarrollo de modelos IFPMTL para la predicción de compuestos anti-leishmania. El mismo conjunto de datos que contiene n = 109.389 ensayos preclínicos fue seleccionado y reprocesado con el software SOFT.PTML. La figura 13 muestra la interfaz de fácil manejo del software con las etapas de FI, PT y ML integradas en una sola aplicación. Esto nos permitió explorar diferentes técnicas de ML no estudiadas antes en este problema de forma más automática. En concreto, se estudiaron los algoritmos de Regresión Logística (LOGR), Máquina de Vectores de Apoyo (SVM) y Bosques Aleatorios (RF) (Cortes *et al*, 1995; Urbanowicz *et al*, 2009; Bzdoc *et al*, 1995). Los resultados obtenidos con los distintos algoritmos se resumen en la Tabla 11.

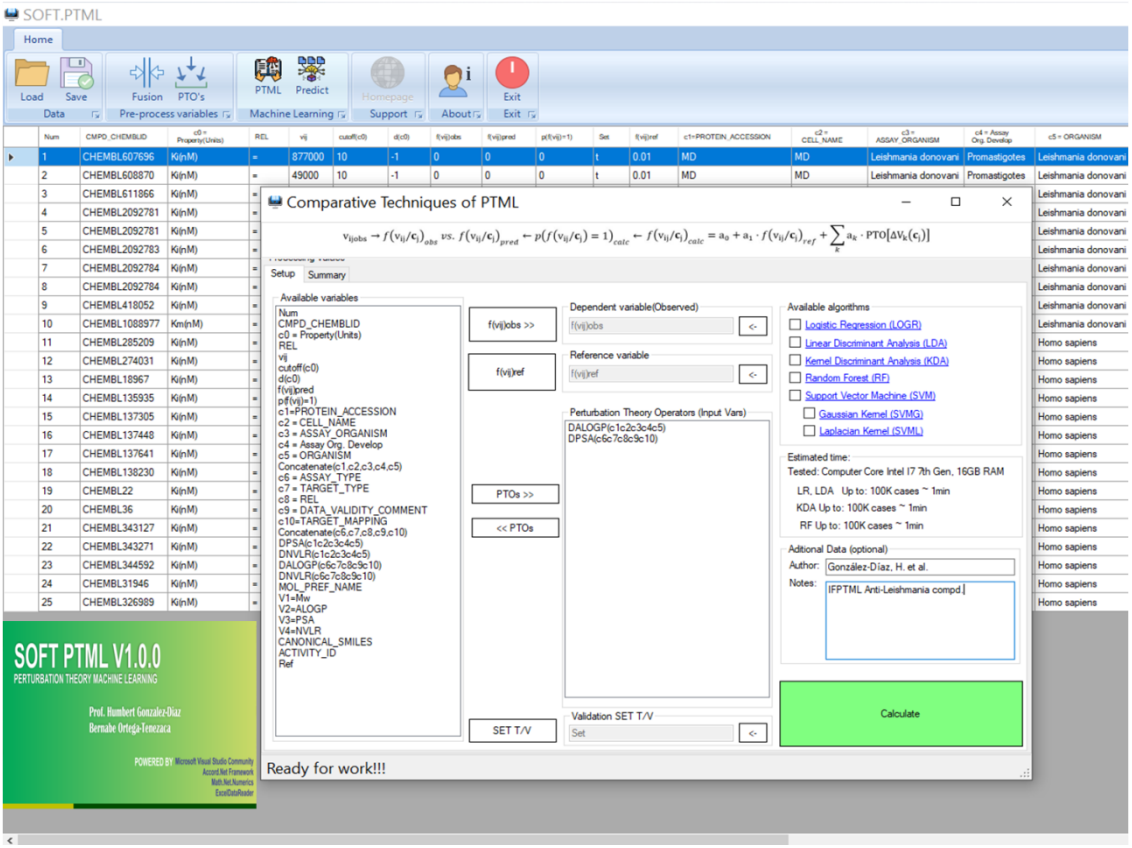


Figura 13. SOFT.PTML - Análisis de los datos de los compuestos antileishmania de ChEMBL

Tabla 11. Resultados del análisis de SOFT.PTML

Modelo PTML	ML	Datos	Stat	Param.	Set	$f(v_{ij})_{obs}$	
		Conjunto de datos	param	valor	$f(v_{ij})_{pred}$	0	1
1	LOGR	Entrenamiento	Sp	0.9885	0	101,493	1,233
			Sn	0.9885	1	1,185	5,478
			Ac	0.9779	Total		
		Validación	Sp	0.9888	0	33,828	407
			Sn	0.8191	1	384	1,843
			Ac	0.9783	Total		
2	RF	Entrenamiento	Sp	0.9985	0	102,519	503
			Sn	0.9985	1	159	6,208
			Ac	0.9939	Total		
		Validación	Sp	0.9934	0	33,985	315
			Sn	0.8600	1	227	1,935
			Ac	0.9851	Total		
3	SVM	Entrenamiento	Sp	0.9972	0	102,388	988
			Sn	0.9972	1	290	5,723
			Ac	0.9883	Total		
		Validación	Sp	0.9944	0	34,021	757
			Sn	0.6636	1	191	1,493
			Ac	0.974	Total		

A continuación, estudiamos los valores de Sensibilidad (Sn) y Especificidad (Sp) obtenidos para estos algoritmos utilizando la estrategia IFPTML tanto en series de entrenamiento como de validación externa de ensayos preclínicos. Cabe destacar que todos los algoritmos dieron resultados interesantes. Sin embargo, se descartó el modelo IFPTML-SVM, porque tenía un valor muy bajo de Sn = 0,6636 en la serie de validación. También se descartó el modelo IFPTML-RF, porque, aunque los resultados son muy prometedores (rango Sp ≈ Sn = 0,8-0,98), el modelo en sí es notablemente más complejo en comparación con los modelos lineales. Por lo tanto, basándose en la navaja de Ocam o en el principio de parsimonia, se seleccionó el modelo

lineal IFPTML-LOGR como el más adecuado (Hoffman *et al*, 1997; Rabinowitz *et al*, 2006). La ecuación del modelo IFPTML-LOGR es la siguiente

$$f(v_{ij})_{calc} = 8.350971 \cdot f(\nu_{ij})_{ref} + 0.885564 \cdot \Delta D_1(\mathbf{c}_I) + 0.024487 \cdot \Delta D_2(\mathbf{c}_I) - 1.388516 \cdot \Delta D_3(\mathbf{c}_I) - 0.853978 \cdot \Delta D_1(\mathbf{c}_{II}) - 0.025694 \cdot \Delta D_2(\mathbf{c}_{II}) + 1.517539 \cdot \Delta D_3(\mathbf{c}_{II}) + 5.090163 \quad (21)$$

n=109389 $\chi^2 = 135169.7$ $p<0.05$

Estrategia "todo en uno" frente a estrategia "multisoftware". SOFT.PTML utiliza una estrategia todo-en-uno que implementa todas las etapas (IF, PT y ML) del algoritmo IFPTML en la misma plataforma. Para validar la estrategia all-in-one, necesitábamos demostrar que este software era capaz de reproducir los resultados obtenidos con la estrategia multi-software. Seleccionamos el modelo IFPTML-LDA con estrategia multisoftware encontrado en este trabajo (Ecuación 22) para compararlo con el modelo IFPTML-LOGR, porque ambos modelos son ecuaciones lineales con la misma forma, a pesar de utilizar diferentes técnicas de ML (LDA y LOGR).

$$f(v_{ij})_{calc} = 59.918599 \cdot f(\nu_{ij})_{ref} + 1.631537 \cdot \Delta D_1(\mathbf{c}_I) + 0.041494 \cdot \Delta D_2(\mathbf{c}_I) - 2.675709 \cdot \Delta D_3(\mathbf{c}_I) \quad (3) - 1.562187 \cdot \Delta D_1(\mathbf{c}_{II}) - 0.041886 \cdot \Delta D_2(\mathbf{c}_{II}) + 2.649511 \cdot \Delta D_3(\mathbf{c}_{II}) - 25.182671 \quad (22)$$

n=109389 χ^2= 135169.7 $p<0.05$

Ambos modelos IFPTML comienzan con una función de referencia $f(v_{ij})_{ref}$, cuyo valor representa la probabilidad *a priori* con la que un compuesto seleccionado al azar puede dar un resultado positivo $f(v_{ij}) = 1$ del parámetro específico c_0 en las condiciones c_j. Esta función se utiliza para especificar dentro de la ecuación qué parámetro biológico (IC_{50}, K_i, MIC*, etc*.) de los muchos posibles queremos estudiar. A continuación, se añaden los valores $PTO_{ki}(D_{ki},c_j)=\Delta D_{ki}(c_j)$ para medir las variaciones/perturbaciones. Cabe destacar que tanto los modelos IFPTML-LOGR como IFPTML-LDA alcanzaron valores similares de Sn y Sp en las series de entrenamiento y validación. También pudimos comprobar que la relación entre los coeficientes (a_k) de las variables $PTO_{ki}(D_{ki}, c_j)$ en ambos modelos es exactamente constante = 2,0. Esto indica que, salvo un factor de escala de 2, ambas ecuaciones dan el mismo peso a las distintas variables y deberían dar resultados similares. De hecho, encontramos un coeficiente de correlación de R = 0,98 para los valores de $f(v_{ij})_{calc}$ obtenidos con la ecuación (21) frente a la ecuación (22). Este resultado demuestra que la estrategia all-in-one implementada en SOFT.PTML es capaz de reproducir los resultados obtenidos con la estrategia multi-software, pero utilizando un único programa con una interfaz amigable, lo que facilita y agiliza notablemente el trabajo.

Estudio computacional y experimental de los 2-acilpirroles. Se presenta un caso de estudio para ilustrar el uso de los modelos SOFT.PTML para el descubrimiento de compuestos antileishmanios en la práctica. Como ya se ha dicho, nos centramos en los 2-heteroarilpirroles **5a** y **5b**, cuya síntesis ya habíamos descrito anteriormente (Santgao, 2020), porque combinaban características estructurales de derivados de pirrol relacionados (Alloco, *et al*., 2006, Baiocco, *et al*, 2013, Viera, *et al*., 2017), con una prometedora actividad antileishmanial. Hasta donde

sabemos, no se han realizado estudios previos sobre su actividad antileishmania. Aquí describimos los primeros ensayos *in vitro* y el cribado computacional en profundidad de estos compuestos. En primer lugar, se llevó a cabo la síntesis de nuevas muestras de los 2-heteroarilpirroles 5a y 5b, con calidad de ensayo preclínico. A continuación, estos compuestos se ensayaron por primera vez contra dos especies de *Leishmania* en diferentes fases de desarrollo. Estos estudios experimentales incluyeron dos parámetros de actividad biológica (IC_{50} y CC_{50}) para dos especies de *Leishmania*. En consecuencia, cerramos el estudio con un amplio cribado computacional de estos compuestos frente a muchas especies de *Leishmania* en diferentes estadios y múltiples proteínas diana.

Síntesis orgánica preparativa. Nuestro grupo ha informado recientemente de la síntesis de una variedad de 2-(hetero)aroilpirroles a través de una acilación catalizada por Pd(II) de pirrol con aldehídos (Kumar *et al*., 2020) en presencia de un oxidante, utilizando 2-metilpiridinilo y 2-pirimidilo como grupos directores (Santiago, *et al*., 2020). Esta reacción de activación radical C-H (Gensch *et al*, 2016; Santiago *et al*, 2020; Hunja *et al*, 2021; Sun *et al*, 2021; Joshi *et al*, 2021) es una buena alternativa catalítica a los métodos clásicos de acilación (reacciones de acilación tipo Friedel-Crafts, Vilsmeier-Haack o Houben-Hoesch), que minimiza la producción de residuos al no requerir el uso de cantidades estequiométricas de ácidos Lewis o próticos. Así, habíamos demostrado que el uso del grupo director 2-pirimidinilo conducía a la metalización C-2 del pirrol utilizando $Pd(OAc)_2$ como precatalizador en tolueno, que se acilaba con aldehídos en presencia de TBHP como oxidante y ácido piválico como aditivo. El procedimiento pudo extenderse eficientemente a una serie de aldehídos con diferentes patrones de sustitución en el anillo aromático, obteniendo los 2-acilpirroles **5aa-an** (Figura 14), aunque la diacilación no pudo evitarse completamente. Sin embargo, en las mismas condiciones experimentales, el uso del

grupo director 3-metil-2-piridinilo condujo a la formación de los derivados de 2-acilpirrol **5ba-bs** en rendimientos de moderados a buenos, excepto cuando había sustituyentes que retiraban electrones en el anillo aromático (Figura 14).

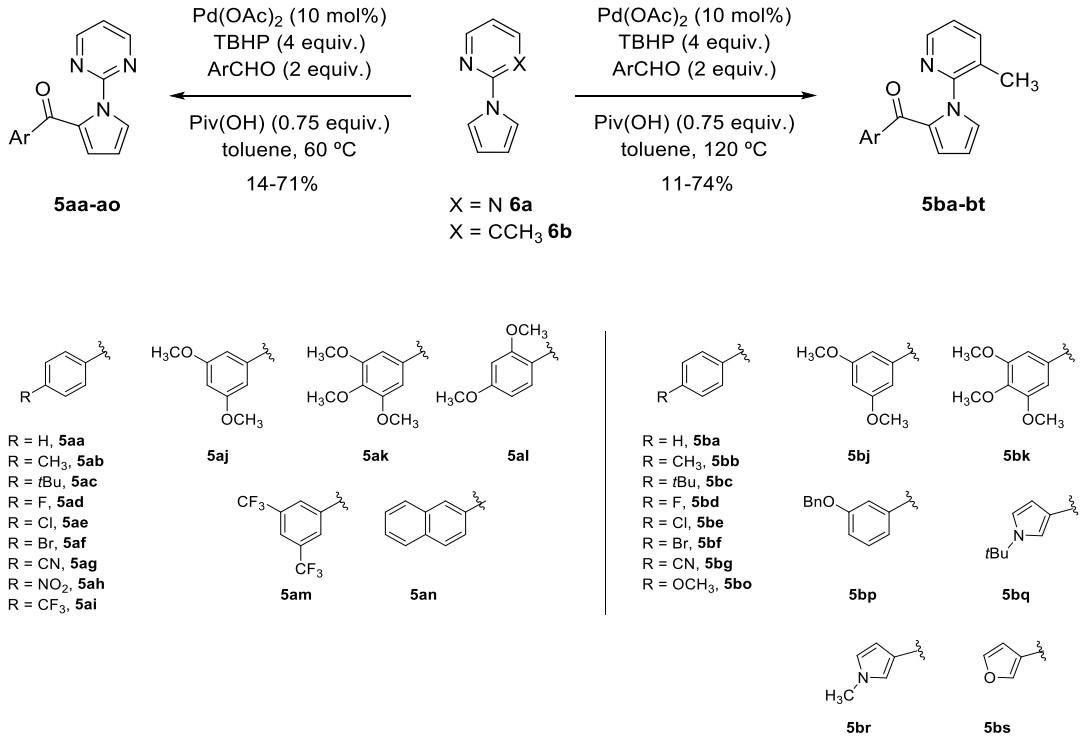


Figura 14. Síntesis de los 2-acilpirroles 5a y 5b examinados contra L. amazonensis y L. donovani.

Ensayo preclínico de actividad antileishmanial. Los derivados de 2-(hetero) aroilpirrol 5a y 5b se ensayaron contra *L. amazonensis* y *L. donovani*, responsables de las dos principales formas clínicas de esta enfermedad tropical desatendida, la leishmaniasis cutánea y la visceral, respectivamente (Tabla 12). Se realizaron ensayos de susceptibilidad *in vitro* de promastigotes y amastigotes intracelulares (IC_{50}), y ensayos de citotoxicidad (CC_{50}) en la línea celular de macrófagos J774 utilizando miltefosina como fármaco de referencia (véase Materiales y Métodos), y se calcularon los correspondientes índices de selectividad (IS). El rendimiento de

cada N-pirimidin-2-il pirrol acilado **5a** se comparó con el del correspondiente *N*-(3-metilpiridin-2-il) derivado **5b** (Tabla 12). La bioactividad de algunos compuestos de ambas series se compara bien en términos de actividad y selectividad contra los promastigotes de L. amazonensis. El patrón de sustitución aromática del grupo acilo juega un papel importante en la actividad antileishmania de estos derivados del pirrol. En algunos casos, observamos tendencias similares en el perfil de bioactividad de los derivados de pirimidina 5a y las correspondientes piridinas 5b. Por ejemplo, las 4-tbutilfenilpirrolmetanonas 5ac/5bc y las 3,5-disustituidas fenilpirrolmetanonas 5aj/5bj, con sustituyentes donadores de electrones (MeO), mostraron IC_{50} en un rango micromolar similar al de la miltefosina (Tabla 12, entradas 3 *vs.* 17 y 10 *vs.* 22). El comportamiento paralelo se mantuvo también para los derivados trisustituidos 5ak/5bk, que fueron ambos inactivos en nuestras condiciones de bioensayo (Tabla 10, entradas 11 *vs.* 23). Sin embargo, hubo diferencias significativas en los derivados de 2-(hetero) aroilpirroles con anillos aromáticos halogenados. En particular, en la serie de piridinas, 5bd (R =F) resultó ser más activo y selectivo que el fármaco de referencia (miltefosina) (IC_{50} = 16,87 ± 0,73 μM, SI > 10,67), mientras que el correspondiente derivado de pirimidina 5ad (R =F) fue inactivo (Tabla 12, entrada 17 *vs.* entrada 4). También hay que señalar que el compuesto 5an, en el que el anillo de fenilo se había cambiado por un anillo de naftilo, mostró una actividad similar a la del fármaco de referencia con una mejor selectividad (Tabla 12, entrada 14).

El mismo conjunto de 2-(hetero) aroilpirroles 5a,b también se probó en formas promastigotes de *L. donovani* (Tabla 12). Todos los compuestos fueron considerablemente menos activos y selectivos que la miltefosina. Los derivados halogenados de piridina 5bd-5bf presentaron los mejores perfiles, siendo 5bd de nuevo el más activo y selectivo de todos los 2-acilpirroles (IC_{50} = 7,78 ± 0,27 μM y SI > 23,15). Sin embargo, cabe destacar que todos los derivados de pirrol ensayados resultaron menos tóxicos que la miltefosina, con valores de concentración del compuesto que produce una reducción del 50% de la viabilidad celular

(Concentración Citotóxica, CC_{50}) en el rango 87 - 401 µM en células J774. Este es un resultado prometedor, teniendo en cuenta la alta toxicidad (baja selectividad) de los fármacos disponibles en el mercado (Brindha, *et al.*, 2021).

Tabla 12. Efectos leishmanicidas y citotóxicos IC_{50} de las series de 2-acilpirrol 5a y 5b (expresados en mM) en ensayos de promastigotes*in vitro*.

Entr.	Comp.	L. amazonensis		L. donovani		Macrophages J774
		IC_{50} ± SD (µM)[a]	SI[b]	IC_{50} ± SD (µM)[a]	SI[b]	CC_{50} ± SD (µM)[c]
1	**5aa**	259.58 ± 40.72	>1.55	N/Ad		401.17e
2	**5ab**	N/A[d]		N/Ad		381.24e
3	**5ac**	32.88 ± 0.74	>2.77	58.54 ± 7.04	>1.55	91.02 ± 6.55
4	**5ad**	N/Ad		N/Ad		270.19 ± 26.30
5	**5ae**	117.09 ± 10.92	>3.01	189.77 ± 6.76	>1.86	352,46e
6	**5af**	67.18 ± 4.94	>4.54	118.18 ± 25.81	>2.58	304.72e
7	**5ag**	N/A[d]		N/A[d]		364.59e
8	**5ah**	N/A[d]		N/A[d]		339.82e
9	**5ai**	48.02 ± 1.61	>3.13	42.82 ± 0.61	>3.51	150.36 ± 49.40e
10	**5aj**	32.55 ± 0.64	>2.71	210.90 ± 32.77	>0.42	88.29 ± 3.36e
11	**5ak**	N/A[d]		N/A[d]		87.27 ± 7.37e
12	**5al**	N/A[d]		N/A[d]		224.28 ± 46.89[e]
13	**5am**	57.34 ± 0.25	>4.53	198.54 ± 25.36	>1.31	259.56[e]
14	**5an**	36.11± 1.23	>3.12	58.50 ± 3.34	>1.93	112.82 ± 39.19e
15	**5ba**	152.77 ± 21.65	>2.50	N/A[d]		381.23[e]
16	**5bb**	149.20 ± 3.84	>2.43	119.64 ± 14.98	>3.02	361.87[e]
17	**5bc**	30.87 ± 3.11	>10.17	221.31 ± 36.90	>1.42	314.05[e]
18	**5bd**	16.87 ± 0.73	>10.67	7.78 ± 0.27	>23.15	136.96 ± 36.42
19	**5be**	51.46 ± 4.72	>2.89	20.69 ± 0.98	>7.19	148.74 ± 19.33
20	**5bf**	38.10 ± 0.85	>3.14	19.87 ± 1.47	>6.01	119.51 ± 37.36[e]
21	**5bg**	191.77 ± 12.04	>1.81	315.43 ± 60.80	>1.10	348.04[e]
22	**5bj**	71.06 ± 7.81	>1.53	43.51 ± 0.59	>0.80	108.45 ± 5.76[e]
23	**5bk**	N/A[d]		N/A[d]		283.78
24	**5bo**	224.26 ± 14.71	>1.53	172.92 ± 60.34	>1.98	342.07[e]
25	**5bp**	207.96 ± 23.62	>1.31	184.22 ± 8.67	>1.47	356.76[e]
26	**5bq**	209.16 ± 4.69	>1.56	92.81 ± 7.68	>1.08	325.31[e]
27	**5br**	226.54 ±13.42	>1.66	N/A[d]		376.90[e]
28	**5bs**	N/A[d]		N/A[d]		396.40[e]
29	miltefosina	30.67 ± 8.80	1.80	0.24 ± 0.02	230.83	55.40 ± 4.19

[a] IC_{50}: 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]CC_{50}: Concentración del compuesto que produjo una reducción del 50% de la viabilidad celular en las células de cultivo tratadas respecto a las no tratadas.[d]N/A: no activo a la dosis máxima

ensayada (100 µg/mL).[e]Los valores de CC_{50}, expresados como µM, corresponden a 100 µg/mL, que fue la dosis más alta ensayada.

A continuación, un compuesto de cada serie se sometió a más pruebas *in vitro* sobre los amastigotes de *L. amazonensis* y *L. donovani* (Tabla 13). El derivado de pirimidina **5bc** mostró un buen rendimiento con una actividad similar a la de la miltefosina y una mejor selectividad (IC_{50} = 60,55 ± 7,88 µM, SI > 5,19) contra *L. amazonensis*. Sin embargo, el derivado de piridina **5bc** presentó malos resultados en términos de actividad y selectividad (IC_{50} = 153,27 ± 9,11 µM, SI > 1,99).

Tabla 13. Efectos leishmanicidas y citotóxicos IC_{50} de los 2-acilpirroles 5af y 5bc (expresados en µM) en el ensayo *in vitro* de amastigotes.

Entrada	Compuesto	L. amazonensis		L. donovani		Macrófagos J774
		IC_{50} ± SD (µM)[a]	SI[b]	IC_{50} ± SD (µM)[a]	SI[b]	CC_{50} ± SD (µM)c
1	**5af**	153.27±9.11	>1.99	210.87±30.26	>1.45	304.72[e]
2	**5bc**	60.55±7.88	>5.19	N/A[d]		314.05[e]
3	miltefosina	47.55±7.04	2.85	0.44±0.05	307.70	135.93±10.19

[a] IC_{50}: 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] 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] N/A: no activo a la dosis máxima ensayada (100 µg/mL).[e]Los valores CC_{50}, expresados en µM, corresponden a 100 µg/mL, que fue la dosis más alta ensayada.

Cribado computacional de nuevos compuestos basado en el IFPTML. Para este estudio predictivo, seleccionamos 28 compuestos previamente sintetizados por nuestro grupo (ver estructuras en la figura 14), cuya actividad biológica *in vitro* (valores IC_{50}) frente a dos especies

de *Leishmania* (*L. donovani* (Blanchard *et al*, 1991) y *L. amazonensis* (Magaraci*et al*, 2003)) y citotoxicidad frente a una línea celular (línea J774 de macrófagos de ratones BALB/c (Hu*et al*, 2008)) se ha llevado a cabo (Tablas 10 y 11). Sin embargo, hay más de 20 especies de *Leishmania* clínicamente relevantes, como: *L. major* (Booth*et al*, 1987) *L. mexicana* (Bressi*et al*, 2001), *L. aethiopica* (Mäntyläet *al*, 2004), *L. braziliensis*, *L. amazonensis* y *L. donovani* (Del Olmo *et al*, 2001), *L. infantum* (Girault*et al*, 2008)*, etc*. Por lo tanto, podría ser muy interesante conocer (a) otros parámetros (K_i, K_m*, etc*.) de actividad biológica *in vitro* frente a proteínas diana específicas; (b) la citotoxicidad de estos compuestos frente a otras líneas celulares humanas y animales como Jurkat (Gamage*et al*, 1997), Vero(Da Silva *et al*, 2007), THP-1 (Cohelho*et al*, 2007), HEK293 (Changtam*et al*, 2010), HeLa (Marhadour*et al*, 2012), HL-60 (Gehrke*et al*, 2013), Sf9 (Syrjänen*et al*, 2013)*, etc*. En consecuencia, decidimos utilizar nuestro modelo IFPTML de múltiples salidas para realizar un cribado computacional en profundidad de la actividad biológica de estos compuestos en todo el espacio de ensayos biológicos. Así, realizamos un experimento de cribado computacional que incluía el cálculo de 20.704 puntuaciones de actividad para 29 compuestos (28 compuestos + miltefosina como referencia) en 647 ensayos preclínicos diferentes. Estos 647 ensayos preclínicos de referencia presentan combinaciones únicas de las condiciones del ensayo biológico c_0 = parámetro (K_i, IC_{50}, K_m*, etc*.), c_1 = proteína diana, c_2 = línea celular (J774, HeLa, HL_{60}*, etc*.), c_3 = organismo (*L. major, L. mexicana, etc*.) ó c_4 = etapa de desarrollo del organismo*, etc*.

Se realizaron los siguientes pasos. En primer lugar, se utilizó el software DRAGON25 para calcular las entradas del vector de descriptores moleculares de cada compuesto. A continuación, sustituimos los valores de los descriptores moleculares D_{ki} en el modelo, obteniendo como salida las puntuaciones de actividad biológica $f(v_{ij})_{calc}$ para el i-ésimo

compuesto en el j-ésimo ensayo. Finalmente, expresamos las puntuaciones de actividad biológica $f(v_{ij})_{calc}$ en términos de desviación relativa $\Delta f(v_{ij})\%_{calc} = 100\text{-}[f(v_{ij})calc - f(v_{mtfj})_{calc}]/f(v_{mtfj})_{calc}$. Estas puntuaciones relativas expresan la desviación del i-ésimo compuesto de consulta con respecto a la miltefosina de referencia (mtf). Las predicciones muestran, en general, una mayor puntuación relativa de actividad biológica $\Delta f(v_{ij})\%calc$ para la serie de compuestos **5b** que para la serie **5a** en comparación con el fármaco de referencia (miltefosina). En concreto, los compuestos **5bc** y **5bp** muestran valores de puntuación de actividad biológica relativa entre 1 y 4 veces superiores en comparación con la miltefosina. La predicción es consistente con nuestros hallazgos experimentales para los ensayos de actividad IC_{50} frente a *L. amazonensis* y *L. donovani* y el índice de selectividad en la línea celular J774 con nuestros hallazgos experimentales (Tabla 12).

Tabla 14. Puntuación de la actividad biológica relativa con respecto a la miltefosina

Condiciones de ensayo c_j[a]				Variación relativa de la puntuación de la actividad biológica						
c_0	c_1	c_2	c_3							
Propiedad	Objetivo	Célula	Ensayo	$\Delta f(v_{ij})\%_{calc}$ valores						
(Unidades)	Proteína	Línea	Organismo	**5am**	**5af**	**5bc**	**5bd**	**5bp**	n_j	$f(v_{ij})_{ref}$
Principales especies predichas con el modelo										
IC_{50}(nM)	**MD**	**MD**	***L. amazonensis***	**1.70**	**-0.06**	**2.09**	**0.30**	**2.02**	**724**	**0**
IC_{50}(nM)	MD	THP-1	*H. sapiens*	3.17	-0.12	3.89	0.56	3.77	546	0
Inh.(%)	MD	MD	*L. donovani*	3.17	-0.12	3.89	0.56	3.77	2654	0.01
IC_{50}(u)	MD	MD	*L. aethiopica*	3.13	-0.12	3.84	0.56	3.73	81	0.82
EC_{50}(nM)	MD	MD	*L. major*	3.13	-0.11	3.84	0.56	3.72	482	0.01
GI(%)	MD	MD	*L. tropica*	3.00	-0.11	3.68	0.53	3.57	279	0.51
IC_{50}(nM)	MD	MD	*L. chagasi*	2.79	-0.10	3.42	0.50	3.31	85	0
Líneas celulares más predecibles con el modelo										
Ratio	**MD**	**J774.A1**	***MD***	**1.32**	**-0.05**	**1.62**	**0.23**	**1.57**	**84**	**0.91**
SI	MD	L6	*MD*	3.39	-0.12	4.16	0.60	4.03	181	0.89
IC_{50}(u)	MD	J774.A1	*L. donovani*	3.38	-0.12	4.15	0.60	4.02	144	0.82
SI	MD	J774	*MD*	3.23	-0.12	3.97	0.58	3.85	65	0.91
IC_{50}(nM)	MD	THP-1	*H. sapiens*	3.15	-0.12	3.87	0.56	3.75	637	0

SI	MD	LLC-MK2	*MD*	3.02	-0.11	3.71	0.54	3.60	61	0.91
SI	MD	Vero	*MD*	2.98	-0.11	3.65	0.53	3.54	51	0.91
SI	MD	HepG2	*MD*	1.41	-0.05	1.72	0.25	1.67	63	0.91
IC_{50}(u)	MD	J774.A1	*L. donovani*	3.38	-0.12	4.15	0.60	4.02	144	0.82
Principales proteínas objetivo predichas con el modelo										
K_i(nM)	P07382	MD	*L. major*	3.75	-0.14	4.60	0.67	4.46	48	0.01
IC_{50}(nM)	O61059	MD	*L. mexicana*	3.29	-0.12	4.03	0.59	3.91	372	0
IC_{50}(nM)	P11166	MD	*L. mexicana*	3.27	-0.12	4.01	0.58	3.89	13658	0
IC_{50}(nM)	O97467	MD	*L. mexicana*	3.27	-0.12	4.01	0.58	3.89	13642	0
IC_{50}(nM)	Q0GKD7	MD	*L. major*	2.95	-0.11	3.62	0.53	3.51	76	0
Inh.(%)	P39050	MD	*L. donovani*	2.81	-0.10	3.44	0.50	3.34	53	0.01
Act.(%)	Q27686	MD	*L. mexicana*	2.36	-0.09	2.90	0.42	2.81	62	0.1
K_i(nM)	Q01782	MD	*L. major*	1.52	-0.06	1.86	0.27	1.81	46	0.01
Inh.(%)	E9BF75	MD	*L. donovani*	1.32	-0.05	1.62	0.24	1.58	68577	0.01

[a]Act.(%) = Actividad(%), Inh.(%) = Inhibición(%), GI(%) = Inhibición del crecimiento(%),Cyt.(%) = Citotoxicidad(%),SI = Índice de selectividad = Relación CC_{50}/IC_{50},IC_{50}(u) = IC_{50}(ug. mL-1),[b]$\Delta f(v_{ij})\%_{calc} = 100\text{-}[f(v_{ij})_{calc} - f(v_{mtfj})_{calc}]/f(v_{mtfj})_{calc}$, $f(v_{ij})_{calc}$ = Puntuación de actividad biológica calculada del ésimo compuesto, $f(v_{mtfj})_{calc}$ = Puntuación de actividad biológica calculada del fármaco de referencia miltefosina, MD = Datos perdidos.

La Tabla 12 resume los resultados de las puntuaciones de actividad biológica $f(v_{ij})_{calc}$ calculadas para algunos de los 28 compuestos (**5am, 5af, 5bc, 5bd** y **5bp**), considerando los organismos, líneas celulares y proteínas diana más relevantes. El compuesto **5bd** se predijo con un valor positivo $\Delta f(v_{ij})\%_{calc}$ = 0,30 para el IC_{50} frente a los promastigotes de *L. amazonensis*. Esto significa que el modelo predice con alta probabilidad que este compuesto se encuentra en el mismo rango de actividad que la miltefosina. Este resultado concuerda con los hallazgos experimentales reportados en la sección anterior, IC_{50} = 16.87 µM para el compuesto 5bd e IC_{50} = 30.67 para la miltefosina*vs.L. amazonensis*promastigotes. El compuesto 5bc también se predice con un valor positivo de Δf(vij)%calc = 2,09 para la IC_{50} frente a los promastigotes de L. *amazonensis*, que también coincide con nuestro hallazgo experimental (IC_{50} = 30,87 µM para el

compuesto 5bc frente a IC_{50} = 30,67 para la miltefosina en el ensayo con promastigotes de *L. amazonensis*). La misma tendencia se ha observado para otros compuestos (por ejemplo, 5am) que también se predijeron con valores positivos de Δf(vij)%calc aproximadamente en el mismo rango que la miltefosina. Curiosamente, el compuesto 5af se predice con valores Δf(vij)%calc inferiores a los de la miltefosina (valores negativos de Δf(vij)%calc) tanto frente a los promastigotes de *L. amazonensis* como frente a los de L. donovani, tal y como se observa en los resultados experimentales (véase la Tabla 12).

Además, se predijeron las puntuaciones de la actividad biológica de estos compuestos frente a diferentes líneas celulares. Para esta serie, el punto de corte para las puntuaciones de actividad biológica fue de 1,62 y 50 para n_j. Sólo se ha mostrado un ensayo por línea celular. En primer lugar, nos centramos en el Índice de Selectividad (IS = RatioCC_{50}/IC_{50}) de los compuestos frente a las líneas celulares J774 porque fueron las líneas experimentales utilizadas. En concreto, el compuesto 5bc tiene un valor de $\Delta f(v_{ij})$%calc = 1,62 para la línea celular J774.A1, lo que significa que se espera que este compuesto muestre una probabilidad entre similar y mayor que la miltefosina de presentar un IS positivo, de acuerdo con nuestros resultados experimentales. De hecho, se encontró que el compuesto 5bc tenía un IS >10,17 que es aproximadamente 6 veces el valor de la miltefosina con un IS = 1,80. El modelo fue capaz de reproducir, en general, las tendencias de los valores de IS de todos los compuestos de ambas series para las líneas celulares J774 y/o J774.A1. Se predijeron resultados similares de SI positivo con el modelo IFPTML para otras líneas celulares no probadas experimentalmente aquí. Los valores más altos se calcularon para las líneas celulares L6, J774.A1, J774 THP-1, LLC-MK2, Vero, HepG2 y J774.A1. Esto apunta a estas líneas como objetivos interesantes para seguir probando la seguridad de estos compuestos en el futuro.

Por último, también seleccionamos aquellas proteínas con el mayor incremento en la puntuación de actividad biológica con respecto a la miltefosina de referencia. Además, filtramos las proteínas por el número de ensayos (n_j) para esta proteína presente en el conjunto de datos ChEMBL. Las proteínas con mayor n_j son las más estudiadas y probablemente las más relevantes debido a la mayor atención que reciben. Para seleccionar los casos que incluyen las proteínas más relevantes, el corte para las puntuaciones de actividad biológica fue $\Delta f(v_{ij})\%_{calc}$ =1,62 y el corte para n_j= 45. Sólo se muestra un ensayo por proteína. Según los resultados obtenidos con el modelo IFPTML, las proteínas objetivo más plausibles son: Dihidrofolato reductasa-timidilato sintasa bifuncional (P07382)(Ferrari *et al*, 2011), Transportador de glucosa (O61059) (Burchmore *et al*, 1998), Transportador de glucosa facilitado miembro 1 de la familia de transportadores de solutos 2 (P11166) (Mueckler *et al*, 1985), Transportador de hexosa 1 (Q0GKD7)(Gosh *et al*, 2004), Farnesil pirofosfato sintasa (O97467) (Louw, 1998), Tripanotión reductasa (P39050)(Shukla *et al*, 2012), Piruvato quinasa (Q27686)(Nowicki *et al*, 2008), Pteridina reductasa 1 (Q01782)(Tulloch *et al*, 2010), y Metionina - ARNt ligasa (E9BF75)(Downing *et al*, 2011). Tras una inspección detallada de todos los compuestos activos frente a estas proteínas en nuestra base de datos ChEMBL, no hemos encontrado una similitud significativa entre los compuestos previamente reportados y los derivados de 2-acilpirrol probados aquí. En consecuencia, los derivados de 2-acilpirrol pueden considerarse como una nueva clase de compuesto líder antileishmanial digno de mayor investigación. En el archivo de información de apoyo SI02.xlsx publicamos los resultados detallados del estudio computacional.

3.4. CONCLUSIÓN

Hemos demostrado que SOFT.PTML es una herramienta útil para desarrollar modelos predictivos para el descubrimiento de fármacos. El software implementa algoritmos IFPTML en

una interfaz fácil de usar sin necesidad de depender de múltiples paquetes de software para ejecutar las diferentes etapas (IF, PT y ML) del algoritmo. Y lo que es más importante, es capaz de procesar conjuntos de datos complejos con características de Big Data (gran volumen, múltiples salidas, múltiples proteínas objetivo, líneas celulares, especies de patógenos, datos perdidos*, etc*.). En concreto, hemos ilustrado el uso de este software procesando un conjunto de datos ChEMBL muy grande (>145.000 casos) de ensayos preclínicos antileishmania. Hemos explorado diferentes algoritmos de ML (SVM, LOGR y RF), no utilizados anteriormente para estudiar este conjunto de datos. El mejor modelo encontrado, IFPTML-LOGR, estima la probabilidad con la que múltiples parámetros (IC_{50}, CC_{50}, SI*, etc*.) de un nuevo compuesto llegan a un nivel deseado en ensayos preclínicos con especificidad y sensibilidad (80-98%) tanto en las series de entrenamiento como en las de validación. Este resultado demuestra que la estrategia "todo en uno" implementada en SOFT.PTML es capaz de reproducir los resultados obtenidos con la estrategia de varios programas, pero utilizando un único programa con una interfaz fácil de usar que hace el trabajo notablemente más fácil y rápido. Los ensayos preclínicos estudiados incluyen diferentes especies de *Leishmania* y líneas celulares, así como múltiples proteínas diana. También hemos ilustrado el uso de la nueva herramienta en un caso práctico de estudio de derivados de 2-acilpirrol. La evaluación *in vitro* de la actividad leishmanicida de las series de 2-acilpirroles **5a** y **5b** contra la leishmaniasis visceral (*L. donovani*) y cutánea (*L. amazonensis*) reveló que todos los 2-acilpirroles probados mostraron una citotoxicidad muy baja, CC_{50}> 100 µg/mL en las 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 particular, 5bd (IC_{50} = 16,87 µM, SI > 10,67) fue aproximadamente 6 veces más potente y selectivo que el fármaco de referencia (miltefosina) en los ensayos con

promastigotes de L. amazonensis. Se trata de un resultado importante, ya que señala a los 2-acilpirroles como una nueva clase de compuestos principales dignos de una mayor optimización como éxitos antileishmanios.

Capítulo 4

CAPÍTULO 4: ANÁLISIS NIFPTML DE LA ESTABILIDAD DE LA LEY GENERAL TRIBUTARIA ESPAÑOLA.

4. Articulo

Ref. Análisis NIFPTML de la Estabilidad de la Ley General Tributaria Española. Ortega-Tenezaca B, Duardo-Sánchez, A., Munetanu, C.R., and González-Díaz H,. *Complex Network*. 2022, *in preparation*.4.1. Herramientas de análisis basadas en teoría de grafos y redes complejas aplicadas a las Ciencias Sociales

Frente a la metodología tradicional, en el estudio de las relaciones sociales en general y jurídicas en particular, las herramientas de análisis basadas en la teoría de grafos y redes complejas aportan una nueva perspectiva científica. Éstas, combinadas con las herramientas de las Tecnologías de la Información y la Comunicación (TIC), permiten interconectar la información estructural de la materia a muy diferentes niveles: desde las pequeñas moléculas hasta las redes macroscópicas. Ejemplos relevantes de ello son los genomas (redes de genes), las redes metabólicas complejas, las redes neuronales biológicas (conectomas del cerebro humano), las redes sociales y las redes TIC-sociales. Estas últimas incluyen, entre otras, las redes de transmisión de enfermedades, las redes de telefonía inalámbrica, las redes de comercio electrónico internacional, Internet y la World Wide Web (www), Facebook, Twitter, etc. (Bornholdt y Schuster, 2003). En todos estos casos, es posible calcular un tipo de parámetros denominados índices topológicos (IT) y centralidades de vértices que describen numéricamente los patrones de conectividad existentes entre los nodos o actores de una red (representada como un gráfico matemático). Estos índices pueden utilizarse como variables de entrada para realizar modelos estadísticos sobre las propiedades de las redes que dependen de su estructura.

En particular, el análisis de redes sociales (más conocido por su denominación en inglés como Social Network Analysis (SNA)) se ha extendido al estudio de las relaciones entre los interlocutores sociales a diferentes niveles de análisis: individuos, grupos, estructuras políticas y otras estructuras sociales susceptibles de representación en redes (Abercrombie, Hill y Turner 2000; Craig 2002; White 1976; Wellman y Berkowitz 1988). El trabajo de Newman: *The Structure and Function of Complex Networks* (Newman, 2003), nos ilustra que es posible representar las relaciones sociales como redes en las que los nodos pueden ser actores individuales y las aristas las relaciones entre estos actores.

Esta combinación del análisis de redes con las TIC puede ser útil para estudiar la interrelación entre los diferentes tipos reglas en Derecho, y predecir las consecuencias y la eficacia de las nuevas normas jurídicas. Uno de los ejemplos más relevantes del uso del SNA en las Ciencias jurídicas es el estudio realizado por el profesor Fowler y sus colegas, que han utilizado la teoría de redes para estudiar la relevancia de una decisión judicial basándose en los patrones de citas a sentencias previas (precedente judicial) del Tribunal Supremo de los Estados Unidos (Fowler y Jeon, 2008). En dicho trabajo se construyó una red de sentencias dictadas por el Tribunal Supremo de EE.UU. y los casos que citan, describiendo un método para establecer una medida cuantitativa de la relevancia de un caso judicial, utilizando los datos de la red para identificar los precedentes judiciales más importantes. Los autores demostraron que los IT y las centralidades de los nodos son una forma ideal de medir la importancia de un precedente (Fowler, *et al.,* 2007). En otro estudio, Boulet *et al* analizaron la estructura del sistema de códigos francés, seleccionaron cincuenta y dos códigos legislativos, a los que consideraron vértices con sus citas (remisiones) mutuas formando las aristas en una red cuyas propiedades fueron analizadas apoyándose en la teoría de grafos (Boulet, 2011). Otro precursor directo del

presente estudio es el trabajo de Bommarito y Katz relacionado con la estructura del Código de Derecho Federal de Estados Unidos (Bommarito, 2010). En dicha red, los nodos son las diferentes partes del código interconectadas por relaciones estructurales de remisión de unas normas a otras (cita) y jerárquicas: título > subtítulos > capítulos > subcapítulos > partes > subpartes > secciones > subsecciones > párrafos > subpárrafos > cláusulas > subcláusulas. En trabajos anteriores Duardo *et al*, han utilizado el SNA para analizar el sistema jurídico tributario español; e esta ocasión los nodos de la red están constituidos por normas jurídicas del sistema tributario español aprobadas en un determinado periodo de tiempo. En esta red dos nodos están conectados si ambas leyes tienen la misma jerarquía y fueron aprobadas en un intervalo de tiempo predeterminado (Duardo-Sánchez *et al*. 2014). También en este campo se han descrito redes que discriminan las con-causas de un delito de la causa principal (Duardo-Sánchez 2010, 2011).

En este capítulo construimos por primera vez, una red compleja para analizar la Ley General Tributaria (LGTN). La Ley 58/2003, de 17 de diciembre, General Tributaria, en adelante LGT, es una norma con vocación codificadora que recoge y establece los principios y normas jurídicas generales del sistema tributario español. El objetivo declarado de esta Ley es adaptar las normas de distribución de competencias derivadas de la Constitución Española (CE) de 1978. En este sentido, la adecuación a la Carta Magna implica el establecimiento de las condiciones básicas que garanticen la igualdad en el cumplimiento del deber constitucional de contribuir al sostenimiento de las cargas públicas; la aplicación y eficacia de las normas jurídicas, y la identificación de las fuentes del derecho tributario; el establecimiento de los conceptos, principios y reglas básicas del sistema tributario en el marco de la Hacienda General; y la adaptación de las particularidades de los procedimientos tributarios al procedimiento

administrativo común, de modo que se asegure a los contribuyentes un tratamiento similar en todas las administraciones tributarias. Así, en un contexto en el que concurren sistemas fiscales de diversas estructuras políticas (el Estado y las diversas Comunidades Autónomas), el estudio de la LGT y la estabilidad en el tiempo de esta norma resulta de gran importancia.

Nuestra red codifica numéricamente la estructura de la Ley 58/2003, de 17 de diciembre, General Tributaria. Para ello, se analiza la estructura jerárquica de la misma, las citas entre los distintos elementos de la norma y las alteraciones que se han producido durante su vigencia. A continuación, calculamos los parámetros numéricos de la LGTN basándonos en la teoría de redes complejas (centralidades de vértices). Por último, utilizamos estas centralidades de vértice como variables de entrada para un análisis NIFPTML. Todo ello ha permitido construir un modelo estadístico capaz de predecir la dinámica de futuras redes, incluidas las modificaciones a lo largo del tiempo (previsión de la estabilidad de la norma legal).

4.2. Materiales y Métodos

4.2.1. Red de la Ley General Tributaria (LGTN).

Construimos esta LGTN en diferentes etapas. En primer lugar, recopilamos toda la información sobre la estructura de la ley y las citas (remisiones) entre artículos LGT. Después, consideramos en un archivo Excell que los nodos ni (i = 0, 1, 2, ... ntot) del LGTN son las diferentes partes de la ley con un total de ntot partes (nodos). A continuación, interconectamos los nodos mediante relaciones estructurales jerárquicas (enlaces estructurales). Estos nodos están vinculados de la siguiente manera: Ley > títulos > capítulos > artículos > párrafos > incisos (o subsección, según el observatorio legislativo del Parlamento Europeo) (*Parlamento Europeo, 2013*). En total, el conjunto de datos LGTN contiene ntot = 1923 nodos. Cada nodo que

representa un artículo se conecta al menos con un párrafo. Para el caso específico de los párrafos indivisibles (los que no tienen subsecciones), asignamos toda la información a un nodo terminal. A esta primera etapa la llamamos Red de Estructura (SN). Posteriormente, añadimos a la SN las citas entre artículos de la propia norma. En este caso, conectamos las citas hacia fuera con las citas hacia dentro obteniendo una nueva red de 2400 nodos: la Red de Estructura-Citaciones (SNA). Además, hicimos dos redes aleatorias (RNDN), con un número de nodos y enlaces similar al de la SN y la SNA respectivamente. Utilizamos estas RNDN con fines comparativos. Véase la tabla 15.

Tabla 15: Parámetros numéricos usados en este trabajo para describir redes complejas

Centralidad de nodo	Formula [a]
Diameter	$D = max\big(dist(v,w)\big)$
HITS-Hubs	$C\text{hubs} = AC\text{auths}$
HITS-Authority	$C\text{auths} = AT\ C\text{hubs}$
Degree	$C(v)\, deg(v)_{deg}$
Closeness vitality	$C_{clv}(v) = W(G) - W(G\backslash\{v\})$
Closeness	$C_{clo}(v) = \left(\sum_{w\in V} dist(v,w)\right)^{-1}$
Closeness centrality	$C_C(v) = \frac{n-1}{\sum_{v\in V} dist(w,v)}$
Eccentricity	$C_{ecc}(v) = \frac{1}{max\{dist(v,w): w \in V\}}$
Radiality	$C_{rad}(v) = \frac{\sum_{w\in V}\big(\Delta_G + 1 - dist(v,w)\big)}{n-1}$
Current-flow closeness	$C_{cfc}(v) = \frac{n-1}{\sum_{t\notin V} p_{vt}(v) - p_{vt}(t)}$
Current-flow betweenness	$C_{cfb}(v) = \frac{1}{(n-1)(n-2)}\sum_{s,t\in V} \tau_{st}(v)$
Stress	$C_{spb} = \sum_{s\notin v\in V}\sum_{t\notin v\in V} \sigma_{st}(v)$

Short-path betweenness	$C_{spb} = \sum_{s \notin v \in V} \sum_{t \notin v \in V} \delta_{st}(v)$
Betweenness centrality	$BC(k) = \sum_{t} \sum_{w} \frac{Q(t,v,w)}{Q(t,w)}$
Eigenvector centrality	$EC(v) = e_1(v)$
Information centrality	$IC(v) = \left[\frac{1}{N} \sum_{v \in V} \frac{1}{I_{vw}}\right]^{-1}$
Katz-status index	$C_{katz} = \sum_{k=1}^{\infty} \alpha^k \cdot (A^t)^k \cdot u$

[a] Todos los símbolos utilizados en estas fórmulas son muy comunes en la literatura sobre PINs y no pueden ser explicados en detalle aquí. Sin embargo, $G = (V, E)$ es un grafo no dirigido o dirigido, conectado (fuerte) con $n = |V|$ vértices; $\delta(v)$ denota el grado del vértice v en un grafo no dirigido; dist(v, w) denota la longitud de un camino más corto entre los vértices v y w; σst denota el número de caminos más cortos de s a t and $\sigma_{st}(v)$ denota el número de caminos más cortos desde s hasta t que utilizan el vértice v. **D** y **A** son la distancia topológica y la matriz de adyacencia del grafo g. Por favor, para más detalles ver las referencias citadas. Software: CB = CentiBin.

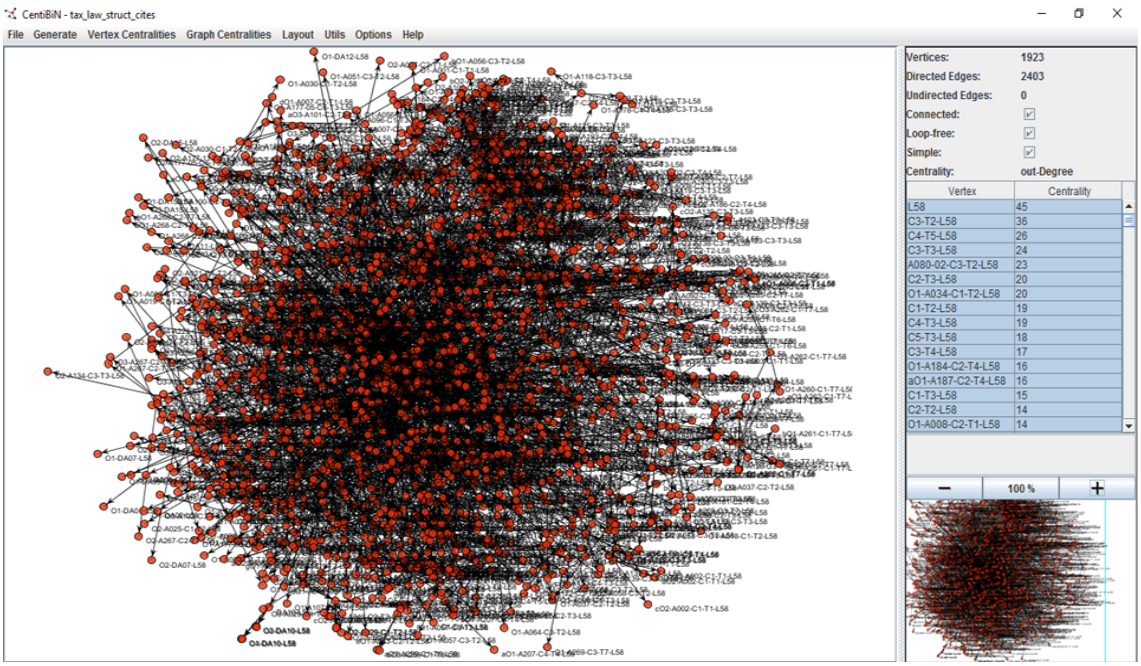


Figura 15: Software CentiBiN - Ilustración de la LGTN

4.3. Métodos computacionales

Centralidades de los nodos de LGTN.

En el caso particular de las redes jurídicas y otras redes sociales abordadas en el SNA, las centralidades de los vértices pueden desempeñar un papel muy importante. La centralidad de un nodo o vértice, en teoría de grafos y redes complejas, se refiere a parámetros numéricos que miden de alguna manera la importancia relativa del nodo dentro de la red. En SNA, el valor de la centralidad de un nodo es útil, por ejemplo, para detectar personas relevantes dentro de una red social. En la Tabla 2, resumimos algunas centralidades de vértice calculadas por CentiBiN, software utilizado en este trabajo. CentiBiN soporta centralidades de vértices tanto para redes dirigidas como no dirigidas. Calcula 17 centralidades de red diferentes, que van desde medidas locales. Estas medidas sólo consideran la vecindad directa de un elemento de la red, hasta medidas globales (Koschützki *et al*., 2005; Jacob *et al*., 2005).

Tabla 16: Red de la Ley General Tributaria (LGTN) vs. Modelos de red aleatoria (RND)

Networka	Connectivity	Networka
S-LGTN	Parameter	C-LGTN
1923	n	1923
1922	L	2400
7.4	D	6.7
0.00052	$<C_{auth}>$	0.00052
Networka	Connectivity	Networka
S-RNDN	Parameter	C-RNDN
1882	n	1949
1932	L	2339
20.8	D	9.7
0.00053	$<C_{auth}>$	0.00046

[a] S-LGTN = LGTN con información structural pero sin citas incluidas. S-RNDN = S-LGTN Modelo de red aleatorio, C-LGTN = S-LGTN con citas incluidas. C-RNDN = C-LGTN Modelo de red aleatorio.

4.3.1. Modelo NIFPTML

Como hemos mencionado anteriormente lo modelos NIFPTML se descomponen en las fases NI + IF + PT + ML. En este ejemplo, hemos usado las fases NI, PT y ML. La fase IF ha sido omitida porque toda la información provenía de una misma fuente. En primer lugar, fase NI, hemos calculado las invariantes de red (network invariants) en forma de centralidades de los nodos ${}^{i}C_{k}$ del tipo k para cada nodo ni (considerando sólo los artículos o subparágrafos) en la LGTN. Hemos considerado que los artículos y los párrafos son los más importantes en términos de contenido. Posteriormente, en la fase PT, hemos definido las variables de entrada incluyendo la función de referencia f(mij)ref, las centralidades, el tiempo, y las desviaciones u operadores de perturbación (PT). La función f(mij)ref es igual a la probabilidad a priori con que una norma relativa a la misma materia puede ser modificada. Se calcula como el cociente entre el número de artículos modificados y el número de artículos totales dedicados a la misma materia. Los operadores de perturbación o desviaciones fueron calculados como $\Delta\ {}^{i}C_{k} = {}^{i}C_{k} - < {}^{i}C_{k}>_{materia}$. Donde ${}^{i}C_{k}$ es la centralidad de la norma estudiada en la red y $< {}^{i}C_{k}>_{materia}$ es el valor promedio de dicha centralidad para todas las normas que se ocupaban de la misma materia.

A continuación, en la fase ML, utilizamos los valores de las variables como entradas para realizar un Análisis Discriminante Lineal (LDA). El objetivo del LDA era buscar un nuevo modelo NIFPTML para los cambios a lo largo del tiempo de LGTN. Expresamos el tiempo ti en términos de número de meses desde el 17 de diciembre de 2003 (fecha de aprobación del LGTN original). La variable a predecir es el parámetro de modificación de la ley mij. El parámetro mij = 1 (modificación) si el artículo ni con centralidades iCk ha sido modificado en el momento tj; mij = 0 en caso contrario. La salida de la ecuación LDA no es m_{ij} sino $f(m_{ij})_{calc}$, que es una puntuación de valor real de mij. Podemos escribir la ecuación LDA con los parámetros

mencionados anteriormente de la siguiente forma, véase también la Figura 1. Utilizamos el algoritmo LDA implementado en el software PTML desarrollado en esta tesis. Podemos utilizar diferentes parámetros estadísticos para evaluar la significación estadística y validar la bondad de ajuste de la ecuación LDA. Estos parámetros son: n = número de casos, $\chi 2$ = Chi-cuadrado, p = el nivel de error, así como la Exactitud, Especificidad y Sensibilidad de las series de entrenamiento y validación externa (Hill y Lewicki, 2006). La ecuación general del modelo buscado es la siguiente.

$$f(m_{ij})_{calc} = a_0 \cdot f(m_{ij})_{ref} + \sum_{k=1}^{7} a_k \cdot \ {}^{i}C_k + \sum_{k=1}^{7} b_k \cdot \Delta \ {}^{i}C_k + d_j \cdot t_j + e_0 \quad (1) \qquad (23)$$

4.4. Resultados y discusión

4.4.1. Modelo NIFPTML predictivo de red compleja

El ideal clásico del Derecho se inspiraba en los principios de generalidad, universalidad, abstracción, estabilidad y permanencia. Hoy en día, estos rasgos están siendo cuestionados. Hay una explosión de leyes diversas según los grupos sociales, con la consiguiente crisis del principio de generalidad (Zagrebelsky, 1992). También en relación con dichos grupos, las nuevas situaciones y las necesidades cambiantes exigen normas jurídicas adecuadas para cada caso, que pone en jaque al principio de abstracción. En cuanto a la permanencia y estabilidad de las normas, en lugar de estos ideales, los criterios que ahora prevalecen son la temporalidad y la oportunidad (Javier, 2015). En consecuencia, los destinatarios de las leyes son cada vez más vulnerables. En este contexto, enmarcamos la utilidad de un modelo capaz de predecir los cambios futuros de una norma jurídica en particular. Aquí reportamos el primer modelo

NIFPTML para la estabilidad a lo largo del tiempo (dinámica) de la LGTN. El mejor modelo encontrado con LDA fue el siguiente:

$$f(m_{ij})_{calc} = 105549.7253 \cdot {}^{i}C_{auth} + 0.0153 \cdot t_j - 2.7138$$
$$n = 36856 \quad \chi^2 = 131.62 \quad p < 0.05 \tag{24}$$

Donde, $f(m_{ij})$ es una puntuación de valor real que mide la propensión del nodo n_i (artículo o párrafo) a ser modificado ($m_{ij} = 1$) o no ($m_{ij} = 0$) en el momento t_j desde que la norma entró en vigor. Cuando el valor de $f(m_{ij}) > 0$ podemos esperar que el nodo n_i (artículo o párrafo) tenga una mayor propensión a ser modificado en el momento tj. Sin embargo, los valores preferidos para hacer esta clasificación son los valores de probabilidad de modificación de la norma pij. Recopilamos información detallada sobre el código, el iCauth, el tiempo tj, el mij observado, el mij predicho y el pij para todos los >40000 casos estudiados en este trabajo. Los parámetros estadísticos de este modelo son N = número de modificaciones de nodos + casos de nodos no modificados a lo largo del tiempo, χ2 es el estadístico Chi-cuadrado, y p es la p-vela l de error. Como se puede observar en la ecuación 2, los términos de la función de referencia y las desviaciones $\Delta\ {}^{i}C_k$ tienen coeficiente 0 y no influyen en la respuesta. En este caso el modelo NIFPTML quedó reducido a un modelo ML lineal clásico.

Nuestro modelo funciona como una serie temporal incrustada en una red compleja (Riera-Fernández, *et al.,* 2012). Esto es así porque predice la modificación (mij) de diferentes artículos o párrafos en el LGT español cuando las condiciones externas (socioeconómicas, por ejemplo) cambian en el momento tj dado que se ha utilizado una norma conocida en el pasado. El modelo predijo correctamente el registro histórico de enmiendas de la LGT con alta precisión,

especificidad y sensibilidad (Tabla 3). La variable de red compleja más importante de acuerdo con este modelo fue la Autoridad de los Impactos iCauth. El Hypertext Induced Topics Search (HITS), es un algoritmo de análisis de enlaces introducido por primera vez por Jon Kleinberg a finales de los noventa (Kleinberg, 1999). Este algoritmo se basa en la idea de que hay dos tipos de nodos importantes -hubs y autoridades- en todas las colecciones de documentos que pueden ser representados por redes dirigidas (Benzi, Estrada, y Klymko, 2013). Las autoridades son estos nodos importantes. Véase la fórmula HITS-Autoridades utilizada en el software Centibin; suponiendo que se conozca Chubs (Kleinberg, 1999), los hubs son nodos considerados importantes.

La contribución de los Cauths a la puntuación de modificación de un artículo o párrafo, según el modelo, es muy alta y positiva (coeficiente a1 = 105549.7253). Esto podría deberse a que las variables más relevantes en el presente problema son el tema de la regulación y el efecto a lo largo del tiempo de los cambios en los factores políticos, sociales y económicos. En este caso, es muy importante tener en cuenta la pertenencia de España a la Unión Europea, una entidad supranacional. Los principios comunitarios de primacía y efecto directo obligan al Estado español a adaptar el ordenamiento jurídico interno a la legislación de la Unión Europea. Por otro lado, la contribución del tiempo tj a la puntuación de la modificación de un artículo o párrafo, según el modelo, es también positiva pero pequeña.

4.4.2. Estudio de la red compleja de la Ley General Tributaria

También realizamos un análisis comparativo de la red compleja observada con los modelos de red aleatoria para estudiar la primera con más detalle. Para ello, construimos dos modelos de redes aleatorias de Erdös-Renyi lo más parecidos posible a la Ley General Tributaria

y a la Ley General Tributaria más citas (SGATC) respectivamente (ver Tabla 1). La atención, en el estudio descriptivo, se centró en la centralidad de Cauths porque esta variable era la mejor entrada en el modelo. De la comparación entre ambos, STGA real y aleatoria, podemos concluir que el valor de la distancia media es muy diferente en uno y otro (7,4 LGTN vs. 20,8 RND). Mientras que los valores de Cauths son muy similares (0,00052 LGTN vs. 0,000 53 RND). Por otro lado, la comparación entre el STGANC real y el aleatoria, muestra más similitud en los valores de distancia media (6,7 SGTANC vs. 9,7 RND) y, una diferencia no significativa entre los valores de Cauhts, aunque no son tan similares como en la comparación anterior (0,00052 SGTANC vs. 0,00046 RND). Quizá lo más llamativo de este tema no sea la similitud o disimilitud entre las redes reales y las aleatorias sino la coincidencia entre los valores de iCauths entre la LGTN y el SGTANC, exactamente 0,00052 en ambos casos. Esto demuestra que estos valores no se ven alterados por la inclusión en la red original de citas entre las diferentes partes de la norma jurídica.

Tabla 17: Top 10 Hits-Authority nodes in C-LGTN (structure + citations network)

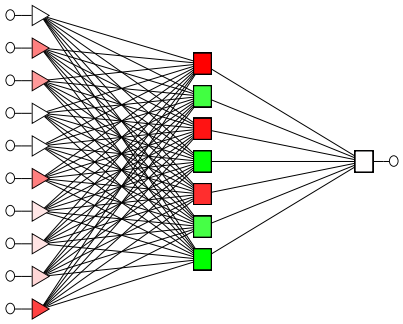
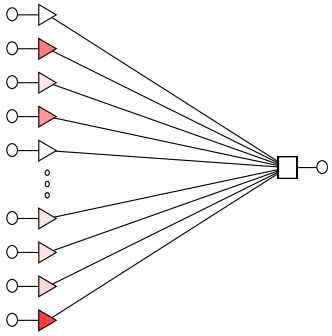
Node [a]	Subject of regulation	m_{ijobs}	m_{ijpred}	t_j	Month / Year	${}^{i}C_{auth}$	p_{ij}
A187-C2-T4-L58	Tax Penalties Determination	0	0	9	09/2004	0.0398	0.185
A184-C2-T4-L58	Tax Infringement	0	0	23	11/2005	0.0345	0.219
A191-C3-T4-L58	Tax Infringement	0	0	29	5/2006	0.0257	0.148
A193-C3-T4-L58	Tax Infringement	0	0	58	10/2008	0.0220	0.381
A192-C3-T4-L58	Tax Infringement	0	0	135	03/2015	0.0215	0.391
A194-C3-T4-L58	Tax Infringement	0	0	48	12/2007	0.0209	0.316
A195-C3-T4-L58	Tax Infringement	0	0	150	06/2016	0.0207	0.448
A188-C2-T4-L58	Reduction of penalties	0	0	34	10/2006	0.0190	0.293

A203-C3-T4-L58	Tax Infringement	1	1	106	10/2012	0.0180	0.575
A199-C3-T4-L58	Tax infringement	1	1	141	09/2015	0.0176	0.664

[a] A = Article-number, C = Chapter number, T = Title number, L58 = Law 58. Title 4 = The power to impose penalties, Chapter 2 = Tax infringements and penalties: general provisions, Chapter 3 = Tax infringements and penalties: types.

Aún cuando la comparación estructural de las redes observadas con los modelos de redes aleatorias puede tener un interés teórico, el objetivo principal de nuestro estudio es encontrar los nodos más relevantes (ver Tabla 4). En este sentido, es bastante significativo que los nodos "más importantes" en términos de iCauths en el LGTN, pertenezcan al mismo título (Título II), relativo a la cuantificación y pago de la deuda tributaria. Además, en la red de estructura y citación, los nodos más importantes también pertenecen al mismo Título, en este caso el IV. El análisis de los valores de centralidad revela la preeminencia de los Títulos relativos al pago de las deudas tributarias y las consecuencias de su infracción. Dichas consecuencias, en última instancia, también se traducen en sanciones económicas. Además, es muy interesante observar que los artículos más importantes apenas se han modificado. En nuestra tabla, sólo los tres últimos ejemplos han sido modificados una sola vez en el plazo de diez años. El contenido de estos artículos afecta directamente a los intereses de los contribuyentes y; cualquier modificación de estas normas afecta a los derechos fundamentales de las personas.

Tabla 18: Resultados modelos NIFPTML (LDA y ANN)

ML models	Obs. m_{ij}	Param.[a]	%	Predicted values 0	1	Param. [a]	%	0	1
LDA 2:2-1:1	0	Sp	70.2	25799	22	Sp	70.3	8605	4
	1	Sn	81.0	10941	94	Sn	89.7	3641	35
MLP 10:10-7-1:1 	0	Sp	89.4	32846	18	Sp	89.5	10955	1
	1	Sn	84.5	3894	98	Sn	97.4	1291	38
LNN 11:11-1:1 	0	Sp	73.4	26963	33	Sp	73.4	8988	11
	1	Sn	71.6	9777	83	Sn	71.8	3258	28

[a]Sn = Sensitivity and Sp = Specificity

4.5. Conclusiones

En este trabajo hemos demostrado cómo construir por primera vez una representación en red compleja de la Red de la Ley General Tributaria (LGTN). Hemos reportado el primer modelo NIFPTML para la estabilidad a lo largo del tiempo de una norma legal (lifetime legal norm dynamics). Específicamente, desarrollamos un NIFPTML lineal utilizando los valores de

centralidad de Hits-Authority para el LGTN. Esto abre una nueva puerta al estudio de las normas legales combinando los estudios teóricos clásicos con el Análisis de Redes Sociales y los métodos ML para tener una mejor comprensión de los sistemas legales o políticos en general.

Capítulo 5

CAPÍTULO 5. APLICACIÓN VERSIÓN FINAL

5. Artículo

Ref. IFPTML Studio a new software for Information Fusion, Perturbation Theory, Machine Learning Analysis in Molecular, Biomedical, and Social Sciences. Ortega-Tenezaca B, Duardo-Sánchez, A., Munteanu, C.R., and González-Díaz H,. J Chem Inf Model. 2022, *in preparation*.

IFPTML Studio, es un nuevo software para el desarrollo de modelos IFPTML. Contiene una interfaz de fácil uso, diseñado para sistemas operativos Windows. Fue creado en el entorno de desarrollo Microsoft Visual Studio 2019 Comunitario. Integra herramientas de terceros con licencia basada en software libre, disponibles mediante el administrador de paquetes Nuget, *véase* Tabla 19. IFPTML Studio se construye bajo el paradigma de programación orientada a objetos POO en el lenguaje C#.

Tabla 19: Paquetes instalados mediante el administrador de paquetes Nuget

Paquete	Versión	Descripción
Accord	3.8.2-alpha	Entorno central para el funcionamiento de Accord
Accord Controls		Proporciona controles gráficos que se pueden añadir a los formularios
Accord MachineLearning		Permite usar los principales algoritmos de Machine Learning
Accord Math		Contiene una librería de extensión para el manejo de matrices
Accord Statistic		Permite realizar cálculos estadísticos y probabilidad
ExcelData	3.7.0	Librería ligera para el manejo y lectura de archivos de Excel
FontAwsome.Sharp	5.15.4	Librería que permite embeber los iconos de Font Awesome
RibbonWinForms	5.0.11	Un paquete que ofrece controles Ribbon Office

IFPTML Studio es una solución de Visual Studio basada en dos Proyectos: FRAMA y PTML STUDIO bajo plataforma .NET Framework 4.7.2. FRAMA es una solución de tipo

"Biblioteca de clases", conformada por varias clases, que agrupan funcionalidades según su área de desarrollo, *véase* Tabla 20.

Tabla 20: Biblioteca de clases.

Clase	Descripción	Funciones principales
ClsAlgorithm	Clase estática que permite calcular probabilidades sigmoidales	Sigmoid
ClsConfusionTerms	Clase que brinda la estructura de los resultados obtenidos de cada técnica aplicada para la obtención de modelos tales como: TP, TN, FP, FN, sensibilidad, especificidad, precisión, Coeficiente de Matthews, Chi cuadrado	ClsConfusionTerms
ClsData	Clase con métodos estáticos que permiten realizar cálculos y manipular valores dentro de una DataTable. Permite cálculos como Discretización, PTOs, Generar columna con etiquetas para valores de entrenamiento y validación, *etc*.	AverageLevel CountLevel Discretice TrainValSetToTable AddPTMLNewTecnique ReplaceNullValuesSync
ClsMath	Clase con métodos estáticos que permiten calcular el valor de p de ChiCuadrado a partir de la Matrix de confusión	ChiSquarePval
ClsModel	Clase Serializable que representa una estructura para el algoritmo	Algorithm
ClsPTML	Clase con métodos estáticos que permiten obtener el modelo, matriz de confusión, afinación del modelo mediante la aplicación de probabilidades	GetModel ConfusionMatrix

Las clases y funcionalidades que realizan cálculos tienen como base la clase tabla de datos DataTable, su representación es similar a la de una hoja de cálculo. Un DataTable se almacena en memoria, los datos que contiene, son accedidos por su ubicación dentro de una fila y columna. El DataTable no tiene un esquema predeterminado. El dataTable principal sobre el que funcióna IFPTML se denomina DTEXCEL, que es accesible de manera global.

IFPTML STUDIO es un proyecto de tipo “Aplicación Windows” desarrollado con una interfaz similar a Microsoft Office *véase* Figura 16. Se encuentra dividido en cuatro módulos funcionales, y un módulo informativo.

Tabla 21: Descripción de los módulos del sistema

Módulo	Descripción
Source File	Permite cargar un archivo de Excel con la información base y convertirlo a DataTable
IFPreprocess	Permite enriquecer información a partir de dos archivos. Contiene funcionalidades que permiten tratar los datos anómalos, y la fusión de información, que permite generar más datos a partir de la concatenación del contenido de dos o más columnas
PT Processing	Permite realizar la discretización multifuncional y el cálculo de Perturbadores PTO´s
ML Analisys	Permite generar modelos IFPTML o PTML y el cálculo de nuevos casos
Support	Es un módulo informativo acerca del desarrollo de IFPTML Studio

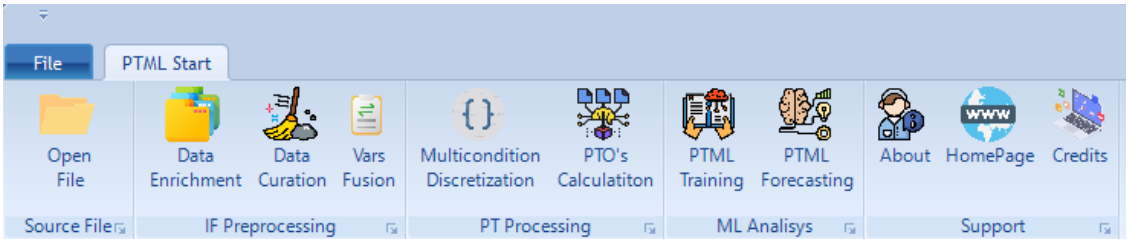


Figura 16: PTML Studio - Menú

5.1. Estructura de IFPTML Studio

En una visión general, IFPTML Studio permite la carga de archivos de Excel que provee información estructurada y organizada en variables categóricas y continuas. En la figura 17, se puede observar el diagrama de flujo de PTML Studio desde la carga de información hasta la construcción del modelo PTML.

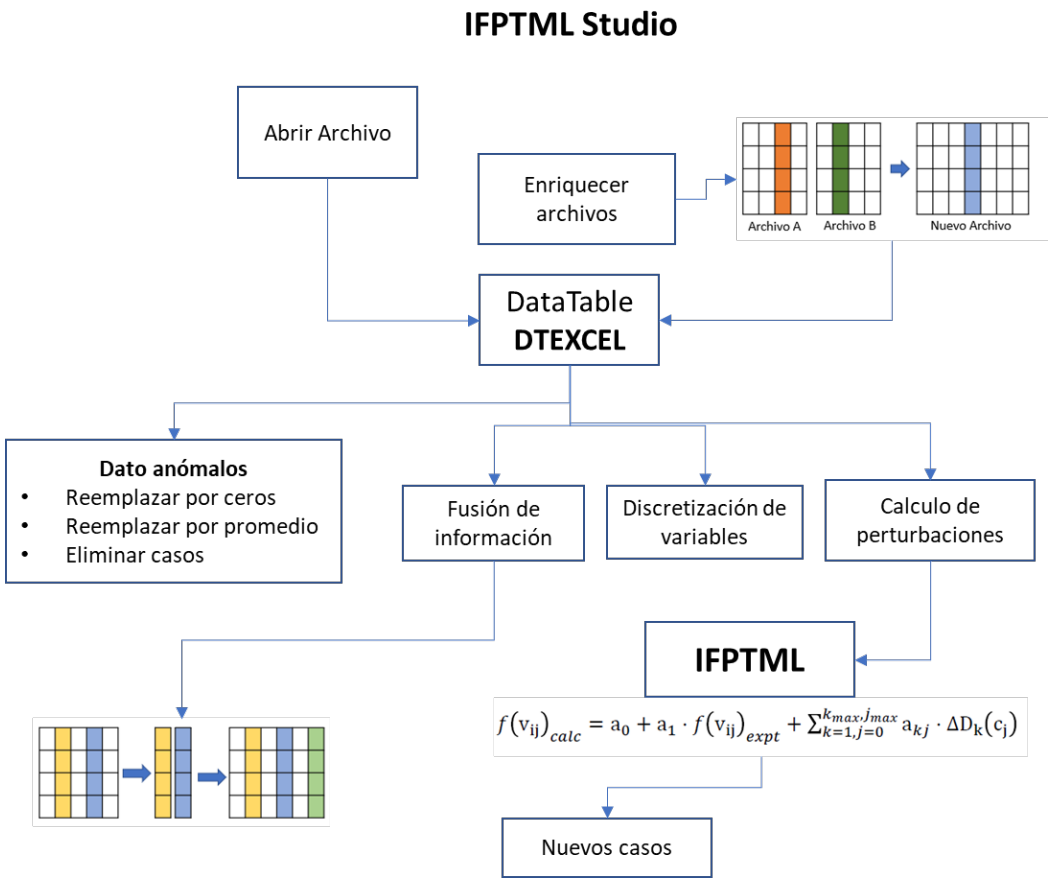


Figura 17: IFPTML Studio - diagrama de flujo general

5.2. Archivo de origen

5.2.1. Carga de Archivos

IFPTML Studio, muestra la opción abrir archivo, que permite al usuario seleccionar un archivo de Excel en formato XLS o XLSX. Se carga automáticamente la primera hoja del archivo seleccionado. El tiempo de carga es proporcional a la cantidad de información del fichero que se ha subido al sistema. Para evitar el bloqueo de la interfaz de IFPTML Studio durante el proceso de carga del archivo, se utiliza una función de tipo asíncrona y un indicador visual durante el proceso. Para la carga de información se invoca a la función *GetDataTableAsync(filename)*que pertenece a la clase ClsData del proyecto FRAMA. Ha sido

escrita como una tarea que devuelve un DataTable. En este contexto se utiliza programación asincrónica basada en Tareas. El nuevo DataTable es el elemento principal del sistema debido a que mantiene inicialmente una copia directa del archivo en memoria, conservando sus nombres de columnas y la totalidad de la información. Cada fila del DataTable se considera como un caso.

```
public static async Task<DataTable>GetDataTableAsync(string filename)
{
    try
    {
        return await Task.Run(() =>
            {
                using (var stream = File.Open(filename, FileMode.Open,
                       FileAccess.Read))
                {
                    using (var reader =
                              ExcelReaderFactory.CreateReader(stream))
                    {
                        var result = reader.AsDataSet(new ExcelDataSetConfiguration
                        {
                              ConfigureDataTable = (tableReader) => new
                                          ExcelDataTableConfiguration()
                            {
                                UseHeaderRow = true,
                            }
                        });
                        return result.Tables[0];
                    }
                }
            });
    }
    catch (Exception ex)
    {
        throw ex;
    }
}
```

Figura 18: Función GetDataTableAsync

El equipo informático genérico donde se realizan las pruebas se conforma de un procesador Core I7 de 7ma generación con memoria de 16GB de RAM. Se realiza la prueba de carga de un archivo con 794 mil casos y 72 columnas, con un tamaño de 587.197 MB. Durante la carga de archivos superiores a 200MB de información se obtienen tiempos modestos de carga, sin llegar a usar ni el CPU ni la memoria de forma crítica.

5.3. IF Preprocessing

El grupo de funcionalidades IF Preprocessing del software IFPTML Studio, presenta opciones fundamentales que dan el nombre a este tipo de modelos.

5.4. Data Enrichment

Se convierte en la función principal de los modelos IFPTML puesto que su utilización permite enriquecer la información cargada inicialmente, a partir una hoja adicional de cálculo, con información estructurada. Para subir la información de archivos, se reutiliza el procedimiento descrito en la carga de archivos y la función descrita en la Figura 19. Una vez cargada la información, se obtendrá la información básica del número de casos y columnas que conforma cada archivo. El usuario deberá definir la cantidad de casos con los que desea tratar, tomando como referencia el archivo con mayor número de casos con lo cual podrá seleccionar un número mayor de casos, un número menor de casos o dejarlos sin modificación. Para la selección de un mayor número de casos al que se ha cargado, IFPTML Studio completa los casos a partir de la información que ya posee. En el caso de seleccionar un número menor de casos, IFPTML descartará los casos necesarios para cumplir con el requerimiento.

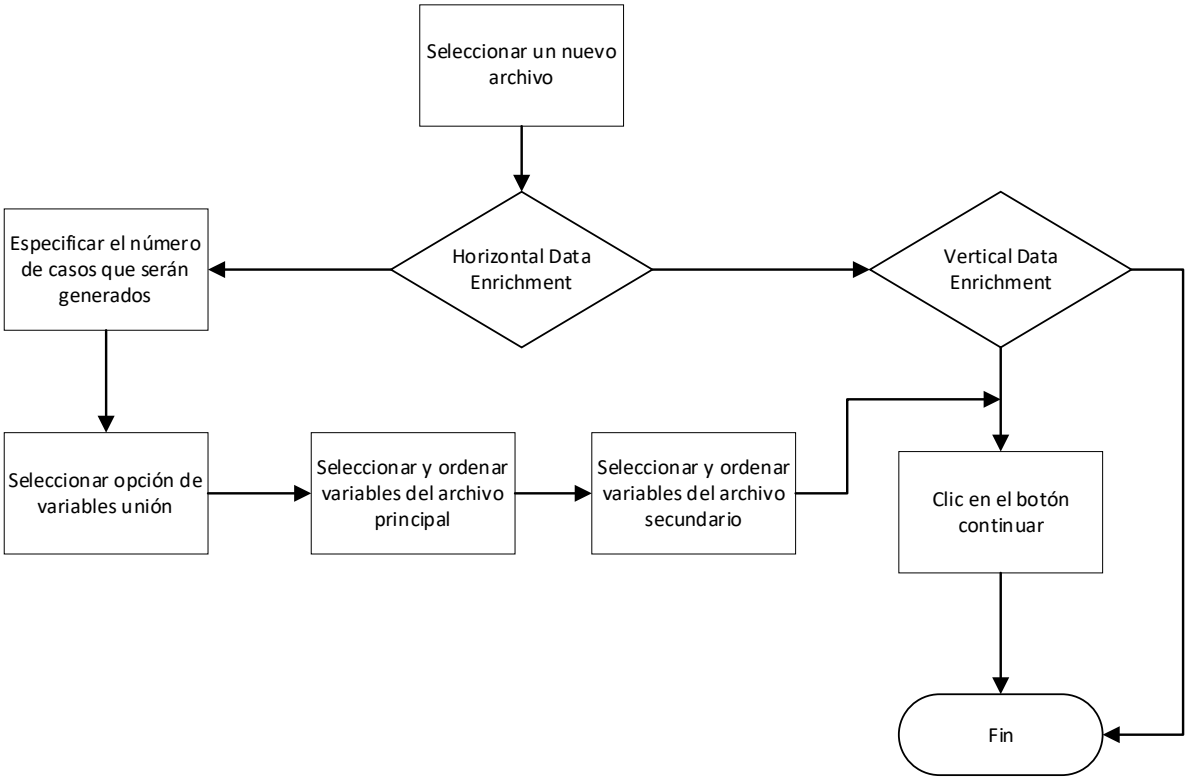


Figura 19: Flujo de trabajo de enriquecimiento de archivos

A continuación, se presenta una ventana con las variables de cada uno de los archivos subidos y la funcionalidad necesaria para seleccionar los elementos que pueden ser comunes en ambos archivos y sobre los cuales se buscará las coincidencias como criterios de unión de información. Si se selecciona más de una variable, se procesa la información bajo un procedimiento denominado fusión de variables que genera una columna de información adicional en cada archivo, en la que se combina la información de las columnas seleccionadas en el orden establecido.

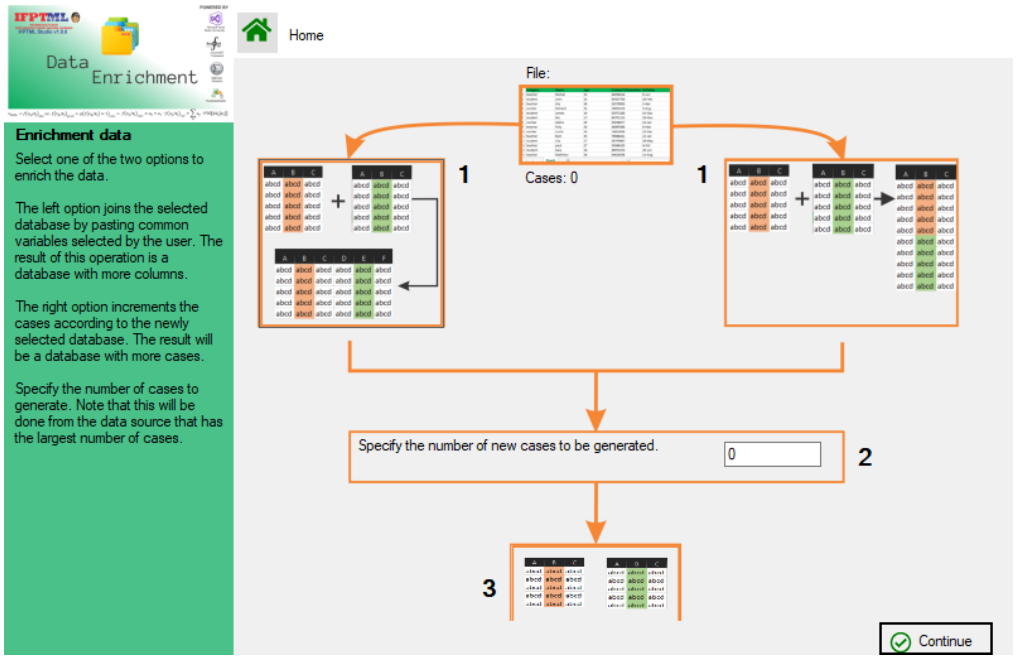


Figura 20: Fusión de información

El resultado de este proceso se refleja en el DataTable principal denominado DTEXCEL, que se almacena en memoria como fuente de datos principal, figura 21.

```
public static async Task DtJoinData(string colGluePrincipal, string colGlueSecondary)
{
    await Task.Run(() => {
        Random random = new Random();
        DataTable tempData = new DataTable();

        List<string> colsDTPrimary = DtGetDistintcValues(Global.DTEXCEL,
                                     colGluePrincipal);
        string[] colNamesDtSecondary = DtGetColsNames(Global.DTSecondary);

        foreach (string item in colsDTPrimary)
        {
            tempData.Clear();
            tempData = DtFilteredDataTable(Global.DTSecondary, colGlueSecondary,item);
            if (tempData.Rows.Count > 0)
            {
                foreach (DataRow dr in Global.DTEXCEL.Select(string.Format("{0}={1}",
                         colGluePrincipal, "'" + item + "'")))
                {
                    int indexRow = random.Next(1, tempData.Rows.Count);
                    foreach (string colName in colNamesDtSecondary)
                    {
                        dr[colName] = tempData.Rows[indexRow][colName];
                    }
                }
            }
        }
    });
}
```

Figura 21: Código de la fuente principal de datos

5.5. Data Curation

Una vez cargada la información mediante las funciones de carga simple de archivo o de enriquecimiento de información, se requiere revisar en primera instancia que los datos sean válidos para su tratamiento.

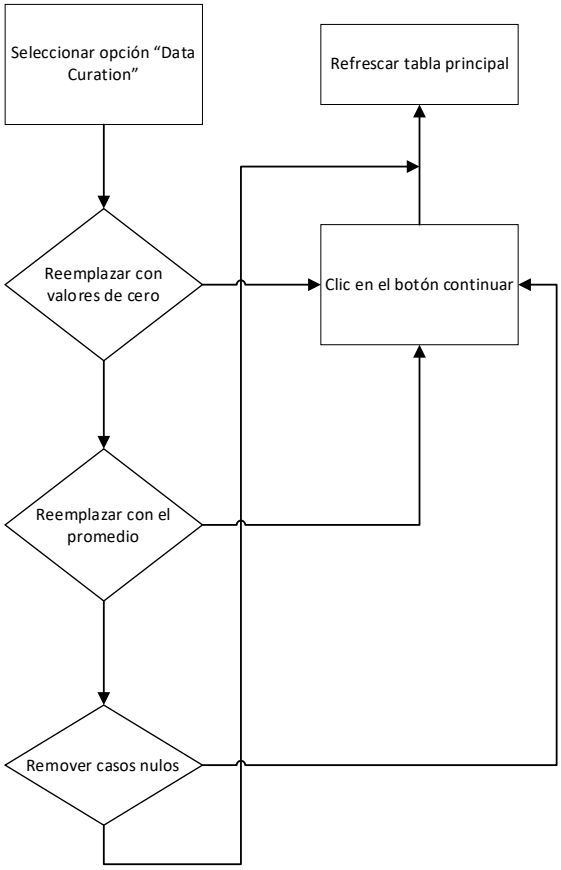


Figura 22: Flujo de información para curación de datos

Data Curation ofrece un resumen de la fuente de datos principal DTEXCEL, para lo cual muestra el tipo de variable, y la cantidad de valores nulos de cada variable. Es posible visualizar la distribución de cada variable númerica y las opciones para el tratamiento de valores nulos. Los

valores nulos de las columnas numéricas pueden ser descartados, rellenados con ceros, o rellenados con el valor promedio de su grupo de datos. Lo valores nulos de las variables categóricas se rellenan con el valor constante “Not Available”(NA)

```
public static async Task ReplaceNullValuesSync(bool average = false)
{
    DataTable temp = GetVarsNull();
    await Task.Run(() =>
    {
        foreach (DataRow dr in temp.Rows)
        {
            if ((int)dr[2] > 0)
            {
                if (dr[0].ToString() == "Double")
                {
                    double replaceValue = average ?
                           Global.DTEXCEL.AsEnumerable().Where(x =>
                                x[dr[1].ToString()] !=
                                       DBNull.Value).Average(x =>
                                x.Field<double>(dr[1].ToString())):0;
                    if (average)
                    {
                        replaceValue =
                            Global.DTEXCEL.AsEnumerable().Where(x =>
                              x[dr[1].ToString()]!=DBNull.Value).Average(x
                                 => x.Field<double>(dr[1].ToString()));
                    }
                }
                else
                {
                    Global.DTEXCEL.Rows.Cast<DataRow>().Where(x =>
                        x[dr[1].ToString()] ==
                                DBNull.Value).ToList().ForEach(x =>
                                       x.SetField(dr[1].ToString(), "ND"));
                }
            }
        }
    });
}
```

Figura 23: Codificación de reemplazo de valores nulos

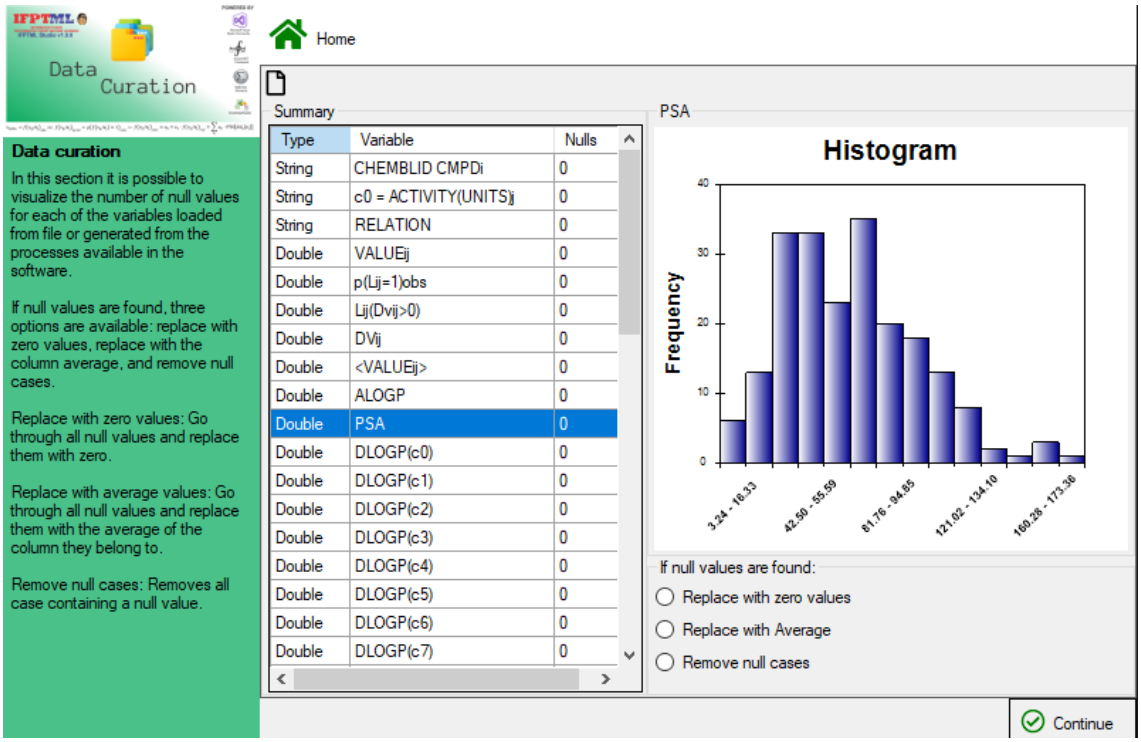


Figura 24: Curation Data (Cambiar D por Δ, LOGP por D_1 y PSA por D_2)

5.6. Vars Fusion

Como parte de la teoría de IFPTML, se puede realizar el proceso de fusionamiento de información, tal como se utiliza en el procedimiento de enriquecimiento de información, sin embargo, esta opción se realiza únicamente en las columnas de la fuente de datos primaria DTEXCEL. Para fusionar la información contenida en las variables, es necesario seleccionar dos o más de ellas, que se encuentran disponibles y establecer el orden en el que deberán ser tratadas. Una vez que se añadan a un listado de variables en espera, se procederá con su ejecución. El resultado de las operaciones de fusión, serán integradas a la fuente de datos principal.

```
public static async Task CreateColFusionInformation(DataTable table, List<string> ListColsToJoin, bool generateName = false, string nameNewcol = "TempColFusionInfo2022")
{
   await Task.Run(() =>
   {
      try
      {
         if (ListColsToJoin is null)
         {
            return;
         }

         if (generateName)
         {
            string nameCol = String.Join(",", ListColsToJoin.ToArray());
            nameNewcol = string.Concat("VarFus(", nameCol, ")");
         }

         if (!table.Columns.Contains(nameNewcol))
         {
            table.Columns.Add(nameNewcol, typeof(string));
         }

         int row = 1;
         for (int i = 0; i < table.Rows.Count; i++)
         {
            string contenido = string.Empty;
            foreach (string item in ListColsToJoin)
            {
               contenido += table.Rows[i][item] + "-";
            }
            contenido = contenido.Remove(contenido.Length - 1);
            table.Rows[i][nameNewcol] = contenido;
         }
         row++;
      }
      catch (Exception ex)
      {
         throw ex;
      }
   });
}
```

Figura 25: Codificación de Fusión de información

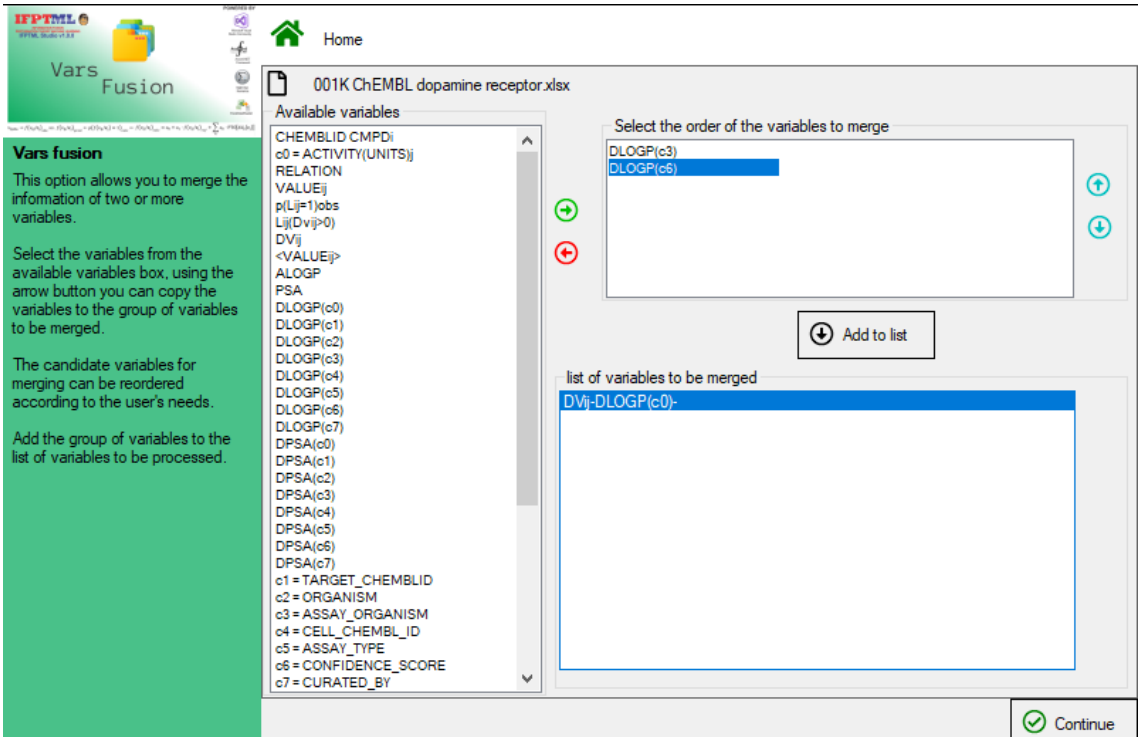


Figura 26: Fusión de información

5.7. Multiconditional Dicretization

Esta opción permite operar una de las variables categóricas sobre una de las numéricas, de donde obtiene los siguientes valores: niveles, n_j, $<v_{ij}>$, d_{ij}, cutOff, n_{ij}. Los valores de las variables d_{ij}, y el cutOff, se pueden modificar y recalcular tantas veces como lo decida el experto. Finalmente, estas variables forman parte de la fuente de datos principal DTEXCEL.

```
public static async Task<DataTable> DiscretizeAsync(string colCategorical, string colNumeric)
{
    try
    {
        return await Task.Run(() =>
        {
            DataTable dt = new DataTable();
            ...........
            var uniData = GetUniqueValues(Global.DTEXCEL, colCategorical);
            var conData = CountLevel(Global.DTEXCEL, colCategorical);
            var avgData = AverageLevel(Global.DTEXCEL, colCategorical, colNumeric);
            ...........
            for (int i = 0; i < uniData.Rows.Count; i++)
            {
                string temp = dt.Rows[i]["Levels"].ToString().Replace("'", "\'");
                string seek = string.Format("[variable] = '{0}'", temp);
```

```
                DataView viewCount = new DataView(conData) RowFilter = seek
                DataView viewAvergare = new DataView(avgData) RowFilter = seek
                if (viewCount.Count > 0) dt.Rows[i][Global.numberJ] = viewCount[0][1];
                if (viewAvergare.Count > 0)
                {
                    dt.Rows[i][avgIJ] = viewAvergare[0][1];
                    dt.Rows[i][cuttoff] = viewAvergare[0][1];
                }
                if (temp.Contains("%") || temp.Contains("(uM)")) dt.Rows[i][desirability] = 1;
            }
            DataColumn colObs = new DataColumn
            {
                DataType = System.Type.GetType("System.Double"),
                ColumnName = observed,
                DefaultValue = 0,
                Expression = "IIF((([" + cuttoff + "]>[" + avgIJ + "] and [" + desirability + "]=1)or([" +
                                         cuttoff + "]>=[" + avgIJ + "] and [" + desirability + "]=-1),1,0)"
            };
            DataColumn colExpt = new DataColumn
            ..................
            return dt;
        });
    }
    catch (Exception ex)
    {
        throw ex;
    }
}
```

Figura 27: Codificación de Discretización

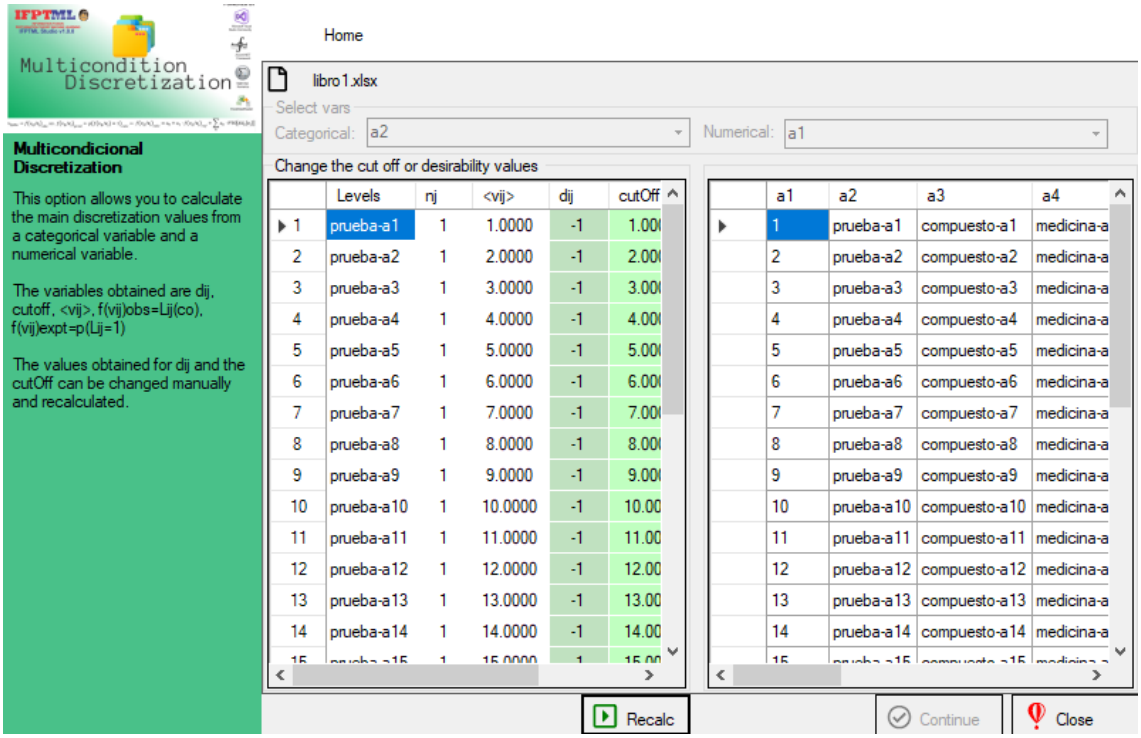


Figura 28: Discretización multicondicional

5.8. PTO´s Calculation

El cálculo de PTO´s se realiza a partir de la selección de una o varias variables categoricas, sobre una o varias variables numericas. Se puede seleccionar entre los operadores de: Moving Average, Sum for Partition, Average(Mean), Standard Deviation, Min-Max probability, Z-Score. Esta selección permite agregar operaciones a un listado de ejecución diferida. Su resultado será incorporado a la fuente de datos principal para su posterior procesamiento.

```
private static void OperatorsMAvg(string[] strToCalc)
{
    string nameCol = GetColNameFormOperPTOStr(strToCalc);

    var avgData = AverageLevel(Global.DTEXCEL, strToCalc[1], strToCalc[2]);

    ......
    checkIfDataNull(avgData.Rows.Count);

    for (int i = 0; i < avgData.Rows.Count; i++)
    {
        Global.DTDISCRETICE.Rows[i][strToCalc[1]] = avgData.Rows[i][0];
        Global.DTDISCRETICE.Rows[i][nameCol] = avgData.Rows[i][1];
    }

    if (!ExistColumnInDatabase(Global.DTEXCEL, nameCol))
    {
        Global.DTEXCEL.Columns.Add(nameCol);
    }
    for (int i = 0; i < avgData.Rows.Count; i++)
    {
        string temp = avgData.Rows[i][0].ToString().Replace("'", "\''");
        string seek = string.Format("[" + strToCalc[1] + "] = '{0}'", temp);

        DataView viewFilter = new DataView(Global.DTEXCEL)
        {
            RowFilter = seek
        };

        if (viewFilter.Count > 0)
        {
            for (int j = 0; j < viewFilter.Count; j++)
            {
                viewFilter[j][nameCol]=(double)viewFilter[j][strToCalc[2]] -
                                        (double)avgData.Rows[i][1];
            }
        }
    }
}
```

Figura 29: Codificación de funcion de cálculo de PTO`s

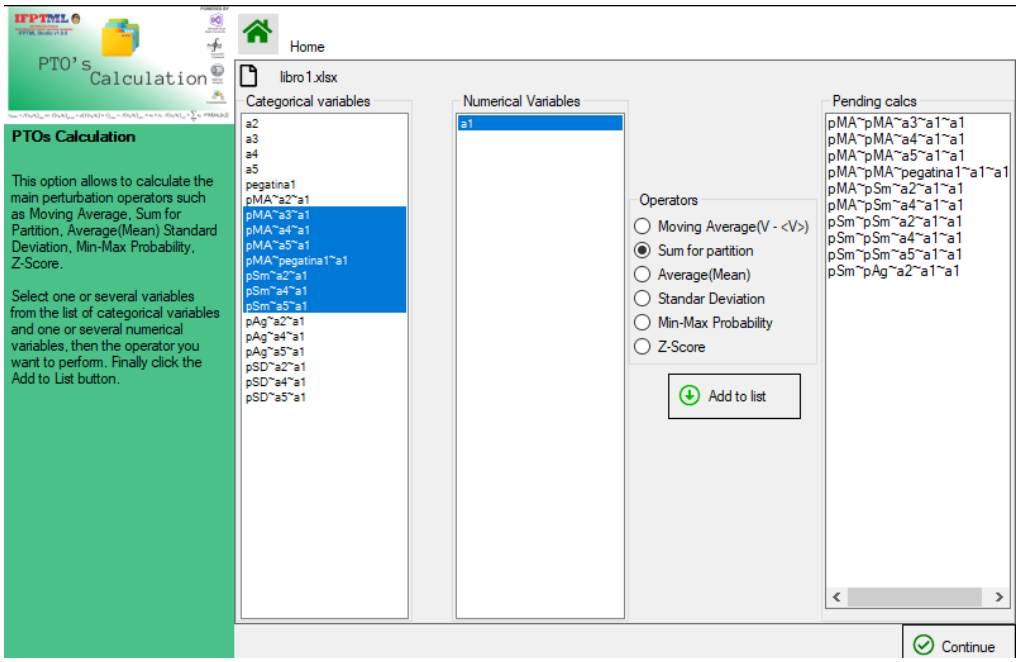


Figura 30: Pantalla de selección de perturbadores

5.9. IFPTML Training

Esta es una de las operaciones principales que se realizan dentro del software, el modelo debe ser cubierto con todos los requerimientos para poder generar un resultado. Se muestra un listado de variables que pueden ser seleccionadas para formar las partes correspondientes del modelo: Variable observada, variable de referencia, Operadores PT, variables de Validación y entrenamiento, selección del algoritmo. Si la fuente de datos no posee variables de entrenamiento y validación es posible crearlas de forma aleatoria a partir de un patrón, porcentaje y necesidad del usuario. Se puede seleccionar y configurar los algoritmos disponibles: Logistic Regression, Linear Discriminant Analysis (LDA), Kernel Discriminant Analysis (KDA), Random Forest (RF), Support Vector Machine (SVM).

```
private async Task<string> ML_LinearDiscriminantAnalysisAsync(clsLDA algorithm)
{
    string equation = "";
    await Task.Run(() => {
        LinearDiscriminantAnalysis _lda = new LinearDiscriminantAnalysis//();
        {
            Threshold = algorithm.Threshold,
            NumberOfOutputs = 2,
        };

        lda = _lda.Learn(inTrainning, outTraining);
        Serializer.Save(lda, getFilenameToSave("_LDA", out string getfile));

        double intercepto = _lda.Eigenvalues[0];
        double[][] weights = _lda.Classifier.First.Weights;
        double[] coef = new double[weights.Length];
        for (int i = 0; i < weights.Length; i++)
        {
            coef[i] = weights[i][0];
        }

        Global.scoreTLda = getScores(cols, intercepto, coef, inTrainning, signalEq);
        Global.scoreVLda = getScores(cols, intercepto, coef, inValidation, signalEq);
        double avg = (Global.scoreTLda.Sum() + Global.scoreVLda.Sum()) /
                     (Global.scoreTLda.Length + Global.scoreVLda.Length);

        Global.scoreTLda = Global.scoreTLda.Select(x => x - avg).ToArray();
        Global.scoreVLda = Global.scoreVLda.Select(x => x - avg).ToArray();

        Global.predTLDA = lda.Decide(inTrainning);
        Global.predVLDA = lda.Decide(inValidation);

        equation = ClsData.getModel(cols, intercepto - avg, coef, signalEq);

        PrintConfusionMatrix(ref PTML, "LDA", Global.predTLDA, Global.predVLDA,
            Global.scoreTLda, outTraining, Global.scoreVLda, outValidation, FileNameToSave:
            getfile);

        sb.AppendLine("LDA       : " + getFilenameToSave("LDA"));
        sb.AppendLine("model: " + equation);
        sb.AppendLine();
});
    return equation;
}
```

Figura 31: Codificación de función de algoritmo LDA

Se agregarán todas las variables que conforman el modelo, para lo cual se contará con el listado de variables disponibles, y la opción se seleccionar o crear el conjunto de valores de entrenamiento y validación

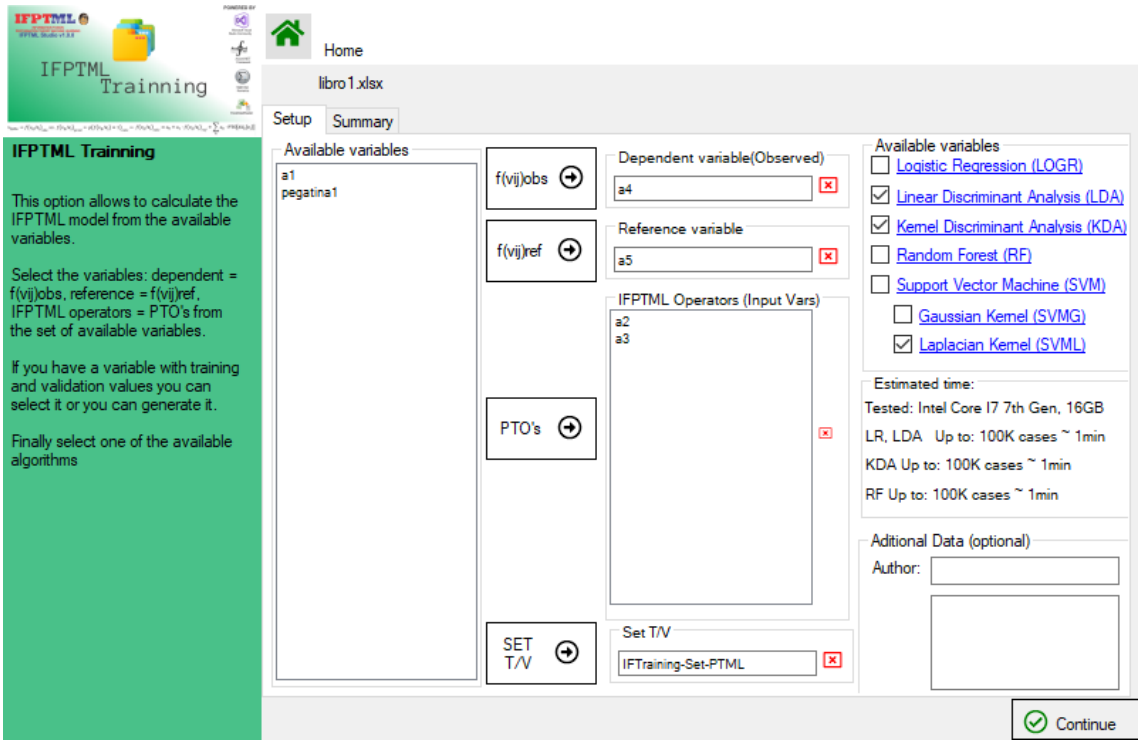


Figura 32: Configuración de aplicación de la teoría IFPTML

Capítulo 6

CAPITULO 6. ASPECTOS JURÍDICOS DEL USO DE PROGRAMAS DE ORDENADOR EN INVESTIGACIÓN CIENTÍFICA

6. Articulo

Ref. Legal issues regarding software uses in scientific research. B. Ortega-Tenezaca, C.R. Munteanu, A. Duardo. Journal of World Intellectual Property, 2022, Submitted

6.1. Relaciones entre las Ciencias de la Información y la Comunicación y Derecho

La relación entre las Ciencias de la Información y la Comunicación, y el mundo de lo jurídico abarca un sinnúmero de cuestiones que condicionan tanto el desarrollo de herramientas informáticas como la investigación científica en este campo. Estas cuestiones van desde el reconocimiento de la propiedad intelectual sobre los programas de ordenador, los contratos y licencias para su explotación comercial, hasta la regulación del desarrollo de productos y los derechos de empresa, por sólo citar algunos ejemplos. Asimismo, con motivo de la influencia del Derecho del Unión Europea sobre los ordenamientos de los distintos Estados miembros, se imponen cada vez más obligaciones a los investigadores en esta región del mundo. Tales obligaciones están relacionadas con la protección de los datos personales usados en investigación, la toma automatizada de decisiones, la seguridad, confidencialidad e integridad de los sistemas inteligentes, etc. ¿Significa, esta creciente interacción, que los científicos, independientemente del área en la que se desenvuelvan, han de ser también expertos en leyes? Por supuesto que no; pero conviene tener presente que el desarrollo de cualquier investigación científica tiene una serie de consecuencias jurídicas importantes que la condicionan; llegando, incluso, a prohibirse determinadas prácticas, que, si bien podrían conducir a un avance científico-

tecnológico importante, supondrían una quiebra para los derechos y libertades tanto individuales como colectivos. Es por ello que dedicamos este capítulo a analizar algunas cuestiones jurídicas del uso de los programas de ordenador en investigación científica. Específicamente, los derechos de propiedad intelectual relacionados con la creación del software, así como la validación y protección de los resultados científicos obtenidos mediante la utilización del mismo. Somos conscientes de que este acápite supone toda una novedad para un informe de investigación de esta naturaleza. Esperamos que contribuya de manera positiva a futuros desarrollos, acordes con las exigencias de índole jurídica que rodean a la investigación científica en relación con la creación del software.

6.2. La protección jurídica de los programas de ordenador.

La primera cuestión a tener en cuenta, cuando se desarrolla un programa informático, es que no existe un único método, desde el punto de vista jurídico, para proteger esta clase de creaciones. De hecho "no existe un régimen internacional único para hacer frente a la protección del programa de ordenador" (Duardo *et al.,* 2015). Si bien lo más común es buscar la protección jurídica por medio de las leyes de derechos de autor/Copyright en distintos territorios; "el alcance y la viabilidad de la aplicación de este tipo de protección varía significativamente de un país a otro" (Duardo *et al.,* 2015).

6.3. Derechos de Propiedad Intelectual. ¿Derechos de autor vs Copyright?

Dado que los programas de ordenador se consideran creaciones de la mente humana[1], ha sido la Propiedad Intelectual la rama del Derecho que se ha ocupado de su protección jurídica. Principalmente, este amparo se ha garantizado por medio de la concesión de derechos de autor o Copyright, en dependencia del territorio donde se quieran hacer valer los mismos. De este modo, nos encontramos con dos modelos, basados en tradiciones jurídicas diferentes: modelo de `Copyright'[2] y el modelo Derechos de autor, que pueden otorgar protección sobre mismo bien en áreas geográficas diferentes.

El primer modelo, propio de la tradición jurídica anglosajona o *Common Law*, utilizado tanto en Reino Unido como en EE. UU; se caracteriza por basarse en un enfoque utilitarista en función del mercado. Así, el modelo de Copyright se asienta en la necesidad práctica de mantener incentivos a particulares y empresas para llevar a cabo nuevas inversiones.

Por su parte, el llamado modelo continental europeo de Derechos de Autor (del francés: *droitd´auteur*), es el propio de la mayoría de los miembros continentales de la Unión Europea y predomina en el resto del mundo. Si bien, existen matices en cuanto al contenido y la duración de los derechos concedidos, en el modelo de derechos de autor, se parte de la base de considerar que el individuo se reserva el derecho de controlar, con respecto a sus creaciones, tanto los derechos de explotación económica –reproducción, distribución etc.-, como los "derechos

[1] En los últimos tiempos, a medida que los sistemas inteligentes han ido ganando en autonomía, se ha abierto el debate acerca la propiedad intelectual sobre las obras "creadas" o los inventos generados por los mismos, en aquellos supuestos donde el resultado final se ha logrado sin intervención humana. Así ha habido varios intentos de reclamar derechos sobre obras o invenciones que no son de autoría humana en el sentido convencional. A pesar de ello, la mayoría de sistemas jurídicos siguen considerando que lo que se protege mediante los derechos de propiedad intelectual son creaciones o invenciones propias del intelecto humano.

[2] "Coloquialmente, este último término se usa para referirse a los dos regímenes por igual. Sin embargo, vale la pena señalar que existe una base filosófica diferente en atención a estos dos distintos regímenes o sistemas internacionales". (Duardo *et al*, p. 101)

morales" inalienables, que se conceden a quien se considere el autor del programa. Esto último, incluye aspectos como: reclamar la paternidad -autoría-, decidir si procede o no su publicación, y mantener la integridad de la obra, es decir autorizar o no la realización de versiones de la misma.

En el sistema de Copyright el registro es ***constitutivo de derechos***, mientras que en el sistema de derechos de autor es ***declarativo***. ¿Qué significa esto? Significa que mientras en el primer modelo es indispensable realizar el registro para reclamar los derechos, en el segundo modelo éstos nacen desde el momento mismo en que se desarrolla el software, sin necesidad de inscripción previa. Sin embargo, las diferencias entre los dos sistemas, no deben sobredimensionarse. En la práctica, también los autores de programas en el sistema continental se apresuran a registrar sus obras para obtener una protección plena y evitar futuras querellas sobre la autoría de las mismas.

En los países firmantes de tratados internacionales como el Convenio de Berna, La Convención Universal sobre los Derechos de Autor y ciertas previsiones del Acuerdo de los ADPICs (Acuerdo de la OMC sobre los Aspectos de los Derechos de Propiedad Intelectual Relacionados con el Comercio), la protección se otorga mediante la concesión de Derechos de Autor, y como se avanzó, está libre de formalidades. Es decir, ésta no se hace depender del cumplimiento de determinados trámites como el registro o el depósito de copias.

El Convenio de Berna[3], que ha sido modelo para muchas legislaciones nacionales, equipara los programas de ordenador a las obras literarias. Esta ficción jurídica, permitió en su momento adaptar el derecho existente y otorgar el mimo tratamiento a ambos tipos de creaciones. No

[3] Así, el Tratado OMPI de 1996 sobre Derecho de autor, señala en su art,4 que «los programas de ordenador están protegidos como obras literarias en el marco de lo dispuesto en el artículo 2 del Convenio de Berna». Asimismo, el art. 10.1 del Acuerdo sobre los Aspectos de los Derechos de Propiedad Intelectual relacionados con el Comercio (Ronda Uruguay-1994), establece que «los programas de ordenador, sean programas fuente o programas objeto, serán protegidos como obras literarias en virtud del Convenio de Berna (1971)».

obstante, la originalidad, requisito indispensable para conceder los derechos de autor a una obra, no puede ser medida de la misma forma en unas y otras.

Los derechos de autor, no conceden exclusividad sobre las ideas o la función, sólo sobre su expresión. Por ejemplo, dos procesadores de textos que realizan la misma función pueden ser protegidos por esta vía siempre que la disposición de comandos en el código fuente sea esencialmente distinta. Así, para constatar la originalidad del software lo decisivo es que resulte ser «una creación intelectual propia de su autor» (STJUE de 23.01.2014 (Asunto C-355/12) y STJUE de 16.07.2009, (AsuntoC-5/08)). La originalidad de los programas de ordenador, ha dicho Erdozain López, "no se mide por parámetros estéticos, sino por el hecho de que el programa haya sido creado por el autor (sea una obra fruto de su «trabajo intelectual» que es como hay que entender la expresión «propia de su autor») y, además, no constituya una copia de otros programas ya creados. Lo que se protege es la expresión concreta que adopta el software (la «secuencia de instrucciones o indicaciones») y no la función estricta que se logra con la misma. (Erdozain López, 2019)"

6.3.1. El marco de protección jurídica en España

El legislador español, como territorio parte de la tradición de Derechos de Autor, ha optado por considerar los programas de ordenador como una forma de creación artística. De este modo, y siguiendo dicha tradición, los programas de ordenador quedan bajo el amparo de los derechos de Propiedad Intelectual y son equiparados a las obras literarias (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia (LPI)). Esta visión del software como "creación intelectual" también puede extraerse de la

Directiva 91/250/CEE, derogada y sustituida por la Directiva 2009/24/CE, del Parlamento Europeo y del Consejo, de 23 de abril; y como se avanzó anteriormente, del Tratado OMPI de 1996 sobre Derecho de autor (art. 4); y del Acuerdo ADPIC (art.10.1), ambos ratificados por España. En contraposición, y de acuerdo con la Ley 24/2015, de 24 de julio, de Patentes, los programas de ordenador no son patentables en territorio español (art 4.4c). No obstante, esta afirmación debería ser matizada, añadiendo que *los programas de ordenador «por sí solos» no son patentables*; pues como aclara la LPI «cuando los programas de ordenador formen parte de una patente o modelo de utilidad gozarán, sin perjuicio de lo dispuesto en la presente Ley, de la protección que pudiera corresponderles por aplicación del régimen jurídico de la propiedad industrial» (art. 96.3).

Desde el punto de vista jurídico, un programa de ordenador es «toda secuencia de instrucciones destinadas a ser utilizadas en un sistema informático para realizar una función o para obtener un resultado determinado». Ello comprende, además del *código fuente*, la documentación preparatoria, la documentación técnica y los manuales de uso (Art. 96.1 LPI). Esto último es relevante, porque la protección es mucho más amplia de lo que se puede pensar, extendiéndose a *bocetos o borradores -esquemas y diagramas relativos a la estructura y a las funciones que desarrolla el programa en cuestión- incluidos en la documentación preparatoria de la solicitud, y los manuales de uso*; elementos estos que podrían facilitar la ingeniería inversa del software. La protección, abarca también las *versiones sucesivas* y otros *programas derivados* del mismo.

De acuerdo con el art. 96.2 LPI, sólo puede otorgase protección a aquellos programas de ordenador que cumplen con el requisito de la «originalidad». Ésta, como se ha dicho, ha de apreciarse, no porque la función que el programa desempeñe sea única o novedosa, sino porque

la «forma de expresión» del mismo es propia del autor (cfr. art. 96.3 LPI). Dicho de otro modo, las palabras clave, sintaxis, comandos o combinaciones de comandos, opciones y valores por defecto o iteraciones compuestas por palabras, cifras o conceptos matemáticos, considerados aisladamente no constituyen una obra susceptible de protección. No obstante, éstos pueden ser «originales» y tener valor creativo «como consecuencia de su especial disposición, elección o combinación» (TJUE de 23.01.2014 (Asunto C-355/12)) y STJUE de 16.07.2009, (AsuntoC-5/08)). Así, si esa disposición o combinación original es objeto de reproducción en otro programa, sin la autorización del titular, estaríamos en presencia de una infracción de los derechos de autor.

Quedan excluidos de la protección, en todo caso, los virus informáticos, los principios, y las ideas en los que se basan cualquiera de los elementos de un programa de ordenador, incluidos los que sirven de fundamento a sus interfaces (cfr. art. 96.4 LPI).

6.5. La titularidad de los programas de ordenador. Anotaciones en el concreto ámbito de la investigación científica.

De acuerdo con el artículo 97 LPI «será considerado autor del programa de ordenador *la persona* o *grupo de personas naturales* que lo hayan creado, o *la persona jurídica* que sea contemplada como titular de los derechos de autor en los casos expresamente previstos por esta Ley».

De este modo, la titularidad puede ser atribuida tanto a *personas físicas* «naturales», como a *personas jurídicas* -empresas -. En cuanto al primer caso contemplado en la ley, -programas creados por una o varias personas físicas- pueden darse a su vez dos supuestos: que éste haya

sido creado en régimen de «obra colectiva» o de «obra en colaboración». Adaptando una vez más, las reglas pensadas para las obras literarias y otras creaciones artísticas. En este sentido, «cuando se trate de una obra colectiva tendrá la consideración de autor, salvo pacto en contrario, la persona natural (…) que la edite y divulgue bajo su nombre» (art. 97.2). Por su parte, «los derechos de autor sobre un programa de ordenador que sea resultado unitario de la colaboración entre varios autores serán propiedad común y corresponderá a todos éstos en la proporción que determinen» (art. 97.3).

En cuanto al hecho de considerar a una persona jurídica como «autora» de un programa de ordenador (art. 97.2 LPI), se trata, una vez más, de una ficción jurídica que intenta adaptar a norma a las exigencias del mercado. En la práctica resulta muy habitual, que las empresas aporten los medios técnicos y humanos necesarios, de modo que, para incentivar su participación e inversión en este tipo de «creaciones», resulta lógico que las que puedan beneficiarse de los derechos, sobre todo de explotación económica, sobre los programas que han financiado. También puede darse el supuesto de cotitularidad entre varias personas jurídicas.

Finalmente, puede darse el caso, contemplado en el art. 97.4 LPI, de que el software haya sido creado en virtud de una relación laboral. En este supuesto particular, «cuando un trabajador asalariado cree un programa de ordenador, *en el ejercicio de las funciones que le han sido confiadas* o *siguiendo las instrucciones de su empresario*, la titularidad de los derechos de explotación correspondientes al programa de ordenador así creado, tanto el programa fuente como el programa objeto, corresponderán, exclusivamente, al empresario, salvo pacto en contrario».

Este último supuesto es muy importante en el caso de los programas creados en el marco de una investigación científica, pues estos pueden llevarse a cabo, *en el ejercicio de las funciones*

que le han sido confiadas al investigador. Dicho de otro modo, la creación de softwares, forma parte de su contenido de trabajo. No obstante, es importante conocer la «política» del centro de investigación en concreto, dado que la mayoría de contratos de trabajo, sobre todo en entidades públicas de investigación, no especifican las tareas concretas que pueden quedar comprendidas dentro del marco genérico de «actividades de investigación», pudiendo ser práctica del centro que los derechos se atribuyan en su totalidad al investigador, por lo general con obligación de incluir el logo de la institución a efectos de distribución, comercialización, etc. Por otro lado, si bien el art. 97.4 otorga derechos de propiedad al empleador o empresario, cabe el pacto en contrario. Por lo que sería perfectamente legal, acordar que el investigador, «autor de la obra» se reserve los derechos sobre esta, o participe de los derechos de explotación comercial de la misma. Del mismo modo, es muy posible, que el centro tenga su propio procedimiento interno para proceder a la solicitud ante el registro correspondiente. En todo caso, aunque la titularidad del programa sea otorgada al empleador, el trabajador asalariado, conserva los derechos morales sobre la misma, es decir, tiene derecho, entre otras cosas, a ser reconocido en todo caso como el autor intelectual de la obra.

Más clara parece la titularidad de los derechos de autor del «empleador», cuando el trabajador haya creado el programa «siguiendo las instrucciones del empresario». Este supuesto, hace referencia a las obras por encargo, donde la creatividad del trabajador está hasta cierto punto limitada. Aquí la LPI está presuponiendo que existe «algún tipo de documento o instrucción escrita referido al hecho de que la contratación tiene por objeto que el trabajador-creador cree el programa informático «por encargo» o «por indicación» del empresario» (Erdozain López, 2019). No obstante, es muy posible que, en la práctica, tal documento, contrato, negocio

jurídico, etc. no exista realmente. También en este caso cabe pacto en contrario, y, por supuesto los derechos morales, seguirán correspondiendo al trabajador/investigador.

6.3.2. Contratos y licencias de programas de ordenador

Como se ha venido reiterando, la creación de un software genera una serie de derechos, tanto de índole moral como económica. Por lo general los derechos económicos se ejercitan, se hacen realidad, por medio de la concesión de licencias. En este sentido, las licencias de uso de software, nos son más que un tipo de contrato o acuerdo por el cual de una parte titular o propietaria de un software (licenciante) concede una licencia (autorización de uso) a un tercero (licenciatario) sobre los derechos de explotación del mismo (su uso) a cambio del pago de un precio o *canon*, o bien de forma gratuita (software no propietario). Las licencias constituyen así la forma principal en que los titulares de derechos de autor, autorizan la utilización, distribución y difusión de sus obras, sin que exista cesión o transferencia de los derechos de propiedad que constituyen los derechos de autor/copyright.

Los contratos, pueden buscar un equilibrio entre autores y usuarios, siempre que se trate de un contrato en el sentido tradicional, que permita a las partes acordar y negociar. Es así que un contrato puede ser más o menos estricto que las previsiones normativas previamente analizadas. En la práctica, en la mayoría de las ocasiones, los contratos propietarios en esta materia constituyen contratos de adhesión, dejando muy poca o ninguna opción de negociar a los usuarios finales (Duardo *et al*, 2015). Ese es el caso de los contratos de licencia de uso más comúnmente utilizados como: la licencias “shrink-wrap” o licencias de envoltura, que han perdido terreno en favor de las licencias “click-wrap” (contrato online, o contratos en línea) y, finalmente las licencia “sharewares”. Todas ellas conocidas como licencias “propietarias”.

La primera, la licencia “shrink-wrap” debe su nombre a la práctica de incluir las condiciones de uso en el envase o paquete que contiene el programa de ordenador, por lo general, bajo una capa de plástico retractilado. Un aviso en el exterior del paquete informa al comprador que, de abrirlo, estaría aceptando implícitamente los términos de la licencia. Se trata, en toda regla, de un contrato de adhesión donde el poder de negociación del usuario se reduce a la posibilidad de rechazar o aceptar la totalidad de los términos impuestos. A pesar de ello, en aras del mercado, se considera un contrato válido, un mal necesario. No obstante, dada la situación de debilidad del consumidor, se pretende su amparo por otras vías como la prohibición de las llamadas cláusulas abusivas o dándole la posibilidad, en algunos casos, de devolver el producto y recuperar su dinero.

Por su parte las licencias en línea, más conocidas por su denominación en lengua inglesa como licencias "click-wrap", son menos problemáticas en cuanto a la incorporación de términos y condiciones, siempre y cuando la página web llame la atención al consumidor sobre existencia de la licencia, señalando sus términos y condiciones antes de la entrada en vigor del contrato. Cuando el comprador hace clic en los botones "De acuerdo", "Sí" o "Acepto" en un formulario de inscripción en línea, ha realizado un negocio jurídico, y ha aceptado los términos y condiciones de la licencia que se muestran en la pantalla quedando vinculado contractualmente. Cada vez es más común que las páginas web contengan un enlace a "términos y condiciones", que, comúnmente, aparecen en la parte inferior de la página.

Por su parte, la licencia shareware constituye otro método común de suministro de los programas de ordenador. Puede ser proporcionada a través de un CD[4], o mediante descarga

[4] Práctica que es previsible que vaya desapareciendo dado el CD, se ha ido quedando obsoleto. Es probable que con el tiempo se cambie a otro tipo de soporte o simplemente se opte en todo caso por el contrato en línea, más sencillo y menos costoso para las empresas.

desde un sitio de Internet. Esta licencia funciona suministrando una prueba gratuita del programa de ordenador por un período limitado de tiempo, al final del cual el usuario puede decidir si paga una cuota de inscripción para obtener una licencia que permita el uso continuado del programa, sujeto a los términos y condiciones de la misma, o simplemente, dejar de usarlo. En la mayoría de los casos, el paquete de programas de ordenador shareware tenderá a dejar de funcionar al final del período de prueba. El programa de ordenador también se puede distribuir como una versión con características limitadas que pueden actualizarse o ampliarse mediante registro. A esto último se le conoce como el sistema de "crippleware", y es una variante del sistema de shareware. Aquí, el programa de ordenador no se distribuye bajo una licencia shareware, sino como una forma incompleta o "inválida", por ejemplo, el programa de ordenador no es capaz de guardar archivos, o imprimirlos. Mediante el pago de una cuota al propietario del Copyright, se enviará al comprador una versión completamente funcional software. Otra variante comúnmente usada es "nagware", bajo esta forma el programa es gratuito durante un período de prueba mientras al usuario se le recuerda con frecuencia en la pantalla que debe registrarse y pagar por el programa si desea seguir utilizándolo cuando el período de prueba haya terminado.

Existen otros modelos de negocio no tradicional y ciertos tipos de licencia menos restrictivas, conocidas como «licencias no propietarias». Estas licencias suelen venir asociadas al llamado movimiento de «software libre». El nombre responde al practica de «liberar» software publicando su código fuente, de este modo el autor del mismo permite de manera consciente y voluntaria que cualquier persona pueda utilizar su creación de forma gratuita. Dicho de otro modo, el uso del «software libre» no está sujeto a una previa autorización de un titular de derechos, aunque ello no significa que se haga dejación de los derechos morales, como el respeto de la autoría del creador. Por otra parte, su uso estará condicionado al hecho de que cualquier

otro programa que se haya basado en un software libre deberá ser liberado a su vez, prohibiendo su uso con fines comerciales. Las licencias asociadas a este tipo de software se conocen como licencias *copyleft*, en contraposición al *copyright*. Así, algunas licencias permiten la ejecución, reproducción y distribución de forma ilimitada (licencia copyleft); otras pueden prohibir la modificación (licencias "semilibres").

En el primer caso la mayor parte del programa de ordenador de "código abierto" se licencia, permitiendo que cualquiera pueda ejecutar, copiar y distribuir dicho programa. También se permiten modificaciones del programa de ordenador para uso personal. Las versiones modificadas del programa de ordenador se pueden dispensar en las condiciones para distribuir tanto el código fuente como el código objeto de modificación. Esta nueva versión, como se avanzó, también debe ser distribuida bajo una licencia "copyleft".

Amén de la filosofía subyacente en el movimiento de software libre- copyleft, asociada a la libertad de creación, la cooperación y el intercambio de conocimientos, ésta no está exenta de ser aprovechada por el mercado. De este modo, el modelo de negocios, por lo general, consiste en liberar el mismo programa de ordenador bajo dos licencias diferentes: la licencia semilibre, que está diseñada para fomentar el uso generalizado y no comercial del programa de ordenador, y una licencia de programas de ordenador propietario para las empresas que, por supuesto, tendrán que pagar.

En investigación científica, es esencial hacer uso de programas bajo licencia y conocer los usos permitidos según los términos de ésta. Ello determina la validez legal y la posibilidad futura de publicación de cualquier resultado científico obtenido con la ayuda de dichos programas de ordenador. También tiene impacto sobre la patentabilidad potencial de cualquier producto o sustancia, cuando los métodos informáticos han sustituido o reemplazado parcialmente los

resultados experimentales, o el software forma parte de la metodología o invento que se pretende quiere patentar.

6.4. Propiedad industrial. Protección mediante Patente

La finalidad de la concesión de derechos de Propiedad Industrial es la de impedir toda utilización no autorizada de las invenciones y signos, así como cualquier otra práctica que pueda inducir a error a los consumidores sobre éstos. En tal sentido, los derechos de propiedad industrial protegen tanto las creaciones intelectuales o invenciones de aplicación industrial (las patentes; los modelos y dibujos industriales; los modelos de utilidad) como los signos que distinguen a los productos (marcas de fábrica y de comercio; los nombres comerciales, las indicaciones de productos) y más recientemente las obtenciones vegetales; los certificados complementarios de protección de medicamentos y de productores fitosanitarios.

En lo que se refiere a la patente, esta otorga derechos exclusivos sobre una invención que, en síntesis, es un producto o un proceso que ofrece una nueva manera de hacer algo o supone una nueva solución técnica a un problema existente. Para que sea posible reclamar una patente son requisitos indispensables la *novedad*, la existencia de una *actividad inventiva* previa, una evaluación de la *aplicabilidad industrial*. Las patentes deben someterse a un exhaustivo proceso de examen.

En resumen, una patente es un título de propiedad que se concede para lo reivindicado en una solicitud si esta reúne los requisitos exigidos por legislación correspondiente. Tal derecho consiste en la explotación en exclusiva y por un tiempo determinado del contenido de lo reivindicado en la misma (en España 20 años desde la fecha de solicitud). De este modo, el titular de la patente, obtiene el "monopolio" de uso y explotación comercial. Al tiempo que se

reducen los riesgos y se aseguran las inversiones de las empresas adquirientes de los derechos mediante licencia.

La normativa relativa a la patentabilidad de los programas de ordenador aún no ha sido armonizada internacionalmente, aunque muchos países han abrazado, en cierta medida, la patentabilidad de las invenciones «relacionadas» con el programa de ordenador (Duardo 2015). Existe una tendencia mundial a favor de la adopción de la patente en la protección de las invenciones relacionadas con el programa de ordenador, principalmente porque una patente concede derechos exclusivos frente a todos. Es decir, es válida contra todo el que hace uso de una invención patentada en un determinado país; ello ocurre, incluso, cuando el infractor llega a la misma invención de forma independiente.

La patente protege la idea subyacente, siempre que dicha idea se encuentre dentro de la categoría estatutaria de "materia patentable" y no sea tan fundamental que constituya una ley de la naturaleza. Por su parte, las normas de derechos de autor/Copyright, como hemos visto, sólo protegen la expresión de una idea, siendo posible que dos programas que realicen la misma función sean dignos de protección.

En el ámbito internacional, el Tribunal Supremo norteamericano (TS), se ha pronunciado en varias ocasiones sobre la patentabilidad de los programas de ordenador. Los sucesivos pronunciamientos se reafirman en la posición de que las ideas abstractas no son patentables. Así en el famoso caso Alice (Alice Corp. v. CLS Bank International 573US, 134 S Ct 2347-2014), el TS consideró que el programa objeto del procedimiento no era patentable dado que el solicitante pretendía la protección de una idea abstracta. *Sensu contrario*, la sentencia, siguió la misma línea de pronunciamientos anteriores, ratificando el «concepto inventivo» como requisito determinante para reclamar una patente sobre que un programa de ordenador. De este modo, en EE.UU, la

patentabilidad del software no queda excluida; pero si supeditada al concepto inventivo. Lo que no se aleja mucho de la concepción del Tribunal de Justicia de la Unión Europea.

En el ámbito europeo la doctrina más extendida en relación con las patentes para invenciones relacionadas con el programa de ordenador, es la doctrina del "efecto técnico". Establecida por primera vez por la Oficina de la Patente Europea (OPE), esta doctrina sostiene que el programa de ordenador es patentable, en general, si su aplicación tiene un "efecto técnico". Así, por ejemplo, el programa que controla la sincronización de un motor electrónico es patentable bajo este canon, mientras que el programa de ordenador que detecta y corrige errores gramaticales en un texto no lo es.

Ello es plenamente aplicable al caso español donde, de acuerdo tanto con la ley de patentes nacional como la normativa comunitaria en la materia, los programas de ordenador no son patentables por sí solos. Sin embargo, es posible que la protección brindada por la patente les alcance cuando forman parte de un invento con efecto técnico, donde se incorporan metodologías e invenciones que cumplen con los requisitos de patentabilidad. Así, la protección concedida como obra intelectual se entiende «sin perjuicio de cualesquiera otras disposiciones legales, tales como las relativas a los derechos de patente, marcas, competencia desleal, secretos comerciales, protección de semiconductores o derecho de obligaciones» (art. 104 LPI). Ello resulta relevante también, para conseguir la protección plena del software, como se verá a continuación.

Finalmente, es importante que los investigadores tengan claro, si pretenden reivindicar una patente, que publicar los resultados científicos antes de realizar la solicitud de la misma, resulta contraproducente. De esta suerte antes de proceder a comunicar los resultados de la investigación por cualquier medio, es conveniente haber realizado previamente la solicitud en una oficina de patentes. Una vez realizada la solicitud, no hay que esperar a que la patente sea concedida para

proceder a la publicación de los resultados científicos. Como en el caso de los derechos de autor, es importante que se informen sobre la política de su centro sobre este particular y sobre la existencia o no de procedimientos internos, antes que realizar los trámites por su cuenta; dado que podrían estar infringiendo los reglamentos internos o incumpliendo obligaciones contractuales aun sin saberlo.

6.5. Protección mediante el Secreto comercial.

Teniendo en cuenta que el acceso o no al código fuente es una cuestión clave en materia de programación y que la mayoría de los desarrolladores de software privativos hacen grandes esfuerzos para protegerlo como información confidencial, la pertinencia de los derechos de secreto comercial no es discutida. Este método de protección cubre todo ataque dirigido a apoderarse de los logros que una empresa ha alcanzado en el desarrollo del programa, tanto si proceden de la deslealtad de sus propios empleados o funcionarios, como del comprador del programa vinculado mediante contrato o licencia con la empresa, o si son el resultado del espionaje industrial o cualquier otra conducta semejante.

Para que sea posible optar por esta forma de protección es preciso que se cumpla con los requisitos generales, a saber: que la información que se pretende proteger sea confidencial, es decir no se haya hecho pública en ningún momento, sea novedosa, reproducible y no forme parte de la experiencia de una persona para llevar a cabo una tarea (know-how).

El régimen del secreto comercial también es válido para proteger los manuales de servicios informáticos y de mantenimiento, que pueden facilitar la ingeniería inversa del programa de ordenador. El acuerdo de los ADPIC aborda los secretos comerciales de cualquier nivel de complejidad. En su artículo 39, se establecen los requisitos para los Estados signatarios, así como que todos esos países deben promulgar leyes de secreto comercial, lo que también contemplaría

el programa de ordenador si se cumplen determinados requisitos como los relativos a la confidencialidad.

6.6. La Marca Registrada

El Derecho protege tanto el contenido interno de los programas de ordenador, como sus aspectos externos tales como la marca que les distingue (marca registrada). Por marca registrada se entiende cualquier marca o medio que puede ser representado gráficamente, y que es capaz de distinguir los productos o servicios de una empresa de los de otra. Una marca puede consistir en palabras, dibujos, letras, números, incluso en la forma del producto o de su embalaje.

La protección de las marcas a través de su inscripción viene siendo aceptada internacionalmente desde hace mucho tiempo. En la mayoría de los países, los derechos se fijan inicialmente por el registro y son mantenidos por su uso posterior en un determinado país. Por regla general, el uso de la marca sin registro no proporciona protección. Algunos países exigen la inscripción de la marca antes de su uso, mientras que otros dan prioridad a los derechos basados únicamente en el uso.

Las limitaciones inherentes a la aplicación territorial de las leyes de marcas han sido mitigadas a través diversos tratados sobre propiedad intelectual, el más importante: el Acuerdo de los ADPICs antes mencionado. Este Acuerdo establece la compatibilidad legal entre las diversas jurisdicciones de los países miembros al exigir la armonización de las leyes aplicables. Por ejemplo, su artículo 15 (1) establece una definición para una "señal" que se usa como o forma parte de la definición de "marca" en la legislación de marcas de muchas jurisdicciones de todo el mundo.

Las cuestiones internacionales relativas a las marcas también se rigen por la marca comunitaria (ECT), el Arreglo de Madrid y el Protocolo de Madrid. Las marcas sólo pueden ser protegidas país por país. El plazo de aplicación desde la inscripción varía según la jurisdicción, desde unos pocos meses hasta varios años. Mientras que algunos países examinan el registro existente para evitar registros conflictivos (*verbi gratia* Estados Unidos), otros (Alemania, Francia, Suiza...), dejan en manos de los solicitantes del registro el identificar y resolver este tipo de conflictos.

6.6.1 La Marca Comunitaria

Como primer acto normativo de importancia en este campo, la Directiva 89/104/CEE, de 21 de diciembre de 1988, relativa a la aproximación de las legislaciones de los Estados miembros en materia de marcas, ha constituido un elemento de peso en proceso de "europeización del Derecho de marcas". Ello, naturalmente, sin dejar de lado, la importancia que en este campo ha tenido la jurisprudencia constante del TJCE.

La Directiva se ocupa de las marcas obtenidas «mediante el registro», dejando total libertad a los Estados miembros para establecer el procedimiento de registro de las mismas. En sus disposiciones más relevantes, por su carácter imperativo, establece qué signos pueden constituir una marca, determina las causas de denegación o de nulidad; los derechos conferidos por la marca, la limitación de sus efectos; el agotamiento del derecho; la concesión de licencias; la prescripción, el uso obligado de la marca registrada y entre otras, las causas de caducidad.

Estas disposiciones de la Directiva están integradas, a su vez, en el Reglamento (CE) núm. 40/1994 del Consejo de 20 de diciembre de 1993 sobre la marca comunitaria destinado a garantizar que las marcas nacionales y comunitarias se rigiesen por un Derecho común en las

materias armonizadas por la Directiva. El Reglamento establece un sistema que permite la concesión de marcas comunitarias por la Oficina de Armonización del Mercado Interior (OAMI), sobre la base de una solicitud única presentada ante la misma. La marca obtenida por esta vía tiene carácter unitario, en el sentido de que produce los mismos efectos en todo el territorio comunitario y, salvo disposición en contrario del Reglamento, «sólo podrá ser registrada, cedida, ser objeto de renuncia, de resolución de caducidad o de nulidad, y prohibirse su uso, para el conjunto de la Comunidad» (art. 1.2). De este modo, el principal interés del sistema de marca comunitaria es permitir a las empresas identificar sus productos y servicios de manera idéntica en todo el territorio de la Unión Europea.

El objeto de protección del sistema comunitario lo constituye la marca como signo o señal que puede ser objeto de representación gráfica (en particular palabras, dibujos, letras, cifras, la forma del producto o su acondicionamiento), siempre que tal signo permita distinguir los los productos o los servicios de una empresa de los de otras (arts. 4 Reglamento y 2 Directiva). Se trata de un concepto que define la marca por su función esencial, que no es otra que indicar la procedencia empresarial del producto o servicio. Además de la exigencia formal de que el signo pueda ser objeto de una representación gráfica, y del carácter distintivo, antes enunciados, la marca comunitaria exige la presencia de otros dos requisitos, a saber: la licitud y la apropiabilidad o disponibilidad. La distintividad y la licitud constituyen requisitos de validez absolutos, mientras que la apropiabilidad tiene carácter relativo. Como se verá más adelante la ausencia de estos requisitos desemboca en causas de denegación o de nulidad de la marca.

El nacimiento del derecho sobre la marca comunitaria, viene regido por el principio de inscripción registral. Así, ésta sólo puede adquirirse mediante el registro (art. 6 Reglamento); que constituye el único modo originario de adquisición de una marca comunitaria. No obstante, el

art. 6 bis del CUP, obliga a los Estados miembros a proteger las marcas nacionales notorias no registradas, en correspondencia, ello puede constituir motivo de denegación y nulidad relativa del registro de una marca comunitaria posterior (arts. 8 y 52.1.a) del Reglamento)

Pueden ser titulares de marcas comunitarias las personas físicas y jurídicas, incluidas las entidades de derecho público, que sean nacionales: de los Estados miembros; de otros Estados que sean partes en el Convenio de París para la Protección de la Propiedad Industrial; de Estados que no sean partes en el Convenio de París pero que estén domiciliados o tengan su sede en territorio de la Comunidad o de un Estado que sea parte en el Convenio; de cualquier otro Estado que conceda a los nacionales de los Estados miembros la misma protección en materia de marcas que a sus nacionales (art. 5). Como regla general no se exige representación para actuar ante la OAMI (art. 88.1), salvo que no se tenga ni domicilio ni establecimiento en la Comunidad, en cuyo caso ha de nombrar un representante profesional en cualquier procedimiento, excepto para presentar una solicitud de marca comunitaria (art. 88.2 RMC).

Como se ha adelantado, la ausencia del requisito formal y de los absolutos constituye motivos de denegación absolutos de registro establecidos por el art. 7 Reglamento (art. 3 Directiva). La carencia del requisito relativo justifica los motivos de denegación relativos de registro consagrados por el art. 8 Reglamento (art. 4 Directiva). Así la carencia de distintividad justifica los motivos de denegación absolutos tipificados por las letras a) a d) del art. 7.1 Reglamento, correlativamente art. 3 Directiva. La falta de licitud sirve de base a los motivos de denegación absolutos establecidos por las letras e) a k) del art. 7.1 Reglamento. Por último, el signo no estará disponible cuando haya sido prioritariamente apropiado por un tercero, quien ostentará la titularidad de un derecho anterior, esta falta de disponibilidad sirve de basamento a los motivos de denegación relativos listados por el art. 8 Reglamento (Art. 4 de la Directiva).

En particular, se denegará el registro: de señales que no puedan constituir marcas comunitarias; de marcas aquellas desprovistas de carácter distintivo; constituidas por señales o indicaciones que se hayan convertido en usuales en la lengua corriente o en las prácticas comerciales; contrarias al orden público o a las buenas costumbres; las marcas susceptibles de engañar al público sobre la naturaleza, la calidad o la procedencia geográfica del producto o servicio.

6.7. La aceptación de resultados científicos obtenidos mediante el uso de programas de ordenador: los requisitos para la validación de los modelos QSAR

Hasta el momento, hemos hecho referencia a las distintas vías de protección del software y los derechos que pueden asistir los titulares/creadores. Estos apuntes son plenamente válidos para los programas desarrollados en el marco de la investigación científica.

En este apartado, nos dedicaremos a analizar otra arista del fenómeno: la validación y la aceptación de los resultados y métodos científicos obtenidos con el uso de un programa de ordenador. «Las definiciones legales por sí solas resultan insuficientes para establecer un marco para la protección del conocimiento científico. Por ello son necesarias las aportaciones de especialistas y científicos, aunque ello, por supuesto, añade mayor complejidad al asunto» (Duardo 2015). El impacto del marco regulador de la UE en materia de los modelos de Relaciones Cuantitativas Estructura-Propiedad -conocidos por sus siglas en inglés: Quantitative Structure-Property Relationships QSPR o QSAR- arroja algo luz sobre la forma en que se debe abordar el tema de la validación y aceptación de nuevos enfoques científicos.

Un modelo QSAR es, en esencia, aquel que conecta la información sobre la estructura de un sistema con información sobre propiedades externas del mismo. El énfasis se pone, generalmente, en propiedades externas que no son evidentes después de una inspección visual directa de la estructura del sistema. Estos modelos son ampliamente utilizados en Quimio-

informática y Bioinformática. En particular, los QSARs conectan la información sobre la estructura química de fármacos y dianas moleculares (proteína, gen, ARN, microorganismos, tejidos, enfermedades…, etc.) con la Actividad biológica (P = A) del fármaco sobre sus posibles dianas. Muchos de estos modelos están basados en el uso de parámetros químico-informáticos estructurales. Dichos parámetros son series numéricas que codifican información estructural para predecir correlaciones entre la estructura molecular y propiedades biológicas.

Pues bien, el Reglamento CE Nº 1907/2006 (art.13; 25) establece que los resultados de un QSAR pueden ser usados para sustituir los ensayos en animales a efectos de identificar la presencia o ausencia de ciertas propiedades peligrosas de las sustancias químicas, siempre que el modelo cumpla una serie de condiciones.

Bajo el texto de esta norma, conocida como Reglamento REACH -por sus siglas en inglés (Registration, Evaluation, Authorization and Restriction of Chemicals)-, en el territorio de la Unión Europea se hace necesario evaluar los peligros humanos y ambientales de todas las sustancias químicas producidas o importadas en cantidades superiores a una tonelada por año (tpa). Si esta evaluación se realizase mediante los métodos tradicionales, requeriría de un gran número de animales de laboratorio, a la vez que no se dispondría ni de los recursos ni del tiempo suficiente. Por ello el REACH impulsa soluciones alternativas como el uso de métodos cuantitativos *in vitro*, relaciones estructura-actividad y/o propiedad (QSAR/QSPR) y agrupación química, para ayudar en esta tarea.

Los requisitos exigidos por el REACH para la validación de los modelos, son los siguientes: que el QSAR ha de ser *científicamente válido*; la sustancia ha de estar comprendida en el *ámbito de aplicabilidad del modelo* QSAR; que el resultado QSAR sea *adecuado a los propósitos* de

evaluación de riesgo de clasificación y etiquetado; y finalmente, que se aporte *documentación fiable adecuada*.

En cuanto a la necesidad de demostrar la validez científica del modelo QSAR utilizado el texto legal remite a los principios de la OCDE internacionalmente acordados para la validación QSAR a saber: un punto final definido; un algoritmo inequívoco; un dominio definido de aplicabilidad; medidas adecuadas de la bondad del ajuste, robustez y capacidad de predicción; y, de ser posible, una interpretación mecanicista. Dichos principios se han resumido brevemente en una publicación de la OCDE conocida como la Guía de la OCDE sobre validación de QSAR.

Así, la información generada por los QSAR, validados científicamente. puede utilizarse potencialmente en lugar de los datos experimentales, siempre que se cumplan el resto de condiciones.

Si bien el REACH exige una documentación pertinente y adecuada que justifique el cumplimiento de los requisitos exigidos para la validación de un modelo QSAR, no establece ningún procedimiento formal. La OCDE, sin embargo, ofrece una guía para documentar las características del modelo QSAR y sus predicciones en un formato de informe específico: el Modelo para la elaboración de informes de QSAR (QMRF) o bien el Formato para la elaboración de informes de predicción de QSAR (QPRF) para las estimaciones de QSAR, (para mayor información sobre ambos consúltese: http://ecb.jrc.it/qsar/qsar-tools/index.php?c=QRF).

Si bien, con algunas deficiencias, este binomio REACH- Modelo OCDE, proporciona las pautas para la validación de método científicos obtenidos con la intervención de programas de ordenador, en sustitución de los métodos tradicionales.

6.8 Reflexiones finales

Como se ha puesto de manifiesto a lo largo de este capítulo, existen múltiples interacciones entre las Ciencias de la Información y las Comunicaciones y el Derecho. En este capítulo apenas hemos analizado unos pocos de esos aspectos. Si bien la forma más ampliamente aceptada para dar cobertura legal a la creación del software es la de recurrir a derechos de autor/Copyright, esta no es única solución para proporcionar protección a los programas de ordenador. Dado que por lo general protegen distintos aspectos relacionados con el desarrollo de software, las diversas vías de protección no son, por lo general, incompatibles o excluyentes. Es posible obtener la protección software como propiedad intelectual y al mismo tiempo registrar el logo o la marca creada para identificar el producto. También, como se ha visto, es posible obtener protección en distintos territorios.

En todo caso, es importante que los científicos que desarrollan una investigación relacionada con el software, e informen sobre la política de su centro sobre este particular y sobre la existencia o no de procedimientos internos, antes que realizar los trámites por su cuenta; dado que podrían estar infringiendo los reglamentos internos o incumpliendo obligaciones contractuales aun sin saberlo.

En el caso de que se pretenda reivindicar de una patente, es vital que los investigadores tengan claro, que publicar los resultados científicos antes de realizar la solicitud de la misma, resulta contraproducente. De esta suerte antes de proceder a comunicar los resultados de la investigación por cualquier medio, es conveniente haber realizado previamente la solicitud en una oficina de patentes. Una vez realizada la solicitud, no hay que esperar a que la patente sea concedida para proceder a la publicación de los resultados científicos.

Por otra parte, en cuanto a la validación y la aceptación con fines regulatorios de los modelos y métodos derivados del uso de programas de ordenador, específicamente en lo que a las técnicas QSAR respecta, han recibido un gran impulso con la implementación del programa REACH. Mostrando una creciente, aunque insuficiente, interacción entre el Derecho y la Bioinformática, cambio muy importante para que las herramientas y metodologías informáticas puedan ser reconocidas legalmente y puedan alcanzar, por fin, un reconocimiento similar al de las llamadas "ciencias duras".

Conclusiones

Conclusiones

1. Se ha podido desarrollar (programar) una versión beta de un software, SOFT.PTML, en el que se implementan por primera vez algoritmos NIFPTML en una misma aplicación. Lo cual facilita el desarrollo y uso de estos algoritmos por usuarios de distintas áreas científicas.

2. Se ha demostrado la utilidad de este programa aplicándolo a distintos problemas prácticos en las áreas mencionadas, como son: el diseño de fármacos, descubrimiento de nanomateriales, estudio de sistemas jurídicos. Específicamente, se desarrollaron nuevos modelos NIFPTML para la predicción de fármacos antiparasitarios contra la leishmaniosis, el descubrimiento de sistemas duales nanopartícula-fármaco antibacteriano, y la predicción de la estabilidad en el tiempo de un sistema jurídico-tributario.

3. Se ha aportado un análisis de las implicaciones jurídicas del desarrollo y aplicación de este tipo de algoritmos en investigación. Si bien la forma más ampliamente aceptada para proteger la creación del software es la de recurrir a derechos de autor/Copyright, esta no es única solución. Las diversas vías de protección no son, por lo general, incompatibles o excluyentes. En todo caso, es importante que los científicos que desarrollan una investigación relacionada con el software, se informen sobre la política de su centro sobre este particular y sobre la existencia o no de procedimientos internos antes que realizar los trámites por su cuenta; dado que podrían estar infringiendo los reglamentos internos o incumpliendo obligaciones contractuales sin saberlo. En el caso de que se pretenda reivindicar una patente, es vital que tengan claro que publicar los resultados científicos antes de realizar la solicitud de la misma, resulta contraproducente. Antes de proceder a comunicar los resultados de la

investigación por cualquier medio, es conveniente haber realizado previamente la solicitud en una oficina de patentes.

Futuros desarrollos

FUTUROS DESARROLLOS

1. Recodificar el programa SOFT.PTML usando otros lenguajes de programación y librerías (C++, Phyton, etc.) que lo hagan más compatible.
2. Acoplar el programa SOFT.PTML con el software MI-NODES para tener una mejor integración de las etapas NI+IFPTML ya que el programa actualmente no hace cálculo de invariantes de redes.
3. Agregar otros algoritmos AI/ML no implementados, en especial algoritmos AI de Deep Learning como Deep Learning Networks (DLN), Convolutionary Neural Networks (CNN), etc.
4. Registrar el software SOFT.PTML para proteger los resultados de este trabajo.

Bibliografía

BIBLIOGRAFÍA

Abercrombie, N, S Hill, y B. S Turner. 2000. Social structure. En The Penguin Dictionary of Sociology. London: Penguin.

Ahmadi, S., Toropova, A. P., & Toropov, A. A. (2020). Correlation intensity index: mathematical modeling of cytotoxicity of metal oxide nanoparticles. Nanotoxicology, 1118-1126. https://doi.org/10.1080/17435390.2020.1808252

Ahn, S., Jung, S. Y., Lee, J. P., Kim, H. K., & Lee, S. J. (2010). Gold nanoparticle flow sensors designed for dynamic X-ray imaging in biofluids. ACS Nano, 3753-3762. https://doi.org/10.1021/nn1003293

Alafeef, M., Srivastava, I., & Pan, D. (2020). Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. ACS sensors, 1689-1698. https://doi.org/10.1021/acssensors.0c00329

Alonso, N., Caamaño, O., Romero-Duran, F., Luan, F., D S Cordeiro, M., Yañez, M., González-Díaz, H., & García-Mera, X. (2013). Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates. ACS Chem. Neurosci., 1393-1403. http://dx.doi.org/10.1021/cn400111n

Ambure, P., & Roy, K. (2016). Understanding the structural requirements of cyclic sulfone hydroxyethylamines as hBACE1 inhibitors against Aβ plaques in Alzheimer's disease: a predictive QSAR approach. RSC Advances, 28171-28186. http://dx.doi.org/10.1039/C6RA04104C

Ambure, P., Halder, A., González Díaz, H., & Cordeiro, M. (2019). QSAR-Co: An open source software for developing robust multitasking or multitarget classification-based qsar models. J. Chem. Inf. Model., 2538-2544. http://dx.doi.org/10.1021/acs.jcim.9b00295

Arasoglu, T. D. (2016). Comparative evaluation of antibacterial activity of caffeic acid phenethyl ester and PLGA nanoparticle formulation by different methods. Nanotechnology. https://doi.org/10.1088/0957-4484/27/2/025103

Arrasate, S., & Duardo-Sanchez, A. (2018). Perturbation theory machine learning models: theory, regulatory issues, and applications to organic synthesis, medicinal chemistry, protein research, and technology. Curr. Top. Med. Chem, 1202-1213. http://dx.doi.org/10.2174/1568026618666180810124031

Autor repetido. 2011. Redes de derecho penal, cadenas de markov, entropía de Shannon y redes neuronales artificiales. En Complex Network Entropy: De las moléculas a la biología, la parasitología, la tecnología y las neurociencias sociales y jurídicas, editado por H. González-Díaz. Kerala, India: Bentham.

Azam, A., Ahmed, A. S., Oves, M., Khan, M. S., & Memic, A. (2012). Size-dependent antimicrobial properties of CuO nanoparticles against Gram-positive and -negative bacterial strains. International journal of nanomedicine, 3527-3535. https://doi.org/10.2147/IJN.S29020

Azam, A., Ahmed, A. S., Oves, M., Khan, M. S., Habib, S. S., & Memic, A. (2012). Antimicrobial activity of metal oxide nanoparticles against Gram-positive and Gram-negative bacteria: a comparative study. International journal of nanomedicine, 6003-6009. https://doi.org/10.2147/IJN.S35347

Barnard, A. S., & Opletal, G. (2019). Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. Nanoscale, 23165-23172. https://doi.org/10.1039/c9nr03940f

Barratt, R.; Balls, M. An overall strategy for the testing of chemicals for human hazard and risk assessment under the EU REACH system. ATLA 2003; 31: 19-20.

Bastian, M.; Hetmann, S.; Jacomy, M. Gephi: An Open Source Programas de ordenador for Exploring and Manipulating Networks. In International AAAI Conference on Weblogs and Social Media (ICWSM09), North America, 2009.

Batagelj, V.; Mrvar, A. Pajek 1.15. 2006.

Bavelas, A. 1948. "Un modelo matemático para las estructuras de grupo". Human Organization (7):16-30.

Bean, D. M., Stringer, C., Beeknoo, N., Teo, J., & Dobson, R. J. (2017). Network analysis of patient flow in two UK acute care hospitals identifies key sub-networks for A&E performance. PLoS One, 2017. https://doi.org/10.1371/journal.pone.0185912

Bediaga, H., Arrasate, S., & González-Díaz, H. (2018). PTML combinatorial model of chembl compounds assays for multiple types of cancer. ACS Comb. Sci., 621-632. http://dx.doi.org/10.1021/acscombsci.8b00090

Benfenati, E. The CAESAR project for in silico models for the REACH legislation. Chem Cent J 2010; 4 Suppl 1: I1.

Benfenati, E. The specificity of the QSAR models for regulatory purposes: the example of the DEMETRA project. SAR & QSAR in Environmental Research 2007; 18: 209-20.

Benhamou, Y. (2004). Antiretroviral therapy and HIV/hepatitis B virus coinfection. Clin. Infect. Dis., S98-S103. http://dx.doi.org/10.1086/381451

Benzi, Michele, Ernesto Estrada y Christine Klymko. 2013. Clasificación de centros y autoridades utilizando funciones matriciales. Linear Algebra and its Applications 438 (5):2447-2474.

Berca, M.N.; Duardo-Sanchez, A.; González-Díaz, H.; Pazos, A.; Munteanu, C.R. Markov entropy for biology, parasitology, linguistic, technology, social and law networks. In Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences, González-Díaz, H.; Prado-Prado, F.J.; García-Mera, X., Eds. Transworld Research Network: Kerala, India, 2011; pp 127-42.

Bercovitz, A. (2000) "Propiedad industrial y globalización de los mercados", Actualidad jurídica Aranzadi (436): 1-4.

Bercovitz, A. (ed.) (2007) Derecho de la Competencia y de la Propiedad Industrial en la Unión Europea. Navarra: Aranzadi.

Bernabe Ortega-Tenezaca, V.-T., & González-Díaz, H. (2017). In: FRAMA 1.0: Framework for moving average operators calculation in data analysis. Proceedings of MOL2NET,

International Conference Series on Multidisciplinary Sciences. MDPI Sciforum, Basel,Switzerland, 3.

Betzer, O., Shwartz, A., Motiei, M., Kazimirsky, G., Gispan, I., Damti, E., Brodie, C., Yadid, G., & Popovtzer, R. (2014). Nanoparticle-based CT imaging technique for longitudinal and quantitative stem cell tracking within the brain: application in neuropsychiatric disorders. ACS Nano, 9274-9285. https://doi.org/10.1021/nn503131h

Bian, L., Sorescu, D. C., Chen, L., White, D. L., Burkert, S. C., Khalifa, Y., Zhang, Z., Sejdic, E., & Star, A. (2019). Machine-Learning Identification of the Sensing Descriptors Relevant in Molecular Interactions with Metal Nanoparticle-Decorated Nanotube Field-Effect Transistors. ACS Appl Mater Interfaces, 1219-1227. https://doi.org/10.1021/acsami.8b15785

Blay, V., Yokoi, T., & González-Díaz, H. (2018). Perturbation theorymachine learning study of zeolite materials desilication. perturbation theory-machine learning study of zeolite materials desilication. J. Chem. Inf. Model., 2414-2419. http://dx.doi.org/10.1021/acs.jcim.8b00383

Blazquez-Barbadillo, C., Aranzamendi, E., Coya, E., Lete, E., Sotomayor, N., & Gonzalez-Diaz, H. (2016). Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. RSC Advances, 38602-38610. http://dx.doi.org/10.1039/C6RA08751E

Bommarito, M.J.; Katz, D.M. 2010. Una aproximación matemática al estudio del Código de los Estados Unidos. Physica A (389):4195-4200.

Bornholdt, S., y H.G. Schuster. 2003. Handbook of Graphs and Complex Networks: From the Genome to the Internet. Wheinheim: WILEY-VCH GmbH & CO. KGa.

Botequim, D., Maia, J., Lino, M. M., Lopes, L. M., Simoes, P. N., Ilharco, L. M., & Ferreira, L. (2012). Nanoparticles and surfaces presenting antifungal, antibacterial and antiviral properties. Langmuir, 7646-7656. https://doi.org/10.1021/la300948n

Boulet, R.; Mazzega, P.; Bourcier, D. 2011. Una aproximación en red al sistema francés de códigos legales - parte I: análisis de una red densa. Artificial Intelligence and Law (19):333-355.

Bouyssie, D.; Gonzalez de Peredo, A.; Mouton, E.; Albigot, R.; Roussel, L.; Ortega, N.; Cayrol, C.; Burlet-Schiltz, O.; Girard, J.P.; Monsarrat, B. Mascot file parsing and quantification (MFPaQ), a new programas de ordenador to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Mol Cell Proteomics 2007; 6: 1621-37.

Breiman, L. (2001). Random forests. Mach. Learn, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Brunetti, V., Bouchet, L. M., & Strumia, M. C. (2015). Nanoparticle-cored dendrimers: functional hybrid nanocomposites as a new platform for drug delivery systems. Nanoscale, 3808-3816. https://doi.org/10.1039/c4nr04438j

Cai, H., & Yao, P. (2013). In situ preparation of gold nanoparticle-loaded lysozyme-dextran nanogels and applications for cell imaging and drug delivery. Nanoscale, 2892-2900. https://doi.org/10.1039/c3nr00178d

Caron, W. P., Morgan, K. P., Zamboni, B. A., & Zamboni, W. C. (2013). A review of study designs and outcomes of phase I clinical studies of nanoparticle agents compared with small-molecule anticancer agents. Clin Cancer Res, 3309-3315. https://doi.org/10.1158/1078-0432.CCR-12-3649

Casado Cervino, A.y Blanco Jiménez, A. (2005) El Diseño comunitario. Una aproximación al régimen legal de los Dibujos y Modelos en Europa. Pamplona: Aranzadi.

Casañola-Martin, G., Le-Thi-Thu, H., Pérez-Giménez, F., Marrero-Ponce, Y., Merino-Sanjuán, M., Abad, C., & González-Díaz, H. (2016). Multi-output model with box-jenkins operators of quadratic indices for prediction of malaria and cancer inhibitors targeting

ubiquitin- proteasome pathway (upp) proteins. Curr. Protein Pept. Sci., 220-227. http://dx.doi.org/10.2174/138920371799160226173500

CentiBiN Versión 1.4.2.

Coecke, S.; Ahr, H.; Blaauboer, B.J. Metabolism: a bottleneck in in vitro toxicological test development. The Report and Recommendations of ECVAM Workshop 54. ATLA 2006; 34: 49-8.

Concu, R., D S Cordeiro, M., Munteanu, C., & González-Díaz, H. (2019). PTML model of enzyme subclasses for mining the proteome of biofuel producing microorganisms. J. Proteome Res., 2735-2746. http://dx.doi.org/10.1021/acs.jproteome.8b00949

Concu, R., Kleandrova, V. V., Speck-Planche, A., & Cordeiro, M. (2017). Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. Nanotoxicology, 891-906. https://doi.org/10.1080/17435390.2017.1379567

Constitución, española. 1978. Boletín Oficial del Estado, BOE 311. España.

Craig, Calhoun 2002. Social Structure. En Dictionary of the Social Sciences. Oxford: Oxford University Press.

Cronin, M.T.; Jaworska, J.S.; Walker, J.D.; Comber, M.H.; Watts, C.D.; Worth, A.P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect 2003; 111: 1391-401.

Davidson, S.J.; Bergs, S.J.; Kapsner, M. Open, click, download, send … What have you agreed to? The possibilities seem endless. Computer Associations Law Conference (CLA 2001). 2001.

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res.,. http://dx.doi.org/10.1093/nar/gkv352

De Angelis, F., Pujia, A., Falcone, C., Iaccino, E., Palmieri, C., Liberale, C., Mecarini, F., Candeloro, P., Luberto, L., de Laurentiis, A., Das, G., Scala, G., & Di Fabrizio, E. (s.f.).

De las Heras, T. (2007) "Marca comunitaria", en A. Bercovitz Rodríguez-Cano (ed.) Derecho de la Competencia y Propiedad Industrial en la Unión Europea. Navarra: Aranzadi.

Dizaj, S. M., Jafari, S., & Khosroushahi, A. Y. (2014). A sight on the current nanoparticle-based gene delivery vectors. Nanoscale research letters. https://doi.org/10.1186/1556-276X-9-252

Duardo A., Pazos-Sierra A., González Díaz A., Redes complejas, Inteligencia Artificial, Bioinformática y Derecho: Un camino de ida y vuelta entre las TICS y las ciencias jurídicas y bio-moleculares, Publicia, 2015.

Duardo-Sánchez, A. 2010. Estudio de redes de derecho penal con centralidades de probabilidad de Markov. En Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, editado por H. González-Díaz. Kerala, India: Bentham.

Duardo-Sánchez, A., C. R. Munteanu, P. Riera-Fernández, A. López-Díaz, A. Pazos y H. González-Díaz. 2014. Modelización de reacciones metabólicas complejas, sistemas ecológicos y redes financieras y legales con modelos MIANN basados en descriptores de nodos Markov-Wiener. J Chem Inf Model 54 (1):16-29.

Duardo-Sanchez, A., Munteanu, C. R., Riera-Fernandez, P., Lopez-Diaz, A., Pazos, A., & Gonzalez-Diaz, H. (2014). Modeling complex metabolic reactions, ecological systems, and financial and legal networks with MIANN models based on Markov-Wiener node descriptors. Journal of chemical information and modeling, 16-29. https://doi.org/10.1021/ci400280n

Duardo-Sánchez, A; Patlewicz, G; López-Díaz, A. Current Topics on Software Use in Medicinal Chemistry: Intellectual Property, Taxes, and Regulatory Issues. Current Topics in Medicinal Chemistry, 2008, 8(18).

Duncan, G. A., & Bevan, M. A. (2015). Computational design of nanoparticle drug delivery systems for selective targeting. Nanoscale, 15332-15340. https://doi.org/10.1039/c5nr03691g

Edler, L.; Poirier, K.; Dourson, M.; Kleiner, J.; Mileson, B.; Nordmann, H.; Renwick, A.B.; Slob, W.; Walton, K.; Wurtzen, G. Mathematical modeling and quantitative methods. Food and Chemical Toxicology 2002; 40: 283–326.

Elizabeth, E., Baranwal, G., Krishnan, A. G., Menon, D., & Nair, M. (2014). ZnO nanoparticle incorporated nanostructured metallic titanium for increased mesenchymal stem cell response and antibacterial activity. Nanotechnology. https://doi.org/10.1088/0957-4484/25/11/115101

Erdozain López JC.: "Programas de ordenador", en: Bercovitz Rodríguez-Cano R; *et al.*: Manual de propiedad intelectual , 9ª edición.; Manuales; Tirant lo Blanch: Valencia, 2019. pp 2235- 2252.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect, 1361-1375. http://dx.doi.org/10.1289/ehp.5758

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognit Lett, 861-874. http://dx.doi.org/10.1016/j.patrec.2005.10.010

Fernández Novoa, C., Otero Lastres, J. M.y Botana Agra, M. (2009) Manual de la Propiedad Industrial. Madrid: Marcial Pons.

Ferreira da Costa, J., Caamaño, O., Fernández, F., García-Mera, X., Sampaio-Dias, I., Brea, J., & Cadavid, M. (2013). Synthesis and allosteric modulation of the dopamine receptor by peptide analogs of L-prolyl-L-leucyl-glycinamide (PLG) modified in the L-proline or L-proline and L-leucine scaffolds. Eur. J. Med. Chem., 146-158. http://dx.doi.org/10.1016/j.ejmech.2013.08.001

Ferreira da Costa, J., Silva, D., Caamaño, O., Brea, J., Loza, M., Munteanu, C., Pazos, A., García-Mera, X., & González-Díaz, H. (2018). Perturbation theory/machine learning model of ChEMBL data for dopamine targets: docking, synthesis, and assay of new l-prolyll-leucyl-glycinamide peptidomimetics. ACS Chem. Neurosci., 2572-2587. http://dx.doi.org/10.1021/acschemneuro.8b00083

Findlay, M. R., Freitas, D. N., Mobed-Miremadi, M., & Wheeler, K. E. (2018). Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. Environmental science. Nano, 64-71. https://doi.org/10.1039/C7EN00466D

Fischbach, M. A., & Walsh, C. T. (2009). Antibiotics for emerging pathogens. Science, 1089-1093. https://doi.org/10.1126/science.1176667

Fisher, R. (1937). The design of experiments. Oliver And Boyd: Edinburgh.

Fjodorova, N.; Novich, M.; Vrachko, M.; Kharchevnikova, N.; Zholdakova, Z.; Sinitsyna, O.; Benfenati, E. Regulatory assessment of chemicals within OECD member countries, EU and in Russia. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 2008; 26: 40-88.

Fjodorova, N.; Novich, M.; Vrachko, M.; Smirnov, V.; Kharchevnikova, N.; Zholdakova, Z.; Novikov, S.; Skvortsova, N.; Filimonov, D.; Poroikov, V.; Benfenati, E. Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 2008; 26: 201-36.

Fowler, J H., T R. Johnson, J F. Spriggs II, S. Jeon y P J. Wahlbeck. 2007. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis (15):324-346.

Fowler, J. H., y S. Jeon. 2008. The authority of Supreme Court precedent. Social Networks 30:16-30.

Gajewicz, A. (2017). What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. Nanoscale, 8435-8448. https://doi.org/10.1039/c7nr02211e

Gao, H. (2001). Application of BCUT metrics and genetic algorithm in binary QSAR analysis. J. Chem. Inf. Comput. Sci., 402-407. http://dx.doi.org/10.1021/ci000306p

García Vidal, Á. (2003) "El Derecho Europeo de la Propiedad Industrial", en S. Cámara Lapuente (ed.) Derecho privado europeo, pp. 1063-1098. Madrid: Colex.

Gómez Lozano, M. (2004) "Indicaciones geográficas protegidas", en A. Bercovitz Rodríguez-Cano (ed.) Derecho de la Competencia y Propiedad Industrial en la Unión Europea. Madrid: Thomson-Aranzadi.

Gonzalez-Diaz, H., Arrasate, S., Gomez-SanJuan, A., Sotomayor, N., 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. Current topics in medicinal chemistry, 1713-1741. https://doi.org/10.2174/1568026611313140011

Gonzalez-Diaz, H., Herrera-Ibata, D. M., Duardo-Sanchez, A., Munteanu, C. R., Orbegozo-Medina, R. A., & 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, 744-755. https://doi.org/10.1021/ci400716y

Graham, D. J. (2002). Information and organic molecules: structure considerations via integer statistics. J. Chem. Inf. Comput. Sci, 215-221. https://doi.org/10.1021/ci0102923

Graham, D. J. (2005). Information content in organic molecules: aggregation states and solvent effects. Journal of chemical information and modeling, 1223-1236. https://doi.org/10.1021/ci050101m

Graham, D. J., Malarkey, C., & Schulmerich, M. V. (2004). Information content in organic molecules: quantification and statistical structure via Brownian processing. J. Chem. Inf. Comput. Sci, 1601-1611. https://doi.org/10.1021/ci0400213

Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem., 2121-2132. http://dx.doi.org/10.1002/jcc.23361

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. SIGKDD Explor., 10-18. http://dx.doi.org/10.1145/1656274.1656278

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. Bioinformatics, 822-830. https://doi.org/10.1093/bioinformatics/btq037

Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997). GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. J. Chem. Inf. Comput. Sci., 306-310. http://dx.doi.org/10.1021/ci960047x

He, J., He, C., Zheng, C., Wang, Q., & Ye, J. (2019). Plasmonic nanoparticle simulations and inverse design using machine learning. Nanoscale, 17444-17459. https://doi.org/10.1039/c9nr03450a

Hem, E.; Bordahl, P.E. Max Sanger - father of the modern caesarean section. Gynecologic & Obstetric Investigation 2003; 55: 127-9.

Hemmateenejad, B., Akhond, M., Miri, R., & Shamsipur, M. (2003). Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines(nifedipine analogous). J. Chem. Inf. Comput. Sci., 1328-1334. http://dx.doi.org/10.1021/ci025661p

Herrera-Ibata, D. M., Pazos, A., Orbegozo-Medina, R. A., Romero-Duran, F. J., & 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. https://doi.org/10.1016/j.biosystems.2015.04.007

Hill, T., & Lewicki, P. (2005). Methods and Applications. StatSoft.

Hill, T., & Lewicki, P. (2006). STATISTICS Methods and applications. A comprehensive reference for science, industry and data mining. StatSoft: Tulsa, 813.

Holden, P. A., Nisbet, R. M., Lenihan, H. S., Miller, R. J., Cherr, G. N., Schimel, J. P., & Gardea-Torresdey, J. L. (2013). Ecological nanotoxicology: integrating nanomaterial hazard considerations across the subcellular, population, community, and ecosystems levels. Acc Chem Res, 813-822. https://doi.org/10.1021/ar300069t

Hossain, S. T., & Mukherjee, S. K. (2012). CdO nanoparticle toxicity on growth, morphology, and cell division in Escherichia coli. Langmuir, 16614-16622. https://doi.org/10.1021/la302872y

Hossain, S. T., & Mukherjee, S. K. (2013). Toxicity of cadmium sulfide (CdS) nanoparticles against Escherichia coli and HeLa cells. Journal of hazardous materials, 1073-1082. https://doi.org/10.1016/j.jhazmat.2013.07.005

Hsu, D. D. (10 de 2013). Chemicool Periodic Table. http://www.chemicool.com/

Hsu, J. C., Naha, P. C., Lau, K. C., Chhour, P., Hastings, R., Moon, B. F., Stein, J. M., Witschey, W. T., McDonald, E. S., Maidment, A., & Cormode, D. P. (2018). An all-in-one nanoparticle (AION) contrast agent for breast cancer screening with DEM-CT-MRI-NIRF imaging. Nanoscale, 17236-17248. https://doi.org/10.1039/c8nr03741h

Hu, X., Cook, S., Wang, P., & Hwang, H. M. (2009). In vitro evaluation of cytotoxicity of engineered metal oxide nanoparticles. The Science of the total environment, 3070-3072. https://doi.org/10.1016/j.scitotenv.2009.01.033

Hubble, L. J., Cooper, J. S., Sosa-Pintos, A., Kiiveri, H., Chow, E., Webster, M. S., & . . . Raguse, B. (2015). High-throughput fabrication and screening improves gold nanoparticle chemiresistor sensor performance. ACS combinatorial science, 120-129. https://doi.org/10.1021/co500129v

Huberty, C. J., & Olejnik, S. (2006). Applied MANOVA and discriminant analysis. New Jersey: John Wiley & Sons.

Inbaraj, B. S., Kao, T. H., Tsai, T. Y., Chiu, C. P., Kumar, R., & Chen, B. H. (2011). The synthesis and characterization of poly(gamma-glutamic acid)-coated magnetite nanoparticles and their effects on antibacterial activity and cytotoxicity. Nanotechnology. https://doi.org/10.1088/0957-4484/22/7/075101

International Legal Protection for Programas de ordenador. http://www.programas de ordenadorprotection.com/ (25/9/2007).

IPR-Helpdesk. Programas de ordenador Copyright Licensing. http://www.ipr-helpdesk.org/docs/docs.EN/programas de ordenadorCopyrightLicensing.html (2007).

Jacob, R, D Koschützki, K.A Lehmann, L Peeters, y D Tenfelde-Podehl. 2005. Algorithms for Centrality Indices. En Network Analysis: Methodological Foundations, editado por U. Brandes y T. Erlebach. Berling: Springer.

Jagiello, K., Grzonkowska, M., Swirog, M., Ahmed, L., Rasulev, B., Avramopoulos, A., Papadopoulos, M. G., Leszczynski, J., & Puzyn, T. (2016). Advantages and limitations of classic and 3D QSAR approaches in nano-QSAR studies based on biological activity of fullerene derivatives. Journal of nanoparticle research. https://doi.org/10.1007/s11051-016-3564-1

Jansen, F.K.; Freytag, G. Immune reactions to fractions of crystalline insulin. II. May peri-insulitis be produced by an antigen different from true Sanger insulin? Diabetologia 1973; 9: 191-6.

Javier, Pamparacuatro. 2015. En torno a la crisis del Derecho. UNED. Revista de Derecho Político (92):165-194.

Jaworska, J.; Comber, M.; Auer, C.; Van Leeuwen, C.J. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environmental Health Perspectives Mini Monografo 2003; 10: 1358–60.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. Nature, 651-654. https://doi.org/10.1038/35036627

Junker, B. H., Koschutzki, D., & Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. BMC bioinformatics. https://doi.org/10.1186/1471-2105-7-219

Kalliokoski, T., Kramer, C., Kramer, A., & Peter, G. (2013). Comparability of mixed IC50 data - a statistical analysis. PLoS One. https://doi.org/10.1371/journal.pone.0061007

Kennard, R., & Stone, L. (1969). Computer aided design of experiments. Technometrics, 137-148. http://dx.doi.org/10.1080/00401706.1969.10490666

Kleandrova, V., Luan, F., González-Díaz, H., Ruso, J., Speck-Planche, A., & Cordeiro, M. (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., 14686-14694. http://dx.doi.org/10.1021/es503861x

Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5):604-63.

Koeter, H.B.; Visser, R. Work in OECD on chemical safety: approaches for human risk assessment. Industrial Health 2000; 38: 109-19.

Koschützki, Dirk, Katharina A. Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podeh y Oliver Zlotowski. 2005. Centrality Indices. En Network Analysis: Methodological Foundations, editado por U. Brandes y T. Erlebach. Berlín: Springer.

Larocque, M., Chenard, T., & Najmanovich, R. (2014). A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. BMC systems biology. https://doi.org/10.1186/s12918-014-0117-z

Le, T. C., Yin, H., Chen, R., Chen, Y., Zhao, L., Casey, P. S., Chen, C., & Winkler, D. A. (2016). An Experimental and Computational Approach to the Development of ZnO Nanoparticles that are Safe by Design. Small, 3568-3577. https://doi.org/10.1002/smll.201600597

Levy, V. (2006). Grant, R.M. Antiretroviral therapy for hepatitis B virus-HIV-coinfected patients: Promises and pitfalls. Clin. Infect. Dis., 904-910. http://dx.doi.org/10.1086/507532

Ley, Impuesto General. 2003. Ley 58/2003. España: Boletín Oficial del Estado, BOE 302.

Li, Y., Zhang, X., & Cao, D. (2015). Nanoparticle hardness controls the internalization pathway for drug delivery. Nanoscale, 2758-2769. https://doi.org/10.1039/c4nr05575f

Luan, F., Kleandrova, V. V., Gonzalez-Diaz, H., Ruso, J. M., Melo, A., Speck-Planche, A., & Cordeiro, M. N. (2014). Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. Nanoscale, 10623-10630. https://doi.org/10.1039/c4nr01285b

Luan, F., Kleandrova, V., González-Díaz, H., Ruso, J., Melo, A., Speck-Planche, A., & Cordeiro, M. (2014). Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR perturbation approach. Nanoscale, 10623-10630. http://dx.doi.org/10.1039/C4NR01285B

Macías Martín, J. (2007) "Nociones sobre el Convenio de la patente europea", en A. Bercovitz (ed.) Derecho de la Competencia y de la Propiedad Industrial en la Unión Europea, pp. 203-232. Navarra: Aranzadi.

Mallawaarachchi, S., Liu, Y., Thang, S. H., Cheng, W., & Premaratne, M. (2019). Machine learning based temperature prediction of poly(N-isopropylacrylamide)-capped plasmonic nanoparticle solutions. Physical chemistry chemical physics : PCCP, 24808-24819. https://doi.org/10.1039/c9cp04544a

Manganelli, S., & Benfenati, E. (2017). Nano-QSAR Model for Predicting Cell Viability of Human Embryonic Kidney Cells. Methods in molecular biology, 275-290. https://doi.org/10.1007/978-1-4939-6960-9_22

Manganelli, S., Leone, C., Toropov, A. A., Toropova, A. P., & Benfenati, E. (2016). QSAR model for predicting cell viability of human embryonic kidney cells exposed to SiO(2)

nanoparticles. Chemosphere, 995-1001. https://doi.org/10.1016/j.chemosphere.2015.09.086

Martínez-Arzate, S., Tenorio-Borroto, E., Barbabosa Pliego, A., Díaz-Albiter, H., Vázquez-Chagoyán, J., & 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., 4093-4103. http://dx.doi.org/10.1021/acs.jproteome.7b00477

Morcon, C.; Roughton, A.; Gaham, J. The Modern Law of Trade Marks. Butterworth: London 2005.

Mu, Q., Annapragada, A., Srivastava, M., Li, X., Wu, J., Thiviyanathan, V., H., W., A., W., D., G., A., A., & Vigneswaran, N. (2016). Conjugate-SELEX: A High-throughput Screening of Thioaptamer-liposomal Nanoparticle Conjugates for Targeted Intracellular Delivery of Anticancer Drugs. Molecular therapy. Nucleic acid. https://doi.org/10.1038/mtna.2016.81

Nabil, A., Elshemy, M. M., Asem, M., Abdel-Motaal, M., Gomaa, H. F., Zahran, F., Uto, K., & Ebara, M. (2020). Zinc Oxide Nanoparticle Synergizes Sorafenib Anticancer Efficacy with Minimizing Its Cytotoxicity. Oxidative medicine and cellular longevity. https://doi.org/10.1155/2020/1362104

Nagar, S. D., Aggarwal, B., Joon, S., Bhatnagar, R., & Bhatnagar, S. (2016). A Network Biology Approach to Decipher Stress Response in Bacteria Using Escherichia coli As a Model. OMICS, 310-324. https://doi.org/10.1089/omi.2016.0028

Najer, A., Wu, D., Nussbaumer, M. G., Schwertz, G., Schwab, A., Witschel, M. C., Schafer, A., Diederich, F., Rottmann, M., Palivan, C., Beck, H., & Meier, W. (2016). An amphiphilic graft copolymer-based nanoparticle platform for reduction-responsive anticancer and antimalarial drug delivery. Nanoscale, 14858-14869. https://doi.org/10.1039/c6nr04290b

NCBI Resource Coordinators. (2016). Database resources of the national center for biotechnology information. Nucleic Acids Res, D7-19.

Newman, M. 2003. The Structure and Function of Complex Networks. SIAM Review 56:167-256.

Nocedo-Mena, D., Cornelio, C., Camacho-Corona, M. D., Garza-Gonzalez, E., Waksman de Torres, N., Arrasate, S., & . . . Gonzalez-Diaz, H. (2019). Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. Journal of chemical information and modeling, 1109-1120. https://doi.org/10.1021/acs.jcim.9b00034

Nocedo-Mena, D., Cornelio, C., Camacho-Corona, M., Garza-González, E., Waksman de Torres, N., Arrasate, S., Sotomayor, N., Lete, 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., 1109-1120. http://dx.doi.org/10.1021/acs.jcim.9b00034

OECD. Organization for Economic Co-operation and Development. Articles of The Model Convention Whit Respect to Taxes on Income and on Capital. http://www.oecd.org/dataoecd/50/49/35363840.pdf (15/9/2007).

OECD. Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationchip ((Q)SAR) Models in the Assessment of New and Existing Chemicals. In OECD Environment Health and Safety Publications: Paris, 2006.

Ojha, P. K., Kar, S., Roy, K., & Leszczynski, J. (2019). Toward comprehension of multiple human cells uptake of engineered nano metal oxides: quantitative inter cell line uptake specificity (QICLUS) modeling. Nanotoxicology, 14-34. https://doi.org/10.1080/17435390.2018.1529836

Organization for Economic Co-operation and Development(OECD). (2007). Guidance document on the validation of (quantitative) structure-activity relationship ((Q)SAR) models. In: OECD Series on Testing and Assessment. OECD Publishing: Paris, 55-65.

Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C., Cordeiro, M., & Gonzalez-Diaz, H. (2020). PTML Multi-Label Algorithms: Models, Software, and Applications. Curr Top Med Chem, 2326-2337. https://doi.org/10.2174/1568026620666200916122616

Parlamento Europeo, Consejo y Comisión. 2013. Guía práctica conjunta para las personas que participan en la elaboración de la legislación de la Unión Europea. Bruselas.

Porcelli, C.; Boriani, E.; Roncaglioni, A.; Chana, A.; Benfenati, E. Regulatory perspectives in the use and validation of QSAR. A case study: DEMETRA model for Daphnia toxicity. Environmental Science and Technology 2008; 42: 491-6.

Porcelli, C.; Roncaglioni, A.; Chana, A.; Benfenati, E. A comparison of DEMETRA individual QSARs with an index for evaluation of uncertainty. Chemosphere 2008; 71: 1845-52.

Pramanik, A., Laha, D., Bhattacharya, D., Pramanik, P., & Karmakar, P. (2012). A novel study of antibacterial activity of copper iodide nanoparticle mediated by DNA and membrane damage. Colloids and surfaces. B, Biointerfaces, 50-55. https://doi.org/10.1016/j.colsurfb.2012.03.021

Premanathan, M., Karthikeyan, K., Jeyasubramanian, K., & Manivannan, G. (2011). Selective toxicity of ZnO nanoparticles toward Gram-positive bacteria and cancer cells by apoptosis through lipid peroxidation. Nanomedicine, 184-192. https://doi.org/10.1016/j.nano.2010.10.001

Pundir, S., Martin, M., & O'Donovan, C. (2016). UniProt Tools. Curr. Protoc. Bioinformatics, , 1-15.

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., M., H. H., Toropov, A., & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nature nanotechnology, 175-178. https://doi.org/10.1038/nnano.2011.10

Quevedo-Tumailli, V., Ortega-Tenezaca, B., & González-Díaz, H. (2018). Chromosome gene orientation inversion networks (goins) of plasmodium proteome. J. Proteome Res., 1258-1268. http://dx.doi.org/10.1021/acs.jproteome.7b00861

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science, 1551-1555. https://doi.org/10.1126/science.1073374

Richards, T.W.; Peirce, B.O.; Baxter, G.P. Charles Robert Sanger. Science 1912; 35: 532.

Riera-Fernández, P., C. R. Munteanu, M. Escobar, F. Prado-Prado, R. Martín-Romalde, D. Pereira, K. Villalba, A. Duardo-Sánchez, y H. González-Díaz. 2012. Nuevos modelos de Entropía de Markov-Shannon para evaluar la calidad de la conectividad en redes complejas: de la vía molecular a la celular, de las redes parásito-huésped, neuronales, industriales y jurídico-sociales. J Theor Biol 293:174-88.

Riera-Fernandez, P., Munteanu, C. R., Escobar, M., Prado-Prado, F., Martin-Romalde, R., Pereira, D., K., V., Duardo-Sanchez, A., & 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., 174-188. https://doi.org/10.1016/j.jtbi.2011.10.016

Rogers, D., & Hopfinger, A. (1994). Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J. Chem. Inf. Comput. Sci., 854-866. http://dx.doi.org/10.1021/ci00020a020

Romero-Durán, F., Alonso, N., Yañez, M., Caamaño, O., García-Mera, X., & González-Díaz, H. (2016). Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. Neuropharmacology, 270-278. http://dx.doi.org/10.1016/j.neuropharm.2015.12.019

Rowland, D.; Campbell, A. Supply of Programas de ordenador: Copyright and Contract Issues. International Journal of Law and Information Technolog 2002; 10: 23-40(18).

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst., 22-29. http://dx.doi.org/10.1016/j.chemolab.2015.04.013

Rudnick, P.A.; Wang, Y.; Evans, E.; Lee, C.S.; Balgley, B.M. Large scale analysis of MASCOT results using a Mass Accuracy-based THreshold (MATH) effectively improves data interpretation. J Proteome Res 2005; 4: 1353-60.

Ruparelia, J. P., Chatterjee, A. K., Duttagupta, S. P., & Mukherji, S. (2008). Strain specificity in antimicrobial activity of silver and copper nanoparticles. Acta Biomater, 707-716. https://doi.org/10.1016/j.actbio.2007.11.006

Rybinska-Fryca, A., Mikolajczyk, A., & Puzyn, T. (2020). Structure-activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept. Nanoscale, 20669-20676. https://doi.org/10.1039/d0nr05220e

Santana, R., Zuluaga, R., Ganan, P., Arrasate, S., Onieva, E., & Gonzalez-Diaz, H. (2019). Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. Nanoscale, 21811-21823. https://doi.org/10.1039/c9nr05070a

Santana, R., Zuluaga, R., Ganan, P., Arrasate, S., Onieva, E., & Gonzalez-Diaz, H. (2020). Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. Nanoscale, 13471-13483. https://doi.org/10.1039/d0nr01849j

Santana, R., Zuluaga, R., Ganan, P., Arrasate, S., Onieva, E., Montemore, M. M., & Gonzalez-Diaz, H. (2020). PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. Molecular pharmaceutics, 2612-2627. https://doi.org/10.1021/acs.molpharmaceut.0c00308

Schulze, J., Kuhn, S., Hendrikx, S., Schulz-Siegmund, M., Polte, T., & Aigner, A. (2018). Spray-Dried Nanoparticle-in-Microparticle Delivery Systems (NiMDS) for Gene Delivery, Comprising Polyethylenimine (PEI)-Based Nanoparticles in a Poly(Vinyl Alcohol) Matrix. Small. https://doi.org/10.1002/smll.201701810

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shlar, I., Poverenov, E., Vinokur, Y., Horev, B., Droby, S., & Rodov, V. (2015). High-Throughput Screening of Nanoparticle-Stabilizing Ligands: Application to Preparing

Antimicrobial Curcumin Nanoparticles by Antisolvent Precipitation. Nano-micro letters, 68-79. https://doi.org/10.1007/s40820-014-0020-6

Sizochenko, N., Gajewicz, A., Leszczynski, J., & Puzyn, T. (2018). Reply to the comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models" by D. A. Tasi. Nanoscale, 20867-20868.

Sizochenko, N., Leszczynska, D., & Leszczynski, J. (2017). Modeling of Interactions between the Zebrafish Hatching Enzyme ZHE1 and A Series of Metal Oxide Nanoparticles: Nano-QSAR and Causal Analysis of Inactivation Mechanisms. Nanomaterials. https://doi.org/10.3390/nano7100330

Sizochenko, N., Mikolajczyk, A., Jagiello, K., Puzyn, T., Leszczynski, J., & Rasulev, B. (2018). How the toxicity of nanomaterials towards different species could be simultaneously evaluated: a novel multi-nano-read-across approach. Nanoscale, 582-591. https://doi.org/10.1039/c7nr05618d

Snedecor, G., & Cochran, W. (1967). Statistical Methods. Oxford and IBH Publishing Co: New Delhi, 593.

Speck-Planche, A., & Cordeiro, M. (2013). Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. Curr. Top. Med. Chem., 1656-1665. http://dx.doi.org/10.2174/15680266113139990116

Speck-Planche, A., & Cordeiro, M. (2014). Chemoinformatics for medicinal chemistry: in silico model to enable the discovery of potent and safer anti-cocci agents. Future Med. Chem, 2013-2028. http://dx.doi.org/10.4155/fmc.14.136

Speck-Planche, A., & Cordeiro, M. (2017). De novo computational design of compounds virtually displaying potent antibacterial activity and desirable in vitro ADMET profiles. Med. Chem. Res., 2345-2356. http://dx.doi.org/10.1007/s00044-017-1936-4

Speck-Planche, A., & Cordeiro, M. (2017). Erratum to: Fragment based in silico modeling of multi-target inhibitors against breast cancer-related proteins. Mol. Divers., 525. https://doi.org/PMID: 28766255

Speck-Planche, A., & Cordeiro, M. (2017). Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. Mol. Divers., 511-523. http://dx.doi.org/10.1007/s11030-017-9731-1

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2015). Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. Nanomedicine, 193-204. https://doi.org/10.2217/nnm.14.96

Speck-Planche, A., Kleandrova, V., Ruso, J., & Cordeiro, M. (2016). First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple gram-positive pathogens. J. Chem. Inf. Model., 588-598. http://dx.doi.org/10.1021/acs.jcim.5b00630

Stanwix, H. (2012). DNA nanoparticles could help to treat arthritis and other inflammatory disorders. Nanomedicine, 945-946. https://doi.org/10.2217/nnm.12.93

Steering Committee for Intellectual Property Issues in Programas de ordenador Computer Science and Telecommunications Board Commission on Physical Sciences, M., and Applications National Research Council. Intellectual Property Issues In programas de ordenador. National Academy Press: Washington, D.C 1991.

Stewart, S. M. (1989) International Copyright and Neighbouring Rights. London: Butterworths.

Story, A. Intellectual Property and Computer Programas de ordenador. In Intellectual Property Rights and Sustainable Development, ICTSD-UNCTAD, Ed. Imprimerie Typhon: Chavanod, 2004; p 12.

Strait, B. J., & Dewey, T. G. (1996). The Shannon information entropy of protein sequences. Biophys, 148-155. https://doi.org/10.1016/S0006-3495(96)79210-X

Sun, B., Fernandez, M., & Barnard, A. S. (2017). Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. Journal of chemical information and modeling, 2413-2423. https://doi.org/10.1021/acs.jcim.7b00272

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, Palyulin, V., . . . Tetko, I. (2011). Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des, 533-554. https://doi.org/10.1007/s10822-011-9440-2

Sutherland, J., O'Brien, L., & Weaver, D. (2003). Spline-fitting with a genetic algorithm: a method for developing classification structureactivity relationships. J. Chem. Inf. Comput. Sci., 1906-1915. http://dx.doi.org/10.1021/ci034143r

Taglietti, A., Diaz Fernandez, Y. A., Amato, E., Cucca, L., Dacarro, G., Grisoli, P., Necchi, V., Pallavicini, P., Pasotti, L., & Patrini, M. (2012). Antibacterial activity of glutathione-coated silver nanoparticles against Gram positive and Gram negative bacteria. Langmuir, 8140-8148. https://doi.org/10.1021/la3003838

Tanaka, M., Hikiba, S., Yamashita, K., Muto, M., & Okochi, M. (2017). Array-based functional peptide screening and characterization of gold nanoparticle synthesis. Acta biomaterialia, 495-506. https://doi.org/10.1016/j.actbio.2016.11.037

Tasi, D. A., Csontos, J., Nagy, B., Konya, Z., & Tasi, G. (2018). Comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models. Nanoscale, 20863-20866.

Toropov, A. A., Toropova, A. P., Benfenati, E., Gini, G., Puzyn, T., Leszczynska, D., & Leszczynski, J. (2012). Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria Escherichia coli. Chemosphere, 1098-1102.

Toropova, A. P., Toropov, A. A., Rallo, R., Leszczynska, D., & Leszczynski, J. (2015). Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotoxicol. Environ. Saf., 39-45. https://doi.org/10.1016/j.ecoenv.2014.10.003

Toropova, A. P., Toropov, A. A., Veselinovic, A. M., Veselinovic, J. B., Benfenati, E., Leszczynska, D., & Leszczynski, J. (2016). Nano-QSAR: Model of mutagenicity of fullerene as a mathematical function of different conditions. Ecotoxicol Environ Saf, 32-36. https://doi.org/10.1016/j.ecoenv.2015.09.038

Van Den Berg, H. A. (2018). Occam's razor: from Ockham's via moderna to modern data science. Science progress, 261-272. https://doi.org/10.3184/003685018X15295002645082

Vásquez-Domínguez, E., Armijos-Jaramillo, V., Tejera, E., & González-Díaz, H. (2019). Multioutput perturbation-theory machine learning (ptml) model of chembl data for antiretroviral compounds. Mol. Pharm., 4200-4212. http://dx.doi.org/10.1021/acs.molpharmaceut.9b00538

Venkatasubramanian, V., & Sundaram, A. (2002). Genetic algorithms: introduction and applications. In: Encyclopedia of Computational Chemistry;. Wiley & Sons Inc.:Hoboken, 2.

Vert, M., Doi, Y., Hellwich, K.-H., Hess, M., Hodge, P., Kubisa, P., Rinaudo, M., & François, S. (2012). Terminology for biorelated polymers and applications (IUPAC Recommendations 2012). Pure Appl. Chem, 377-410. https://doi.org/10.1351/PAC-REC-10-12-04

Vidal, M., Cusick, M. E., & Barabasi, A. L. (2011). Interactome networks and human disease. Cell, 986-998. https://doi.org/10.1016/j.cell.2011.02.016

Villaverde, J. J., Sevilla-Moran, B., Lopez-Goti, C., Alonso-Prados, J. L., & Sandin-Espana, P. (2018). Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. The Science of the total environment, 1530-1539. https://doi.org/10.1016/j.scitotenv.2018.04.033

Walker, J.; Jaworska, J.; Comber, M.; Schultz, T.; Dearden, J. Guidelines for developing and using quantitative structure–activity relationships. Environmental Toxicology and Chemistry 2003; 22: 1653–65.

Wang, F., Wang, X., Gao, L., Meng, L. Y., Xie, J. M., Xiong, J. W., & Luo, Y. (2019). Nanoparticle-mediated delivery of siRNA into zebrafish heart: a cell-level investigation on the biodistribution and gene silencing effects. Nanoscale, 18052-18064. https://doi.org/10.1039/c9nr05758g

Wang, H., Liu, K., Chen, K. J., Lu, Y., Wang, S., Lin, W. Y., Guo, F., Kamei, K., Chen, Y., Ohashi, M., Wang, M., Garcia, M., Zhao, X., Shen, C., & Tseng, H. (2010). A rapid pathway toward a superb gene delivery system: programming structural and functional diversity into a supramolecular nanoparticle library. ACS nano, 6235-6243. https://doi.org/10.1021/nn101908e

Wellman, Barry, y Stephen D Berkowitz. 1988. Social Structures: A Network Approach. Cambridge: Cambridge University Press.

Westkamp, G.N. Protección del material biológico mediante derechos de autor. ¿Vuelta de la Bioinformática a los Derechos de autor en la biotecnología? IPR-Helpdesk Bulletin 2005.

White, Harrison, Scott Boorman y Ronald Breiger. . "." 1976. Social Structure from Multiple Networks: I Blockmodels of Roles and Positions. American Journal of Sociology 81:730-780

Wilks, S. (1932). Certain generalizations in the analysis of variance. Biometrika, 471-494. http://dx.doi.org/10.1093/biomet/24.3-4.471

WIPO. Madrid Agreement Concerning the International Registration of Marks of April 14, 1891, as revised at Brussels on December 14, 1900, at Washington on June 2, 1911, at The Hague on November 6, 1925, at London on June 2, 1934, at Nice on June 15, 1957, and at Stockholm on July 14, 1967,1 and as amended on September 28, 1979. http://www.wipo.int/madrid/en/legal_texts/trtdocs_wo015.html

WIPO. Protocol Relating to the Madrid Agreement Concerning the International Registration of Marks adopted at Madrid on June 27, 1989 and amended on October 3, 2006. http://www.wipo.int/madrid/en/legal_texts/trtdocs_wo016.html

Wong, M. S., Chen, C. W., Hsieh, C. C., Hung, S. C., Sun, D. S., & Chang, H. H. (2015). Antibacterial property of Ag nanoparticle-impregnated N-doped titania films under visible light. Scientific reports. https://doi.org/10.1038/srep11978

Worth, A.P.; Hartung, T.; Van Leeuwen, C.J. The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. SAR & QSAR in Environmental Research 2004; 15: 345-58.

Worth, A.P.; Van Leeuwen, C.J.; Hartung, T. The prospects for using (Q)SARs in a changing political environment--high expectations and a key role for the European Commission's joint research centre. SAR & QSAR in Environmental Research 2004; 15: 331-43.

Yan, T., Sun, B., & Barnard, A. S. (2018). Predicting archetypal nanoparticle shapes using a combination of thermodynamic theory and machine learning. Nanoscale, 21818-21826. https://doi.org/10.1039/c8nr07341d

Yang, R., Gui, X., Xiong, Y., Gao, S., & Yan, Y. (2014). Impact of hepatitis B virus infection on HIV response to antiretroviral therapy in a Chinese antiretroviral therapy center. Int. J. Infect. Dis., 29-34. http://dx.doi.org/10.1016/j.ijid.2014.07.018

Yang, Y., Song, W., Chen, Z., Li, Q., & Liu, L. (2019). Ameliorative effect of synthesized silver nanoparticles by green route method from Zingiber zerumbet on mycoplasmal pneumonia in experimental mice. Artificial cells, nanomedicine, and biotechnology, 2146-2154. https://doi.org/10.1080/21691401.2019.1620757

Zagrebelsky, G. 1992. Il diritto mite. Legge, diritti, giustizia. Italia: Einaudi.

Zelepukin, I. V., Yaremenko, A. V., Shipunova, V. O., Babenyshev, A. V., Balalaeva, I. V., Nikitin, P. I., Deyev, S., & Nikitin, M. P. (2019). Nanoparticle-based drug delivery via RBC-hitchhiking for the inhibition of lung metastases growth. Nanoscale, 1636-1646. https://doi.org/10.1039/c8nr07730d

Zhao, Y., Chen, Z., Chen, Y., Xu, J., Li, J., && Jiang, X. (2013). Synergy of non-antibiotic drugs and pyrimidinethiol on gold nanoparticles against superbugs. J Am Chem Soc, 12940-12943. https://doi.org/10.1021/ja4058635

Zhen, J. B., Kang, P. W., Zhao, M. H., & Yang, K. W. (2020). Silver Nanoparticle Conjugated Star PCL-b-AMPs Copolymer as Nanocomposite Exhibits Efficient Antibacterial Properties. Bioconjug Chem, 51-63. https://doi.org/10.1021/acs.bioconjchem.9b00739

Anexos

Anexo I

ANEXOS

ANEXO 1

7. MANUAL DE USUARIO

Ref. IFPTML Studio a new software for Information Fusion, Pertubation Theory, Machine Learning Analysis in Molecular, Biomedical, and Social Sciences. Ortega-Tenezaca B, Duardo-Sánchez, A., Munteanu, C.R., and González-Díaz H, J Chem Inf Model. 2022, *in preparation*.

IFPTML Studio v1.0.1, es un software desarrollado para el cálculo de modelos IFPMTL. El software funciona bajo el sistema operativo Windows y no requiere instalación. La capacidad de procesamiento es directamente proporcional con los recursos computacionales del equipo informático donde se lo ejecute.

La interfaz de usuario está compuesta por cuatro grupos de botones principales y un grupo secundario de información:

- Source File
    - Open File
- IF Preprocessing
    - Data Enrichment
    - Data Curation
    - Vars Fusion
- PT Processing
    - Multicontional Discretization
    - PTO´s Calculation
- ML Analisys
    - IFPTML Trainning
- Support
    - About
    - HomePage

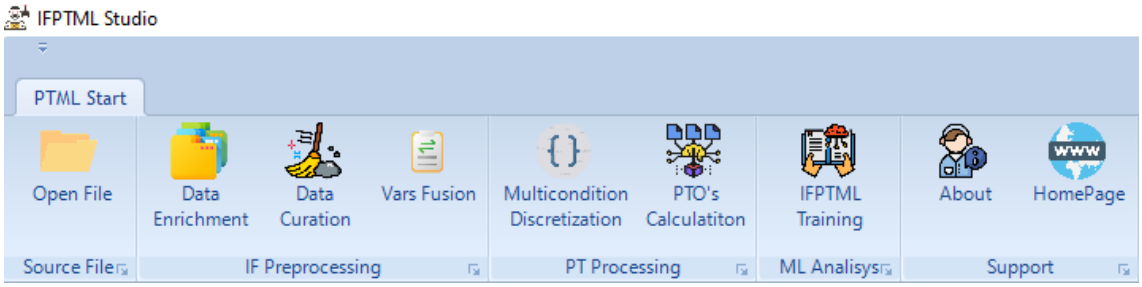


Figura 33: Menú principal del software IFPTML v 1.0.1

7.1. Source File

Contiene un botón para la carga del archivo principal con la información que será procesada dentro del software

7.1.1. Open File - Carga del archivo principal

Open File es un botón que permite la carga de un archivo de Excel con extensión xls o xlsx. Por defecto el software obtendrá los datos de la primera hoja del archivo seleccionado. Se requiere que el archivo no tenga columnas con nombres repetidos.

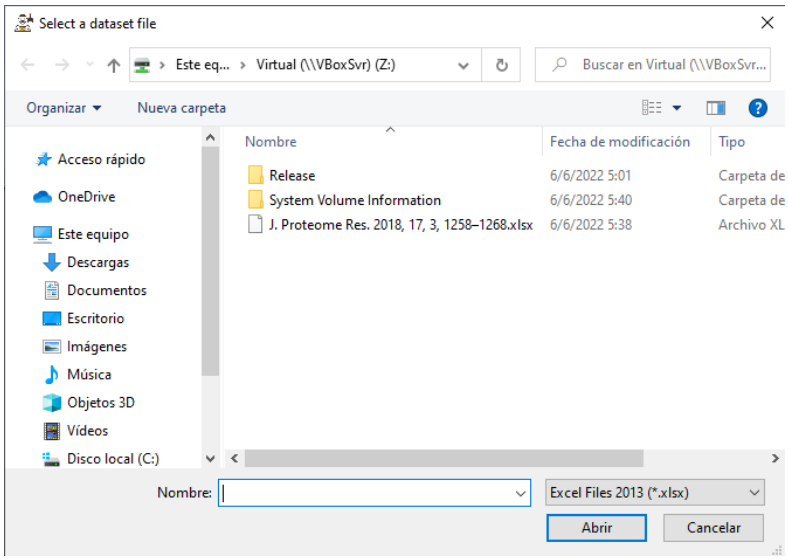


Figura 34: Ventana de selección de archivo principal

El equipo en el que se ejecuta IFPTML Studio v1.0.1 no requiere tener instalado ninguna licencia de Microsoft Excel

El tiempo de carga del archivo dependerá del número de casos que contenga el archivo y de la potencia de cómputo del equipo desde el cual se ejecuta. Una vez cargado el archivo, se mostrará en la pantalla principal del software, una muestra de los datos que contiene el archivo cargado.

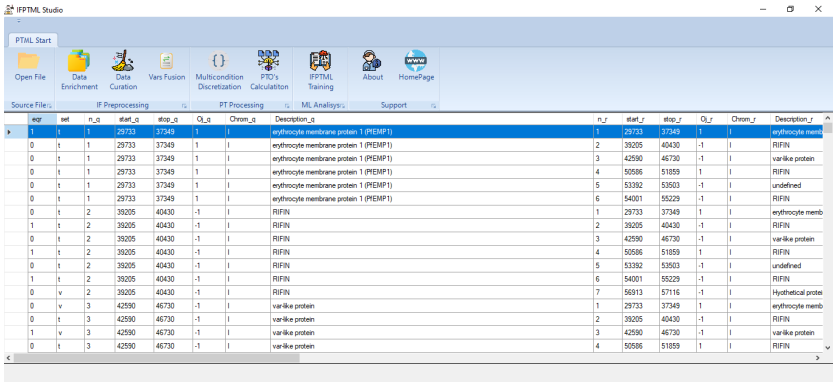


Figura 35: Pantalla principal con carga de datos

Se toma como ejemplo de carga, el archivo de Microsoft Excel de la publicación “Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome” de Journal Research Proteome.

7.2. IF Preprocessing

El menú IF Preprocessing, presenta tres opciones para el tratamiento de la fusión de información.

Data Enrichment (Enriquecimiento de datos)

La opción “Data Enrichment” otorga características que permiten extender el conjunto de datos principal a partir de un archivo adicional al cargado en el menú “source file”.

1. Al ingresar en el menú se muestra la opción de selección de un nuevo archivo similar véase figura 2.
2. Una vez seleccionado el archivo se muestra en la parte superior izquierda el nombre del archivo y el número de casos del archivo principal y en la parte superior derecha el número de casos del archivo recientemente seleccionado.

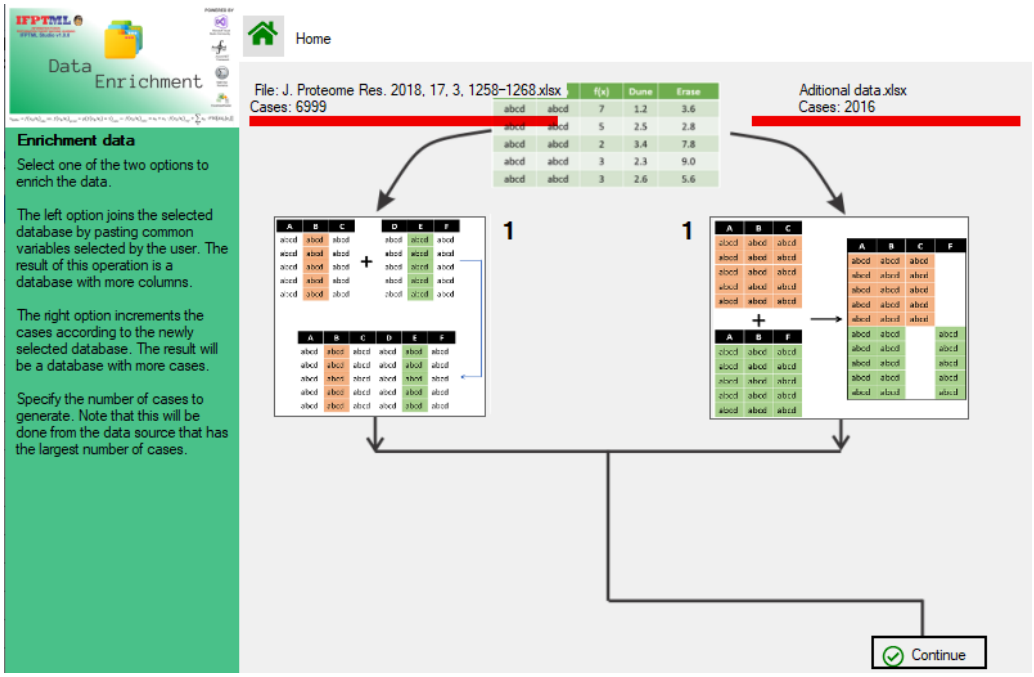


Figura 36: Carga de archivo adicional para enriquecer el archivo principal

En el ejemplo cargado se muestra que el archivo principal tiene 6999 casos y el archivo secundario con 2016 casos

3. A continuación, el usuario deberá elegir la forma en la que se extenderán los datos, de forma horizontal o vertical, para lo cual deberá hacer clic en una de las imágenes marcadas con el número 1

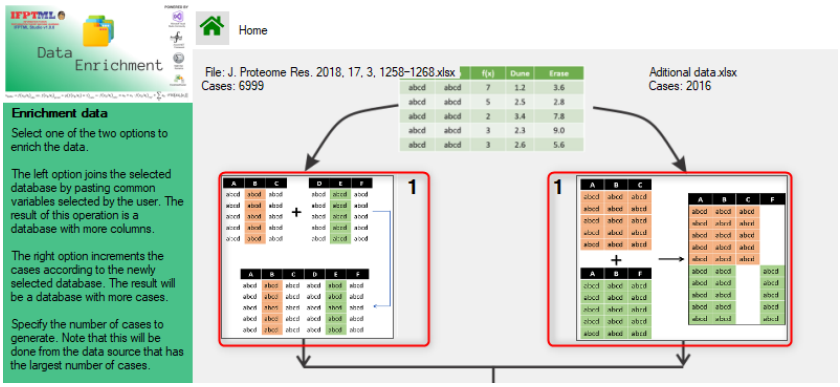


Figura 37: Selección horizontal o vertical de extensión de los datos

Se requiere hacer clic en una de las imágenes según la necesidad del investigador.

7.3. Enriquecimiento de datos horizontales

Esta opción permite agregar nueva información al archivo principal, de acuerdo con uno o varios campos en común entre los dos archivos. Una vez efectuado este proceso el archivo principal puede incrementar sus columnas, de igual manera cabe la posibilidad de establecer el número total de casos bajo las siguientes premisas:

- Si el número de casos seleccionado es mayor al número de casos del archivo principal, se adicionarán casos aleatorios del mismo archivo
- Si el número de casos seleccionado es menor al número de casos del archivo principal, se eliminarán los últimos casos del archivo
- Si el número de casos seleccionado es igual al número de casos del archivo principal, el software no realizará ningún cambio

Para realizar este proceso se realizarán los siguientes pasos:

1. Se seleccionará la imagen de la derecha marcada con el número 1
2. Se mostrará la nueva ventana con un número de casos del archivo principal, el cual puede ser configurado de acuerdo con las necesidades del usuario, el cual se identificará con el número 2.

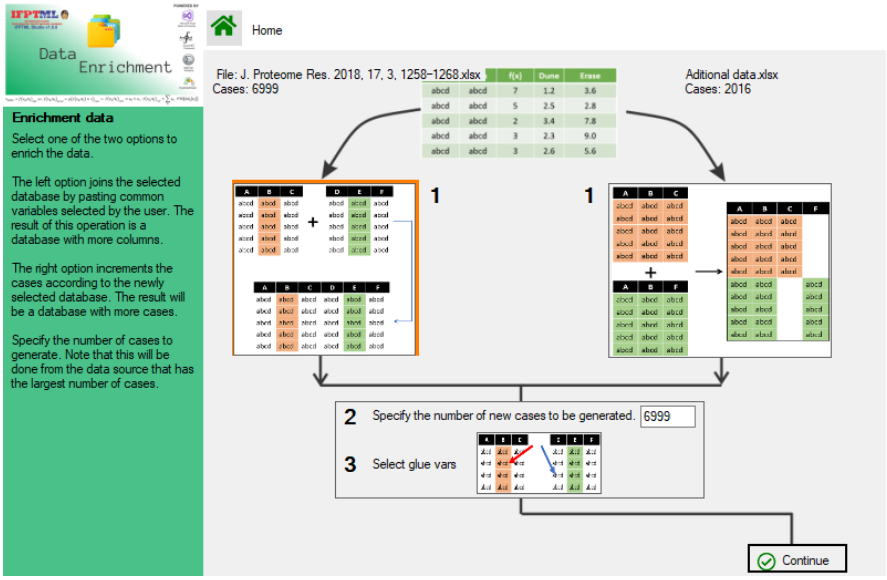


Figura 38: Selección del número de casos resultantes

El número de casos solo acepta números enteros

3. Se deberán seleccionar la o las variables que servirán de referencia para la integración de los dos archivos, para lo cual se hará clic sobre la imagen marcada con el número 3.

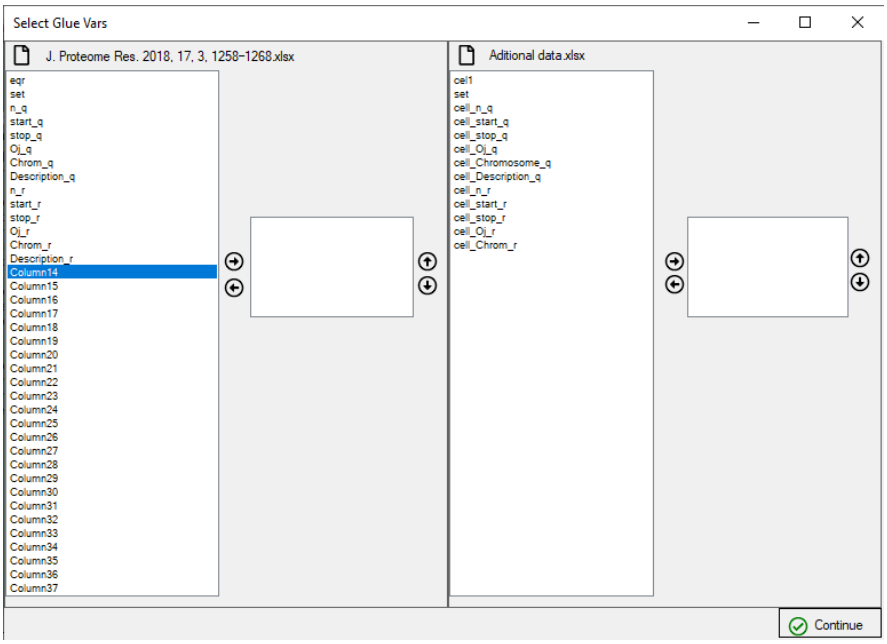


Figura 39: Ventana de selección de variables comunes

Los nombres de las columnas no deben ser necesariamente los mismos

4. Una vez realizados los pasos 2 y 3, se hará clic en el botón continue, luego de lo cual el software procesará la información dando como resultado un solo archivo principal.

7.4. Enriquecimiento de datos verticalmente

Esta opción permite agregar nueva información al archivo principal, de acuerdo con las columnas que llevan el mismo nombre en ambos archivos. Una vez efectuado este proceso el archivo principal puede incrementar sus columnas, y el número de casos. Durante la ejecución de este proceso es muy probable que aparezcan datos nulos, los cuales podrán ser tratados en otra de las opciones del software.

Para realizar este proceso se realizarán los siguientes pasos:

1. Se seleccionará la imagen de la izquierda marcada con el número 1.
2. Luego se hará clic en el botón continue, luego de lo cual el software procesará la información dando como resultado un solo archivo principal.

7.5. Data Curation (Conservación de datos)

Esta opción permite tratar las anomalías presentadas en la información, tales como valores nulos o en blanco, para lo cual se ofrecen tres opciones: reemplazar los datos nulos o en blanco con el promedio de los datos de su columna correspondiente, rellenar con ceros, o eliminar los casos. Para el desarrollo de este proceso se seguirán los siguientes pasos:

1. Seleccionar el botón “Data curation”
2. Se mostrará la ventana con el sumario de las columnas del archivo principal junto con el tipo de dato reconocido por el software (string, double, object) y la cantidad de valores nulos encontrados por cada variable
3. A continuación, se deberá seleccionar una de las opciones:
    a. Replace with zero values
    b. Replace with average
    c. Remove null cases
4. Finalmente se hará clic en el botón “continue”

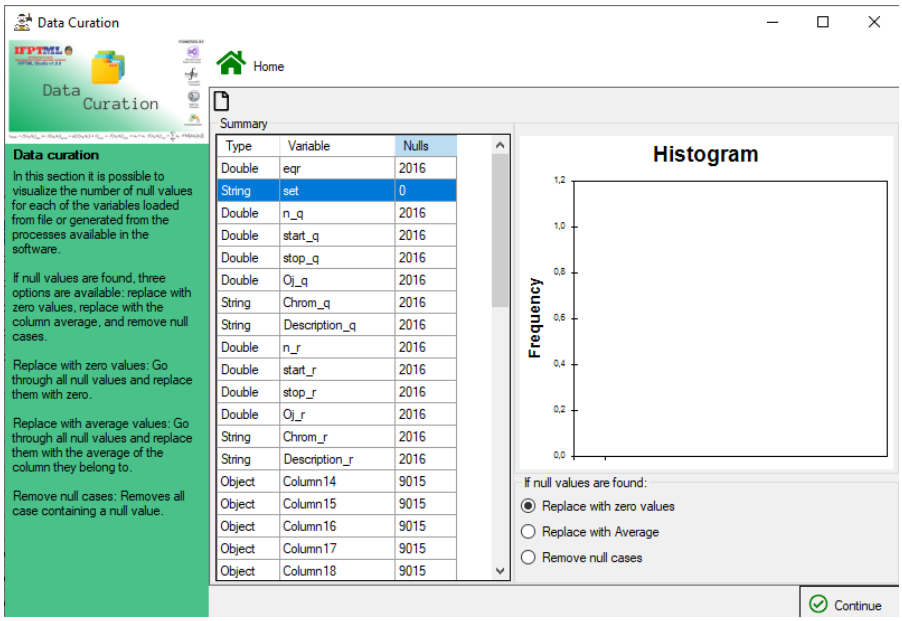


Figura 40: Pantalla de selección de método de tratamiento de valores anómalos

En esta ventana se puede observar el histograma de los datos de tipo Double que no tienen valores nulos

5. El software realizará los cálculos y el reemplazo pertinente.

7.6. Fusión de variables

Fusión de variables es una función por lotes, que permite reunir varias columnas en una nueva. Se deberá seleccionar las columnas para luego ordenarlas y agregarlas a la lista que serán procesadas. Para la realización de este proceso, es necesario realizar los siguientes pasos:

1. Seleccionar el menú “Vars fusión”
2. Se mostrará una ventana que contiene las variables disponibles en el archivo principal

3. Se seleccionarán las variables que se deseen fusionar y se pasarán al listado de columnas a unir, mediante el uso del botón con un símbolo de flecha verde hacia la derecha y si se desea eliminar del listado se usará el botón de color rojo con flecha hacia la izquierda
4. Una vez que se han seleccionado las variables se las debe ordenar mediante el uso de los botones de color verde con flecha hacia arriba o hacia abajo.
5. Una vez que se tenga el orden requerido, mediante el botón “Add to list”
6. Se repetirán los pasos 3, 4 y 5 para agregar nuevas variables al listado.
7. Una vez que se tenga el listado completo de las variables a procesar, se hará clic en el botón “Continue”
8. Se procesará el listado de columnas agregadas, para lo cual se crearán las nuevas columnas con la fusión del contenido de las columnas seleccionadas.
9. Una vez culminado este proceso se deberá cerrar la ventana.

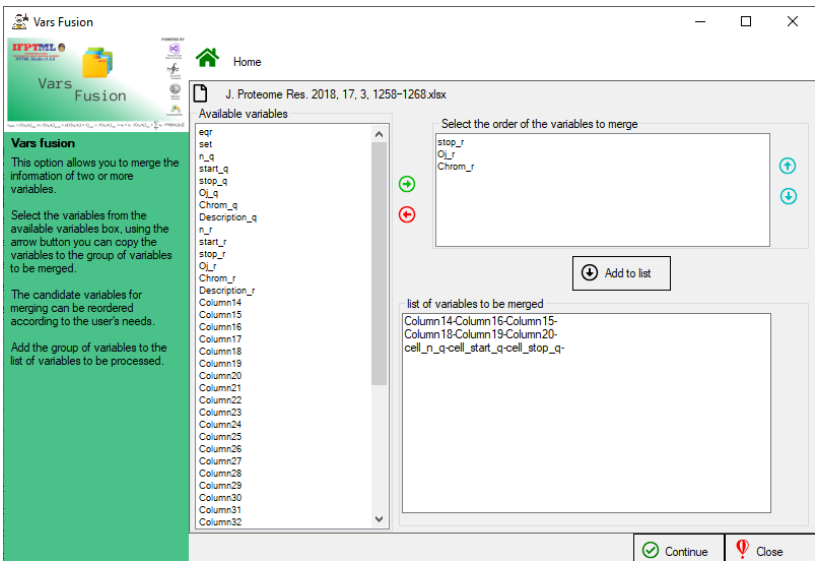


Figura 41: Ventana de fusión de información de variables

EL procesamiento se realiza de acuerdo con el orden en la que fueron agregadas al listado

7.7. PT Processing

Este apartado muestra dos opciones, relacionadas con el cálculo de descriptores y operaciones de perturbación.

7.7.1. Multicondition Discretization (Discretización Multicondicional)

Discretización Multicondicional, permite calcular los principales valores de discretización para el modelo IFPTML. El cálculo de descriptores se basa en la selección de una variable categórica y una numérica. Para este procedimiento, se realizarán los siguientes pasos

1. Seleccionar el menú “Multicondition Discretization”
2. Se seleccionará una variable categórica del listado de variables disponibles
3. Se seleccionará una variable numérica del listado de variables disponibles
4. Se pulsará el botón continuar

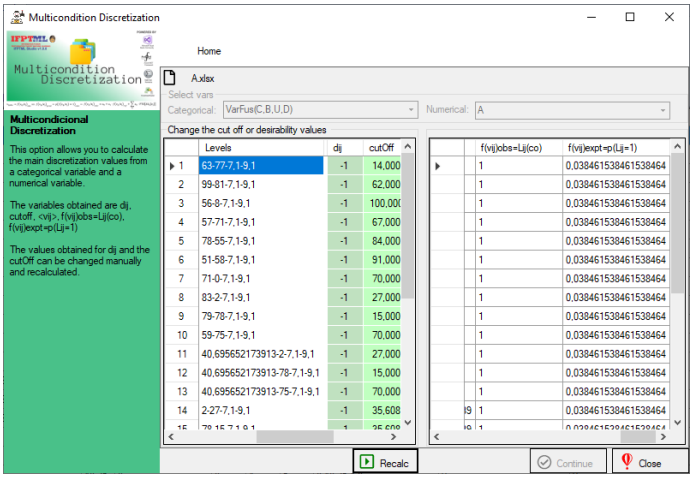


Figura 42: Ventana de para cálculo de valores de discretización

5. Se podrán editar los valores dij, y el cutoff, luego de lo cual se pulsará el botón “Recalc”
6. Una vez obtenidos los cálculos, se podrá cerrar la ventana

Finalmente, los resultados se mostrarán en la pantalla principal.

7.7.2. Operadores de perturbación

PTO´s Calculation es una opción de procesamiento por lotes, que permite listar las operaciones de perturbación necesarios para generar el modelo IFPTML. Para realizar este procedimiento es necesario realizar los siguientes pasos:

1. Seleccionar el menú PTO´s calculation
2. Seleccionar una o varias variables categóricas
3. Seleccionar una o varias variables numéricas
4. Seleccionar el Operador
5. Hacer clic en el botón para agregar a la lista de procesamiento por lotes

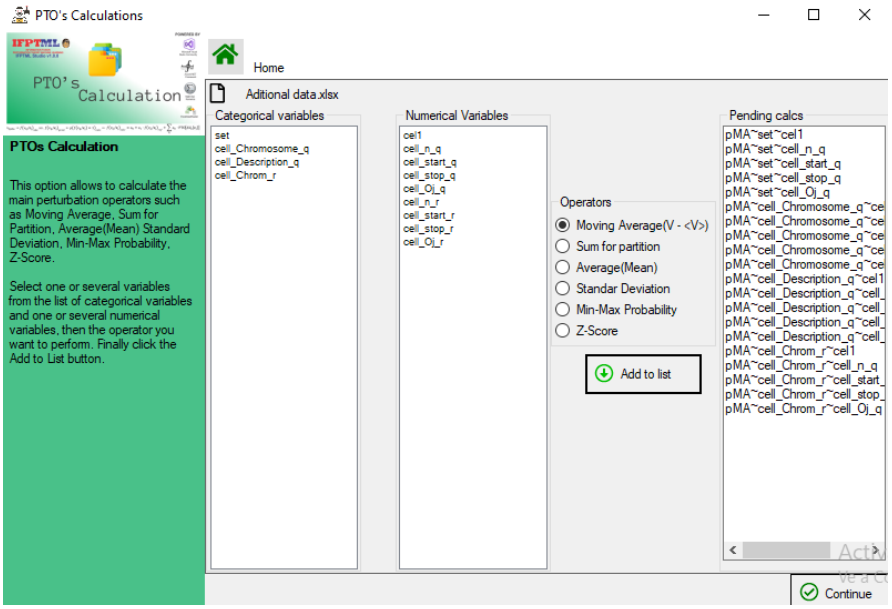


Figura 43: Ventana de cálculo de operaciones de perturbación

El proceso puede tardar varios minutos

6. Finalmente, hacer clic en el botón Continue

7.8. ML Analisys

Es el proceso final de cálculo del modelo IFPTML, que reúne la mayoría de variables calculadas en las opciones anteriores de este software.

7.8.1. IFPTML Training

IFPTML Training permite generar el modelo IFPTML a partir de los datos y cálculos, luego de lo cual es posible ajustar los valores p1 y p2 para mejorar la exactitud de los cálculos. Para el cálculo del modelo se realizarán los siguientes pasos:

1. Seleccionar el menú IFPTML Training
2. Se mostrará el listado de todas las variables disponibles, de la cual se deberá seleccionar la variable observada.
3. Seleccionar la variable de referencia
4. Seleccionar el conjunto de variables de los cálculos de perturbaciones
5. Seleccionar una variable que contiene valores de entrenamiento y validación

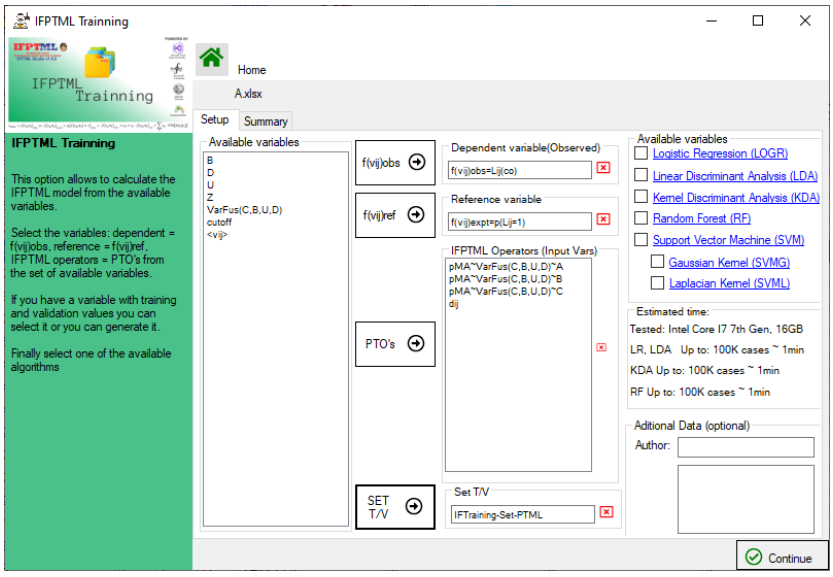


Figura 44: Ventana de configuración de variables que ingresan al cálculo del modelo IFPTML

Cada algoritmo tiene la posibilidad de configurar sus propiedades

6. Seleccionar el algoritmo que aplicará y se pulsará el botón “Continue”

La variable de entrenamiento y validación se puede generar, para lo cual se podrá indicar el porcentaje de los datos que se consideraran como validación y como de entrenamiento.

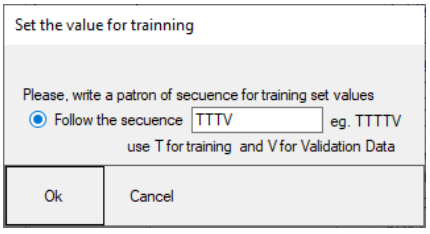


Figura 45: Ventana de generación de valores de entrenamiento y validación

Nota: Los valores ingresados deben ser T o V y la cantidad de ellos establece el porcentaje de valores de entrenamiento y validación por ejemplo TTTV, establecerá el 75% de valores de entrenamiento y 25% de valores de validación.

Antes de generar el modelo se solicitará al usuario que se seleccione una carpeta en la cual se almacenaran los resultados obtenidos, luego de lo cual se podrá obtener los resultados generados.

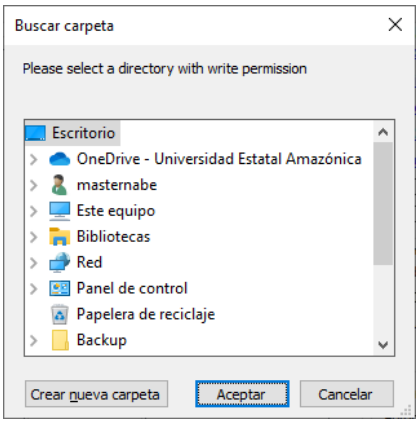


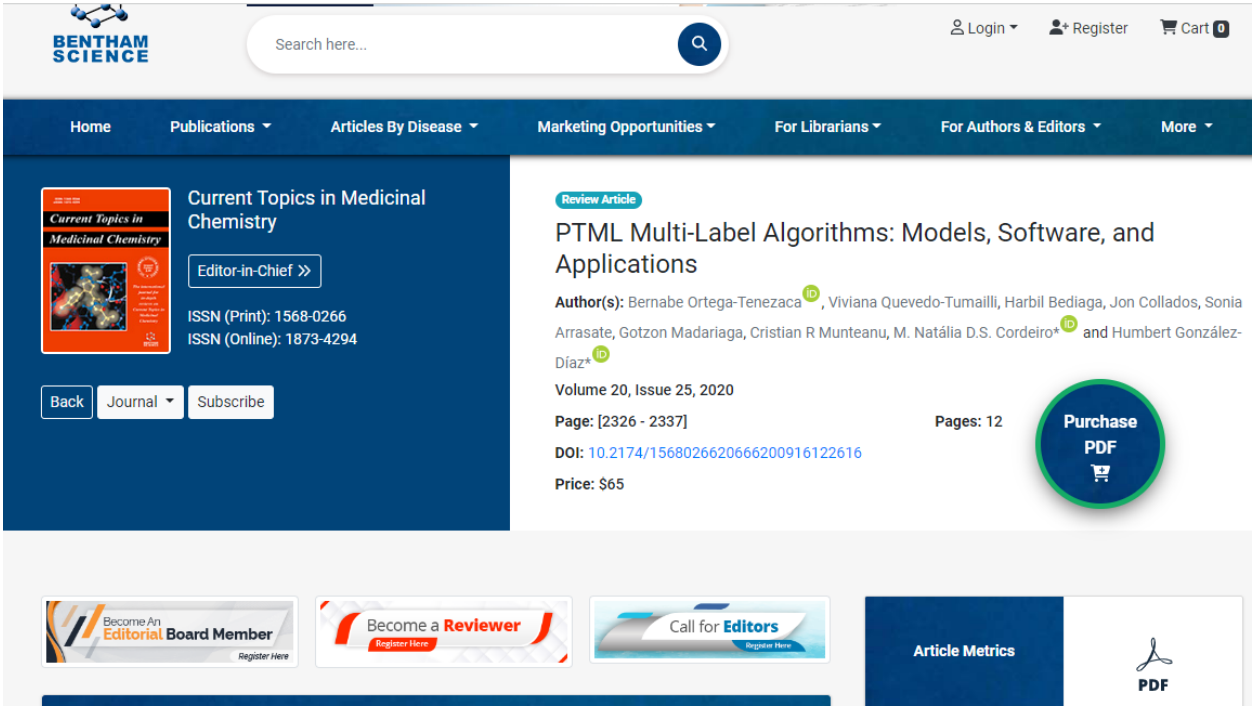
Figura 46: Selección de la carpeta donde se almacenarán los resultados del modelo IFPTML

ANEXO II

8. PUBLICACIONES

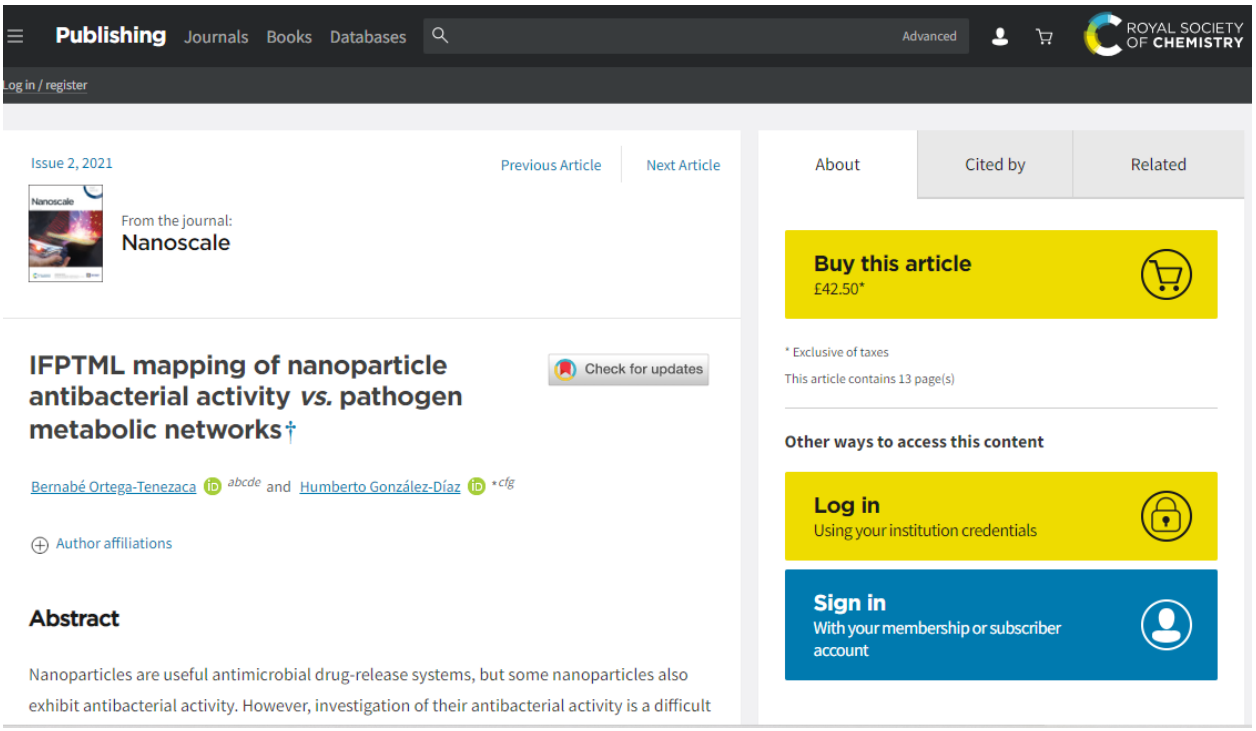
8.1. Publicación 1: PTML Multi-Label Algorithms: Models, Software, and Applications

September 2020 Current Topics in Medicinal Chemistry 20(25)
DOI: 10.2174/1568026620666200916122616
Lab: CHEMPTML.LAB
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli; Harbil Bediaga; Jon Collados Piña; Sonia Arrasate; Gotzon Madariaga; Cristian R Munteanu; M Natália D S Cordeiro; Humbert G. Díaz



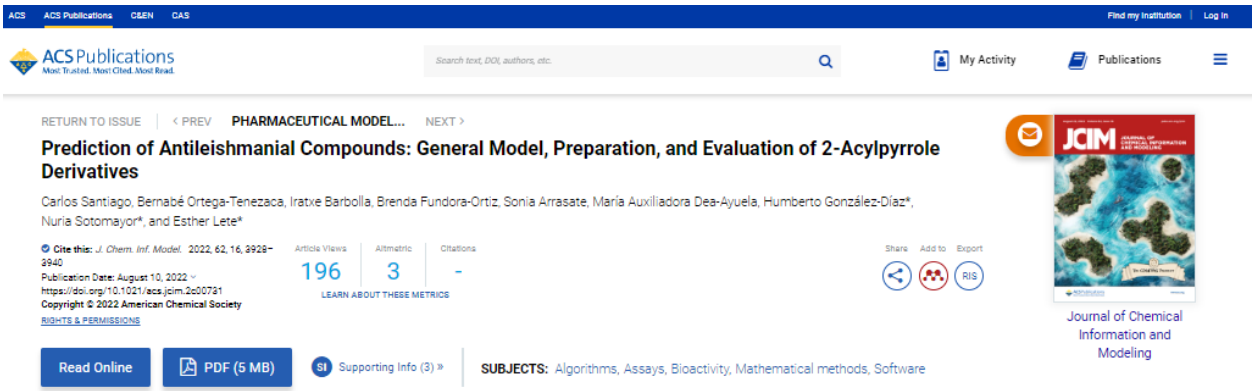
8.2. Publicación 2: IFPTML Mapping of Nanoparticles Antibacterial Activity vs. Pathogens

Metabolic Networks
December 2020 Nanoscale 13(10)
DOI: 10.1039/D0NR07588D
Bernabe Ortega-Tenezaca; Humbert G. Díaz



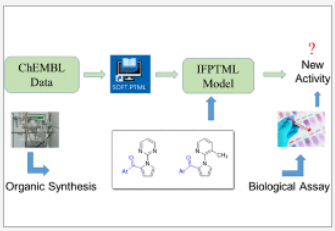
8.3. Publicación 2: Prediction of Antileishmanial Compounds: General Model, Preparation, and Evaluation of 2-Acylpyrrole Derivatives

August 2022 Journal of Chemical Information and Modeling 62(16)
DOI: 10.1021/acs.jcim.2c00731
Labs: Esther Lete's Lab CHEMPTML.LAB
Carlos Santiago; Bernabe Ortega-Tenezaca; Iratxe Barbolla; Brenda Fundora-Ortiz; Sonia Arrasate; Auxiliadora Dea; Humbert G. Díaz; Nuria Sotomayor; Esther Lete



Abstract

In this work, the SOFT.PTML tool has been used to pre-process a ChEMBL dataset of pre-clinical assays of antileishmanial compound candidates. A comparative study of different ML algorithms, such as logistic regression (LOGR), support vector machine (SVM), and random forests (RF), has shown that the IFPTML-LOGR model presents excellent values of specificity and sensitivity (81–98%) in training and validation series. The use of this software has been illustrated with a practical case study focused on a series of 28 derivatives of 2-acylpyrroles **5a**,**b**, obtained through a Pd(II)-catalyzed C–H radical acylation of pyrroles. Their *in vitro* leishmanicidal activity against visceral (*L. donovani*) and cutaneous (*L. amazonensis*) leishmaniasis was evaluated finding that compounds **5bc** (IC_{50} = 30.87 μM, SI > 10.17) and **5bd** (IC_{50} = 16.87 μM, SI > 10.67) were approximately 6-fold more selective than the drug of reference (miltefosine) in *in vitro* assays against *L. amazonensis* promastigotes. In addition, most of the compounds showed low cytotoxicity, CC_{50} > 100 μg/mL in J774 cells. Interestingly, the IFPMTL-LOGR model predicts correctly the relative biological activity of these series of acylpyrroles. A computational high-throughput screening (cHTS) study of 2-acylpyrroles **5a**,**b** has been performed calculating >20,700 activity scores *vs* a large space of 647 assays involving multiple *Leishmania* species, cell lines, and potential target proteins. Overall, the study demonstrates that the SOFT.PTML all-in-one strategy is useful to obtain IFPTML models in a friendly interface making the work easier and faster than before. The present work also points to 2-acylpyrroles as new lead compounds worthy of further optimization as antileishmanial hits.



Anexo III

9. CONGRESOS

9.1. Publicación 1: Mapping Bacterial Metabolic Network topology vs. Nanoparticle antibacterial activity

November 2021
DOI: 10.3390/mol2net-07-11832
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Bernabe Ortega-Tenezaca

MDPI sciforum

MOL2NET, International Conference Series on Multidisciplinary Sciences

Mapping Bacterial Metabolic Network topology vs. Nanoparticle antibacterial activity

Bernabe Ortega-Tenezaca [a, b, c]

[a] RNASA-IMEDIR, Computer Science Faculty, University of A Coruna, 15071 A Coruña, Spain.
[b] Amazon State University UEA, Puyo, Pastaza, Ecuador.
[c] Center for Investigation on Technologies of Information and Communication (CITIC), University of Coruña (UDC), Campus de Elviña s/n, 15071 A Coruña, Spain.

Abstract

The study of the bacterial metabolic structure of MNs with high resistance to the action of Nanoparticles (NPs) could help to design new NPs with specific antibacterial activity. We have used 3 numerical parameters which, in terms of graph theory, N_{nm} represents the number of nodes, $<L_{int}>$ the total number of arrows entering a node and $<L_{outk}>$ is the total number of arrows leaving a node of the complex graph. From another point of view, N_{nm} is the number of metabolites in the MN, $<L_{int}>$ is the number of metabolites that are precursors of the query metabolite and $<L_{outk}>$ is the number of metabolites that are products of a metabolic reaction with the query metabolite as a precursor. Finally, ACUs is a new parameter that represents the fusion of the 3 parameters mentioned above. This study also provides a closer look at the predictive power of the IFPTML-LOGR and IFPTML-RF models.

9.2. Publicación 2: PTML in optimizing preclinical plasmodium assays

November 2021
DOI: 10.3390/mol2net-07-11820
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca

PTML in optimizing preclinical plasmodium assays

Viviana F. Quevedo-Tumailli [a, b], Bernabe Ortega-Tenezaca [a, b]

[a] RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.
[b] Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.

Abstract:

The use of algorithms to predict optimal results in preclinical malaria assays from a data set that includes the experimental conditions of preclinical assays and characteristics of proteins, genes and chromosomes is a breakthrough in research. The process from data collection to data processing takes 70 percent of the time to develop. The process from the creation of preliminary models to the production of the model takes 30 percent of the total research time. There are several databases such as ChEMBL, Uniprot and NCBI-GDV that allow the collection of information on both preclinical assays and characteristics of any species, in this case study is *plasmodium falciparum*. This species is a major public health problem in tropical and subtropical countries. *P. falciparum* usually causes high fever, diarrhea, chills and in a few hours, it can evolve to a severe case causing death. The use of different algorithms such as: Linear Discriminant Analysis (LDA), Classification Tree with Univariate Splits (CTUS), Classification Tree with Linear Combinations (CTLC), and so on. The use of these algorithms and the perturbation theory allows pharmaceutical industries to optimize preclinical testing processes obtaining the most optimal models with a high percentage of specificity and sensitivity.

9.3. Publicación 3: Critical essay on predictive models for anti-sarcoma compounds

October 2021
DOI: 10.3390/mol2net-07-11231
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Bernabe Ortega-Tenezaca

Critical essay on predictive models for anti-sarcoma compounds

Bernabe Ortega-Tenezaca [a,b]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.*

Abstract.

Today, studies are performed from a dataset spanning multiple preclinical assays and different experimental conditions for sarcomas. PTML is a tool that combines Machine Learning (ML) algorithms and Perturbation Theory (PT) principles. With PTML, ML techniques can be used to predict antisarcoma compounds. At the same time, different PT techniques can be applied. One of the most widely used ML techniques is the neural network which showed high accuracy for both training and model validation. It is important to emphasize that the production of the most optimal model would save resources in the pharmaceutical industries. In a recent paper Cabrera *et al.* reported a new model for prediction of anti-sarcoma compounds. The model is very interesting because it can predict the biological activity vs multiple proteins, etc. The authors also explored multiple molecular descriptors of drugs as well as many assay conditions like protein target, cell line, etc. There are some suggestions we can make to improve future versions of this paper. For instance, the authors could calculate also sequence descriptor of target proteins to predict the results for new mutants. On my opinion, it could be very interesting developing a user-friendly software for use of non-expert medicinal chemists. This software could be a desktop or online server application increasing the use of the model worldwide. Another interesting step could be the fusion of the present pre-clinical data with clinical data including variables of patients or population groups. In all case, the paper is very interesting an opens new gates to the authors for future works including new features to the design of antisarcoma compounds.

9.4. Publicación 4: Predictive Modeling with Machine Learning and Perturbation Theory

October 2021
DOI: 10.3390/mol2net-07-11217
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli

CHEMINFOICD3, 2021, 7, ISSN: 2624-5078, https://mol2net-07.sciforum.net/cheminfoicd3-03

CHEMINFOICD3-03: Cheminfo. Int. Conf. Drug Des. & Discov. Alfenas, Brazil-München, Germany-Ch. Hill, Durham, USA, 2021

Predictive Modeling with Machine Learning and Perturbation Theory

Bernabe Ortega-Tenezaca [a, b], *Viviana F. Quevedo-Tumailli* [a, b]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.*

Abstract

PTML is a combination of Machine Learning (ML) and Perturbation Theory (PT) that allows to create prediction models in many areas of knowledge mainly in Medicinal Chemistry to handle large amounts of data representing physical and chemical properties of different organisms and biological systems under different input conditions. PTML allows to establish dispersion measurements on descriptors of physicochemical properties of different organisms with high values of sensitivity, specificity and accuracy higher than 70%.

9.5. Publicación 5: Predictive models for compounds against Plasmodium Falciparum

October 2021
DOI: 10.3390/mol2net-07-11215
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca

CHEMINFOICD3-03: Cheminfo. Int. Conf. Drug Des. & Discov. Alfenas, Brazil-Mittnchen, Germany-Ch. Hill, Durham, USA, 2021

MDPI sciforum

Predictive Modeling with Machine Learning and Perturbation Theory

Bernabe Ortega-Tenezaca [a, b], *Viviana F. Quevedo-Tumailli* [a, b]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.*

Abstract

PTML is a combination of Machine Learning (ML) and Perturbation Theory (PT) that allows to create prediction models in many areas of knowledge mainly in Medicinal Chemistry to handle large amounts of data representing physical and chemical properties of different organisms and biological systems under different input conditions. PTML allows to establish dispersion measurements on descriptors of physicochemical properties of different organisms with high values of sensitivity, specificity and accuracy higher than 70%.

9.6. Publicación 6: Predictive models as a useful tool for preclinical assay optimization in antimalarial compounds

October 2021
DOI: 10.3390/mol2net-07-11216
Conference: MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca

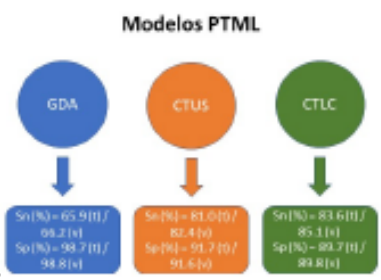
CHEMINFOICD3-03: Cheminfo. Int. Conf. Drug Des. & Discov. Alfenas, Brazil-München, Germany-Ch. Hill, Durham, USA, 2021

Predictive models as a useful tool for preclinical assay optimization in antimalarial compounds.

Viviana F. Quevedo-Tumailli [a, b], *Bernabe Ortega-Tenezaca* [a, b]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador.*

Graphical Abstract



Abstract.

In this study, three Perturbation Theory Machine Learning (PTML) models were created to optimize preclinical assays on antimalarial compounds of the parasitic species of the genus *Plasmodium falciparum*. Between General Discriminant Analysis (GDA), Classification Tree with Univariate Splits (CTUS) and Classification Tree with Linear Combinations (CTLC). The PTML-CTLC presented the best performance with a Sensitivity percentage

9.7. Publicación 7: Modelos PTMLIF en la predicción de sistemas

March 2021
DOI: 10.3390/mol2net-06-08949
https://sciforum.net/paper/view/8949
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca

SciForum
MOL2NET

Modelos PTMLIF en la predicción de sistemas de nanopartículas decoradas con fármacos.

Viviana F. Quevedo-Tumailli [1,2], Bernabé Ortega-Tenezaca [1,2]
[1] RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain;
[2] Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador;

Abstract:
Los modelos PTMLIF (Perturbation Theory, Machine Learning and Information Fusion) son la combinación de la teoría de la perturbación con el aprendizaje automático y la fusión de la información. Estos modelos se usaron en esta investigación para predecir la probabilidad de que los complejos nanopartícula decoradas con fármacos (Drugs-decorated Nanoparticles o DDNP) tengan actividad antipalúdica en este caso de estudio contra el Plasmodium. Esta enfermedad es la causa de la malaria en los seres humanos que se transmite a través de la picadura del mosquito hembra del género Anopheles. Se fusionó 107 características de entrada y 249.990 ejemplos aproximadamente desde la base de datos ChEMBL. El mejor modelo de clasificación fue proporcionado por método Random Forest, con solo 27 características seleccionadas de fármacos / compuestos y nanopartículas en todas las condiciones experimentales consideradas. El alto rendimiento del modelo se demostró mediante el área media bajo las características operativas del receptor (AUC) en un subconjunto de prueba con un valor de 0,9921 ± 0,000244 (validación cruzada de 10 veces). En este trabajo también se demostró el poder de la fusión de información de las características experimentales de fármacos / compuestos y nanopartículas para la predicción de la actividad antipalúdica de nanopartículas-compuestos.

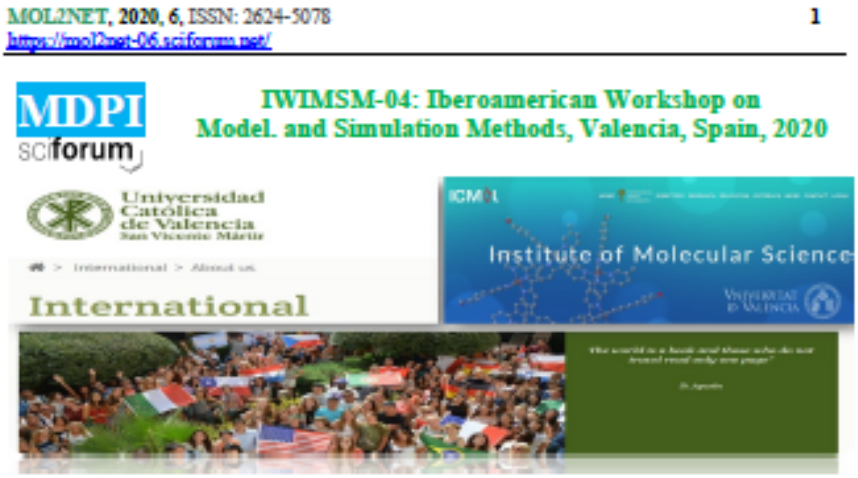
Keywords: Machine Learning, Perturbation Theory, Information Fusion, Nanoparticle, Random Forest.

References:

1. Dutta, P.P.; Bordoloi, M.; Gogoi, K.; Roy, S.; Narzary, B.; Bhattacharyya, D.R.; Mohapatra, P.K.; Mazumder, B. Antimalarial silver and gold nanoparticles: Green synthesis, characterization and In Vitro study. Biomed. Pharmacother. 2017,91, 567–580.
2. Quevedo-Tumailli, V.F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H. Chromosome gene orientation inversion networks (GOINs) of plasmodium proteome. J. Proteome Res. 2018,17, 1258-1268.
3. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J.M.; Loza, M.I.; Munteanu, C.R.; Pazos, A.; Garcia-Mera, X.; Gonzalez-Diaz, H. Perturbation theory/machine learning modelo f ChEMBL data for dopamine targets: Docking, synthesis, and Assay of new L-prolyl-L-leucyl-glycinamide peptidomimetics. ACS Chem. Neurosci. 2018, 9, 2572-2587
4. Kleandrova, V.V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J.M.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation

9.8. Publicación 8: Vision IA Microservice for the detection of ID personal data

August 2020
DOI: 10.3390/mol2net-06-06898
Conference: USEDAT-08: USA-Europe Data Analysis Training Program Workshop, UPV/EHU, Bilbao-MDC, Miami, USA, 2020
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli; Rivadeneira-Ramos Edgar; Luis Alberto Uvidia Armijo

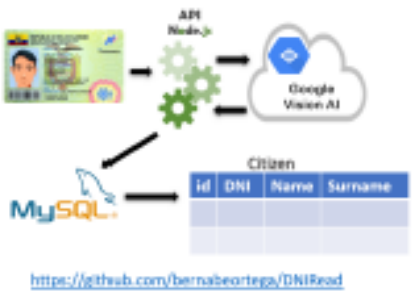


Vision IA Microservice for the detection of ID personal data

Bernabe Ortega-Tenezaca [a, b], Viviana Quevedo-Tumailli [a, c], Edgar Rivadeniera-Ramos [d], Luis Alberto Uvidia Armijo [e]

[a] RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain
[b] Universidad Estatal Amazónica, UTIC, Puyo, Ecuador
[c] Universidad Estatal Amazónica, Sciences of Earth Departament, Puyo, Ecuador
[d] Universidad Estatal de Bolívar, DTIC, Bolívar, Ecuador
[e] Escuela Superior Politécnica de Chimborazo, Sede Morona, Ecuador

Graphical Abstract



https://github.com/bernabeortega/DNIRead

Abstract

We provided a microservice developed in Nodejs connected to Google Cloud services. From a photograph of the ID obtained through a mobile device, its content is analyzed to extract its information. The Vision AI API uses AutoML Vision, which is capable of interpreting text. The information returned by the Vision AI service is processed in the microservice and the confidential information is anonymized, to be stored in a MySQL database server. The microservice is part of a mobile application to register the entry of citizens to a government institution.

9.9. Publicación 9: IF for the dataset of Plasmodium Falciparum

September 2019
DOI: 10.3390/mol2net-05-06253
Conference: MOL2NET 2019, International Conference on Multidisciplinary Sciences, 5th edition
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca

MDPI SCIforum

USEDAT: USA-Europe Data Analysis Training School Workshop

IF for the dataset of Plasmodium Falciparum

Viviana Quevedo-Tumailli[a,b], Bernabe Ortega-Tenezaca[b,c].

[a] RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain
[b] Universidad Estatal Amazónica, Sciences of Earth Departament, Puyo, Ecuador
[c] Universidad Estatal Amazónica, UTIC, Puyo, Ecuador

Graphical Abstract

Abstract.
Build a Dataset of structural and external variables of the Plasmodium Falciparum organism using the information collected from different databases such as ChEMBL, NCBI-GDV and Uniprot to predict the activity of the drug to the gene on the chromosome of this organism and determine if the orientation of the gene it influences this process. In this research work it is important to calculate the Shannon of Entropy in Canonical Smile, Protein Sequences and genes sequences.

References

1. Chao, W., Yin, C., Takahashi, K., & Lin, J. J.-M. (2019). Hydrogen-bonding Mediated Reactions of Criegee Intermediates in the Gas Phase - The Competition between Bimolecular and Termolecular Reactions and the Catalytic Role of Water. *The Journal of Physical Chemistry. A.* https://doi.org/10.1021/acs.jpca.9b07117

9.10. Publicación 10: Batch processing in transformation of continuous variables for PTML Theory

September 2019
DOI: 10.3390/mol2net-05-06252
Conference: MOL2NET 2019, International Conference on Multidisciplinary Sciences, 5th edition
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli

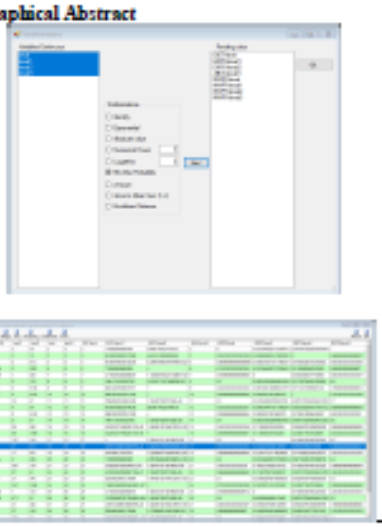
MDPI

USEDAT: USA-Europe Data Analysis Training School Workshop

Batch processing in transformation of continuous variables for PTML Theory

Bernabe Ortega-Tenezaca (bortega@uea.edu.ec) [a,b],
Viviana Quevedo-Tumailli (vquevedo@uea.edu.ec)[a,c].
[a] RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.
[b] Universidad Estatal Amazónica, UTIC, Puyo - Ecuador
[c] Universidad Estatal Amazónica, Science of earth Departament, Puyo – Ecuador

Graphical Abstract



Abstract.

In the present work, a software module has been developed that allows the selection of continuous variables from an Excel file, which are initially subjected to a process of verification and cleaning of the information, allowing the elimination of cases, or otherwise replacing outliers. With the average value, it is then possible to perform transformation operations such as identity, exponential, absolute value, numerical power, logarithm, maximum and minimum probability, z-score, harmonic mean sum, Euclidean distance, in batch processing of continuous variables. In the end, the respective results can be obtained within a dataset that can be stored in CSV format, or in turn continue processing with PTML.

9.11. Publicación 11: MVVM design pattern for asynchronous events in information system

December 2018
DOI: 10.3390/mol2net-04-06081
Conference: MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th edition
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli; Lenin Ochoa Carrión; Luis Alberto Uvidia Armijo; Ronny Rodríguez Cabrera

MOL2NET, 2018, 4, ISSN: 2624-5078, ISBN: 978-3-03842-820-6
http://sciforum.net/conference/mol2net-04

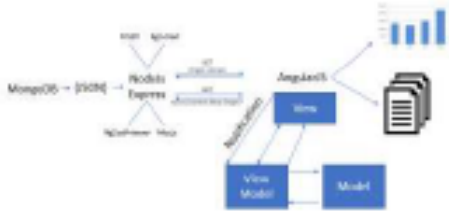
MDPI [redacted]

MOL2NET, International Conference Series on Multidisciplinary Sciences

" MVVM design pattern for asynchronous events in information systems "

Bernabe Ortega-Tenezaca [a, b, c]
Viviana F. Quevedo-Tumailli [a, c]
Lenin Patricio Ochoa Carrión [b, c]
Ronny Fabricio Rodriguez Cabrera [b]
Luis Alberto Uvidia Armijo [d]
[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Regional Autónoma de los Andes Uniandes – Puyo, Pastaza, Ecuador.*
[c] *Universidad Estatal Amazónica – Puyo, Pastaza, Ecuador.*
[d] *Escuela Superior Politécnica de Chimborazo, Riobamba, Chimborazo, Ecuador*

Graphical Abstract



Abstract.
The technological integration of software development components and languages allows the creation of business management platforms that include metrics and standards in data processing[1]. In this work we propose the generation of a scalable and integrable web management application, through the asynchronous use of events and the MVVM development pattern[2,3].

Introduction

The development of business management software and data management is subject to constant maintenance stages depending on internal or legislative requirements[1], therefore, it is important to consider the scalability property of the web applications[4], the degree of cohesion and coupling of their components, as well as the interoperability methods of their modules and programming languages. The MVVM design pattern[2,3] is adopted for its operational advantages measured through performance and productivity tests. For storage of information, a document-oriented database is used whose queries and use of JSON objects[5-8]. In addition, a component for radial transmission was used.[9,10]

Materials and Methods

An ApiREST application was developed that combines development languages such as NodeJS, AngularJS, with Mongoose components, ng2-charts[11,12], ng2-pdf-viewer, bcrypt-nodejs, mrp.js[10], adminLTE II, MongoDB with resource exchange through CORS, under the MVVM design pattern, with asynchronous programming events programming model.

9.12. Publicación 12: Web Application for Real Time Data Visualization of Heat Sensor

December 2018
DOI: 10.3390/mol2net-04-05909
Conference: MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th edition
Bernabe Ortega-Tenezaca; Viviana Fernanda Quevedo Tumailli; Víctor Cerda; Octavio Guijarro Rubí; Estela Guardado Yordi; Amaury Pérez

MDPI

MOL2NET, International Conference Series on Multidisciplinary Sciences

" Web Application for Real Time Data Visualization of Heat Sensors "

Bernabe Ortega-Tenezaca [a, b, c], Viviana F. Quevedo-Tumailli [a, c], Victor Cerda Mejía [c], Octavio Edelberto Guijarro Rubí [d], Estela Guardado Yordi [e] and Amaury Pérez Martínez [a, e]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Regional Autónoma de los Andes Uniandes – Puyo, Pastaza, Ecuador.*
[c] *Universidad Estatal Amazónica – Puyo, Pastaza, Ecuador.*
[d] *Instituto Superior Tecnológico Francisco de Orellana – Puyo, Pastaza, Ecuador.*
[e] *Universidad de Camagüey, Cuba.*

Graphical Abstract

Abstract. Calorimetry [1-5] and real-time monitoring systems [2,6-9], are essential aspects in environmental and agroindustrial processes. In this work, we develop a web application [6] that allows to remotely visualize continuous graphs of data coming from heat sensors connected to an Arduino device [10] with Internet access. The information is initially stored in a MySQL database [11-13], which reactively [14,15] generate the graph and the calculation of descriptive statistics [2,6-9].

Introduction

Agroindustrial processes require real-time monitoring tools[2,6-9] and automated control, which allow for a visual follow-up during the phases of their development, and for a subsequent processing and analysis of data, specifically on processes in the which depends on the temperature. In order to carry out measurements and tests of electronic components, eight sensors connected to an Arduino board, which sends the information obtained through the internet, have been placed on a liquid conduction channel, variable temperature in relation to time[10,16] by means of your Wi-Fi device, to a database for storage, and immediately graphical the output of dynamic Datasets and their descriptive statistical indicators, within a certain configured range of maximum and minimum value in a web application reactive and optimized.[9,14,15,17,18]

9.13. Publicación 13: Prediction of RIFIN proteins with gene orientation network indices

January 2018
DOI: 10.3390/mol2net-03-05124
Conference: MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition
Projects: Mol2Net Conference USEDAT Training Prog
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca; Julio César Vargas-Burgos; Alejandro Pazos; Humbert G. Díaz

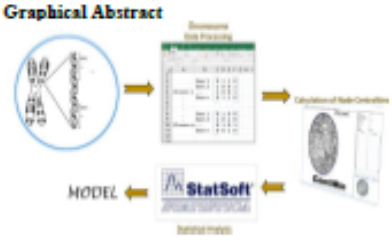
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Prediction of RIFIN proteins with gene orientation network indices

Viviana F. Quevedo-Tumailli [a, b], Bernabé Ortega-Tenezaca [a, b, c], Julio César Vargas-Burgos [b], Alejandro Pazos-Sierra [a] and Humbert Gonzalez-Diaz [d, e, *]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador*
[c] *Universidad Regional Autónoma de los Andes UNIANDES-Puyo, Ecuador*
[d] *Dept. of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain*
[e] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain*

Graphical Abstract



Abstract. Gene orientation may have a direct influence on the expression of genes. In this work, we developed a Linear Discriminant Analysis (LDA) model able to predict RIFIN-like proteins of out of 5365 of *Plasmodium falciparum* proteins with Sensitivity and Specificity 70-80% in training and external validation series.

Keywords: Malaria; *Plasmodium sp.* proteome; Chromosome microstructure; Gene orientation; Complex Networks; Machine Learning

* **Corresponding author:** H.G.D. (humberto.gonzalezdiaz@ehu.es)

References

1. Junker, B. H.; Koschuetzki, D.; Schreiber, F., Exploration of biological network centralities with CentiBiN. BMC Bioinformatics 2006, 7, (1), 219.
2. Hayo, A.; Albert, J.; Rooman, M., Modeling the Drosophila gene cluster regulation network for muscle development. PLoS One 2014, 9, (3), e90285.
3. Inoue, L. Y.; Neira, M.; Nelson, C.; Gleave, M.; Etzioni, R., Cluster-based network model for time-course gene expression data. Biostatistics 2007, 8, (3), 507-25.
4. Lavazec, C.; Sanyal, S.; Templeton, T. J., Hypervariability within the Rifin, Stevor and Pfmc-TM superfamilies in Plasmodium falciparum. Nucleic Acids Res 2006, 34, (22), 6696-707.
5. Hurst, L. D.; Williams, E. J.; Pal, C., Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet 2002, 18, (12), 604-6.
6. Kustatscher, G.; Grabowski, P.; Rappsilber, J., Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol 2017, 13, (8), 937.

9.14. Publicación 14: Notes Towards a Network Approach to Gene Orientation

December 2017
DOI: 10.3390/mol2net-03-05092
Conference: MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition
Projects: Mol2Net Conference USEDAT Training Prog
Viviana Fernanda Quevedo Tumailli; Bernabe Ortega-Tenezaca; Julio César Vargas-Burgos; Alejandro Pazos; Humbert G. Díaz

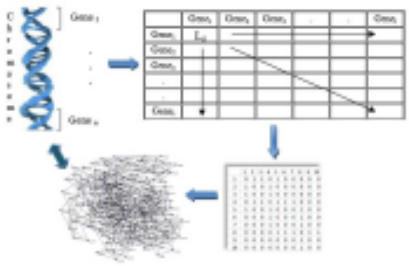
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Notes Towards a Network Approach to Gene Orientation

Viviana F. Quevedo-Tumailli [a, b], Bernabé Ortega-Tenezaca [a, b, c], Julio César Vargas-Burgos [b], Alejandro Pazos-Sierra [a] and Humbert González-Díaz [d, e, *]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador*
[c] *Universidad Regional Autónoma de los Andes UNIANDES-Puyo, Ecuador*
[d] *Dept. of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain*
[e] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain*

Graphical Abstract *(mandatory)*



Abstract. The distribution of the orientation genes in chromosomes is not random and may have function implications. In this work, we used complex networks to study the orientation of genes. We used as a case of study the 14 chromosomes of *Plasmodium falciparum*. We constructed the respective 14 networks with an average of 383 nodes (genes) and 1314 links (pairs of gene with inverse orientation). Node centralities of these nodes were used to study the structure of the network.

Keywords: Malaria; *Plasmodium sp.* proteome; Chromosome microstructure; Gene orientation; Complex Networks; Machine Learning

* **Corresponding author:** H.G.D. (humberto.gonzalezdiaz@ehu.es)

References

1. Hurst, L. D.; Williams, E. J.; Pal, C., Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet 2002, 18, (12), 604-6.
2. Kustatscher, G.; Grabowski, P.; Rappsilber, J., Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol Syst Biol 2017, 13, (8), 937.
3. Newman, M., The Structure and Function of Complex Networks. SIAM Review 2003, 56, 167-256.
4. Lin, H. H.; Zhang, L. L.; Yan, R.; Lu, J. J.; Hu, Y., Network Analysis of Drug-target Interactions: A Study on FDA-approved New Molecular Entities Between 2000 to 2015. Sci Rep 2017, 7, (1), 12230.

9.15. Publicación 15: The KP algorithm for the analysis of the optimal flow of information

December 2017
DOI: 10.3390/mol2net-03-05053
Conference: MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition
Bernabe Ortega-Tenezaca; Carlos Núñez Miranda

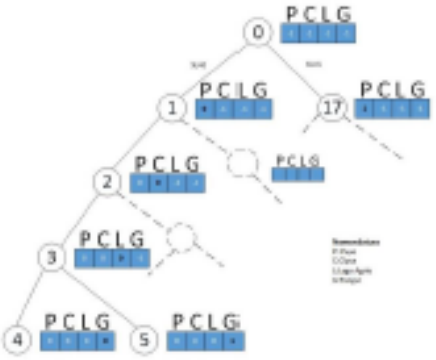
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

The KP Algorithm for the analysis of the optimal flow of information

Bernabe Ortega-Tenezaca (bortega@uea.edu.ec) [a] [b], Carlos Israel Núñez Miranda (ci.nunez@uta.edu.ec)[a].

[a] *MAESTRÍA EN GESTIÓN DE BASE DE DATOS, Ingeniería en Sistemas Electrónica e Industrial, Universidad Técnica de Ambato, Ambato, Ecuador*
[b] *Universidad Estatal Amazónica, Puyo, Pastaza, Ecuador*

Graphical Abstract



Abstract.

The purpose of the present research is to propose a solution to optimize the flow of institutional academic information, given that, for the economic moment in Ecuador, government budgetary allocations do not cover the investment deficit in the renovation, acquisition and technological updating, generating inconveniences in the flow of university information, in compliance with the Law Reforming the Law of Creation of the Amazon State University, which allows to extend the academic offer to different sectors of the region.

The solution proposed is applied to the Academic Information System, where most of the records related to the generation of institutional evidences of university accreditation, internal control of academic processes, research, community ties and management are stored. The optimization analysis of the information flow is based on the application of the algorithm KP.

9.16. Publicación 16: FRAMA 1.0: Framework for Moving Average Calculation in Operators in Data Analysis

November 2017
DOI: 10.3390/mol2net-03-05044
Conference: MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition
Projects: Mol2Net Conference USEDAT Training Prog.
Viviana Fernanda Quevedo Tumailli; Humbert G. Díaz; Bernabe Ortega-Tenezaca

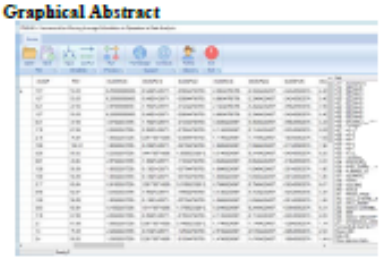
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

FRAMA 1.0: Framework for Moving Average Operators Calculation in Data Analysis

Bernabé Ortega-Tenezaca [a,b], Viviana F. Quevedo-Tumailli [a,b], and Humbert Gonzalez-Diaz [a, b,c]

[a] *RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[b] *Universidad Estatal Amazónica, Puyo, Pastaza, Ecuador*
[c] *Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain*
[d] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain*

Graphical Abstract



Abstract. Moving Average (MA) operators are used in Box-Jenkins's ARIMA models in time series analysis (1). We can used MA operators of structural descriptors are useful to quantify multiple conditions or parameters in complex datasets in Omics, Medicinal Chemistry, Nanotechnology, *etc.* (2-7). Speck-Planche and Cordeiro have also used this kind of models in multiple problems (8-11). In this work, we develop a desktop application that allows applying mathematical and statistical calculations in batches, on input and output variables selected by the user. From the obtained result a percentage sample of data is taken with a random contrast on which Machine Learning algorithms are applied

Introduction

In principle, we can calculate numerical parameters to quantify the structure of chemical compounds, peptides, and/or proteins. We can also use them as input variables for Machine Learning (ML) algorithms in order to predict the biological properties of these drugs, peptides, or proteins (13-29). On the other hand, Perturbation Theory (PT) models allow us to predict the solutions to a query problem (q) based on a previous known solution for a similar problem or problem of reference (r). In a recent works, we outlined a new type of ML method called PTML (PT + ML) based on both kind of models with applications in drug discovery and proteome research (25, 30). The PTML method uses different kind of PT operators to predict the properties of one system based on the properties of a system of reference. For instance, Moving Average (MA) operators used in Box-Jenkins's ARIMA models in time series analysis (31). We have used MA operators of structural descriptors are useful to quantify multiple conditions or parameters in complex datasets in Omics, Medicinal Chemistry,